High-dimensional sparse random networks with covariates

Stefan Stein

A thesis submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy in Statistics

Department of Statistics
University of Warwick

December, 2021

Contents

List of Figures

List of Tables

Acknowledgements

I am truly grateful to my supervisor, Professor Chenlei Leng for his support and guidance throughout the last years. His insightful feedback and advice have allowed me to learn a lot during this time; both in the academic context and beyond.

I thank Steve Thornton and Michel Randrianandrasana from Tata Steel for the fascinating insights into the steel industry that they have provided me with. I would also like to extend special thanks to Dr. Martine J. Barons. She has been most generous with her time and advice and has discussed the applied parts of my thesis with me at length.

Funding for this project was provided by EPSRC Industrial CASE training grant (EP/ R51214X/1) with project reference number 1935144.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

Parts of this thesis have appeared in the following papers:

1. Stein, S., Leng, C. (2020), 'A sparse β-model with covariates for networks', *arXiv preprint, arXiv:2010.13604*,
2. Stein, S., Leng, C. (2021), 'A sparse random graph model for sparse directed networks', *arXiv preprint, arXiv:2108.09504*,
3. Stein, S, Leng. C, Thornton, S., Randrianandrasana, M. (2021), 'A guided analytics tool for feature selection in steel manufacturing with an application to blast furnace top gas efficiency', *Computational Materials Science*, **186**, 110053.

Stefan Stein

Abstract

An increasingly urgent task in analysis of network data is to develop statistical models that include contextual information in the form of covariates while respecting degree heterogeneity and network sparsity. We study various stochastic network models with parameters that explicitly account for these stylized features of real-world networks.

To set the tone of the thesis, we highlight in Chapter 1 the fallacy of *data-selective inference* – a common practice of artificially truncating an observed network by throwing away any nodes that are not well-connected. This constitutes a form of sampling bias, which we quantify theoretically for the Erdős-Rényi model and empirically for the Stochastic Block Model.

We introduce the *sparse β-model with covariates* (SβM-C) in Chapter 2. By assuming sparsity of the degree heterogeneity parameter, SβM-C is capable of fitting sparse, undirected networks, enabling us to avoid data-selective inference. For parameter estimation, we propose the use of a penalized likelihood method with an ℓ_1-penalty on the nodal parameters. This gives rise to a convex optimization formulation which immediately connects our estimation procedure to the LASSO literature. We provide finite sample bounds on the excess risk and the ℓ_1-error of the resulting estimator and develop a central limit theorem for the parameter associated with the covariates.

In Chapter 3 we zoom in on the special case of SβM-C when no degree heterogeneity parameter is present. We call this the *sparse Erdős-Rényi model with covariates* (ER-C) and show that it can model networks of almost arbitrary sparsity.

We extend SβM-C to directed networks by introducing the parameter-Sparse Random Graph Model (SRGM) in Chapter 4. We prove that an ℓ_1-penalized estimator is model selection consistent for SRGM. We further recover results similar to the ones we established for SβM-C. Special focus is placed on the interplay of the network sparsity, the parameter sparsity and the penalty we use. This allows us to paint a nuanced picture of the effect of different sparsity regimes on parameter estimation.

Chapter 6 presents the results of a collaboration with Tata Steel in Europe and can be read independently from the rest of this thesis. In it we present the *initial Guided Analytics for parameter Testing and controlband Extraction* (iGATE) framework, a novel feature selection procedure for industry applications that combines expert knowledge with statistical techniques.

Preface

This PhD project has been funded by an Industrial Cooperative Awards in Science & Technology (ICASE) studentship and funding was partially provided by Tata Steel in Europe. As a result, there are two main outputs from this research, to each of which one part of the present thesis is dedicated. One one hand, I have been working on the application of methods from high-dimensional statistics to sparse random networks with covariates. On the other hand, I have been working with Tata Steel on creating a software application for finding influential parameters in manufacturing processes. The theoretical work on sparse random networks makes up the bulk of the present thesis, both in length as well as statistical novelty. Therefore, the emphasis of this thesis is on the former project. A short overview of the two projects is given below.

Sparse random networks with covariates

This work is presented in Part I and is mostly based on the work in Stein & Leng (2020, 2021). The focus of my theoretical work has been researching the mathematical properties of novel sparse random network models. Particular attention has been given to developing a mathematical theory for network models that capture three stylized features commonly observed in real-world networks:

1. *Sparsity*, which roughly means that the number of observed edges in a network scales sub-quadratically in the number of nodes,
2. *Degree-heterogeneity*, which refers to the phenomenon that real-world networks tend to have heavy-tailed degree distributions, or, broadly speaking, the model allows for the emergence of "hub-nodes" with many connections,
3. *Homophily*, which is the phenomenon that nodes that are similar to one another are more likely to form a connection.

Part I is organized as follows. In Chapter 1 the necessary definitions and notation are provided. Furthermore, Section 1.2 is dedicated to highlighting the commonly observed fallacy of focusing exclusively on dense sub-graphs of observed networks when analysing network data. This fallacy has been named *data-selective inference*

in Stein & Leng (2021). It constitutes a form of sampling bias, which has been quantified mathematically for the Erdős-Rényi model and via simulation for the Stochastic Block Model in Stein & Leng (2021). Section 1.2 sets the tone for the remainder of Part I and showcases the necessity for statistical network models to allow for sparse networks from the outset. In Section 1.3 we will discuss how the main models of interest of this thesis fit into the broader field of stochastic network theory.

Chapters 2 and 3 are in large part from Stein & Leng (2020) and are a step towards tackling the problem of data-selective inference for undirected networks. Chapter 2 introduces the sparse β-model with covariates (SβM-C) an extension of the sparse β-model (SβM) introduced in Chen et al. (2020). SβM is a novel generative network model for sparse random networks with degree heterogeneity. SβM-C extends this model by incorporating covariates on the node or edge level, which allows it to model homophily. The addition of covariates into the likelihood of SβM breaks the estimation procedure advocated by Chen et al. (2020), calling for a novel approach for estimating the model parameters. The sparse Erdős-Rényi model witch covariates (ER-C) is a special case of SβM. Since the theory developed for ER-C can be of independent interest, it is treated in its own chapter (Chapter 3).

Chapter 4 presents an extension of SβM-C to directed networks and is based on Stein & Leng (2021). Some of the results from SβM-C, notably the results on consistency and asymptotic normality, carry over to its directed version without much effort and some of the proofs are – sometimes line by line – the same. Therefore, those proofs have been relegated to the appendix. Methodological novelty is added in Stein & Leng (2021) via a theorem on model selection consistency, which is presented in Section 4.2.1.

It is worth noting that after having read Chapter 1 and having familiarized oneself with the necessary definitions and notation, the reader may choose to read Chapters 2, 3 and 4 in any order as each chapter is self-contained.

In January 2021 the paper Stein & Leng (2020) received an Honourable Mention award in the *2021 Student Paper Competition of the Statistical Learning and Data Science Section* of the American Statistical Association. In March 2021 the author was awarded the *IMS Hannan Graduate Student Travel Award* by the Institute of Mathematical Statistics for the paper Stein & Leng (2020). In September 2021, the author received an Honourable Mention in the *2021 Doctoral Researcher Awards* in the category Natural & Life Sciences for the work in Stein & Leng (2020).

A guided analytics tool for feature selection in steel manufacturing

Part II is a summary of the work I have been doing with Tata Steel in Europe. I have been working very closely with them on several applied data science projects over the course of my PhD. The main output of that work was the creation of the *initial Guided Analytics for parameter Testing and controlband Extraction* (iGATE) tool. The details of iGATE have been published in Stein et al. (2021) and the core functionality has been made freely available in the `igate` package for the R programming language (Stein 2019). The work on iGATE is presented in Chapter 6, the content of which has been taken from Stein et al. (2021).

iGATE is a feature selection framework that finds influential process parameters and their optimal ranges. At Tata Steel in Europe it has been made available to process operators in the form of a graphical user interface hosted on a server, accessible to anybody within the company. Process operators, who usually possess indispensable domain knowledge, but may not have been trained in statistics, can navigate through the analysis, modifying the results according to their expertise. The results of the analysis are verified by them and in the end a report is automatically generated of the conducted analysis. iGATE streamlines data science projects by combining the power of suitable statistical tools with process-expertise, whilst dramatically reducing the time needed for an analysis.

Part I

Sparse random networks with covariates

Chapter 1

Preliminaries: Motivation and definitions

Organization of this chapter

In Section 1.1 we discuss the most commonly observed properties of real-world networks we would like to capture in a stochastic network model. We informally introduce the sparse β-model with covariates (SβM-C) and the parameter-Sparse Random Graph Model (SRGM), which are the main objects of interest of Chapters 2 and 4 respectively. We introduce definitions in Section 1.1.1 and notation in Section 1.1.2. A large portion of this chapter is taken up by Section 1.2 which discusses the fallacy of what was called *data-selective inference* in Stein & Leng (2021): The commonly observed practice of "arguing away" the sparsity in observed networks by discarding any small connected components before fitting a model to the observed data, thus artificially truncating the network sample. We examine the bias resulting from data-selective inference in the Erdős-Rényi model theoretically and in the Stochastic Block Model empirically. This serves as motivating example for why we care about developing stochastic network models that incorporate sparsity from the outset. Sections 1.1 and 1.2 are largely based on the introductory sections of Stein & Leng (2021).

This is followed by a review of various random network models in Section 1.3. Finally, in Section 1.4 we review the most important aspects of LASSO theory, which will provide us with the mathematical tools needed for proving many of the theorems in this thesis. All proofs are relegated to Section 1.5.

1.1 Random networks

The study of relationships between entities in data is taking a central role in modern science and society. Over the past few decades, this trend has been largely driven by the rapid deployment of information systems and measurement technology. This gave rise to increasing availability of data about interacting components. Often, these interactions are conveniently represented as graphs that exhibit entities as nodes and interactions as edges. Biological, social, financial, computer and transportation networks are all examples of this network data deluge. To understand the stochastic nature of these data, statistical analysis of networks has seen an increas-

ing research interest in both theory and applications. We refer to Kolaczyk (2009) for a book-length introduction, and Goldenberg et al. (2009) and Fienberg (2012) for reviews of various network models, as well as Kolaczyk (2017) for emerging challenges and issues.

This thesis concerns a series of new random graph models for describing networks. Informally, a network represents the relationships between entities. We call the entities *nodes* and represent a relationship between two nodes by an *edge* connecting the two. A network is *directed*, if we distinguish between a connection from node i to node j and a connection from j to i. It is called *undirected* of no such distinction is made. The *degree* of node i in an undirected network is the number of edges connected to i. In a directed network the *out-* (*in-*) *degree* of i is the number of edges originating from (ending at) i. See Section 1.1.1 for the formal definitions.

As a motivating example, Figure 1.1 depicts the friendship network between the 71 lawyers of a law firm (Lazega 2001): An edge from node i to node j exists if and only if lawyer i indicated in a survey that they socialized with lawyer j outside of work. While it may seem intuitive to treat friendships as undirected – and indeed we will choose to do so at a later point (c.f. Section 2.6) – not all friendships indicated in the survey were reciprocated. Therefore, Figure 1.1 shows the directed network as obtained from the survey.

Many network datasets come with covariate information. For the lawyer dataset, these covariates include a lawyer's status (partner or associate), their gender (man or woman), which of three offices they worked in, the number of years they had spent with the firm, their age, their practice (litigation or corporate) and the law school they attended (Harvard and Yale, UConn or other). The main interest here is to understand how these covariates affect link formation.

To develop a realistic model for the lawyer network, we start by summarizing its main features. These features are also commonly observed in many real-life networks.

1. *Degree heterogeneity*, which refers to the different tendency that nodes in a network have in forming connections. Degree heterogeneity is a hallmark of networks in observational studies, often manifested as nodes having – sometimes vastly – different degrees. To appreciate degree heterogeneity in the lawyer data, note that the maximum and minimum out-degrees are 25 and zero respectively, with 22 and zero for the in-degrees.
2. *Sparsity.* Real-life networks are typically sparse, in the sense that the observed number of connections does not scale proportionally to the total number of possible links. In Figure 1.1, for example, the average in- (and out-) degree is 8.1, whereas the maximum possible value is 70. The sparsity of a network is often reflected in it having nodes that are not well connected. In the lawyer data there

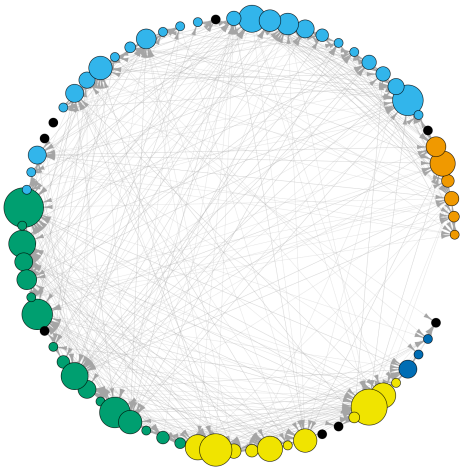


Figure 1.1: Lazega's lawyer friendship network. The area of the nodes is proportional to their in-degrees. For better visibility, all nodes with an in-degree of five or less are plotted with the same size. The 71 lawyers are colour-coded by their age group: The lawyers aged 20–29 are represented in orange at the top-right corner of the plot. Going anti-clockwise, they are followed by those aged 30–39 in light blue, those aged 40–49 in green, those aged 50–59 in yellow and finally those aged 60 or older in dark blue. The eight nodes in black correspond to lawyers with either zero in- or out-degree.

are eight nodes having either no incoming edges or no outgoing edges, as coloured in black in Figure 1.1.

3. *Covariates.* Covariates are often useful for explaining linking patterns. For our motivating data, whether two lawyers are connected by a friendship depends naturally on their covariates. For example, lawyers working in the same office or practice tend to befriend each other. Developing regression models that incorporate covariates is at the core of statistical modelling. In network science, however, we have only started to see statistical models very recently that involve covariates. See the rest of this chapter for references.

We will study in depth two novel random network models that can effectively capture all the above features. For undirected networks we study the *sparse β-model with covariates* (SβM-C) which was introduced in Stein & Leng (2020), in Chapter 2. For directed networks, we study the *parameter-Sparse Random Graph Model* (SRGM) introduced in Stein & Leng (2021), in Chapter 4.

Let us fix some ideas for the SβM-C and the SRGM. Let n denote the number of nodes. For the undirected SβM-C assume that we have observed data organized as $\{A_{ij}, Z_{ij}\}_{i,j=1,i\neq j}^{n}, A_{ij} = A_{ji}, Z_{ij} = Z_{ji}$, where $A = (A_{ij})_{i,j=1}^{n} \in \mathbb{R}^{n\times n}$ is the symmetric adjacency matrix with $A_{ij} = 1$ if node i and j are connected and $A_{ij} = 0$ otherwise, and $Z_{ij} \in \mathbb{R}^p$ are covariates associated with nodes i and j. Our model assumes that, given the Z_{ij}, links are independently made with the probability of a connection between node i and j being

$$P(A_{ij} = 1|Z_{ij}) = p_{ij} = \frac{\exp(\beta_i + \beta_j + \mu + Z_{ij}^T\gamma)}{1 + \exp(\beta_i + \beta_j + \mu + Z_{ij}^T\gamma)}, \quad i < j, \tag{1.1}$$

where $\beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n$ and β_i is associated with the ith node, taking the role of the heterogeneity parameter. Furthermore, $\gamma \in \mathbb{R}^p$ is the parameter for the covariates and $\mu \in \mathbb{R}$ is a parameter common to all the nodes. For identifiability, we assume $\min_i\{\beta_i\} = 0$. Central to our model is the idea that the vector β is sparse, although we do not assume that its support is known. This model is a generalization of the sparse β-model (SβM) proposed in Chen et al. (2020) that does not consider covariates (see Section 1.3.3). As such, we name this model SβM with covariates, or SβM-C. In the directed SRGM, we have a similar setup and directed links are independently made with the probability of observing a link *from* node i *to* node j specified as

$$P(A_{ij} = 1|Z_{ij}) = p_{ij} = \frac{\exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}{1 + \exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}, \quad i \neq j, \tag{1.2}$$

where now we may have $A_{ij} \neq A_{ji}, Z_{ij} \neq Z_{ji}$. For identifiability, we assume $\min_i\{\alpha_i\} = \min_j\{\beta_j\} = 0$. Importantly, we again assume that the two parameters $\alpha = (\alpha_1, \ldots, \alpha_n)^T$ and $\beta = (\beta_1, \ldots, \beta_n)^T$ are sparse. Without the sparsity constraints on α and β, the SRGM becomes the model in Yan et al. (2019) by absorbing μ into α_i and β_j as $\mu/2+\alpha_i$ and $\mu/2+\beta_j$, respectively. The SβM-C in (1.1) and the SRGM in (1.2) possess the following attractive properties.

1. Explicit handling of degree heterogeneity via node-specific parameters β_i and α_i. In particular, in SβM-C we interpret $\beta_i > 0$ as the (excess) popularity to form connections, relative to μ. In SRGM, $\alpha_i > 0$ is interpreted as the (excess) sociability of node i to form outgoing links and $\beta_i > 0$ as its (excess) popularity to attract connections, relative to μ. In this sense, α and β are local parameters.
2. Modelling of sparse networks in two ways. Firstly, they include μ, which can be interpreted as the global density parameter. Allowing $\mu \to -\infty$, as $n \to \infty$ (and thus $p_{ij} \to 0$ at least for some i, j), will generate sparse networks as discussed in Chapter 2. Secondly, by imposing a sparsity assumption on α and β, both models can have a much smaller number of heterogeneity parameters than the maximal possible n (respectively $2n$). This avoids over-parametrization, one of the major bottlenecks for modelling sparse networks, as discussed in Section 1.3. Intuitively, the fewer parameters to estimate, the sparser the network that SβM-C and SRGM can fit.
3. Handling covariates by including the term $\gamma^T Z_{ij}$. When a covariate encodes the similarity of a node attribute, a positive γ implies homophily, the tendency of nodes similar in attributes to connect.
4. When heterogeneity and covariate parameters are zero, the models in (1.1) and (1.2) can be interpreted as null models in which a link between any pair of nodes is formed with the same probability $\exp(\mu)/(1 + \exp(\mu))$. S$\beta$M-C and SRGM

build on this null model by including a regression component in the covariates and node-specific non-negative effects for the vertices.

Regarding point 2 above, in the extreme case when all the heterogeneity parameters are zero, we only have $1 + p$ parameters to estimate. In Chapter 3 we show that statistical inference can be conducted for this sub-model of SβM-C as long as the total number of links of a network is in the order $O(n^{\xi})$ for any $\xi \in (0, 2]$. That is, this sub-model can model networks that are almost arbitrarily sparse.

A major methodological contribution of SβM-C and SRGM is that they are general models that handle all three main features of a real-life network (degree heterogeneity, sparsity and covariates) in the undirected and directed setting. From a computational viewpoint, they are extremely attractive since the estimation of its parameters leverages the fast computation extensively developed for penalized likelihood, as we discuss in Chapters 2 and 4.

From a purely technical point of view, one of the main feats of SβM-C and SRGM is their ability to provide valid inference on γ in the presence of vanishing link probabilities. While allowing for $p_{ij} \to 0$ may seem like a minor modification to existing theory, this has far-reaching consequences. Very broadly speaking, proving the asymptotic normality of an estimator in many cases relies on a Taylor or Mean Value Theorem expansion of the loss function, followed by the inversion of the Hessian of the loss function. This fundamental idea can be found (with some variations) in a myriad of applications, ranging traditional M-estimation (van der Vaart 1998), to LASSO theory (van de Geer et al. 2014), to inference in networks (Yan et al. 2019). For this strategy to succeed, it is routinely assumed that the minimum eigenvalue of the Hessian is bounded away from zero, uniformly in n. In the case random networks, however, the Hessian depends on the link probabilities p_{ij} and if we allow $p_{ij} \to 0$ for many i and j, such a uniform lower bound assumption becomes invalid (see Sections 2.4 and 4.2.3 for details). However, allowing $p_{ij} \to 0$ is a necessary condition for modelling sparse networks, since otherwise each degree will scale in the order of n. This is one of the reasons why inference results in sparse network regimes are scarce. By choosing our rates carefully, we are able to deal with this difficulty and open up our models to sparse settings. The ability to do inference with an asymptotically non-invertible Hessian and vanishing link probabilities is a significant improvement over many existing methods and a prerequisite for dealing with sparse networks. In this vein, our results substantially generalize those in the literature (Ravikumar et al. 2010, van de Geer & Bühlmann 2011) in a different context (network modelling versus regression modelling).

We now briefly recall some of the most common definitions and notation we will need.

1.1.1 Definitions

Definition 1.1 (Undirected graph)**.** An *undirected graph* $G = (V, E)$ on $n \in \mathbb{N}$ nodes is a tuple consisting of a *node set* V with cardinality $|V| = n$ and an *edge set* $E \subseteq \{\{i, j\} : i, j \in V, i \neq j\}$. Hence, an *edge* $e \in E$ between nodes i and j of G is a set $\{i, j\}, i, j \in V$.

Definition 1.2 (Directed graph)**.** A *directed graph* $G = (V, E)$ on $n \in \mathbb{N}$ nodes is a tuple consisting of a *node set* V with cardinality $|V| = n$ and an *edge set* $E \subseteq V \times V \backslash \{(i, i) : i \in V\}$. Hence, an *edge* $e = (i, j) \in E$ from node i to j of G is a tuple and $(i, j) \neq (j, i)$.

Except where stated otherwise, we implicitly assume the number of nodes in the graph is n and for convenience we take $V = [n] \coloneqq \{1, \ldots, n\}$. In particular, by definition of the edge sets E for undirected and directed graphs we are forbidding the existence of self-loops, that is, a node connecting to itself, as well as multi-edges, that is, the existence of several edges between the same pair of nodes. We use *node* and *vertex* interchangeably, as well as *edge*, *link* or *connection*.

Denote by $\mathcal{G}_n$ (respectively $\overrightarrow{\mathcal{G}}_n$) the set of all undirected (respectively directed) graphs on n nodes. Since for any $n \in \mathbb{N}$, $\mathcal{G}_n$ and $\overrightarrow{\mathcal{G}}_n$ are finite sets, we will always consider their power sets as canonical σ-algebra on them. In particular, the use of the power set as σ-algebra is implicit in the following definition.

Definition 1.3 (Random network)**.** An *undirected random network* on n nodes is a random variable taking values in $\mathcal{G}_n$. A *directed random network* on n nodes is a random variable taking values in $\overrightarrow{\mathcal{G}}_n$.

When we refer to a *network* we will always refer to a random network in one of the two senses above. When simply speaking of a network without specifying whether it is directed or undirected, it will either be clear from the context what is meant or the distinction will not matter for the argument being made.

Definition 1.4 (Adjacency matrix)**.** Given a graph $G = (V, E)$ (undirected or directed) we can identify G with its binary adjacency matrix $A \in \mathbb{R}^{n \times n}$, where we let $A_{i,j} = 1$, if $\{i, j\} \in E$ if G is undirected and $A_{ij} = 1$ if $(i, j) \in E$ if G is directed and $A_{i,j} = 0$ otherwise. By definition, $A_{i,i} = 0$ for all i and if G is undirected, the matrix A is symmetric.

We will identify a network with its adjacency matrix without further mention.

Definition 1.5 (Degree, total degree, degree distribution)**.** In an undirected network A the *degree* of node i is denoted $d_i = \sum_{j=1}^{n} A_{i,j}$ and the number of edges, or

total degree of A, by $d_+ = \sum_{i=1}^n d_i/2 = |E|$. The *degree distribution* of A is the distribution of the values d_i and is fully characterized by the vector $d = (d_1, \dots, d_n)^T$. For a directed network A we denote by $b_i = \sum_{j=1, j\neq i}^n A_{ij}$ the *out-degree* of vertex i and by $d_i = \sum_{j=1, j\neq i}^n A_{ji}$ the *in-degree* of vertex i. The vector $d = (d_1, \dots, d_n)^T$ is called *in-degree sequence* and the vector $b = (b_1, \dots, b_n)^T$ is called *out-degree sequence*. Since an out-edge from i to j is also an in-edge of j coming from i, it is immediate that $d_+ \coloneqq \sum_{i=1}^n d_i = \sum_{i=1} b_i =: b_+$.

Definition 1.6 (Edge density)**.** The *edge density* of a random network is the proportion of all possible edges that are observed. That is, for a network G with total degree d_+, the edge density of G is given by d_+/N, where $N = \binom{n}{2}$ in case of undirected networks and $N = n(n-1)$ for directed networks. The edge density of a random network is itself a random variable.

We are interested in the regime where the number of nodes n becomes large. We define the notion of *sparse* and *dense* networks with respect to the limit $n \to \infty$. Before defining sparse networks, recall the well-known Landau notation.

Definition 1.7 ((Stochastic) Landau notation)**.** Let $X_1, X_2, \dots$ be real-valued random variables. Let $x_1, x_2, \dots$ be real numbers. We write

1. $x_n = o(1)$, if $x_n \to 0$, as $n \to \infty$.
2. $x_n = O(1)$, if $\sup_{n\in\mathbb{N}} |x_n| < \infty$.
3. $X_n = o_P(1)$, if $X_n \xrightarrow{P} 0$, as $n \to \infty$.
4. $X_n = O_P(1)$, if $(X_n)_{n\in\mathbb{N}}$ is bounded in probability, that is, if for any $\epsilon > 0$ there exists a compact set $K_\epsilon \subseteq \mathbb{R}$, such that $P(X_n \in K_\epsilon) \geq 1 - \epsilon$, for all $n \in \mathbb{N}$.

More generally, let $a_1, a_2, \dots$ be strictly positive real numbers. We write $x_n = o(a_n)$, if $x_n/a_n = o(1)$ and $X_n = o_P(a_n)$, if $X_n/a_n = o_P(1)$. Analogously for the big-O notation.

Definition 1.8 (Sparse networks)**.** A sequence of networks $(G_n)_{n\in\mathbb{N}}$ with total degree $d_+ = d_+(n)$ is called *sparse*, if there is a $\xi \in [0, 2)$ such that $\mathbb{E}[d_+] = o(n^\xi)$.

Remark. In words, we call a networks sparse if the expected total degree scales "properly" sub-quadratically with respect to n. By "properly" we mean the following: For example, consider a network in which $\mathbb{E}[d_+]$ scales as $n^2 \cdot \log(n)^{-C}$ for some $C > 0$. Even though technically $\mathbb{E}[d_+]$ scales sub-quadratically in n in the sense that $\mathbb{E}[d_+] = o(n^2)$, we would still consider such a network dense, because $\mathbb{E}[d_+]$ "essentiality" still behaves like n^2. For a network to be considered sparse, the order in n has to be strictly less than two.

The above definition is asymptotic in n. Thus, it does not apply to any single realization of a network and when we speak of *sparse/ dense networks*, we implicitly

mean their behaviour in the limit $n \to \infty$. Also, any network model with link probabilities $p_{ij} > \epsilon > 0$, where ϵ is a constant independent of n, will inevitably be a dense model. Indeed, in that case $\mathbb{E}[d_+] \geq \epsilon \cdot N$, where $N = \binom{n}{2}$ for undirected and $N = n(n-1)$ for directed networks, which scales as n^2.

We are primarily concerned with inference about model parameters. Aside from being sparse, the "shape" of the networks, in particular the global connectivity structure of it, is of secondary interest. Nonetheless, we will touch upon these things in Sections 1.2 and 3.3 and it will be useful to have the following definitions at our disposal when we do. We only focus on the undirected case here.

Definition 1.9 (Connected nodes and connected components)**.** In an undirected graph $G = (V, E)$ we say that two nodes $i, j \in V$ are *connected* if there exists a *path* from i to j in E. That is, there are edges $e_k = \{v_{k,1}, v_{k,2}\} \in E, k = 1, \ldots, K$ for some $K \in \mathbb{N}$, with $v_{k,2} = v_{k+1,1}, k = 1, \ldots, K-1$ and $i = v_{1,1}, j = v_{K,2}$. The *connected component* $\mathcal{C}(i)$ of i consists of i and all nodes connected to i:

$$\mathcal{C}(i) \coloneqq \{i\} \cup \{j \in V : j \text{ is connected to } i\}.$$

We call $G = (V, E)$ *connected*, if for any pair $i, j \in V$ there exists a path in E from i to j. We call G *disconnected* otherwise.

A commonly observed phenomenon in real world networks is the existence of a unique *giant component*; that is, a unique largest connected component, the size of which grows linearly in n. For the Erdős-Rényi model (defined in Section 1.2) the behaviour of the giant component is well understood, see Section 1.2 and the references therein. We may leverage the existing results for the Erdős-Rényi model to show that – under weak general conditions – in any undirected random network model with independent edges, a giant component will exist. This is formalized in the following Lemma, which is proved in Section 1.5.

Lemma 1.10. *Let M be any undirected random network model on n nodes with independent edges for which the probability of observing an edge between nodes i and j is denoted p_{ij}. Denote by $\mathcal{C}^M$ the largest connected component of M. Assume that there exists some $\lambda > 1$ such that, for all i, j, almost surely,*

$$p_{ij} \geq \frac{\lambda}{n}.$$

Let ζ_λ be the survival probability of a Poisson branching process with mean offspring λ. Then, for every $\epsilon > 0$, as $n \to \infty$,

$$P\left(|\mathcal{C}^M| > (\zeta_\lambda - \epsilon)n\right) \to 1.$$

This claim remains true if links are only made independently, conditionally on having observed covariates Z_{ij} as long as the resulting probabilities are almost surely bounded below by the same rate. That is, as long as $P(A_{ij} = 1|Z_{ij}) \geq \lambda/n$, almost surely.

Put differently: Lemma 1.10 tells us that in any undirected random network with (conditionally) independent edges and link probabilities not dropping to zero too quickly, we will observe a giant component with size scaling linearly in n with high probability.

1.1.2 Mathematical notation

For any $n \in \mathbb{N}$, let $[n] \coloneqq \{1, \ldots, n\}$. For any set S, denote its cardinality by $|S|$.

Unless stated otherwise all vectors are column vectors and for a vector $v \in \mathbb{R}^n$, we use v^T to denote its transpose and denote its entries as $v = (v_1, \ldots, v_n)^T$. $S(v)$ denotes its support, that is, $S(v) = \{i : v_i \neq 0\}$. Let $\|\,.\,\|_1, \|\,.\,\|_2 . \|\,.\,\|_\infty$ denote the vector ℓ_1-, ℓ_2- and ℓ_∞-norm respectively and $\|\,.\,\|_0$ denotes the ℓ_0-"norm", $\|v\|_0 = |S(v)|$. For any subset $S \subset [n]$, we denote by v_S the vector v with components not belonging to S set to zero. That is, $v_{S,i} = v_i \mathbb{1}(i \in S)$, where $\mathbb{1}$ is the indicator function. For convenience of notation, when dealing with a vector $v \in \mathbb{R}^{\binom{n}{2}}$, we will number its elements as $v = (v_{ij})_{i<j}$. Similarly, when dealing with a vector $v \in \mathbb{R}^N, N = n(n-1)$, we will number its elements as $v = (v_{ij})_{i \neq j}$.

For a matrix $A \in \mathbb{R}^{d \times d}$ and sets $S, T \subseteq [d]$, let $A_{S,T} \in R^{|S| \times |T|}$ be the sub-matrix of A obtained by only taking the rows belonging to S and columns belonging to T. Define $A_{-,S} \coloneqq A_{[d],S} \in \mathbb{R}^{d \times |S|}$ and $A_{S,-} \coloneqq A_{S,[d]} \in \mathbb{R}^{|S| \times d}$. For any square matrix A, maxeval(A) and mineval(A) denote its maximum and minimum eigenvalue.

We use C for some generic, strictly positive constant that may change between displays.

By $a_n \sim b_n$ we mean $0 < \liminf_{n\to\infty} a_n/b_n \leq \limsup_{n\to\infty} a_n/b_n < \infty$ for two sequences of positive numbers a_n and b_n. For any $a, b \in \mathbb{R}$ we use the notation $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Write $\mathbb{R}_+ = [0, \infty)$.

1.2 Data-selective inference

In this section we seek to demonstrate why it is essential to allow for sparsity in stochastic network models. This section is meant to serve as a motivation for the study of SβM-C and SRGM.

1.2.1 Theory of data-selective inference

Statistical models are motivated by data in real life. For many statistical network models a common practice has emerged in the way scientists apply them to network data: Before the analysis they will often discard the (usually non-negligible) fraction of nodes that are either disconnected from the largest connected component or do not have very many connections. This may be done due to convenience or, more significantly, due to the nature of the model and its associated algorithms not working in case of disconnected or very sparse networks. Often it is argued that the remaining well-connected nodes and edges are the only nodes and edges that matter.

Take the lawyer network from Section 1.1 as an example: The eight nodes with zero in- or out-degrees (c.f. Figure 1.1) were excluded from the analysis in Yan et al. (2019). More examples can be found in Section 1.2.3. The resulting *data-selective inference* – the exercise of fitting a model to a sub-network excluding nodes based on their links – is a special case of biased sampling, because nodes in a giant component or well-connected nodes are systematically favoured over other nodes. This immediately raises the following fundamental question:

Does data-selective inference provide valid inference?

An argument to avoid the question above would be to simply assume that the intended statistical model only works for the nodes in a giant component or those nodes that are well connected. While this argument is acceptable for mathematical convenience, it is not logically coherent or correct from a statistical or practical point of view. The selection of nodes is based entirely on the links – the response variable in a network model – and thus is non-random. Intuitively, biased sampling in such data-selective inference may produce artificial signal that does not exist at all or mitigate existing signals or both, leading to problematic or even completely wrong findings. The fact that a non-negligible fraction of non-random nodes are removed before the analysis suggests that systematic bias will occur as a result.

The practice of ignoring selected nodes for modelling a network appears to originate from physics and computer science communities where the intention was to find meaningful clusters of nodes and hence is not based on statistical models (Girvan & Newman 2002, Newman 2006). Later, statisticians injected rigour into this line of research, notably by introducing likelihood-based estimators for statistical network models (Bickel & Chen 2009). On one hand, statistical modelling is extremely attractive because it provides a proper probabilistic framework for statistical inference and allows easy generalization of a model to more complex situations. On the other hand issues inevitably arise, such as sampling and asymptotics, including

consistency and limiting distributions as the size of a network grows. In particular, it is no longer appropriate for a statistical framework to ignore the non-random sampling issue in data-selective inference that removes nodes based on their links.

We now quantify the bias caused by data-selective inference. In what follows, we assume that an observed network is the realization of some statistical network model $f \in \mathcal{F}$, with $\mathcal{F}$ a family of candidate models. Crucially, we will assume that f would have produced the whole network, *including* any isolated vertices or small components. Given a realized network from the model, we want to quantify the bias of the estimator of the unknown parameter(s) in f, if we only use those nodes in the giant component.

Motivated by the lawyer data and several other widely used datasets in the literature as discussed in Section 1.2.3, we will specify the parameter(s) in f such that a fixed proportion of nodes are not in the giant component. We will derive theoretically the bias of data-selective inference in the Erdős-Rényi model (Erdős & Rényi 1959, 1960) and use simulation to study the bias in estimating the parameters in a simple Stochastic Block Model (Holland et al. 1983). The Erdős-Rényi model is interesting because the insights we gain into this model provide the foundation for the study of more general graphs. On the other hand, the Stochastic Block Model is one of the most popular network models that is widely applied, studied and extended in the literature, see for example Abbe (2018) for an overview of recent developments and the references in Section 1.2.3.

The Erdős-Rényi model. In the Erdős-Rényi model undirected edges are independently formed with the same probability p. A sufficient condition for a realized network from this model to have a giant component with probability tending to one is to take $p = p(n) = \lambda/n$ for some fixed $\lambda > 1$ (van der Hofstad 2016, Chapter 4). We will denote this family of models by $\mathrm{ER}(\lambda/n)$ and we are interested in estimating p, given a network generated from this model. The expected degree of each node is $\lambda \cdot (n-1)/n$ and consequently $\mathrm{ER}(\lambda/n)$ produces sparse networks with the expected total degree scaling linearly in n.

Denote by η_λ the unique solution with $\eta_\lambda < 1$ to the equation

$$\eta_\lambda = e^{\lambda \cdot (\eta_\lambda - 1)}, \tag{1.3}$$

which exists if and only if $\lambda > 1$. We can interpret $\zeta_\lambda = 1 - \eta_\lambda$ as the survival probability of $\mathcal{P}(\lambda)$, the Poisson branching process with mean offspring λ, whose behaviour is closely linked to the connectivity behaviour of $\mathrm{ER}(\lambda/n)$ (van der Hofstad 2016, Chapters 3 and 4). In particular, it is known that for $\lambda > 1$, $\mathrm{ER}(\lambda/n)$ will produce a unique giant component with size tightly concentrating around $\zeta_\lambda \cdot n$

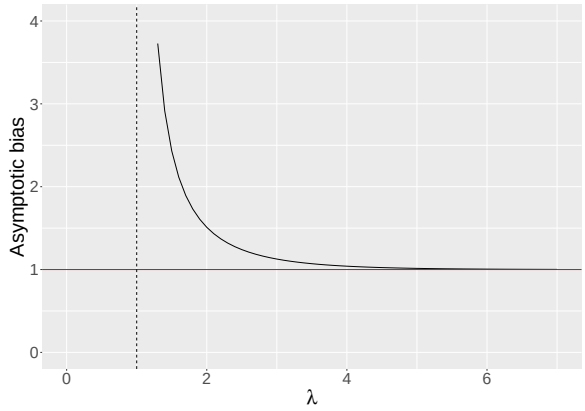


Figure 1.2: Asymptotic bias $(1+\eta_\lambda)/(1-\eta_\lambda)$ of $\hat{p}_{\max}$ as a function of λ. For better visibility we only display values of λ ranging from 1.3 to 7 since the bias diverges to $+\infty$ when λ approaches 1.

(van der Hofstad 2016, Theorem 4.8).

Consider the estimation of p by based solely on the sub-graph $G_{\max}$ of $\mathrm{ER}(\lambda/n)$ induced by the giant component. That is, $G_{\max} = (\mathcal{C}_{\max}, E(\mathcal{C}_{\max}))$, where $\mathcal{C}_{\max}$ denotes the giant component and $E(\mathcal{C}_{\max})$ contains only those edges between the nodes in $\mathcal{C}_{\max}$. Denote by $\hat{p}_{\max}$ the maximum likelihood estimator of p based on $G_{\max}$:

$$\hat{p}_{\max} = \frac{|E(\mathcal{C}_{\max})|}{\binom{|\mathcal{C}_{\max}|}{2}}. \tag{1.4}$$

We have the following result, which is proved in Section 1.5.

Proposition 1.11. *Fix any $\lambda > 1$ and consider the models $\mathrm{ER}(\lambda/n)$. Let $G_{\max} = (\mathcal{C}_{\max}, E(\mathcal{C}_{\max}))$ be the sub-graph of $\mathrm{ER}(\lambda/n)$ induced by the giant component. Let $p = p(n) = \lambda/n$, $\hat{p}_{\max}$ as in (1.4) and η_λ as in (1.3). Then,*

$$\frac{\hat{p}_{\max}}{p} \xrightarrow{P} \frac{1+\eta_\lambda}{1-\eta_\lambda}.$$

A few remarks are in order. Firstly, $(1+\eta_\lambda)/(1-\eta_\lambda) > 1$ for any fixed $\lambda > 1$. Thus, the incurred bias, that is, the asymptotic factor by which we are overestimating p, will not disappear as n grows large. Figure 1.2 shows how the asymptotic bias $(1+\eta_\lambda)/(1-\eta_\lambda)$ deviates from one as a function of λ. Secondly, we see that the bias increases when λ decreases. Indeed, the larger λ, the more nodes in the giant component and the smaller the probability η_λ and thus the smaller the bias. On the other hand, as $\lambda \to 1$, it is easy to see that $\eta_\lambda \to 1$, making the bias in Proposition 1.11 approach $+\infty$. For the limit case $\lambda = 1$, $\mathcal{C}_{\max}$ will have size of order $n^{2/3}$ (van der Hofstad 2016, Chapter 5), which in light of the proposition means that we must abandon all hope of recovering p if we only focus on the giant component.

Proposition 1.11 follows readily from results in the random network literature. However, to the best of our knowledge, in the present form it has been stated for the first time in Stein & Leng (2021). In particular, the results we draw upon for

its proof are mostly rooted in probability theory and appear to not have been used before to explicitly quantify the biases incurred by statistical procedures.

The Stochastic Block Model. This model, introduced in Holland et al. (1983), postulates that nodes in a network can be grouped into communities where the probability of any pair of nodes making connections depends only on their community membership. We focus on what is called the *symmetric Stochastic Block Model* with two communities, for which the probability matrix of making connections is

$$P = \frac{1}{n}\begin{pmatrix} a & b \\ b & a \end{pmatrix},$$

where $a, b > 0$ are constants. For this model, a pair of nodes link with probability a/n within the same community and with probability b/n between communities. The scaling $1/n$ ensures that a resulting network from this model will have a giant component with smaller-than-one proportion of the nodes with high probability. See Figure 1.3a for the proportion of the nodes in the giant component produced under this parametrization. The symmetric Stochastic Block Model is widely studied and relatively well understood as reviewed by Abbe (2018). Because of the sparsity of any resulting network, many clustering methods for community detection including spectral methods based on the adjacency matrix or the graph Laplacian, as well as their semi-definite relaxations, do not work well under this parametrization. Indeed, Zhang & Zhou (2016) showed that under our scaling no consistent algorithm exists that achieves vanishingly small misclassification. In view of this, we take an *oracle* approach by assuming that the community membership of each node is known *a priori*, and focus on what happens if we estimate P when nodes not in the giant component are removed as in data-selective inference.

In our simulation, we fix the number of nodes to be $n = 10{,}000$ and set the size of each community as $n/2$. We consider a fine grid of values (a, b) by taking their values from 0.05 to 8.05 in steps of 0.05, resulting in 25,921 distinct combinations. For each such pair (a, b) we sample a network from the symmetric Stochastic Block Model and calculate the maximum likelihood estimate for P using only the nodes in the giant component. We repeat this process $M = 1{,}000$ times for every pair (a, b).

Denote the estimator as $\hat{P}$. We measure the bias of $\hat{P}$ as the ratio $\rho = \|\hat{P}\|_2/\|P\|_2$, where $\|P\|_2 = (a+b)/n$ is the spectral norm of the 2×2 matrix. We have purposefully chosen a large n and M so that the resulting averages of the estimates will be close to their true limit. Figure 1.3a shows the average proportion of nodes in the giant component for each pair (a, b) and Figure 1.3b the average value of ρ. When the proportion in Figure 1.3a is one, the giant component contains all the nodes. The closer the average value of ρ is to one, the less biased the estimates are.

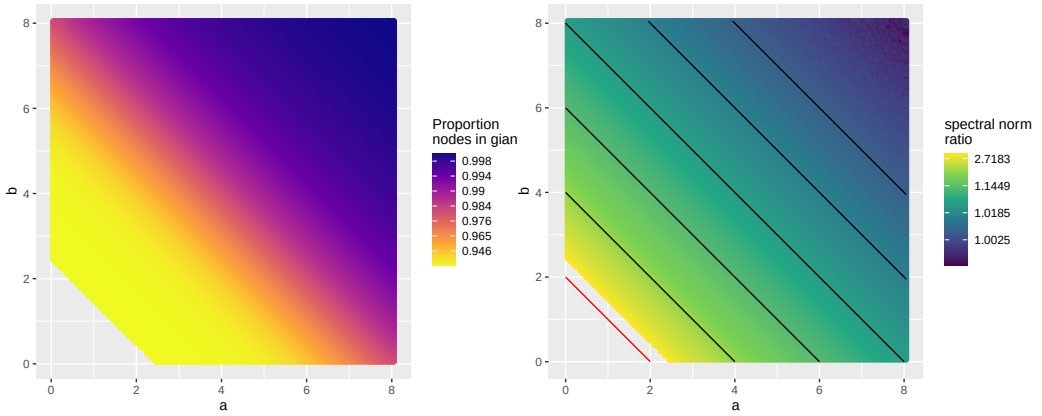


Figure 1.3: For better visibility we have truncated by only including points for which $a + b \geq 2.5$. Also note the use of an exponential colour scaling in both plots for better visibility. (a): Mean proportion of nodes in the giant component for each pair (a, b), averaged over $M = 1{,}000$ repetitions. On the border $a + b = 2.5$ the proportion of nodes in the giant is 37%. (b): Mean spectral ratio ρ when estimates are based on the giant component only, averaged over $M = 1{,}000$ repetitions. The red line corresponds to $a + b = 2$ and an average bias factor of 60.57. The black lines correspond to (from bottom left to top right) $a + b = 4, 6, 8, 10, 12$ with bias factors $\rho = 1.51, 1.13, 1.04, 1.014, 1.005$ respectively.

The simulations show that the incurred bias and the size of the giant component behave similarly when $a + b$ is a constant. We highlight the bias of the parameter estimates more closely. When $a + b = 2.5$ the giant component on average contains 37% of all nodes and we overestimate $\|P\|_2$ by a factor of 4.4. This may not exactly come as a surprise: If we discard a large proportion of nodes, it can be expected that the resulting estimates are inaccurate. A more interesting and critical behaviour is observed as we make our way from the bottom left to the top right corner of the plots in Figure 1.3.

For $a+b = 4$ (bottom black line in Figure 1.3b), the giant component on average contains around 80% of all nodes, with an average bias $\rho = 1.5$. For $a+b = 6$ (second black line in Figure 1.3b), the giant component contains on average 94% of all nodes, while the average estimated ρ is 1.13, still significantly larger than one. Even for $a + b = 8$ (middle black line in Figure 1.3b), when the giant component contains on average 98% of all nodes, we still overestimate $\|P\|_2$ by a factor of 1.04. Only once $a + b \geq 10.70$, where the giant component contains 99.5% of all nodes, is the incurred bias smaller than 1%.

The above results illustrate that even if the statistician has perfect knowledge of the underlying communities, and even if only a seemingly insignificant fraction of, say, 1% of the nodes is removed, parameter estimation based solely on the giant component will be biased regardless of the network size. This casts severe doubt on the suitability of the Stochastic Block Model and its variants for fitting many popular datasets if only the nodes in giant components are retained. Having biased estimators will have consequences in all aspects of any downstream statistical

inference including consistency, model selection, hypothesis testing, and so on; see Section 1.2.3 for some results that may be affected.

1.2.2 Model-selective inference for the lawyer data

We return to our motivating example by comparing our estimates of the regression coefficients with those in Yan et al. (2019). For the seven covariates in this dataset, we followed Yan et al. (2019) in using the absolute differences of the continuous variables and the indicators whether the categorical variables are equal as our covariates.

Practically, to fit their model, the authors in Yan et al. (2019) had to remove the eight nodes in black in Figure 1.1 that have zero in-degree or out-degree. Otherwise their maximum likelihood estimates (MLEs) would be $-\infty$ for α_i if node i has no outgoing connections or for β_i if the node has no incoming connections. We remark that this means that their estimates can be biased, as discussed. Another interesting aspect of the model in Yan et al. (2019) lies in the inference for the fixed-dimensional parameter γ. Because the rate of convergence of its MLE is slowed down by the MLE of the growing-dimensional heterogeneity parameters α and β, the estimator of γ requires a bias correction to be asymptotically normal. In contrast, by making a sparsity assumption on α and β in SRGM, we estimate the parameters via penalized likelihood and prove that inference for the estimated γ can be read off after fitting the model via a model selection procedure, as seen in Theorem 4.5 and thus is straightforward. This result is remarkable, because in high-dimensional statistics, it is often found that a different debiasing procedure must be conducted for the valid inference for the estimated parameters due to the bias incurred by regularization (Zhang & Zhang 2014).

To summarize: For SRGM, rather than throwing away data at the beginning of the modelling process as in data-selective inference, we fit a model to all the data by judiciously assigning heterogeneity parameters to the nodes using a model selection criterion. For this reason, we refer to our modelling framework as *model-selective inference*.

When the Bayesian information criterion (BIC) is used to choose the tuning parameter in the penalized likelihood estimation, SRGM gives a model with 7 non-zero α_i's and 7 non-zero β_i's. Four pairs of these non-zeros come from the same nodes. In Table 1.1 we present the estimated γ and their standard errors when our model and the model in Yan et al. (2019) are fitted to the lawyer dataset. Although generally similar, we see a few differences. Firstly, we see that the standard errors of our estimates are smaller than those in Yan et al. (2019), reflecting that our

estimates are based on a larger sample size (a network with 71 nodes compared to one with 63 nodes in the latter paper) with fewer parameters (22 versus 132). Secondly, the effect of age difference is not significant in our model whereas it is in the model in Yan et al. (2019). To explore the age effect visually, we colour-coded the lawyers by their age group in Figure 1.1. Plenty of connections were made between age groups and "across the circle", i.e. between lawyers with a large difference in age, suggesting that age may not have played an important role. Indeed, a third (33.9%) of all friendships were formed between lawyers with an age difference of ten or more years. Thirdly, we estimate the effect of attending the same law school as positive, while Yan et al.'s model states the opposite. The former conforms better to our intuition about social networks.

	SRGM		Yan et al. (2019)	
Covariate	Estimate	SE	Estimate	SE
Same status	1.52	0.10	1.76	0.16
Same gender	0.44	0.09	0.96	0.14
Same office	2.02	0.10	3.23	0.18
Same practice	0.58	0.09	1.11	0.12
Same law school	0.29	0.10	−0.48	0.12
Difference in years with firm	−0.01	0.006	−0.064	0.014
Difference in age	0.003	0.006	−0.027	0.011

Table 1.1: Estimated regression coefficients and their standard errors (SE) for Lazega's lawyer friendship network.

1.2.3 Data-selective inference in the literature

Many papers in the literature chose to ignore the modelling of those nodes in smaller components, or isolated nodes, or nodes with small degrees. Many real-life networks have such nodes. Sometimes, ignoring these nodes is due to restrictions on a model from the outset, for example, the degree-corrected Stochastic Block Model (Karrer & Newman 2011) and the β-model (Chatterjee et al. 2011) cannot handle nodes with zero degree. As we have argued in Section 1.2.1, this practice is highly problematic. Here we list several popular datasets in which data-selective inference is routinely performed. The first is the lawyer dataset previously discussed, for which Yan et al. (2019) chose to work with 63 out of 71 nodes (i.e. 11% of the nodes are removed). Below are two of statisticians' favourite datasets:

- Political blog data. This is a dataset recorded during the 2004 U.S. Presidential Election in the form of a directed network of hyperlinks between 1,494 political blogs (Adamic & Glance 2005). Depending on their political views, these blogs can be liberal or conservative. Often converted to an undirected graph for analysis, this dataset has become a testbed for many network models especially the Stochastic Block Model and its generalizations. In practice, most papers chose to focus on 1,222 blogs that appear in a giant component or 1,224 nodes which have at least

one connection. Either way, this amounts to removing about 18% of the nodes in this network. See Amini et al. (2013), Olhede & Wolfe (2014), Jin (2015), Cai & Li (2015), Caron & Fox (2017), Chen & Lei (2018), Huang & Feng (2018), Ma, Ma & Yuan (2020), among many others.

- Statistician citation network. This dataset, collected by Ji & Jin (2016), contains rich citation information about all papers published between 2003 and 2012 in four statistics journals. The original dataset has 3,607 authors or nodes based on which various networks have been constructed, but almost all attempts to use this data have chosen to examine subnetworks with fewer than 3,607 nodes. Ji & Jin (2016) applied various community detection methods to three networks constructed from this dataset. The first one is a co-authorship network with 236 nodes (7% of all nodes), in which a link is formed between two authors if they wrote at least 2 papers together. See also Jin et al. (2021). The second one is another co-authorship network with 2,236 nodes (63% of all nodes), in which a link is formed between two authors if they wrote at least 1 paper together. The third one is a directed citation network with 2,654 authors (74% of all nodes). See also Zhang et al. (2021). Other attempts to use this dataset include Li et al. (2020) in which a network with 706 authors (20% of all nodes) was formed by repeatedly deleting nodes with less than 15 mutual citations and their corresponding edges. Jin et al. (2021) examined a citee network with 1,790 (50% of all nodes) constructed by tying an edge between two authors if they have been cited at least once by the same author other than themselves.

In addition to the datasets above, there is a growing body of works opting for data-selective inference by removing nodes before their analysis. Among many others, see Chen et al. (2018) and Ma, Ma & Yuan (2020) for the Simmons College and Caltech data, two datasets on friendship networks in universities, Sengupta & Chen (2018) for the British MPs network (where 329 out of 360 MPs belonging to the giant component were analysed), and Ma, Su & Zhang (2020) for Pokec social network for which only those nodes with no fewer than 10 links were retained for analysis. We emphasize that a notable feature of the analyses in these papers is that non-negligible portions of the nodes are excluded.

We now illustrate the fallacy of data-selective inference with the Stochastic Block Model when it is applied to detecting communities in the political blogs network, to highlight a wider problem in statistical modelling of networks. If one assumes that this model generates all the 1,494 nodes, the mere existence of more than 200 nodes not in the giant component suggests imposing a scaling of $1/n$ on the connectivity probability matrix of the data generating process. This is similarly done in Section 1.2.1 to ensure that the resulting giant component contains a positive fraction of

the vertices. Under this regime, however, there is no way to separate all the vertices and thus no algorithm can provide consistent community detection or parameter estimation. If instead one focuses on the giant component and assumes consistent community estimation for the nodes in the giant component, the connectivity probability matrix will be estimated with bias, as we have illustrated. The estimation bias of course is just the tip of a larger problem in biased sampling. By focusing on a non-random sample, we have no idea whether an intended model truly reflects the data generating process, or rather is merely an artefact of biased sampling. Equally importantly, the bias problem incurred via this data-selective inference will have knock-on effects of all aspects of any downstream analysis, including goodness-of-fit measures of a model, hypothesis testing and model selection; see, for example, Bickel & Sarkar (2016), Lei (2016), Wang & Bickel (2017), and Hu et al. (2020), for additional use of data-selective inference for data analysis.

Thus, there is a fundamental choice that we statisticians need to make. If we assume the Stochastic Block Model generates the whole network including all the nodes in the political blog, consistent community detection and parameter estimation will be impossible. On the other hand, if we make the unrealistic assumption that the model generates only a subnetwork consisting of those nodes in the giant component, we face the problem of data-selective inference. Although this fallacy seems ubiquitous in statistical applications of many network models, to the best of our knowledge Stein & Leng (2021) was the first attempt at a systematic study of the effect of omitting nodes due to their degrees on model fitting.

1.3 Different random network models

In this section we give a review of various random network models and show how SβM-C and SRGM can be seen as natural extensions from existing models.

1.3.1 The Erdős-Rényi model

The simplest random network model is the Erdős-Rényi model (Erdős & Rényi 1959, 1960, Gilbert 1959), which we already encountered in Section 1.2.1. It is arguably the most studied random graph model in probability theory. In it, the presence of each edge is modelled as an independent Bernoulli random variable with success probability $p \in [0, 1]$. We denote the law of the Erdős-Rényi model on n nodes with link probability p by $\mathrm{ER}(n, p)$. For most theoretical investigations $p = p(n)$ is assumed to be a function in n and the behaviour of $\mathrm{ER}(n, p)$ is studied as $n \to \infty$. Despite its apparent simplicity, the Erdős-Rényi model has a very rich and complex theory and even today is still subject of active research. In particular, we have a

theory for it in the dense as well as sparse graph regimes. We refer to van der Hofstad (2016), especially Chapters 4 and 5, for a thorough discussion of the most important properties. To be accurate, it should be pointed out that Erdős and Rényi did not actually study ER(n, p), but the closely related model in which the number of edges m is fixed first and then a graph is drawn uniformly at random from all graphs on n nodes with m edges. It can be shown that the two models are equivalent in a certain sense, see the aforementioned references.

Perhaps unsurprisingly, the Erdős-Rényi model is not very good at modelling real-world networks. One of its most glaring shortcomings is that the degree distributions it produces will look almost like realizations from a Poisson random variable with parameter $\lambda = p \cdot n$, see van der Hofstad (2016), Theorem 5.12, for a precise mathematical statement[1]. This is problematic, because Poisson distributions have very thin tails. As we have seen in the previous section, one of the distinctive features of real-world networks, is that they exhibit *degree heterogeneity*. That is, their empirical degree distribution usually has heavy tails and frequently contains *hub* nodes with many connections, which cannot be modelled by the Erdős-Rényi model.

Example. Let us consider an example to illustrate this point. Consider once more Lazega's friendship network (Lazega 2001), that we already encountered in the previous sections. Since each lawyer answered the survey questions individually, the original network is directed in nature and in some cases, an outgoing friendship arrow is not reciprocated by the recipient. For our purposes, we only keep those connections in which both lawyers indicated being friends with one another. This leaves us with an undirected network between $n = 71$ nodes with an edge density of 0.07. The minimum degree is zero and the maximum degree is 16. On the other hand, if we sample from an Erdős-Rényi model with the same edge density, $p = 0.07$, the degree distribution we obtain is much more concentrated with much thinner tails, as illustrated in Figure 1.4. We plotted the number of lawyers with d friends for $d = 0, \ldots, 16$ (yellow line). For comparison, we drew 10,000 realizations from the Erdős-Rényi model with parameters $n = 71, p = 0.07$ and recorded the average number of nodes with degree $d, d = 0, \ldots, 17$ (red line; 17 was the largest observed degree). Finally, we considered the Poisson distribution with parameter $\lambda = p \cdot n \approx 5.03$, $Poi(\lambda)$, which is the theoretical limiting distribution of the degree distribution of ER(n, p) in the sense of van der Hofstad (2016), Theorem 5.12. We indicated the number of times we would expect to see the value $d, d = 0, \ldots, 17$, when drawing $n = 71$ times from $Poi(\lambda)$. As we can see, the empirical degrees from ER(n, p) indeed behave very similarly to the distribution $Poi(\lambda)$. The observed

[1] In a bit more detail: Theorem 5.12 says, if we pick a node at random, the empirical probability of this node having degree k will converge to the probability of observing k when drawing from a $Poi(\lambda)$ random variable, for all k.

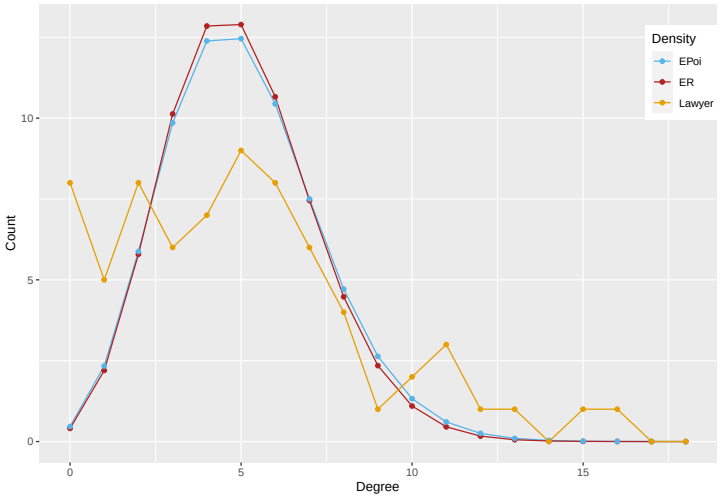


Figure 1.4: Comparison of the degree distribution of the lawyer friendship network (yellow, labelled "Lawyer"), the empirical degree distribution of ER(n, p), averaged over 10,000 draws from that model (red, "ER") and the theoretical limiting distribution $Poi(\lambda)$ (blue, "EPoi"). The degree distribution of ER(n, p) aligns very well with the theoretical limit. Real-world networks tend to have heavier tails, however.

degrees in the Lawyer network, however, behave nothing like these distributions, exhibiting much heavier tails.

1.3.2 Towards degree heterogeneity: The β-model

To generalize the Erdős-Rényi model to include degree heterogeneity, one intuitive idea is to assign node-specific parameters, one for each node. This gives rise to the β-model. Although this model can be dated back to Holland & Leinhardt (1981), the name "β-model" was coined much more recently in Chatterjee et al. (2011). Given a degree heterogeneity parameter $\beta \in \mathbb{R}^n$, links between nodes i and j in the β-model are made independently with probability

$$p_{ij} = \frac{\exp(\beta_i + \beta_j)}{1 + \exp(\beta_i + \beta_j)}, \quad i < j.$$

Chatterjee et al. (2011) showed that the maximum likelihood estimator (MLE) of the heterogeneity parameter β in this model is consistent, when the observed network is dense. Since then, the β-model has attracted a lot of attention. Yan & Xu (2013) proved the asymptotic normality of that MLE and Rinaldo et al. (2013) derived necessary and sufficient conditions for the existence of the MLE of the β-model parameters based on the polytope of degree sequences. Karwa & Slavković (2016) proved that inference in the β-model is possible under privacy constraints and provided a quadratic-time algorithm for checking the existence of the MLE. Yan, Qin & Wang (2016) derived the asymptotic properties of an estimator based on moment equations. Yan, Leng & Zhu (2016) studied a directed version of the β-model and provided asymptotic normality results for the MLE in that model. Their

model assumes that each node i has two parameters, an *outgoingness* parameter α_i and in *incomingness* parameter β_i. In their model, directed links are formed independently with the probability of observing a directed edge from i to j given as

$$p_{ij} = \frac{\exp(\alpha_i + \beta_j)}{1 + \exp(\alpha_i + \beta_j)}, \quad i \neq j.$$

Thus, their model is the canonical extension of the β-model to directed networks.

Lemma 1.12 below, originally found in Chatterjee et al. (2011), might explain why researchers are so interested in the β-model and versions thereof. Observe the following. For given β, denote the law of the β-model by P_β. Then, given a graph G with degree sequence $d = (d_1, \ldots, d_n)^T$, the probability of observing G under P_β is

$$P_\beta(G) = \frac{\exp\left(\sum_{i=1}^n \beta_i d_i\right)}{\prod_{i<j}(1 + \exp(\beta_i + \beta_j))}.$$

That is, the β-model has the form of an exponential family with the degree sequence as sufficient statistic. Indeed, the β-model is arguably the simplest member of the so-called *Exponential random graph* models, which are random graph models with exponential family form. The above equation tells us immediately that all the information about the parameter of interest, β, is contained in the degree sequence of the observed graph, see Chatterjee et al. (2011), Chatterjee & Diaconis (2013) for an in-depth treatment.

Lemma 1.12 (Theorem 1.4 in Chatterjee et al. (2011))**.** *Fix $n \in \mathbb{N}$. Let $\mathcal{R}$ be the the set of all the expected degree sequences of random graphs following the law P_β as β ranges over $\mathbb{R}^n$. Let $\mathcal{D}$ denote the set of all possible degree sequences of undirected graphs on n nodes. Then,*

$$\text{conv}(D) = \bar{\mathcal{R}},$$

where conv(D) *denotes the convex hull of $\mathcal{D}$ and $\bar{\mathcal{R}}$ is the topological closure of $\mathcal{R}$.*

Keep in mind that $\mathcal{D}$ is a strict subset of $\{0, 1, \ldots, n-1\}^n$.[2] Lemma 1.12 tells us that for fixed n and any possible degree sequence $d \in \mathcal{D}$, we can always find a $\beta \in \mathbb{R}^n$, such that the expected degree sequence under P_β is arbitrarily close to d. From a probability theory point of view, if we are solely concerned about the flexibility of our model to produce samples which exhibit degree heterogeneity for a fixed value of n, we cannot really ask for more.

There is a caveat, however. Lemma 1.12 holds for a fixed n only. From the point of view of a statistician tasked with inferring properties of the data generating process

[2]Indeed, it is for example immediate that the sum of all degrees must always be an even number, such that any vector in $\{0, 1, \ldots, n-1\}^n$ whose elements sum to an odd number cannot be contained in $\mathcal{D}$. Thus $\mathcal{D}$ is a slightly more complex object than simply the hypercube grid $\{0, 1, \ldots, n-1\}^n$. See, for example, Rinaldo et al. (2013) and Karwa & Slavković (2016) for discussions of the polytope of degree sequences.

given a sample, we are most frequently concerned with the limiting behaviour of our procedures as n tends to infinity. Since the β-model and its variants associate each node with its own parameters, they are high-dimensional and over-parametrized in nature. As a result, consistency and asymptotic normality for the MLE in the β-model is only known to hold for dense networks (Chatterjee et al. 2011, Yan & Xu 2013). In particular, the MLE for β_i will be $-\infty$ if node i has degree zero. This makes it necessary to remove isolated nodes before fitting the β-model to network data, immediately calling into question any findings obtained from such a procedure due to the issue of data-selective inference, as we have argued in Section 1.2.1.

Remark. Another popular idea for incorporating degree heterogeneity is to assume that nodes in a network can be clustered into communities and that the probability of two nodes making connections only depends on their corresponding communities. This gives rise to the Stochastic Block Model we encountered in Section 1.2. Community detection is a vast field of random network theory in its own right and is of no further concern to us.

1.3.3 Towards sparsity: The sparse β-model

For sparse networks, Chen et al. (2020) recently proposed a sparse β-model (SβM) by assuming that the heterogeneity parameter β has entries equal to zero after introducing a location-shift global sparsity parameter μ. In detail, they considered the model in which links between nodes are formed independently with probability

$$p_{ij} = \frac{\exp(\beta_i + \beta_j + \mu)}{1 + \exp(\beta_i + \beta_j + \mu)}, \quad i < j,$$

where μ is a global sparsity parameter and we allow $\mu \to -\infty$, as $n \to \infty$. For reasons of identifiability $\min_i \beta_i = 0$ was imposed. Notice that, for $\beta = 0$, SβM becomes the Erdős-Rényi model with parameter $\exp(\mu)/(1 + \exp(\mu))$, for which a theory of sparse networks exists, but which cannot model degree heterogeneity. On the other hand, if we absorb μ into β as $\beta_i + \mu/2$ for all i, it becomes the β-model, which can model arbitrary degree sequences, but for which inference is only possible in the dense graph regime. Thus, we may interpret SβM as interpolating these two models, resulting in a network model that can model sparse networks as well as degree heterogeneity.

Recall that a necessary condition for a model to produce sparse networks is that some p_{ij} go to zero (cf. the remark after Definition 1.8). Implicit in the assumption $\min_i \beta_i = 0$ is that many entries of β may be zero. In fact, Chen et al. (2020) assume β in SβM to be sparse. As a result, if $\beta_i = \beta_j = 0$ for some nodes i, j and $\mu \to -\infty$, then $p_{ij} \to 0$. Hence, if β is sparse enough and $\mu \to -\infty$ fast enough, the resulting

networks will be sparse. See Chen et al. (2020) for the details.

An interesting aspect about SβM is the procedure advocated in Chen et al. (2020) for fitting it to data and how they propose to achieve the sparsity of their estimator in practice. They employed a penalized likelihood method for estimating the parameters in their model using an ℓ_0-penalty on β. The use of ℓ_0-penalties is generally not feasible, because they require an optimization over all subsets of possible indices, which is computationally intractable. For SβM, however, Chen et al. (2020) observed that the MLEs of the β_i are roughly ranked according to the corresponding degrees of nodes, meaning that for a given sparsity level s, they simply have to assign non-zero β to the s nodes with largest degrees. Thus, they only had to iterate over the $n-1$ possible values for s to find the model that gave the best fit. See the original paper, especially the *monotonicity lemma* (Lemma 1 in Chen et al. (2020)) for the details.

1.3.4 Understanding the drivers of network formation: Homophily

Aside from degree heterogeneity, homophily is another stylized feature of many real-world networks (Kolaczyk 2009, Newman 2018). It refers to similar nodes being more likely to connect to one another than dissimilar ones, based on node attributes or covariates. In the eyes of a more application oriented statistician one of the most pressing questions when analysing real-world networks usually is assessing the influence of specific covariates on network formation by including a regression component into their network models.

Thus, when one has a powerful model such as the β-model at one's disposal, it seems natural to want to include covariates into it. Graham (2017) was the first to include nodal-covariates into the β-model. In our notation, his model is equivalent to the model where we observe some covariates $Z_{ij} \in \mathbb{R}^p$ between nodes i and j and given these covariates, links are formed independently with probability

$$P(A_{ij} = 1 | Z_{ij}) = \frac{\exp(\beta_i + \beta_j + Z_{ij}^T \gamma)}{1 + \exp(\beta_i + \beta_j + Z_{ij}^\gamma)}.$$

It was shown that the MLE for the heterogeneity and covariate parameters are only consistent under dense graph sequences, though a separate estimator for the covariate parameter based on conditioning is consistent for dense and sparse graphs. Jochmans (2018) derived results for estimating the parameter associated with covariates in directed networks by profiling out the degree heterogeneity parameter. Yan et al. (2019) investigated the issue of statistical inference for these two sets of parameters in dense directed networks when covariates are present.

1.3.5 Combining sparsity, degree heterogeneity and homophily

Given the discussion in the previous sections, it is easy to see why we want to study a model like SβM-C. Recall its definition in (1.1): We see how SβM-C emerges naturally from the models discussed in this section so far: We want to have a heavy-tailed degree distribution, which is why we include a degree heterogeneity parameter β. We also want to have sparsity, for which a necessary condition is for some of the p_{ij} to go to zero, which is why we introduce the global sparsity parameter μ, for which we allow $\mu \to -\infty$, effectively interpolating the Erdős-Rényi model and the β-model. Finally, we want our model to be able to capture homophily, which is why we include covariates and weigh them by γ. Analogously, SRGM arises naturally as directed extension from the models discussed.

1.4 LASSO theory

We briefly review the literature on using the LASSO for logistic regression relevant for this thesis. For reasons of space we restrict ourselves to presenting the main ideas, referring the reader to the books van de Geer & Bühlmann (2011) and Wainwright (2019) for a thorough introduction to this vast topic.

LASSO stands for *least absolute shrinkage and selection operator* and is now loosely referred to as a general procedure for simultaneous variable selection and parameter estimation when a loss function is regularized by constraining the ℓ_1-norm of the parameters. The usual setup for most LASSO type problems is that we observe some data $(y_i, X_i)_{i=1}^n$, where $y_i \in \mathbb{R}$ is a univariate *outcome* or *response* variable and $X_i \in \mathbb{R}^p$ are p-dimensional *covariates* or *predictors*, which can be either fixed or random. It is generally assumed that the samples $(y_i, X_i^T)^T$ are independent.

It is assumed that the response y_i is related to the predictors X_i *somehow* via a parametric model. The simplest such model is the linear model, in which we assume

$$y_i = X_i^T \beta + \epsilon_i, \quad i = 1, \dots, n,$$

where ϵ_i are i.i.d. mean-zero, error terms independent of $\{X_i\}_{i=1}^n$ and $\beta \in \mathbb{R}^p$ is the parameter of interest. Extensions to generalized linear models or more complex parametric models are possible. Of special interest to us will be the case of logistic regression, in which the y_i take values in $\{0, 1\}$ and

$$P(y_i = 1 | X_i) = \frac{\exp(X_i^T \beta)}{1 + \exp(X_i^T \beta)}.$$

In standard statistical theory, an estimator $\hat{\beta}$ for β to these problems is found

by minimizing some convex loss function $\mathcal{L}$, typically the squared error loss, the negative log-likelihood etc. In LASSO theory it is assumed that the statistician has some additional information that leads them to believe that some (often: many) of the entries β_i of β are zero. However, they do not know which ones, nor how many. Assuming a sparse β also allows treatment of the case where there are many more predictors than observations, $p \gg n$, in which case many standard algorithms fail. Thus, what the statistician would like to have is some procedure that reliably tells them which β_i are unequal to zero and what their value is. What they really would like to do is restrict the number of non-zero entries, for example via employing an ℓ_0-penalty in their minimization procedure:

$$\min_{\beta} \mathcal{L}(y, X, \beta), \quad \text{subject to } \|\beta\|_0 \leq s,$$

for some $s < p, s \in \mathbb{N}$. The problem with this approach is that the parameter space constructed via the constraint $\|\beta\|_0 \leq s$ is not convex, making the above problem computationally intractable. It can be shown that if we replace the ℓ_0-constraint by a constraint on the ℓ_1-norm,

$$\min_{\beta} \mathcal{L}(y, X, \beta), \quad \text{subject to } \|\beta\|_1 \leq \tilde{s},$$

for some possibly different parameter $\tilde{s}$, we do obtain a convex optimization problem, while retaining the property that a solution to above problem will have some of its entries set to zero, see the aforementioned references. The ℓ_1-constraint thus is a convex proxy for the ℓ_0-constraint, which we actually would like to solve, but cannot. It follows from standard results in convex optimization (Bertsekas 1995, Chapter 5.3), that this ℓ_1-constrained problem is equivalent to solving the problem

$$\min_{\beta} \mathcal{L}(y, X, \beta) + \lambda \|\beta\|_1, \tag{1.5}$$

for an appropriate penalty parameter $\lambda > 0$. Equation (1.5) is the typical form of a LASSO problem and we call a solution to (1.5) *LASSO estimator*.

An advantage of the LASSO constraint is that estimated parameters will be automatically sparse for a suitably chosen tuning parameter λ. Thus, LASSO performs model selection and parameter estimation simultaneously. Originally developed for Gaussian linear regression in Tibshirani (1996), the LASSO methodology was studied for generalized linear models by van de Geer (2008) in which a non-asymptotic oracle inequality for the empirical risk minimizer was provided. Among others, Buena (2008) studied the asymptotic consistency of variable selection for LASSO logistic regression by providing sufficient conditions to identify the true model at a given level of confidence. The LASSO methodology now represents a

widely used toolbox for high-dimensional data analysis for data sets that can have more variables than observations.

A caveat of LASSO methodology is that due to the shrinkage incurred by using the ℓ_1-penalty, we are at the same time biasing our estimates for the non-zero entries of β. Thus, it is notoriously difficult to derive the limiting distribution of our estimates, making it hard to arrive at inference results for our parameter estimates such as confidence intervals. Zhang & Zhang (2014) and van de Geer et al. (2014) proposed to overcome the bias due to shrinkage in estimation by debiasing. The main tool for this is to construct an approximate inverse of the Gram matrix using node-wise LASSO regression (Meinshausen & Bühlmann (2006)). See also Javanmard & Montanari (2014*a*) and Javanmard & Montanari (2014*b*), as well as Kock & Tang (2019), a related paper to our context that derived uniform inference results for high-dimensional dynamic panel data models.

1.5 Proofs of Section 1.2

We prove Lemma 1.10 and Proposition 1.11. Both follow from the following deep result on the phase transition of $\mathrm{ER}(\lambda/n)$, which can be found in van der Hofstad (2021).

Theorem 1.13 (Phase transition in Erdős-Rényi random graphs, Theorem 2.33 in van der Hofstad (2021), abbreviated)**.** *Fix $\lambda > 0$ and let $\mathcal{C}_{\max}$ be the largest connected component of the Erdős-Rényi graph $\mathrm{ER}(\lambda/n)$. Then,*

$$\frac{|\mathcal{C}_{\max}|}{n} \xrightarrow{P} \zeta_\lambda, \tag{1.6}$$

where ζ_λ is the survival probability of a Poisson branching process with mean offspring λ. In particular $\zeta_\lambda > 0$ precisely when $\lambda > 1$. Further, for $\lambda > 0$, with $\eta_\lambda = 1 - \zeta_\lambda$,

$$\frac{|E(\mathcal{C}_{\max})|}{n} \xrightarrow{P} \frac{1}{2}\lambda(1 - \eta_\lambda^2). \tag{1.7}$$

The proof of Lemma 1.10 follows from Theorem 1.13 together with a coupling argument.

Proof of Lemma 1.10. Denote the law of M by $P = (p_{ij})_{ij}$. We construct a coupling between P and the law of $\mathrm{ER}(\lambda/n)$ as follows. Start out with n nodes, numbered $1, \ldots, n$, without any connections between them.

1. For each pair of nodes $i < j$, draw an independent uniform random variable, $U_{ij} \sim \mathcal{U}([0,1])$.
2. Place a link between nodes i and j precisely when $U_{ij} \leq \lambda/n$. Thus, $P(i \leftrightarrow j) = P(U_{ij} \leq \lambda/n) = \lambda/n$ and the resulting graph has distribution $\mathrm{ER}(\lambda/n)$.
3. On another copy of the set $\{1, \ldots, n\}$ place a link between nodes i and j precisely when $U_{ij} \leq p_{ij}$, using the same realizations of U_{ij}. The resulting graph has distribution P.

By construction of the coupling, the realization of M will contain the same edges as the realization of $\mathrm{ER}(\lambda/n)$ and possibly some more edges. This implies $|\mathcal{C}^M| \geq |\mathcal{C}_{\max}|$, almost surely. Let $\epsilon > 0$. By (1.6),

$$\begin{aligned} P\left(|\mathcal{C}^M| \leq (\zeta_\lambda - \epsilon)n\right) &= P\left(\zeta_\lambda - \frac{|\mathcal{C}^M|}{n} \geq \epsilon\right) \\ &\leq P\left(\zeta_\lambda - \frac{|\mathcal{C}_{\max}|}{n} \geq \epsilon\right) \\ &\leq P\left(\left|\zeta_\lambda - \frac{|\mathcal{C}_{\max}|}{n}\right| \geq \epsilon\right) \to 0. \end{aligned}$$

For edges formed conditionally independent given some covariates Z_{ij} the proof

remains the same since the lower bound of λ/n holds almost surely, independent of the value of Z_{ij}. This proves the claim. □

We now prove Proposition 1.11, which also follows from Theorem 1.13.

Proof of Proposition 1.11. Proposition 1.11 follows from Theorem 1.13 by repeated application of Slutzky's Theorem. By (1.6) and Slutzky,

$$\frac{|\mathcal{C}_{\max}| - 1}{n} \xrightarrow{P} \zeta_\lambda.$$

Thus, by Slutzky,

$$\binom{|\mathcal{C}_{\max}|}{2} \cdot \frac{1}{n^2} \xrightarrow{P} \frac{1}{2} \cdot \zeta_\lambda^2.$$

Now, a final application of Slutzky's Theorem together with (1.6) and (1.7) yields,

$$\frac{\hat{p}_{\max}}{p} = \frac{|E(\mathcal{C}_{\max})|}{\binom{|\mathcal{C}_{\max}|}{2}} \cdot \frac{n}{\lambda} = \frac{|E(\mathcal{C}_{\max})|}{n} \cdot \frac{n^2}{\binom{|\mathcal{C}_{\max}|}{2}} \cdot \frac{1}{\lambda} \xrightarrow{P} \frac{(1-\eta_\lambda^2)}{\zeta_\lambda^2} = \frac{1+\eta_\lambda}{1-\eta_\lambda},$$

where we used the definition of ζ_λ in the last step. □

Chapter 2

A sparse β-model with covariates

Organization of this chapter

In Section 2.1 we formally introduce the sparse β-model with covariates (SβM-C). In Section 2.2 we propose the use of a penalized likelihood method with an ℓ_1-penalty on the nodal parameters for parameter estimation. We study the finite-sample error bounds on the excess risk and the ℓ_1-error of our estimator (Theorem 2.4).

We then zoom in on the first of two special cases of SβM-C. In Section 2.3, we show how the results from Section 2.2 can be applied to the SβM without covariates, i.e. the model studied in Chen et al. (2020). We compare our rates of convergence to theirs. In Section 2.4, we derive a central limit theorem for our estimator of the homophily parameter γ. We present extensive simulation results in Section 2.5 and apply our model to Lazega's lawyer friendship data and the world trade network in Section 2.6. All proofs are relegated to Section 2.7. The content of this chapter is taken from Stein & Leng (2020).

2.1 Random networks and high-dimensional statistics

We formally introduce the sparse β-model with covariates (SβM-C). By allowing its link probabilities to go to zero, SβM-C can model sparse networks. On the computational side we will set up our model in the language of LASSO estimation, making our estimation approach very fast and scalable. Crucially, this estimation procedure can handle networks that are disconnected or even have isolated nodes. This allows us to perform model-selective inference, fitting our model to the entire network, thus avoiding the biases incurred by data-selective inference (c.f. Section 1.2).

Recall the model definition (1.1): We observe data $\{A_{ij}, Z_{ij}\}_{i,j=1,i<j}^{n}$, where $A = (A_{ij})_{ij} \in \mathbb{R}^{n\times n}$ is a symmetric adjacency matrix and $Z_{ij} \in \mathbb{R}^p$ are p-dimensional covariates associated with nodes i and j. Given the covariates, undirected links are independently made with the probability of a connection between node i and j being

$$P(A_{ij} = 1|Z_{ij}) = p_{ij} = \frac{\exp(\beta_i + \beta_j + \mu + Z_{ij}^T\gamma)}{1 + \exp(\beta_i + \beta_j + \mu + Z_{ij}^T\gamma)},$$

where $\beta = (\beta_1, \dots, \beta_n)^T \in \mathbb{R}^n$ is the heterogeneity parameter, $\gamma \in \mathbb{R}^p$ is the parameter for the covariates, and $\mu \in \mathbb{R}$ is the global sparsity parameter for which we allow $\mu \to -\infty$, as $n \to \infty$. For identifiability we assume $\min_i \beta_i = 0$.

For brevity, we denote the parameters collectively as $\theta = (\beta^T, \mu, \gamma^T)^T$ and its true value as $\theta_0 = (\beta_0^T, \mu_0, \gamma_0^T)^T$. We write $S_0 = S(\beta_0)$ for the support of β_0. For ease of presentation, we introduce the shorthand notation $s_0 = |S_0|$ and $S_{0,+} \coloneqq S_0 \cup \{n+1, n+2, \dots n+1+p\}$ with cardinality $s_{0,+} = |S_{0,+}| = s_0 + p + 1$ to refer to all active indices including those of μ and γ.

We focus on the finite-dimensional covariate case by assuming that p, the dimension of the covariates Z_{ij}, is fixed. We assume that Z_{ij} are independent realizations from centred, uniformly bounded random variables. We do not require Z_{ij} to be *i.i.d.* and Z_{ij} may have correlated entries. These assumptions imply in particular the existence of constants $\kappa, c > 0$ such that $|Z_{ij}^T \gamma_0| \leq \kappa$ and $|Z_{ij,k}| \leq c$ for all $1 \leq i < j \leq n, k = 1, \dots, p$. We assume further that γ_0 lies in a compact, convex set $\Gamma \subset \mathbb{R}^p$, which means we may choose a universal κ independent of γ_0. We let $\Theta \coloneqq \mathbb{R}_+^n \times \mathbb{R} \times \Gamma$ denote the parameter space.

The reader may have spotted a potential issue with this model setup: On one hand, we are assuming that the Z_{ij} are independent random variables. On the other hand, we frequently would like to consider covariates of the form $Z_{ij} = g(X_i, X_j)$, with $g(X_i, X_j) = -\|X_i - X_j\|$ or similar, where X_i, X_j are nodal covariates. This would entail that the Z_{ij} are, in fact, dependent. To get out of this predicament, many authors (Graham 2017, Yan et al. 2019, Ma, Ma & Yuan 2020, to name but a few) opt for assuming that the Z_{ij} are fixed design points, which allows them to gracefully avoid having to deal with this issue. To make the present work a bit more interesting and different from existing approaches, we have chosen to treat the Z_{ij} as random and independent. We would like to point out, however, that in case of fixed design, we would obtain the same results, error rates, etc. and many of the proofs would remain exactly the same or even become simpler.

We highlight that the model in (1.1) as well as the original SβM introduced in Chen et al. (2020) is sparse in terms of its parametrization and the density of the resulting network. The latter is a natural consequence of the former and thus the word "sparse" in SβM-C refers to the former. For example, when β is sparse with finite support, by allowing $\mu \to -\infty$ at appropriate rates, the networks generated from this model will be sparse. When $\beta = 0$, model (1.1) becomes what we call a sparse Erdős-Rényi model with covariates (ER-C). We show in Chapter 3 that this special version of SβM-C can model any network whose expected number of edges scales as $O(n^{2-\xi})$ with $\xi \in [0, 2)$. That is, the network modelled by this special case of SβM-C can be almost arbitrarily sparse.

2.1.1 Main results

The main methodological contribution comes from the SβM-C as the first model capable of capturing node heterogeneity differentially while accounting for covariates. In the literature, closely related models allowing node-specific parameters either ignore covariates and thus homophily (Chen et al. 2020), or overly parametrize by assigning parameters indistinguishably to all the nodes (Graham 2017, Yan et al. 2019, e.g.), leading to theoretical and practical difficulties in applying these models, as we have argued in Chapter 1. In particular, by associating each node with its own parameter(s), these overly parametrized models require a network to be dense for the purposes of estimation and statistical inference. By differentially modelling node-specific parameters, SβM-C can drastically reduce the number of parameters needed and thus model networks that are sparse.

The first main result we will show is the consistency of an ℓ_1-penalized estimator (2.2) for SβM-C in terms of excess risk and ℓ_1-norm. Despite the somewhat superficial similarity of our estimator to the penalized logistic regression with an ℓ_1-penalty, great care needs to be taken when applying results from LASSO theory to our estimator. Firstly, the parts of the design matrix of our model associated with β and μ are deterministic while those for γ are random, making some assumptions on the eigenvalues of the design matrix typically seen in LASSO type problems invalid. Secondly, our approach differs from classical LASSO theory for logistic regression insofar that we do not assume that the linking probabilities p_{ij} between two nodes stay uniformly bounded away from zero. This is often assumed in LASSO theory for easier derivations; see, for example, van de Geer & Bühlmann (2011), Theorem 6.4; Buena (2008), Theorem 2.4; or van de Geer (2008), Theorem 2.1. Were we to impose such a condition, however, the expected degree of each node would scale linearly in the number of nodes, automatically putting us in the dense graph regime (c.f. the remark after Definition 1.8). Instead, we allow the link probability for any dyad to go to zero at a certain rate as the number of nodes tends to infinity. This has far reaching consequences, especially for the derivation of the limiting distribution of our estimator for γ (Theorem 2.7), as we have indicated in Section 1.1 and discuss in detail in Section 2.4.

Importantly, our approach also differs from classical LASSO theory in that the various parameters in SβM-C have differing effective sample sizes, resulting in different rates of convergence. Loosely speaking, the effective sample size for each β_i depends on the number of possible connections that the ith node has, while μ and γ are global parameters relevant to all edges. Remarkably, we are still able to recover almost the classical LASSO rate of convergence for excess risk and ℓ_1-error, up to an additional factor having an explicit relation to the expected edge density of the

network. This factor is the price we pay for allowing the link probabilities to tend to zero.

The second main result is to prove a central limit theorem for our estimator of the homophily parameter γ, which, for statistical inference, is often of major interest, as the heterogeneity parameter β can be seen as a nuisance parameter. Due to the shrinkage of the parameter estimates, approaches using the LASSO procedure generally produce biased estimators, which makes deriving limiting distributions and inference results difficult. A great deal of work has resorted to debiasing the LASSO estimator, see the discussion in Section 1.4. Our results show that, quite remarkably, for inference on the covariate parameter γ, no debiasing is necessary. Specifically, the columns of the design matrix for β and those for μ and γ become asymptotically orthogonal as n increases. As a result, the bias incurred by shrinkage estimation does not affect the derivation of standard central limit theory for the estimated γ. See Section 2.4 for details.

As byproducts of our theory, we develop the theory for the first of two special cases of SβM-C. We provide results analogous to Chen et al. (2020) in Section 2.3 when covariates are not considered, by replacing the ℓ_0-penalty on β used by Chen et al. (2020) by an ℓ_1-penalty. The second special case, a simplified model of SβM-C when the heterogeneity parameter is not present, i.e. when $\beta = 0$, is treated in Chapter 3.

2.2 Sparse β-model with covariates

Given an observed adjacency matrix A and the associated covariates $\{Z_{ij}\}_{i\neq j}$, the negative log-likelihood of SβM-C at $\theta = (\beta^T, \mu, \gamma^T)$ is easily seen to be

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \beta_i d_i - d_+\mu - \sum_{i<j}(Z_{ij}^T\gamma)A_{ij} + \sum_{i<j}\log(1+\exp(\beta_i+\beta_j+\mu+Z_{ij}^T\gamma)). \quad (2.1)$$

It is easily seen by differentiating that $\theta_0 = \arg\min_{\theta\in\Theta}\mathbb{E}[\mathcal{L}(\theta)]$.

Our model is high-dimensional with $n+p+1$ unknown parameters, where the heterogeneity parameter β admits a sparse representation with an unknown support. This fact immediately motivates the use of a penalized likelihood approach for estimation. As discussed in Section 1.4 in a situation like this it would be tempting to estimate the parameters of the model via regularized likelihood by penalizing the ℓ_0-norm of β. For the sparse β-model without covariates, Chen et al. (2020) found that this non-convex optimization problem is computationally tractable, thanks to a key monotonicity lemma stating that the elements of the estimated β are ranked according to the degrees of the nodes. The arguments leading to the conclusion of this

lemma, however, do not extend to the current setting where covariates are included. This effectively means that the ℓ_0-norm penalized likelihood becomes a combinatorial problem for the SβM-C and an exhaustive search, which is computationally intractable, in the model space is inevitable.

As discussed in Section 1.4, one approach popular in high-dimensional data analysis is to replace the ℓ_0-penalty on β by an ℓ_1-penalty, which serves as a convex proxy for the ℓ_0-penalty. This leads us the following problem, where our estimator is obtained by solving

$$\min_{\beta \in R_+^n, \mu \in \mathbb{R}, \gamma \in \mathbb{R}^p} \frac{1}{\binom{n}{2}} \mathcal{L}(\beta, \mu, \gamma) + \lambda \|\beta\|_1, \tag{2.2}$$

where $\mathcal{L}(\beta, \mu, \gamma)$ is the negative log-likelihood defined in (2.1) and λ is a tuning parameter. This formulation immediately connects our approach to the LASSO methodology (Tibshirani 1996, c.f. Section 1.4), enabling us to draw upon the vast literature on high-dimensional data analysis, especially for logistic regression.

On the computational side, the formulation in (2.2) is the same as penalized logistic regression with the LASSO penalty. Thus, to solve it in practice, we can invoke standard algorithms developed for LASSO and thus the estimation of the parameters of SβM-C can be done extremely fast. In particular, we can use the functions in the `glmnet` R package (Friedman et al. 2010) by properly setting up the design matrix and the constraints on β. Experience shows that this algorithm can effectively compute the estimator for a network with the number of nodes up to a few thousand.

2.2.1 Theory

Since we aim to develop a theory for sparse networks, we allow $\mu_0 \to -\infty$ as $n \to \infty$. As a result, some link probabilities may go to zero as $n \to \infty$. In order to perform consistent estimation, it is clear that we need to restrict the rate at which this may happen. Therefore, we assume there is a non-random sequence $1/2 \geq \rho_{n,0} > 0, \rho_{n,0} \to 0$, as $n \to \infty$, such that almost surely for all i, j:

$$1 - \rho_{n,0} \geq p_{ij} \geq \rho_{n,0}.$$

Since a smaller $\rho_{n,0}$ allows sparser networks, we refer to $\rho_{n,0}$ as the *network sparsity parameter*. Applying $\text{logit}(x) = \log(x/(1-x))$ to the inequality above we get for all i, j,

$$-\text{logit}(\rho_{n,0}) = \text{logit}(1 - \rho_{n,0}) \geq \beta_{0,i} + \beta_{0,j} + \mu_0 + \gamma_0^T Z_{ij} \geq \text{logit}(\rho_{n,0}),$$

which is equivalent to

$$|\beta_{0,i} + \beta_{0,j} + \mu_0 + \gamma_0^T Z_{ij}| \leq -\text{logit}(\rho_{n,0}) =: r_{n,0}, \quad \forall i, j.$$

Since $\rho_n \leq 1/2$, we have $r_{n,0} \geq 0$. The previous inequality can also be expressed in terms of the design matrix D associated with the corresponding logistic regression problem (see below for the exact formulation) and is equivalent to $\|D\theta_0\|_\infty \leq r_{n,0}$. This motivates the following estimation procedure: Given a sufficiently large constant r_n, we define the local parameter space

$$\Theta_{\text{loc}} = \Theta_{\text{loc}}(r_n) \coloneqq \{\theta \in \Theta : \|D\theta\|_\infty \leq r_n\} \tag{2.3}$$

and propose to perform estimation via

$$\hat{\theta} = (\hat{\beta}^T, \hat{\mu}, \hat{\gamma}^T)^T = \underset{\theta=(\beta^T,\mu,\gamma^T)^T \in \Theta_{\text{loc}}}{\arg\min} \frac{1}{\binom{n}{2}} \mathcal{L}(\beta, \mu, \gamma) + \lambda \|\beta\|_1. \tag{2.4}$$

As we have seen in the equations above, any $r_n > 0$ used in the definition of Θ_{loc} corresponds to some ρ_n which uniformly lower bounds the connection probability and thus can be seen as a proxy for the permissible sparsity of our network. This type of restriction on the parameter space is similar to what was done in Chen et al. (2020), although they restricted the parameter values of β and μ directly. The condition in (2.3) is slightly more general and somewhat more natural. Noting that Θ_{loc} is convex, we have a convex optimization problem in (2.4).

In (2.4) we replaced the condition $\min_i \beta_i = 0$ by the less strict condition $\beta \in \mathbb{R}^n_+$. The following Lemma shows that this is viable: As long as the observed graph is neither empty nor complete and $\lambda > 0$, a solution $\hat{\beta}$ to (2.4) always exists and automatically fulfils $\min_{1 \leq i \leq n} \hat{\beta}_i = 0$.

Lemma 2.1. *Assume that $0 < d_+ < \binom{n}{2}$. Then, for any $0 < \lambda < \infty$ there exists a minimizer for the optimization problem (2.4) and any solution $\hat{\theta} = (\hat{\beta}^T, \hat{\mu}, \hat{\gamma}^T)^T$ of (2.4) must satisfy $\min_{1 \leq i \leq n} \hat{\beta}_i = 0$.*

Following the empirical risk literature (cf. Greenshtein & Ritov (2004), Koltchinskii (2011)) we will analyse the performance of our estimator in terms of excess risk. We define the (global) excess risk as

$$\mathcal{E}(\theta) \coloneqq \frac{1}{\binom{n}{2}} \mathbb{E}[\mathcal{L}(\theta) - \mathcal{L}(\theta_0)].$$

Since we define the local parameter space Θ_{loc} with respect to some rate r_n, in our derivations we must account for the fact that this r_n may be smaller than the true $r_{n,0}$. In that case there is no way for us to find the true parameter θ_0 and the best

we can hope to achieve is to find the best local approximation θ^* of the truth θ_0, which we define as

$$\theta^* = \arg\min_{\theta \in \Theta_{\text{loc}}} \frac{1}{\binom{n}{2}} \mathbb{E}[\mathcal{L}(\theta)].$$

Note that the truth θ_0 fulfils

$$\theta_0 = \arg\min_{\theta \in \Theta} \frac{1}{\binom{n}{2}} \mathbb{E}[\mathcal{L}(\theta)] = \arg\min_{\theta \in \Theta_{\text{loc}}(r_{n,0})} \frac{1}{\binom{n}{2}} \mathbb{E}[\mathcal{L}(\theta)].$$

Hence, if $r_{n,0} \leq r_n$, $\theta^* = \theta_0$. In general, however, estimating θ^* is the best we can achieve when solving (2.4). Thus, we introduce the notion of local excess risk as in Chen et al. (2020), which measures how close a parameter θ is to the best local approximation θ^* in terms of excess risk:

$$\mathcal{E}_{\text{loc}}(\theta) \coloneqq \mathcal{E}(\theta) - \mathcal{E}(\theta^*).$$

Clearly, θ^* also fulfils $\theta^* = \arg\min_{\theta \in \Theta_{\text{loc}}} \mathcal{E}(\theta)$ and we may consider the excess risk of the best local approximation, $\mathcal{E}(\theta^*)$, as the approximation error of our model. It accounts for the fact that our model might be misspecified, in the sense that the parameter r_n is not large enough. As is usual in LASSO theory (cf. van de Geer & Bühlmann (2011), Chapter 6), it is tacitly assumed that this approximation error is small, i.e. we assume that r_n is sufficiently large. Note that the global excess risk of our estimator $\hat{\theta}$ decomposes as

$$\mathcal{E}(\hat{\theta}) = \mathcal{E}(\theta^*) + \mathcal{E}_{\text{loc}}(\hat{\theta}),$$

where we can consider the approximation error $\mathcal{E}(\theta^*)$ as a deterministic bias.

As is commonly assumed in LASSO theory (cf. van de Geer & Bühlmann (2011), Chapter 6), we assume that the unpenalized parameters of θ^* are active. That is, $\mu^* \neq 0, \gamma_i^* \neq 0, i = 1, \ldots, p$. Denote the set of true active indices by $S^* = S(\beta^*) = \{i : \beta_i^* > 0\}$ with cardinality $s^* = |S^*|$. For ease of notation, we introduce the set $S_+^* = S^* \cup \{n+1, n+2, \ldots n+1+p\}$ with cardinality $s_+^* = |S_+^*| = s^* + p + 1$ to refer to all active indices including those of μ and γ.

We set up our problem in the language of LASSO theory for logistic regression. For each pair $i < j$, denote by $X_{ij} \in \mathbb{R}^n$ the vector containing one at the ith and jth position and zeros everywhere else. Define the matrices

$$X = \begin{bmatrix} X_{12}^T \\ \ldots \\ X_{ij}^T \\ \ldots \\ X_{(n-1),n}^T \end{bmatrix} \in \mathbb{R}^{\binom{n}{2} \times n}, \; Z = \begin{bmatrix} Z_{12}^T \\ \ldots \\ Z_{ij}^T \\ \ldots \\ Z_{n-1,n}^T \end{bmatrix} \in \mathbb{R}^{\binom{n}{2} \times p}.$$

Let $\mathbf{1} \in \mathbb{R}^{\binom{n}{2}}$ be the vector containing only ones. Then the design matrix of (1.1) can be written as

$$D = \left[\begin{array}{c|c|c} X & \mathbf{1} & Z \end{array} \right] \in \mathbb{R}^{\binom{n}{2} \times (n+p+1)},$$

where D, consisting of the matrices $X, \mathbf{1}$ and Z written next to each other, is the analogue to the design matrix in logistic regression. We number the rows of D as $D = (D_{ij}^T)_{i<j}$. Here we see a crucial feature of our design matrix D: While each column corresponding to the parameters μ and γ appears in the link probability of all $\binom{n}{2}$ node pairs, each β_i only appears in $(n-1)$ such probabilities. That means, while the effective sample size for μ and γ is $\binom{n}{2}$, it is only $n-1$ for each entry of β, i.e. of order n smaller. This is also reflected in the different rates of convergence we obtain in Theorem 2.4 below. See Figure 2.1 for an example.

$$X = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad Z = \begin{bmatrix} 0.1530 & 0.1301 \\ -0.1851 & 0.1384 \\ 0.0603 & 0.2401 \\ 0.0872 & 0.0784 \\ -0.0318 & -0.1978 \\ 0.2454 & 0.0473 \\ -0.0628 & 0.2462 \\ -0.3601 & 0.2064 \\ -0.0665 & -0.1280 \\ -0.0461 & -0.0814 \end{bmatrix}$$

Figure 2.1: Example of the blocks in the design matrix D for $n = 5$ and $p = 2$. While the columns in X, associated with β, have $n-1$ non-zero entries, those columns in $\mathbf{1}$ and Z, associated with μ and γ, have $\binom{n}{2}$.

2.2.2 A compatibility condition

A crucial assumption in LASSO theory is the so-called compatibility condition (van de Geer & Bühlmann 2011, van de Geer et al. 2014). It relates the quantities $\|(\hat{\theta} - \theta^*)_{S_+^*}\|_1$ and

$$\frac{1}{\binom{n}{2}} \sum_{i<j} \mathbb{E}[(\hat{\beta}_i - \beta_i^* + \hat{\beta}_j - \beta_j^* + \hat{\mu} - \mu^* + (\hat{\gamma} - \gamma^*)^T Z_{ij})^2]$$

in a suitable sense made precise below and is crucial for deriving consistency results. Notice that the above quantity can be written as

$$(\hat{\theta} - \theta^*)^T \left(\frac{1}{\binom{n}{2}} \mathbb{E}[D^T D] \right) (\hat{\theta} - \theta^*),$$

where $\frac{1}{\binom{n}{2}} \mathbb{E}[D^T D]$ is the population Gram matrix of our design matrix D. In the SβM-C however, the classical compatibility condition as for example defined for generalized linear models in van de Geer et al. (2014) does not hold. The reason for this is that β and $(\mu, \gamma^T)^T$ have different effective sample sizes. We need to account

for this fact and therefore have to use a sample size adjusted Gram matrix. To that end, we introduce the matrix

$$T = \begin{bmatrix} \sqrt{n-1} I_n & \mathbf{0} \\ \mathbf{0} & \sqrt{\binom{n}{2}} I_{p+1} \end{bmatrix},$$

where I_m is the $m \times m$ identity matrix and we use $\mathbf{0}$ to denote the zero block matrix of appropriate dimensions. We define the sample size adjusted Gram matrix Σ as

$$\Sigma \coloneqq T^{-1}\mathbb{E}[D^T D]T^{-1} = \frac{1}{\binom{n}{2}} \begin{bmatrix} \frac{n}{2} X^T X & \frac{\sqrt{n}}{\sqrt{2}} X^T \mathbf{1} & \mathbf{0} \\ \frac{\sqrt{n}}{\sqrt{2}} \mathbf{1}^T X & \mathbf{1}^T \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{E}[Z^T Z] \end{bmatrix}.$$

We consider the limit of the matrix Σ entrywise.

Definition 2.2 (Compatibility Condition)**.** We say the compatibility condition holds if the sample size adjusted Gram matrix Σ has the following property: There is a constant b such that for every $\theta \in \mathbb{R}^{n+1+p}$ with $\|\theta_{S_+^{*c}}\|_1 \leq 3\|\theta_{S_+^*}\|_1$ we have

$$\|\theta_{S_+^*}\|_1^2 \leq \frac{s_+^*}{b} \theta^T \Sigma \theta.$$

To prove that Σ has this property, we will use techniques similar to the ones used in Kock & Tang (2019). Their matrix structure is somewhat simpler than ours as they obtain an identity matrix where we obtain a special Toeplitz matrix. More precisely, we will first show that the compatibility condition holds for the matrix

$$\Sigma_A \coloneqq \begin{bmatrix} \frac{1}{n-1} X^T X & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{E}[Z^T Z / \binom{n}{2}] \end{bmatrix} \in \mathbb{R}^{(n+1+p)\times(n+1+p)}.$$

To show that the compatibility condition also holds with high probability for Σ, it will then suffice to show that Σ and Σ_A are sufficiently close to each other in an appropriate sense. To this end, it is sufficient to impose the following eigenvalue restriction, which effectively quantifies how strongly the columns of Z may be correlated.

Assumption 2.1. There are universal constants $C > c_{\min} > 0$, independent of n, such that for all $n \in \mathbb{N}$, the minimum eigenvalue $\lambda_{\min} = \lambda_{\min}(n)$ and the maximum eigenvalue $\lambda_{\max} = \lambda_{\max}(n)$ of $\frac{1}{\binom{n}{2}} E[Z^T Z]$ fulfil $c_{\min} \leq \lambda_{\min} \leq \lambda_{\max} \leq C < \infty$. Without loss of generality we assume $c_{\min} < 1/2$.

We summarize these results in the following proposition the proof of which is given in Section 2.7.1.2.

Proposition 2.3. *Under Assumption 2.1, for $s^* = o(\sqrt{n})$ and n large enough, we have for every $\theta \in \mathbb{R}^{n+1+p}$ with $\|\theta_{S_+^{*c}}\|_1 \leq 3\|\theta_{S_+^*}\|_1$, that*

$$\|\theta_{S_+^*}\|_1^2 \leq \frac{2s_+^*}{c_{\min}}\theta^T\Sigma\theta.$$

Proposition 2.3 requires $s_+^* = o(\sqrt{n})$. The "n large enough"-condition is made precise in the proof and requires that n be such that $1/\sqrt{n} < 1/s_+^*$, which is implied by $s_+^* = o(\sqrt{n})$ and sufficiently large n. Let us put this in the context of general LASSO theory. In general LASSO theory, to show that the ℓ_1-error goes to zero in probability for increasing n, it is imposed that the sparsity s of the true parameter fulfils

$$s \cdot \sqrt{\frac{\log(\text{ number of columns of design matrix })}{\text{effective sample size}}} \overset{n\to\infty}{\longrightarrow} 0,$$

see for example van de Geer & Bühlmann (2011), Chapter 6. In our case the sparsity refers to β and we thus should expect that the restrictions we have to impose on s^* are based on the sample size associated with β. To make our conditions on s^* precise, define $\eta \coloneqq 2r_n + 2\|\beta^* - \beta_0\|_\infty + |\mu^* - \mu_0| + 2\kappa$ and let

$$K_n = K_n(\eta) = \frac{2(1+\exp(r_{n,0}+\eta))^2}{\exp(r_{n,0}+\eta)}. \tag{2.5}$$

Notice that η essentially quantifies the approximation error we commit. We make the following assumption on s^*.

Assumption 2.2. $s^* = o\left(\frac{\sqrt{n}}{\sqrt{\log(n)}\cdot K_n}\right)$.

That means, up to an additional factor K_n – which is the price we have to pay for allowing our link probabilities to go to zero – the permissible sparsity for β^* is the permissible sparsity in classical LASSO theory for an effective sample size of order n. Clearly, Assumption 2.2 is stronger than the condition $s = o(\sqrt{n})$ in Proposition 2.3, which thus is not a major restriction.

2.2.3 Consistency

In our proof for consistency of the estimator $\hat{\theta}$, we reformulate our likelihood problem in the language of the sample size adjusted design matrix. This new formulation is entirely equivalent to the previous one in (2.4), but gives a different interpretation to the sample size adjusted Gram matrix. We use it mostly to ease notation in our

proofs. In particular, we introduce vectors $\bar{X}_{ij} = \frac{\sqrt{n}}{\sqrt{2}} X_{ij}$ and define

$$\bar{X} = \frac{\sqrt{n}}{\sqrt{2}} X = \begin{bmatrix} \bar{X}_{12}^T \\ \dots \\ \bar{X}_{ij}^T \\ \dots \\ \bar{X}_{(n-1),n}^T \end{bmatrix} \in \mathbb{R}^{\binom{n}{2} \times (n+1)}, \quad \bar{D} = \left[\bar{X} | \mathbf{1} | Z\right].$$

We may consider $\bar{D}$ as a sample size adjusted design matrix, in the sense that

$$\Sigma = \frac{1}{\binom{n}{2}} \mathbb{E}[\bar{D}^T \bar{D}].$$

Likewise, we introduce sample size adjusted parameters. Here, we are effectively blowing up those columns of the design matrix corresponding to β to compensate for the fact that β has effective sample size of order n smaller than μ and γ. The details can be found in Section 2.7.1.3. Naturally, these changes will also result in a sample size adjusted penalty parameter $\bar{\lambda}$. For now, we simply remark that $\bar{\lambda} = \frac{\sqrt{n}}{\sqrt{2}} \lambda$ and refer the reader to Section 2.7.1.3 for the details.

We now state our first main theorem. Its proof is developed in Sections 2.7.1.2–2.7.1.8.

Theorem 2.4. *Assume Assumptions 2.1 and 2.2. Fix a confidence level t and let*

$$a_n \coloneqq \sqrt{\frac{2 \log(2(n+p+1))}{\binom{n}{2}}} (1 \vee c).$$

Choose $\lambda_0 = \lambda_0(t, n)$ as

$$\lambda_0 = 8 a_n + 2 \sqrt{\frac{t}{\binom{n}{2}} (11(1 \vee (c^2 p)) + 8\sqrt{2}(1 \vee c)\sqrt{n} a_n)} + \frac{2\sqrt{2} t (1 \vee c) \sqrt{n}}{3\binom{n}{2}}.$$

Let $\bar{\lambda} = \frac{\sqrt{n}}{\sqrt{2}} \lambda \geq 8\lambda_0$ and let K_n be defined as in (2.5). Then, with probability at least $1 - \exp(-t)$ we have

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda} \left(\frac{\sqrt{2}}{\sqrt{n}} \|\hat{\beta} - \beta^*\|_1 + |\hat{\mu} - \mu^*| + \|\hat{\gamma} - \gamma^*\|_1 \right) \leq 6\mathcal{E}(\theta^*) + 32 \frac{s_+^* K_n \bar{\lambda}^2}{c_{\min}}.$$

Theorem 2.4 has especially interesting implications if no approximation error is committed, that is in the case that $\theta^* = \theta_0$, for which it is sufficient that $r_{n,0} \leq r_n$.

Corollary 2.5. *Under the assumptions and with the definitions in Theorem 2.4, assume that no approximation error is made, i.e. $\theta^* = \theta_0$. Then, with probability at*

least $1 - \exp(-t)$ *we have*

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{\sqrt{2}}{\sqrt{n}}\|\hat{\beta} - \beta^*\|_1 + |\hat{\mu} - \mu^*| + \|\hat{\gamma} - \gamma^*\|_1\right) \leq C\frac{s_+^*\bar{\lambda}^2}{\rho_{n,0}}$$

with constant $C = 128/c_{\min}$.

Corollary 2.5 is proved in Section 2.7.1.8. It gives us an explicit formula for how the sparsity of our network will affect our rate of convergence, which is particularly nice, since in many related works the conditions on network density enter the rate of convergence only indirectly as assumptions on the norm of the true parameter vector, see for example Chatterjee et al. (2011), Yan & Xu (2013). Also, notice that this is essentially the rate of convergence we would expect in the classical LASSO setting for logistic regression up to an additional factor $\rho_{n,0}^{-1}$. Let us consider the implications of Theorem 2.4 in more detail.

Note that $\lambda_0 \asymp \sqrt{\log(n)/\binom{n}{2}}$. Hence, we may choose $\bar{\lambda}$ also of order $\sqrt{\log(n)/\binom{n}{2}}$. Recall that in the classical LASSO setting for logistic regression (cf. van de Geer & Bühlmann (2011)), when no approximation error is committed, when probabilities stay bounded away from zero and when we have the same effective sample size for each parameter, we obtain the rates

$$O_P\left(\text{sparsity} \cdot \frac{\log(\text{number of columns of design matrix})}{\text{effective sample size}}\right)$$

for the excess risk and

$$O_P\left(\text{sparsity} \cdot \sqrt{\frac{\log(\text{number of columns of design matrix})}{\text{effective sample size}}}\right)$$

for the ℓ_1-error. In the setting of Corollary 2.5, we obtain

$$\begin{aligned}
\mathcal{E}(\hat{\theta}) &= O_P\left(s_+^* \cdot \frac{1}{\rho_{n,0}} \cdot \frac{\log(n)}{\binom{n}{2}}\right), \\
\frac{\sqrt{2}}{\sqrt{n}}\|\hat{\beta} - \beta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1 &= O_P\left(s_+^* \cdot \frac{1}{\rho_{n,0}} \cdot \sqrt{\frac{\log(n)}{\binom{n}{2}}}\right), \\
\|\hat{\beta} - \beta_0\|_1 &= O_P\left(s_+^* \cdot \frac{1}{\rho_{n,0}} \cdot \frac{\sqrt{\log(n)}}{\sqrt{n-1}}\right).
\end{aligned}$$

That is, up to an additional factor $1/\rho_{n,0}$, we obtain the LASSO rate of convergence for sample size $\binom{n}{2}$ for the global excess risk. By the second line of the display above, we have immediately $\hat{\mu} \xrightarrow{P} \mu_0$ and $\hat{\gamma} \xrightarrow{P} \gamma_0$ at the rate expected from a LASSO type estimator with effective sample size $\binom{n}{2}$ (up to an additional factor). Furthermore, the third line implies that, again, up to an additional factor, for the error of $\hat{\beta}$, we obtain the rate of convergence we would expect for a LASSO type estimator

with sample size $n-1$. In particular, the assumptions we have to impose to obtain ℓ_1-consistency include the case $\|\beta_0\|_\infty = o(\log(\log(n)))$, which is the condition that had to be imposed in the original β-model for their strong consistency result (cf. Yan & Xu (2013), Theorem 1).

2.3 Sparse β-model without covariates

By letting $p = 0, \gamma = 0$ and consequently $\kappa = 0$, the results for the SβM-C derived in the previous section have implications for the SβM without covariates introduced in Chen et al. (2020). In SβM, the negative log-likelihood is given by

$$\mathcal{L}(\beta, \mu) = -\sum_i \beta_i d_i - d_+ \mu + \sum_{i<j} \log(1 + e^{\beta_i + \beta_j + \mu})$$

and our design matrix is simply $D = \Big[X|\mathbf{1}\Big] \in \mathbb{R}^{\binom{n}{2} \times (n+1)}$. The definitions of $\rho_{n,0}$ and $r_{n,0}$ do not change, as we can simply set $\gamma = 0$ in their original definitions. In this section we will abuse notation slightly by reusing the names from SβM-C, but redefining them to have the components corresponding to γ removed. For example, we will use $\theta = (\beta^T, \mu)^T$ for a generic parameter, $\theta_0 = (\beta_0^T, \mu_0)^T$ to denote the truth, $S_+^* = S^* \cup \{n+1\}$ to denote the sparsity including the μ component etc. This slight abuse of notation is justified as it makes the connection to the respective objects in the model with covariates clearer. Our estimator reduces to

$$\hat{\theta} = (\hat{\beta}^T, \hat{\mu})^T = \underset{(\beta^T, \mu)^T \in \Theta_{\text{loc}}}{\arg\min} \frac{1}{\binom{n}{2}} \mathcal{L}(\beta, \mu) + \lambda \|\beta\|_1,$$

where by slight abuse of notation, for this section only, we define $\Theta_{\text{loc}} = \Theta_{\text{loc}}(r_n) \coloneqq \{\theta = (\beta^T, \mu)^T \in \mathbb{R}_+^n \times \mathbb{R} : \|D\theta\|_\infty \leq r_n\}$, for the reduced design matrix D defined above and a rate r_n.

We make definitions completely analogous to the case in which we observe covariates. We adapt the definitions of the excess risk $\mathcal{E}(\theta)$ in the canonical way by letting the components corresponding to γ and Z_{ij} equal zero. We define the best local approximation θ^* as

$$\theta^* = \underset{\theta \in \Theta_{\text{loc}}}{\arg\min}\, \mathcal{E}(\theta)$$

and as before, we assume that all unpenalized parameters, i.e. μ^* in this case, are active. Since the sparsity assumptions of our parameter only concern β, it is natural that we should need the same assumptions on s_+^* as before, most notably Assumption 2.2. We have the analogue to Theorem 2.4.

Theorem 2.6. *Assume Assumption 2.2. Fix a confidence level t and let*

$$a_n = \sqrt{\frac{\log(2(n+1))}{\binom{n}{2}}}$$

and

$$\lambda_0 = 8a_n + 2\sqrt{\frac{t}{\binom{n}{2}}(9 + 8\sqrt{2n}a_n)} + \frac{2\sqrt{2}t\sqrt{n}}{3\binom{n}{2}}.$$

Let $\bar{\lambda} = \frac{\sqrt{n}}{\sqrt{2}}\lambda \geq 8\lambda_0$ and define η and K_n as in (2.5) with κ set to zero. Then, with probability at least $1 - \exp(-t)$ we have

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{\sqrt{2}}{\sqrt{n}}\|\hat{\beta} - \beta^*\|_1 + |\hat{\mu} - \mu^*|\right) \leq 6\mathcal{E}(\theta^*) + 4s_+^* K_n \bar{\lambda}^2. \tag{2.6}$$

A proof, which follows almost immediately from the case in which we do observe covariates, is given in Section 2.7.1.8. It is interesting to put this result into context by comparing it with Theorem 2 in Chen et al. (2020). The parameter space over which Chen et al. (2020) are optimizing is not convex and the analogous notion of best local approximation we are using need not be well-defined in their setting. Thus, it is not possible to derive ℓ_1-error bounds for their estimator, as we do in Theorem 2.6. Nonetheless and quite remarkably, they are able to prove an existence criterion for their ℓ_0-constrained estimator and a high-probability, finite sample bound on its excess risk. To compare their results to ours, we consider a special case that they discuss at length. They consider the situation in which $\mu_0 = -\xi \cdot \log(n) + O(1)$ for some $\xi \in [0, 2)$ and $\beta_{0,i} = \alpha \cdot \log(n) + O(1)$ for some $\alpha \in [0, 1)$ and all $i \in S_0$. It is easy to see that under these assumptions we have $\rho_{n,0} \sim n^{-\xi}$. Consider the regime in which no approximation error is committed. Then, using an analogous argument as in the proof of Corollary 2.5, K_n is of order $\rho_{n,0}^{-1}$. Recalling Assumption 2.2, we see that to obtain ℓ_1-consistency of our estimator, we need $\xi < 1/2$, which restricts the degree of network sparsity that our estimator can handle. Chen et al. (2020) need no such condition and only need to balance the global sparsity parameter ξ with the local density parameter α, by imposing $0 \leq \xi - \alpha < 1$, to have convergence of their excess risk to zero. This illustrates that to obtain our more refined consistency result in terms of ℓ_1-error, we understandably need to impose stricter assumptions on the permissible sparsity. We now compare the bounds on the excess risk. Note that Chen et al. (2020) scale their excess risk by $\mathbb{E}[d_+]^{-1} \sim n^{-2+\xi}$, rather than $\binom{n}{2} \sim n^2$ as we do. To put the excess risk on the same scale, we denote by $\mathcal{E}^{(\mathrm{r})}(\hat{\theta}) = n^{\xi}\mathcal{E}(\hat{\theta})$ the excess risk rescaled to their setting. With this notation, we see that by Theorem 2.6 the error rate for the rescaled excess risk of our ℓ_1-constrained estimator becomes

$$\mathcal{E}^{(r)}(\hat{\theta}) = O_P(s_+^* \cdot \log(n) \cdot n^{-2+2\xi}),$$

which by Assumption 2.2 is $o_p(\sqrt{\log(n)} \cdot n^{-3/2+\xi})$. From Chen et al. (2020), Theorem 2, it is seen that the rate for the excess risk of their ℓ_0-constrained estimator is

$$O_P(\log(n) \cdot n^{-1+\xi/2}).$$

Thus, in the regime $\xi \in [0, 1/2)$ necessary for ℓ_1-consistent parameter estimation, our estimator will always achieve a rate faster than the one in Chen et al. (2020). When we leave this regime, however, consistent estimation with respect to the ℓ_1-norm may no longer be possible and the estimator in Chen et al. (2020) can outperform our estimator with regards to the excess risk.

2.4 Inference for the homophily parameter

We derive the asymptotic normality of our estimator $\hat{\gamma}$ when $\theta^* = \theta_0$. We will see that the same arguments used for deriving the limiting distribution for $\hat{\gamma}$ also work for $\hat{\mu}$ and as a by-product we also obtain an analogous result for $\hat{\mu}$.

Our strategy will be inverting the Karush-Kuhn-Tucker (KKT) conditions, similar to van de Geer et al. (2014). The estimation in (2.4) is a convex optimization problem. Hence, by subdifferential calculus, we know 0 has to be contained in the subdifferential of $\frac{1}{\binom{n}{2}}\mathcal{L}(\theta) + \lambda\|\beta\|_1$ at $\hat{\theta}$. That is, there exists some $v \in \mathbb{R}^{n+1+p}$ such that

$$0 = \frac{1}{\binom{n}{2}} \nabla \mathcal{L}(\theta)|_{\theta=\hat{\theta}} + \lambda v, \tag{2.7}$$

where $\nabla \mathcal{L}(\theta)|_{\theta=\hat{\theta}}$ is the gradient of $\mathcal{L}(\theta)$ evaluated at $\hat{\theta}$ and for $i = 1, \ldots, n, v_i = 1$ if $\hat{\beta}_i > 0$ and $v_i \in [-1, 1]$ if $\hat{\beta}_i = 0$, and $v_i = 0$ for $i = n+1, \ldots, n+1+p$.

To ease notation a little we write $\vartheta = (\mu, \gamma^T)^T$. Thus, denoting $\nabla_\vartheta \mathcal{L}(\theta)|_{\theta=\hat{\theta}} \in \mathbb{R}^{p+1}$ the gradient of $\mathcal{L}$ with respect to the unpenalized parameters $(\mu, \gamma^T)^T$ only, evaluated at $\hat{\theta}$, we have

$$0 = \nabla_\vartheta \mathcal{L}(\theta)|_{\theta=\hat{\theta}}\,. \tag{2.8}$$

Denote by $H(\hat{\theta}) \coloneqq H_{\vartheta\times\vartheta}(\theta)|_{\theta=\hat{\theta}}$ the Hessian of $\frac{1}{\binom{n}{2}}\mathcal{L}(\theta)$ with respect to ϑ only, evaluated at $\hat{\theta}$. Denote $p_{ij}(\theta) = \frac{\exp(\beta_i+\beta_j+\mu+\gamma^T Z_{ij})}{1+\exp(\beta_i+\beta_j+\mu+\gamma^T Z_{ij})}$. Consider the entries of $H(\hat{\theta})$. For all $k, l = 1, \ldots, (p+1)$,

$$H(\hat{\theta})_{k,l} = \frac{1}{\binom{n}{2}} \partial_{\vartheta_k\vartheta_l} \mathcal{L}(\hat{\theta}) = \frac{1}{\binom{n}{2}} \sum_{i<j} D_{ij,n+k} D_{ij,n+l} p_{ij}(\hat{\theta})(1 - p_{ij}(\hat{\theta})),$$

where D_{ij}^T is the (i,j)-th row of the design matrix D. In particular $D_{ij,n+k} = 1$ if $k = 1$ and $D_{ij,n+k} = Z_{ij,k-1}$ for $k = 2, \ldots, (p+1)$. We have the following matrix representation of $H(\hat{\theta})$. Let $D_\vartheta = [\mathbf{1}|Z]$ be the part of D corresponding to ϑ with

rows $D_{\vartheta,ij}^T = (1, Z_{ij}^T), i < j$. Let $\hat{W} = \text{diag}(\sqrt{p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))}, i < j) \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}$. Then

$$H(\hat{\theta}) = \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 D_\vartheta.$$

Let $W_0 = \text{diag}(\sqrt{p_{ij}(\theta_0)(1-p_{ij}(\theta_0))}, i < j)$ and define the population version:

$$\mathbb{E}[H(\theta_0)] = \frac{1}{\binom{n}{2}} \mathbb{E}[D_\vartheta^T W_0^2 D_\vartheta].$$

To be consistent with commonly used notation (e.g. van de Geer et al. (2014)), call $\hat{\Sigma}_\vartheta = H(\hat{\theta})$ and $\Sigma_\vartheta = \mathbb{E}[H(\theta_0)]$ and $\hat{\Theta}_\vartheta \coloneqq \hat{\Sigma}_\vartheta^{-1}, \Theta_\vartheta \coloneqq \Sigma_\vartheta^{-1}$.

We will need to invert $\hat{\Sigma}_\vartheta$ and Σ_ϑ and show that these inverses are close to each other in an appropriate sense. It is commonly assumed in LASSO theory (cf. van de Geer et al. (2014)) that the minimum eigenvalues of these matrices stay bounded away from zero. In our case, however, such an assumption is invalid.

Indeed, using $\rho_n \leq 1/2$, we find that for all $i < j$, $p_{ij}(\theta_0)(1-p_{ij}(\theta_0)) \geq 1/2 \cdot \rho_n$. Furthermore, by Assumption 2.1, the minimum eigenvalue $\lambda_{\min}$ of $\mathbb{E}[Z^T Z/\binom{n}{2}]$ stays uniformly bounded away from zero for all n. Then, for any $n \in \mathbb{N}$ and $v \in \mathbb{R}^{p+1}\backslash\{0\}$ with components $v = (v_1, v_R^T)^T, v_1 \in \mathbb{R}, v_R \in \mathbb{R}^p$, we have

$$\begin{aligned} v^T \Sigma_\vartheta v &\geq \frac{1}{2}\rho_n v^T \frac{1}{\binom{n}{2}} \mathbb{E}[D_\vartheta^T D_\vartheta] v = \frac{1}{2}\rho_n v^T \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \frac{1}{\binom{n}{2}}\mathbb{E}[Z^T Z] \end{pmatrix} v \\ &= \frac{1}{2}\rho_n \left(v_1^2 + v_R^T \frac{1}{\binom{n}{2}} \mathbb{E}[Z^T Z] v_R \right) \\ &\geq \frac{1}{2}\rho_n (v_1^2 + \lambda_{\min}\|v_R\|_2^2) \geq \frac{1}{2}\rho_n (1 \wedge \lambda_{\min})\|v\|_2^2 > 0. \end{aligned}$$

Hence, for finite n all eigenvalues of Σ_ϑ are strictly positive and consequently Σ_ϑ is invertible. However, since we allow $\rho_{n,0} \to 0$, we are unable to achieve a strictly positive lower bound that is uniform in n.

Using similar techniques as in the proof of Proposition 2.3 in Section 2.7.1.2 we can show that with high probability the minimum eigenvalue of $D_\vartheta^T D_\vartheta/\binom{n}{2}$ is also strictly larger than zero and thus for any $v \in \mathbb{R}^{p+1}\backslash\{0\}$ and any finite n (the exact derivations are given in Section 2.7.2.1),

$$\frac{1}{\binom{n}{2}} v^T D_\vartheta^T \hat{W}^2 D_\vartheta v \geq C\rho_n \text{mineval}\left(\frac{1}{\binom{n}{2}} Z^T Z\right) \|v\|_2^2 > 0.$$

Thus, for every finite n, $\hat{\Sigma}_\vartheta$ is invertible with high probability. Since these lower bounds tend to zero with increasing n, a careful argument is needed and we have to impose stricter assumptions than for our consistency result alone.

Assumption 2.3. $s_+^* \frac{\sqrt{\log(n)}}{\sqrt{n}\rho_n^2} \to 0, n \to \infty$.

Assumption 2.3 is a slightly stricter version of the previously imposed Assumption 2.2. Previously we only needed a factor of $1/\rho_n$ to ensure that the ℓ_1-error for $\hat{\beta}$ in Theorem 2.13 goes to zero. Notice, though, that Assumption 2.3 still allows sparsity rates for $\rho_{n,0}$ of small polynomial order. More precisely, up to a log-factor and depending on the rate of s_+^*, $\rho_{n,0}$ may still go to zero at a speed of order up to $n^{-1/4}$.

Theorem 2.7. *Under Assumptions 2.1 and 2.3, when $\theta^* = \theta_0$ and with λ chosen as in Theorem 2.4, we have for any $k = 1, \ldots, p$, as $n \to \infty$,*

$$\sqrt{\binom{n}{2}} \frac{\hat{\gamma}_k - \gamma_{0,k}}{\sqrt{\hat{\Theta}_{\vartheta,k+1,k+1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

We also have for our estimator of the global sparsity parameter, $\hat{\mu}$, as $n \to \infty$,

$$\sqrt{\binom{n}{2}} \frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\Theta}_{\vartheta,1,1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Theorem 2.7 does not require a debiasing of the parameters $\hat{\mu}$ and $\hat{\gamma}$. It is well known in the LASSO literature, that using an ℓ_1-penalized likelihood approach will produce biased estimates for the penalized parameters due to the shrinking of the parameter estimates enforced by the penalty (c.f. Section 1.4). Due to this bias, it is generally not possible to derive limiting distributions for LASSO type estimators and a debiasing procedure is needed in order derive distributional limit results (van de Geer et al. 2014). This bias is made explicit in equation (2.7): The penalized parameter values do not fulfil the first-order estimating equations exactly, but rather a bias of the form λv is incurred as prescribed by subdifferential calculus. While the unpenalized parameter estimates $\hat{\vartheta} = (\hat{\mu}, \hat{\gamma}^T)^T$ do fulfil the first-order estimating equations exactly, in standard settings this alone would still not be enough to ensure the asymptotic normality of $\hat{\vartheta}$. However, in our special case, this is enough to allow us to derive a limiting distribution without a debiasing step. More precisely, deriving the limiting distribution of ϑ relies on a Taylor expansion of the negative log-likelihood $\mathcal{L}$. To derive Theorem 2.7 it is necessary that in said Taylor expansion the bias incurred from the part of the likelihood relating to β vanishes in probability. This essentially is a condition on the asymptotic correlation between the columns of the design matrix D corresponding to β and those corresponding to μ and γ. Due to each column in X – the deterministic part of the design matrix relating to β – being very sparse and having only $n-1$ non-zero entries, this bias vanishes fast enough to allow Theorem 2.7. The exact details of this are given in Section 2.7.2.5.

2.5 Simulation: SβM-C

We illustrate the finite sample performance of our estimator (2.4) with an extensive set of Monte Carlo simulations. We only show results for SβM-C and the estimator (2.4), as – where applicable – the results in the case without covariates are very similar. We check the ℓ_1-convergence of our parameter estimates to the true parameter, as well as the asymptotic normality of $\hat{\gamma}$.

Since our estimation involves the choice of a tuning parameter, we explored the use of the Bayesian Information Criterion (BIC) as well as a heuristic based on the theory developed in the previous sections for model selection. While the former criterion is purely data-driven, the use of latter is to ensure that our theoretical results are about right in terms of the rates. To make the dependence of our estimator (2.4) on the penalty parameter explicit, we denote the solution of (2.4) when using penalty λ by $\hat{\theta}(\lambda) = (\hat{\beta}(\lambda)^T, \hat{\mu}(\lambda), \hat{\gamma}(\lambda)^T)^T$ and write $s(\lambda) = |\{i : \hat{\beta}_i(\lambda) > 0\}|$ for its sparsity. The value of the BIC at λ is given by

$$\text{BIC} = 2\mathcal{L}(\hat{\theta}(\lambda)) + s(\lambda)\log(n(n-1)/2)$$

and the penalty λ was chosen to minimize BIC.

To motivate the heuristic approach to tuning parameter selection, recall that Theorem 2.4 suggests that based on a confidence level t picked by us, we should first define a λ_0. The consistency results derived hold for any $\bar{\lambda} \geq 8\lambda_0$, where $\bar{\lambda}$ is the penalty in the rescaled penalized likelihood problem, which relates to the penalty λ in the original penalized problem (2.4) as $\bar{\lambda} = \sqrt{n}/\sqrt{2} \cdot \lambda$. Looking at the proof of Theorem 2.4, we see that the factor eight in the relation between λ_0 and $\bar{\lambda}$ is a technical artefact we had to introduce to prove that the sample size adjusted estimator $\hat{\bar{\theta}}$ as defined in Section 2.7.1.3 was close enough to the sample size adjusted best local approximation $\bar{\theta}^*$ (c.f. Section 2.7.1.6). If we assume that our estimator is close enough to the truth, we may ignore that factor and set $\lambda = \frac{\sqrt{2}}{\sqrt{n}}\lambda_0$. We pick $t = 2$ and set c to the maximum observed covariate value. It is known that in high-dimensional settings the penalty values prescribed by mathematical theory in practice tend to over-penalize the parameter values, see, for example, Yu et al. (2019). Decreasing the penalty by removing the factor eight is thus in line with these empirical findings.

We fixed $p = 2$, set $\gamma_0 = (1, 0.8)^T$ and generated the covariates from a centred Beta $(2, 2)$ distribution as $Z_{ij,k} \sim \text{Beta}(2, 2) - 1/2$. We considered networks of sizes $n = 300, 500, 800$ and 1,000 in which the sparsity of β_0 is set as 7, 9, 10, and 12 respectively. We tested our estimator on three different model configurations with different combinations of β_0 and μ_0, resulting in networks with varying de-

grees of sparsity. For each simulation configuration, 1,000 data sets were simulated. Specifically,

Model 1: We picked $\beta_0 = (1.2, 0.8, 1, \ldots, 1, 0, \ldots, 0)^T$, where the number of ones increases with the network size to match the aforementioned sparsity level, and set $\mu_0 = -0.5\log(\log(n))$;

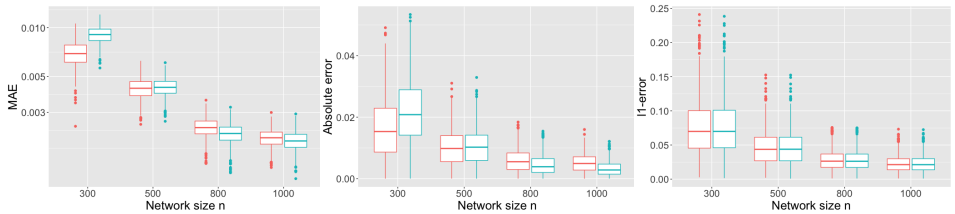
Model 2: $\beta_0 = \log(\log(n)) \cdot (1.2, 0.8, 1, \ldots, 1, 0, \ldots, 0)^T$ and $\mu_0 = -1.2 \cdot \log(\log(n))$;

Model 3: $\beta_0 = \log(\log(n)) \cdot (2, 0.8, 1, \ldots, 1, 0, \ldots, 0)^T$ and $\mu_0 = -0.5 \cdot \log(n)$.

In these three models, we allow μ_0 to get progressively more negative to generate networks that are increasingly sparse, and allow the sparsity of β to increase with network size n. The median edge density and the minimum and maximum link probabilities p_{ij} for each model and network size are reported in Table 2.1. All three models get progressively sparser with increasing n. Model 3 gives the sparsest networks when $n = 1{,}000$, with only around 3.6% of all possible edges being present on average.

	n	Median edge density	$\min p_{ij}$	$\max p_{ij}$
Model 1				
	300	0.309	0.145	0.903
	500	0.298	0.140	0.899
	800	0.288	0.136	0.896
	1000	0.285	0.134	0.894
Model 2				
	300	0.127	0.048	0.933
	500	0.115	0.043	0.938
	800	0.103	0.040	0.943
	1000	0.099	0.038	0.944
Model 3				
	300	0.068	0.023	0.963
	500	0.052	0.018	0.964
	800	0.040	0.014	0.963
	1000	0.036	0.013	0.962

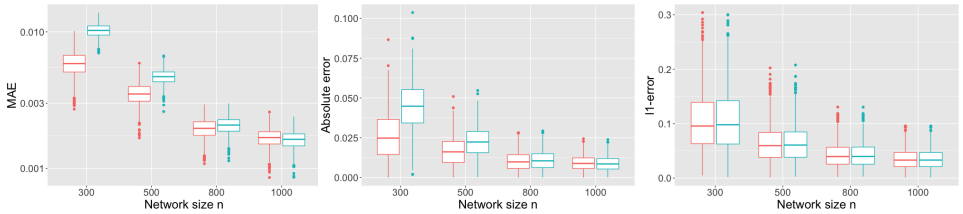
Table 2.1: Network summary statistics for networks sampled from models 1 - 3



(a) MAE for β_0 **(b)** Absolute error for μ_0. **(c)** ℓ_1-error for γ_0.

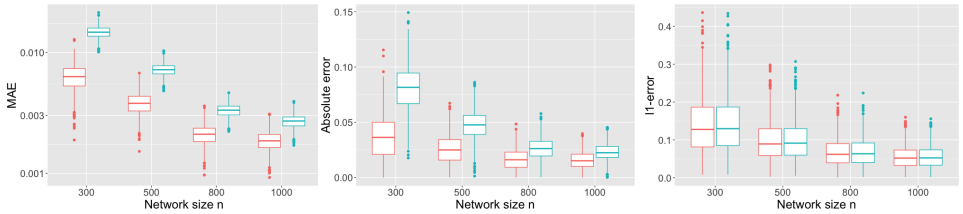
Figure 2.2: Errors for estimating θ_0 in Model 1 across various network sizes and 1,000 repetitions. Comparison between model selection via BIC and a heuristic approach. The results for BIC are displayed in red (left boxes), those for the pre-determined λ in green (right boxes).

Consistency. We calculated the mean absolute error (MAE) for estimating β_0, the absolute error for estimating μ_0 and the ℓ_1-error for estimating γ_0. For Model 1 the results are shown in Figures 2.2a–2.2c. While BIC performs slightly better for estimating β_0 and μ_0 for smaller network sizes, our heuristic performs better



(a) MAE for β_0 **(b)** Absolute error for μ_0. **(c)** ℓ_1-error for γ_0.

Figure 2.3: Errors for estimating θ_0 in Model 2. BIC in red (left boxes), heuristic in green (right boxes).



(a) MAE for β_0 **(b)** Absolute error for μ_0. **(c)** ℓ_1-error for γ_0.

Figure 2.4: Errors for estimating θ_0 in Model 3. BIC in red (left boxes), heuristic in green (right boxes).

for larger network sizes. The ℓ_1-error for estimating γ is almost the same between both model selection schemes across all network sizes. For both methods we see that the errors decrease with increasing network size. Model 2 gives similar results with slightly smaller errors produced by BIC for β_0 and μ_0 for smaller networks and similar or slightly better errors produced by the heuristic for large networks (Figures 2.3a, 2.3b). The error for γ_0 is similar between both methods (Figure 2.3c). For Model 3, the errors for parameter estimation are shown in Figures 2.4a – 2.4c. The errors are generally larger than in the other network models, which is to be expected due to the much higher sparsity of the network. For this very sparse case, BIC is performing better than the heuristic. The heuristic consistently selects higher penalty values than BIC and we can see how this results in worse estimates for very sparse networks. Also, for the heuristic we choose one predefined penalty value for any network of a given size n, while BIC can adapt to the observed sparsity. This illustrates the point made by Yu et al. (2019), that the penalty prescribed by mathematical theory tends to over-penalize the model. It is to be noted, though, that even in this very sparse regime both model selection techniques produce reasonable estimates that are close to the truth.

Asymptotic normality. We calculated the standardized γ-values

$$\sqrt{\binom{n}{2}} \frac{\hat{\gamma}_k - \gamma_{0,k}}{\sqrt{\hat{\Theta}_{\vartheta,k+1,k+1}}}, \quad k = 1, 2,$$

which by Theorem 2.7 asymptotically follow a $\mathcal{N}(0, 1)$ distribution. This allowed us

to construct approximate 95%-confidence intervals for $\gamma_{0,k}$ as

$$CI_k = \left(\hat{\gamma}_k - z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\Theta}_{\vartheta,k+1,k+1}}{\binom{n}{2}}}, \hat{\gamma}_k + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\Theta}_{\vartheta,k+1,k+1}}{\binom{n}{2}}} \right), \ k = 1, 2,$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard-normal distribution and we used $\alpha = 0.05$. We present the empirical coverage of these intervals and their median length for the different network sizes. Table 2.2 shows the results for $\gamma_{0,1}$ across the different models and sample sizes. The results for $\gamma_{0,2}$ are similar and are omitted to save space. The coverage is very close to the 95%-level across all network sizes and all models and independent of which model selection criterion we use. This empirically illustrates the validity of the asymptotic results derived in Theorem 2.7. The median length of the confidence interval decreases with increasing network size and is similar between BIC and the heuristic. This is what we would expect since the estimates for γ_0 are very similar between both methods as shown in Figures 2.2c, 2.3c, and 2.4c. Comparing the length of the confidence intervals between Models 1, 2 and 3, we see that as the models become sparser, the median length increases, which is also to be expected.

	n	Coverage	CI	Coverage	CI
		Pre-determined λ		BIC	
Model 1					
	300	0.949	0.182	0.950	0.182
	500	0.944	0.110	0.944	0.110
	800	0.953	0.069	0.954	0.069
	1000	0.945	0.056	0.942	0.056
Model 2					
	300	0.927	0.251	0.937	0.252
	500	0.958	0.158	0.961	0.158
	800	0.940	0.103	0.940	0.103
	1000	0.945	0.083	0.947	0.083
Model 3					
	300	0.931	0.333	0.939	0.335
	500	0.937	0.225	0.942	0.226
	800	0.939	0.159	0.942	0.159
	1000	0.941	0.133	0.942	0.134

Table 2.2: Empirical coverage under nominal 95% coverage and median lengths of confidence intervals.

2.6 Data analysis

We illustrate our results by applying our estimator to two real world data sets.

Lazega's lawyer friendship data. We already encountered this dataset in Chapter 1. As a reminder, the 71 lawyers of a law firm were asked to indicate with whom in the firm they regularly socialized outside of work. This is a frequently

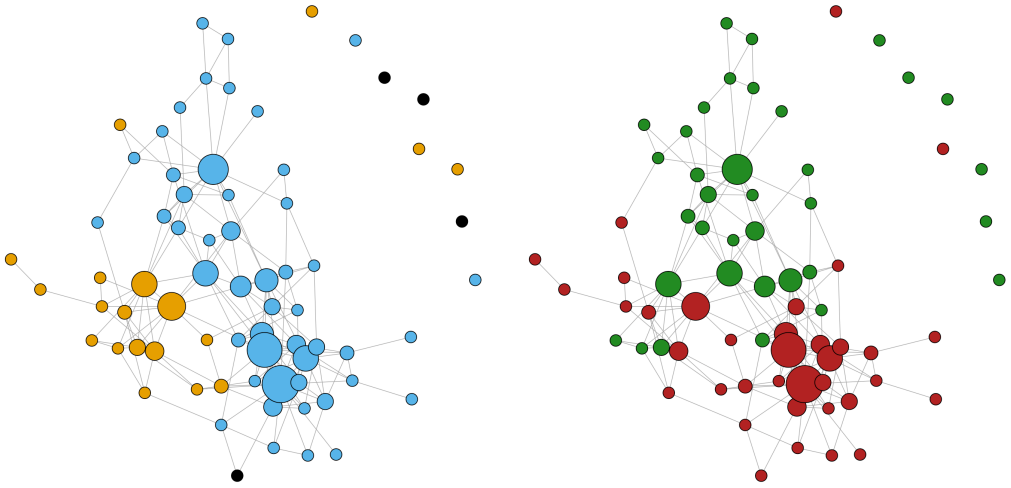
used network data set that was also analysed, for example, in Snijders et al. (2006), Jochmans (2018) and Yan et al. (2019). For our analysis we focus on mutual friendships between lawyers as in Snijders et al. (2006), that is, we place an undirected edge between two lawyers when they both indicated to socialize with one another. The degrees of the resulting network range from 0 to 16, with eight isolated nodes. The average degree is 4.96 and the edge density is 7%. It is to note that we did not remove the isolated nodes before doing inference. Alongside the network, the following variables were collected: The status of the lawyer (partner or associate), their gender (man or woman), which of three offices they worked in, the years they had spent with the firm, their age, their practice (litigation or corporate) and the law school they had visited (Harvard and Yale, UConn or other).

We fitted the SβM-C to this data set, by using as covariates between two nodes the positive absolute difference between these seven variables, where for categorical variables the difference is defined as the indicator whether the values are equal. Since our simulations suggest that BIC performs better for smaller networks, we used it for model selection. Model selection with the heuristic results in a slightly larger penalty and slightly different estimates, but overall very similar results. We constructed confidence intervals for the estimated covariate values at the 95%-level. The estimates and confidence intervals for the covariates are shown in Table 2.3.

Covariate	Point estimate	Confidence Interval
Same status	0.91	$(0.54, 1.28)$
Same gender	0.46	$(0.12, 0.81)$
Same office	2.21	$(1.81, 2.60)$
Difference years with firm	-0.073	$(-0.11, -0.04)$
Difference age	-0.031	$(-0.060, -0.002)$
Same practice	0.57	$(0.25, 0.89)$
Same law school	0.30	$(-0.03, 0.62)$

Table 2.3: Covariate weights for Lazega's lawyer friendship network and 95% confidence intervals.

In terms of magnitude of estimated weight as well as, more importantly, the sign of each weight, these findings are in line with what we would expect and with the results in the aforementioned papers. In order of importance, working in the same office, having the same status, being of the same practice and having the same gender have a positive effect on friendship formations, whereas a big difference in tenure or age has a negative effect on friendship formation. While our point estimate for having gone to the same law school is positive, its confidence interval extends to the negative real line and we thus cannot make a definite statement about its effect on friendship formation. This effect is also present when doing model selection with our heuristic. To appreciate how the covariates influence the connection pattern, we visualize the network in Figure 2.5 by examining the effect of office in Figure 2.5a and that of status in Figure 2.5b respectively. We can see indeed that these two



(a) Lawyer network by office. **(b)** Lawyer network by status.

Figure 2.5: Lazega's friendship network among 71 lawyers. The size of the nodes is proportional to their degree. For better visibility we set the size of all nodes with a degree of five or lower to the size corresponding to a degree of five. In 2.5a the different colours indicate different offices (blue: Boston, yellow: Hartford, black: Providence; only four lawyers are based in the small Providence office) and in 2.5b different statuses (red: partner, green: associate). The positions of the vertices are the same in both plots.

covariates have played important roles in shaping how connections were made.

Trade partnerships network. For our second data set, we analysed mutually important trade partnerships between 136 countries/regions in 1990. This data was originally analysed by Silva & Tenreyro (2006) and further analysed in Jochmans (2018). Even back in 1990 almost every country would trade with every other country, which is why we focus only on those trade partnerships in which the trade volume exceeds a certain limit. More precisely, we place an undirected edge between two countries if the trade volume makes up at least 3% of the importing countries total imports or if it makes up at least 3% of the exporting countries total exports. This leaves us with an undirected network with 136 nodes and 1,279 edges (edge density of 13.9%). The minimum degree of the resulting network was 3 (Dominican Republic), the maximum degree was 126 (USA), and the median degree was 13. We visualize the resulting network in Figure 2.6.

We analyse the same covariates as Jochmans (2018). That is, we have indicator variables common language and common border that take the value one if countries i and j share a common language or border and zero otherwise, log distance which is the log of the geographic distance between the countries, colonial ties which is one if at some point i colonized j or vice versa and zero otherwise, and preferential trade agreement which is an indicator whether or not a preferential trade agreement exists between the countries. Again, we chose BIC for model selection for the reasons outlined above. The results are summarized in Table 2.4. These results are in line with what one would expect. Having a preferential trade agreement has the strongest

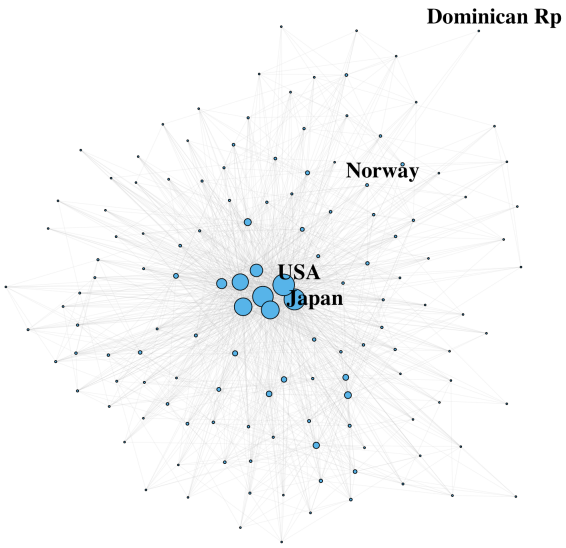


Figure 2.6: The world trade network in 1990 between 136 countries/regions. The size of the nodes is proportional to their degree. For better visibility we set the size of all nodes with a degree of 10 or lower to the size corresponding to a degree of 10. The six large, highly connected nodes in the middle correspond to: USA, Japan, Germany, France, United Kingdom and Italy.

positive effect on mutual trade between countries. Speaking the same language, sharing a border or having colonial ties also has a positive effect, while a large geographical distance has a strong negative effect.

Covariate	Estimated weight	Confidence Interval
Log distance	-1.03	$(-1.04, -1.02)$
Common border	0.45	$(0.10, 0.79)$
Common language	0.31	$(0.09, 0.54)$
Colonial ties	0.42	$(0.17, 0.66)$
Preferential trade agreement	0.81	$(0.36, 1.27)$

Table 2.4: Covariate estimation for world trade data and 95% confidence intervals.

The confidence intervals for the categorical variables are all much larger than the one for the continuous variable log distance between countries. This is due to the fact that all the columns corresponding to categorical covariates are quite sparse, while the column corresponding to log distance contains only non-zero entries. Out of 9,180 dyads only 142 are part of a preferential trade agreement and only 180 share a common border. Consequently the confidence intervals corresponding to these covariates are largest. Furthermore, 1,565 node pairs have colonial ties with one another and 1,925 speak a common language. While the columns corresponding to these covariates are thus much more populated, they are still relatively sparse when compared to the total number of dyads.

BIC selected 32 active β-entries, which are visualized on a map in Figure 2.7. We presented the top half of these countries/regions with their degree and GDP in Table 2.5. The ranking of the β values correlates with our intuition of the economic power of the countries. However, we also pick up underlying network formation mechanisms that go beyond sheer economic power and that are neither explainable

by only looking at network summary statistics (such as degree of a node) nor by only looking at economic metrics such as a country's GDP. More precisely, we note that the top six positions are occupied by six of the G7 countries, which serves to show that SβM-C works well for identifying the most important nodes in a network. Note however, that Japan has the largest β, albeit having a smaller degree (122) and a significantly smaller GDP than the USA (degree = 126), which comes in second place. In general, the order of degrees no longer aligns exactly with the order of the β-values as would have been predicted by the monotonicity lemma for the SβM without covariates in Chen et al. (2020). We observe this pattern for non-active β-values as well. Norway, for example, has a β-value of zero, even though its degree of 17 and GDP of US\$$1.22 \times 10^{11}$ exceeds the degree and the GDP of several nodes with positive β-value. Looking at Norway's neighbouring nodes, we see that it was trading mostly with countries that either are close geographically or have a large β-value themselves (such as USA and Japan), meaning that the observed covariates are sufficient to explain the linking behaviour of Norway. This illustrates that the SβM-C is able to pick up subtleties in network formation that one might miss if one relied solely on network summary statistics such as the degree of a node or solely on non-relational summary statistics such as a country's GDP.

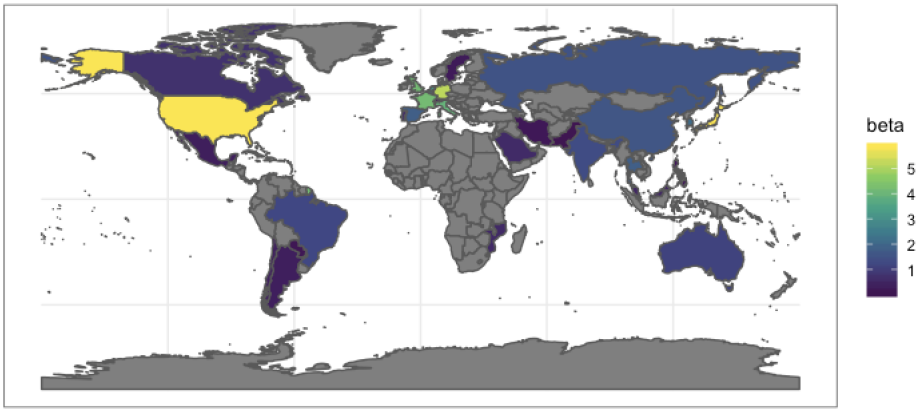


Figure 2.7: World map of the estimated β values in the world trade network in 1990 between 136 countries. The colour of the country corresponds to the magnitude of the estimated β. Countries in grey either have an estimated β value of zero or were not present in the data set

Country/Region	$\hat{\beta}$	Degree	GDP (US$)
Japan	5.85	122	4.95e+12
USA	5.82	126	6.51e+12
Germany	5.17	120	2.27e+12
France	4.16	103	1.47e+12
UK	4.15	104	1.04e+12
Italy	3.92	95	1.03e+12
Netherlands	3.15	73	3.75e+11
Belgium-Lux	2.60	59	2.56e+11
Korea Republic	2.11	34	3.42e+11
Singapore	2.06	37	5.39e+10
Hong Kong	2.05	40	1.07e+11
Spain	1.80	41	5.46e+11
Thailand	1.78	33	1.11e+11
China	1.58	30	3.98e+11
Russian Federation	1.53	28	5.43e+11
India	1.33	32	2.75e+11

Table 2.5: The top 16 active β values for the world trade network.

2.7 Proofs of Chapter 2

2.7.1 Proof of consistency results

2.7.1.1 Proof of Lemma 2.1

Proof of Lemma 2.1. We first show that a solution exists. Using duality theory from convex optimization (cf. Bertsekas (1995), Chapter 5), we know that for any $\lambda > 0$ there exists a finite $s > 0$ such that the penalized likelihood problem is equivalent to the primal optimization problem

$$\begin{aligned} &\min_{\beta,\mu,\gamma} \frac{1}{\binom{n}{2}} \mathcal{L}(\beta, \mu, \gamma), \\ &\text{subject to: } (\beta^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}}, \sum_{i=1}^{n} |\beta_i| \leq s. \end{aligned} \tag{2.9}$$

Let $\beta = (\beta_1, \ldots, \beta_n)^T, \gamma = (\gamma_1, \ldots, \gamma_p)^T$ be fixed. To obtain an estimate for μ, we minimize the function

$$\begin{aligned} g_{\beta,\gamma}(\mu) &= \frac{1}{\binom{n}{2}} \mathcal{L}(\beta, \mu, \gamma) \\ &= \frac{1}{\binom{n}{2}} \Bigg(- \sum_{i=1}^{n} \beta_i d_i - d_+ \mu - \sum_{i<j} (Z_{ij}^T \gamma) A_{ij} \\ &\qquad + \sum_{i<j} \log(1 + \exp(\beta_i + \beta_j + \mu + Z_{ij}^T \gamma)) \Bigg). \end{aligned}$$

It has derivative

$$g'_{\beta,\gamma}(\mu) = \frac{1}{\binom{n}{2}} \left(-d_+ + \sum_{i<j} \frac{e^{\beta_i + \beta_j + \mu + Z_{ij}^T \gamma}}{1 + e^{\beta_i + \beta_j + \mu + Z_{ij}^T \gamma}} \right).$$

We observe that

$$\lim_{\mu \to \infty} g'_{\beta,\gamma}(\mu) = \frac{1}{\binom{n}{2}} \left(-d_+ + \binom{n}{2} \right) > 0$$

and

$$\lim_{\mu \to -\infty} g'_{\beta,\gamma} = -d_+ \frac{1}{\binom{n}{2}} < 0.$$

Furthermore, $g'_{\beta,\gamma}$ is continuous and strictly increasing in μ. Hence, there exists a unique value $\mu^* = \mu^*(\beta, \gamma)$, such that $g'_{\beta,\gamma}(\mu^*) = 0$. Since

$$g''_{\beta,\gamma}(\mu) = \sum_{i<j} \frac{e^{\beta_i + \beta_j + \mu + Z_{ij}^T \gamma}}{(1 + e^{\beta_i + \beta_j + \mu + Z_{ij}^T \gamma})^2} > 0$$

for all μ, μ^* is a minimizer of $g_{\beta,\gamma}$. Since $g''_{\beta,\gamma}$ is invertible, we can apply the implicit function theorem with function $F(\beta, \gamma, \mu) = g'_{\beta,\gamma}(\mu)$, which gives us that the corre-

sponding function $\mu^* = \mu^*(\beta, \gamma)$ is continuously differentiable. Plugging in $\mu^*(\beta, \gamma)$ for μ in (2.9), we are left with the minimization problem

$$\begin{aligned} &\min_{\beta,\gamma} \mathcal{L}(\beta, \mu^*(\beta, \gamma), \gamma), \\ &\text{s.t.: } (\beta^T, \mu^*(\beta, \mu), \gamma^T)^T \in \Theta_{\text{loc}}, \sum_{i=1}^{n} |\beta_i| \leq s. \end{aligned} \tag{2.10}$$

Since Γ is compact, we are minimizing a continuous function over a compact set in (2.10). Hence it attains a minimum $\mathcal{L}^*$. By the definition of μ^*, $\mathcal{L}^*$ must also be a solution of (2.9).

For the second claim of the Lemma, suppose there is an $1 \leq i_0 \leq n$ such that $\hat{\beta}_{i_0} = \min_{1 \leq i \leq n} \hat{\beta}_i > 0$. Consider the following vector $\tilde{\theta} = (\tilde{\beta}^T, \tilde{\mu}, \tilde{\gamma}^T)^T$: for all k let $\tilde{\beta}_k = \hat{\beta}_k - \hat{\beta}_{i_0}$ and $\tilde{\mu} = \hat{\mu} + 2\hat{\beta}_{i_0}$, while keeping $\tilde{\gamma} = \hat{\gamma}$. Then, $\tilde{\beta}_k \geq 0$ for all k, i.e. $\tilde{\theta}$ is a feasible point for the penalized likelihood problem (2.4). Furthermore $\min_k \tilde{\beta}_k = 0$ and $\mathcal{L}(\tilde{\theta}) = \mathcal{L}(\hat{\theta})$. However,

$$\|\tilde{\beta}\|_1 = \sum_{i=1}^{n} |\hat{\beta}_i - \hat{\beta}_{i_0}| < \|\hat{\beta}\|_1,$$

where the inequality follows from the minimality of $\hat{\beta}_{i_0}$. This gives

$$\mathcal{L}(\tilde{\theta}) + \|\tilde{\beta}\|_1 < \mathcal{L}(\hat{\theta}) + \|\hat{\beta}\|_1.$$

A contradiction to the optimality of $\hat{\theta}$. □

2.7.1.2 Proof of Proposition 2.3

The compatibility condition is clearly equivalent to the condition that

$$\kappa^2(\Sigma, s^*) := \min_{\substack{\theta \in \mathbb{R}^{n+1+p} \setminus \{0\} \\ \|\theta_{S_+^{*c}}\|_1 \leq 3\|\theta_{S_+^*}\|_1}} \frac{\theta^T \Sigma \theta}{\frac{1}{s_+^*}\|\theta_{S_+^*}\|_1^2} \geq C > 0,$$

for all n and some universal C. Recall the definition of Σ_A. The key to proving Proposition 2.3 is to show that Σ is close to Σ_A in an appropriate sense and that Σ_A fulfils $\kappa^2(\Sigma_A, s^*) \geq C > 0$ for all n and some universal $C > 0$. We then show that $\kappa^2(\Sigma, s^*)$ is also bounded away from zero. Let us analyse the top left block matrix of Σ_A, i.e. $1/(n-1) \cdot X^T X$:

$$\frac{1}{n-1} X^T X = \begin{bmatrix} 1 & \frac{1}{n-1} & \frac{1}{n-1} & \cdots & \frac{1}{n-1} \\ \frac{1}{n-1} & 1 & \frac{1}{n-1} & \cdots & \frac{1}{n-1} \\ \vdots & \ddots & \ddots & \cdots & \vdots \\ \frac{1}{n-1} & \frac{1}{n-1} & \cdots & \cdots & 1 \end{bmatrix},$$

that is, $1/(n-1) \cdot X^T X$ has all ones on the diagonal and $1/(n-1)$ everywhere else. This is a special kind of Toeplitz matrix; a circulant matrix to be precise. It is known (c.f. Kra & Simanca (2012)), that every circulant matrix M has an associated polynomial p and that the eigenvalues of M are given by $p(\xi_j), j = 0, \dots, n-1$, where $\xi_j, j = 0, \dots, n-1$, denote the nth roots of unity, $\xi_j = \exp(\iota 2\pi j/n)$, where ι is the imaginary unit and $\xi_0 = 1$. The associated polynomial of the matrix $1/(n-1)X^T X$ is

$$p(x) = 1 + \frac{1}{n-1}(x + x^2 + \cdots + x^{n-1})$$

and thus the eigenvalues of $1/(n-1)X^T X$ are

$$p(1) = 2, \quad p(\xi_j) = 1 + \frac{1}{n-1}(-1) = \frac{n-2}{n-1}, \ j = 1, \dots, n-1,$$

where the eigenvalue $(n-2)/(n-1)$ has multiplicity $n-1$. Hence, for any vector $\theta = (\beta^T, \mu, \gamma^T)^T$,

$$\begin{aligned}
\theta^T \Sigma_A \theta &= \beta^T \left(\frac{1}{n-1} X^T X \right) \beta + \mu^2 + \frac{1}{\binom{n}{2}} \gamma^T \mathbb{E}[Z^T Z] \gamma \\
&\geq \frac{n-2}{n-1} \beta^T \beta + \mu^2 + \frac{1}{\binom{n}{2}} \gamma^T \mathbb{E}[Z^T Z] \gamma,
\end{aligned}$$

where for the inequality we have used that for any positive definite, symmetric matrix M with smallest eigenvalue λ and any vector $x \neq 0$ of appropriate dimension, we have $x^T M x \geq \lambda x^T x$. Thus, for any θ,

$$\begin{aligned}
\frac{\theta^T \Sigma_A \theta}{\frac{1}{s_+^*} \|\theta_{S_+^*}\|_1^2} &\geq \frac{\frac{n-2}{n-1} \beta^T \beta + \mu^2 + \frac{1}{\binom{n}{2}} \gamma^T \mathbb{E}[Z^T Z] \gamma}{\frac{1}{s_+^*} \|\theta_{S_+^*}\|_1^2} \\
&\geq \frac{\frac{n-2}{n-1} \|\beta\|_2^2 + \mu^2 + \frac{1}{\binom{n}{2}} \gamma^T \mathbb{E}[Z^T Z] \gamma}{\|\beta\|_2^2 + \mu^2 + \|\gamma\|_2^2}, \qquad \text{by Cauchy-Schwarz} \\
&\geq \frac{\frac{n-2}{n-1} (\|\beta\|_2^2 + \mu^2) + \frac{1}{\binom{n}{2}} \gamma^T \mathbb{E}[Z^T Z] \gamma}{\|\beta\|_2^2 + \mu^2 + \|\gamma\|_2^2}, \qquad \text{since } 1 \geq (n-2)/(n-1) \\
&= \frac{n-2}{n-1} \cdot \frac{\|\beta\|_2^2 + \mu^2 + \frac{n-1}{n-2} \frac{1}{\binom{n}{2}} \gamma^T \mathbb{E}[Z^T Z] \gamma}{\|\beta\|_2^2 + \mu^2 + \|\gamma\|_2^2}.
\end{aligned}$$

We now use that for any $a, b, c \in \mathbb{R}_+$, we have $\frac{a+b}{a+c} \geq \min\{1, b/c\}$. This is easily seen by considering the cases $\min\{1, b/c\} = 1$ and $\min\{1, b/c\} = b/c$ separately and rearranging. Thus,

$$\begin{aligned}
\frac{\theta^T \Sigma_A \theta}{\frac{1}{s_+^*} \|\theta_{S_+^*}\|_1^2} &\geq \frac{n-2}{n-1} \min \left\{ 1, \frac{n-1}{n-2} \frac{1}{\binom{n}{2}} \frac{\gamma^T \mathbb{E}[Z^T Z] \gamma}{\|\gamma\|_2^2} \right\} \\
&= \min \left\{ \frac{n-2}{n-1}, \frac{\gamma^T \left(\frac{1}{\binom{n}{2}} \mathbb{E}[Z^T Z] \right) \gamma}{\|\gamma\|_2^2} \right\} \geq \min \left\{ \frac{n-2}{n-1}, \lambda_{\min} \right\},
\end{aligned}$$

where $\lambda_{\min}$ is the minimum eigenvalue of $\frac{1}{\binom{n}{2}}\mathbb{E}[Z^T Z]$. By Assumption 2.1, for $n \geq 3$,

$$\kappa^2(\Sigma_A, s^*) = \min_{\substack{\theta \in \mathbb{R}^{n+1+p} \setminus \{0\} \\ \|\theta_{S_+^{*c}}\|_1 \leq 3\|\theta_{S_+^*}\|_1}} \frac{\theta^T \Sigma_A \theta}{\frac{1}{s_+^*}\|\theta_{S_+^*}\|_1^2} \geq c_{\min} > 0. \tag{2.11}$$

Now, we show that with high probability $\kappa(\Sigma, s^*) \geq \kappa(\Sigma_A, s^*)$, which implies that the compatibility condition holds with high probability for Σ. To that end, we have the following auxiliary Lemma found in Kock & Tang (2019). For completeness, we give the short proof of it. The notation is adapted to our setting.

Lemma 2.8 (Lemma 6 in Kock & Tang (2019))**.** *Let A and B be two positive semi-definite $(n+1+p) \times (n+1+p)$ matrices and $\delta = \max_{ij} |A_{ij} - B_{ij}|$. For any set $S^* \subset \{1, \ldots, n\}$ with cardinality s^*, one has*

$$\kappa^2(B, s^*) \geq \kappa^2(A, s^*) - 16\delta(s^* + p + 1).$$

Proof. Denote by $S_+^* = S^* \cup \{n+1, \ldots, n+1+p\}$ and $s_+^* = s^* + (1+p)$. Let $\theta \in \mathbb{R}^{n+1+p} \setminus \{0\}$, with $\|\theta_{S_+^{*c}}\|_1 \leq 3\|\theta_{S_+^*}\|_1$. Then,

$$\begin{aligned} |\theta^T A \theta - \theta^T B \theta| &= |\theta^T (A - B)\theta| \leq \|\theta\|_1 \|(A-B)\theta\|_\infty \leq \delta \|\theta\|_1^2 \\ &= \delta(\|\theta_{S_+^*}\|_1 + \|\theta_{S_+^{*c}}\|_1)^2 \leq \delta(\|\theta_{S_+^*}\|_1 + 3\|\theta_{S_+^*}\|_1)^2 \\ &\leq 16\delta \|\theta_{S_+^*}\|_1^2. \end{aligned}$$

Hence, $\theta^T B \theta \geq \theta^T A \theta - 16\delta \|\theta_{S_+^*}\|_1^2$ and thus

$$\frac{\theta^T B \theta}{\frac{1}{s_+^*}\|\theta_{S_+^*}\|_1^2} \geq \frac{\theta^T A \theta}{\frac{1}{s_+^*}\|\theta_{S_+^*}\|_1^2} - 16\delta s_+^* \geq \kappa^2(A, s^*) - 16\delta s_+^*.$$

Minimizing the left-hand side over all $\theta \neq 0$ with $\|\theta_{S_+^{*c}}\|_1 \leq 3\|\theta_{S_+^*}\|_1$ proves the claim. □

Thus, to control $\kappa^2(\Sigma, s^*)$, we need to control the maximum element-wise distance between Σ and Σ_A: $\max_{ij} |\Sigma_{ij} - \Sigma_{A,ij}|$. We will show that in the setting of Proposition 2.3,

$$\max_{ij} |\Sigma_{ij} - \Sigma_{A,ij}| \leq \frac{c_{\min}}{32 s_+^*},$$

and thus, by Lemma 2.8, we have $\kappa^2(\Sigma, s^*) \geq \kappa^2(\Sigma_A, s^*) - \frac{c_{\min}}{2} \geq \frac{c_{\min}}{2} > 0$, i.e. the compatibility condition holds for Σ.

Proof of Proposition 2.3. To make referencing of sections of Σ easier, we number its

blocks as follows

$$\Sigma = \frac{1}{\binom{n}{2}} \begin{bmatrix} \underbrace{\frac{n}{2} X^T X}_{①} & \underbrace{\frac{\sqrt{n}}{\sqrt{2}} X^T \mathbf{1}}_{②} & \underbrace{\mathbf{0}}_{③} \\ \underbrace{\frac{\sqrt{n}}{\sqrt{2}} \mathbf{1}^T X}_{④} & \underbrace{\mathbf{1}^T \mathbf{1}}_{⑤} & \underbrace{\mathbf{0}}_{⑥} \\ \underbrace{\mathbf{0}}_{⑦} & \underbrace{\mathbf{0}}_{⑧} & \underbrace{\mathbb{E}[Z^T Z]}_{⑨} \end{bmatrix}$$

For $i, j = 1, \ldots, n$, we have $\Sigma_{ij} = \Sigma_{A,ij}$ (block ①). The entry at position $(n+1), (n+1)$ (block ⑤) is also equal and so are blocks ③, ⑥, ⑦, ⑧ and ⑨. For the entries at positions i, j with $i = n+1$ and $j = 1, \ldots, n$ as well as positions with $i = 1, \ldots, n$ and $j = n+1$ (blocks ② and ④), we have:

$$\Sigma_{ij} - \Sigma_{A,ij} = \Sigma_{ij} = \frac{(n-1)\sqrt{2}}{(n-1)\sqrt{n}} = \frac{\sqrt{2}}{\sqrt{n}} \leq \frac{c_{\min}}{32 s_+^*}$$

for $n \gg 0$, since we assume that $s^* = o(\sqrt{n})$. The claim now follows from Lemma 2.8. □

In the SβM without covariates we define the matrices Σ and Σ_A completely analogously to the SβM-C by removing the blocks corresponding to the covariates Z.

$$\Sigma = \frac{1}{\binom{n}{2}} \begin{bmatrix} \frac{n}{2} X^T X & \frac{\sqrt{n}}{\sqrt{2}} X^T \mathbf{1} \\ \frac{\sqrt{n}}{\sqrt{2}} \mathbf{1}^T X & \mathbf{1}^T \mathbf{1} \end{bmatrix}, \quad \Sigma_A \coloneqq \begin{bmatrix} \frac{1}{n-1} X^T X & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}.$$

To be consistent with our numbering scheme from before, we number these four blocks from top left to bottom right as ①, ②, ④ and ⑤, skipping ③.

Lemma 2.9 (Compatibility condition for SβM)**.** *Under Assumption 2.2 and for n large enough, the compatibility condition holds for the sample size adjusted Gram matrix Σ in SβM. That is, for every $\theta \in \mathbb{R}^{n+1}$ with $\|\theta_{S_+^{*c}}\|_1 \leq 3\|\theta_{S_+^*}\|_1$,*

$$\|\theta_{S_+^*}\|_1^2 \leq \frac{1}{4} s_+^* \theta^T \Sigma \theta.$$

Proof. Following the exact same steps as above for the SβM-C, we find for any $\theta = (\beta^T, \mu)^T$

$$\kappa^2(\Sigma_A, s_+^*) = \frac{\theta^T \Sigma_A \theta}{\frac{1}{s_+^*} \|\theta_{S_+^*}\|_1^2} = \frac{n-2}{n-1} > \frac{1}{2} > 0, \tag{2.12}$$

for any $n \geq 3$. Using line by line the same arguments as in the proof of Proposition 2.3 we prove that the compatibility condition holds for the sparse β model without

covariates. More precisely, by Lemma 2.8, we know that $\kappa^2(\Sigma, s_+^*) \geq \kappa^2(\Sigma_A, s_+^*) - 16\delta s_+^*$, where $\delta = \max_{ij} |\Sigma_{ij} - \Sigma_{A,ij}|$. Looking back at the proof of Proposition 2.3, we only need the part of it that deals with blocks ② and ④, since Σ and Σ_A coincide in the blocks ① and ⑤. Doing the same calculation as in said proof, we see that for any entry (i, j) in blocks ② or ⑤,

$$|\Sigma_{ij} - \Sigma_{A,ij}| \leq \frac{\sqrt{2}}{\sqrt{n}}.$$

Under Assumption 2.2, that is $s^* = o\left(\frac{\sqrt{n}}{\sqrt{\log(n)} \cdot K_n}\right)$ where we define K_n with the components corresponding to γ set to zero, this expression will be smaller than $\frac{1}{64 s_+^*}$ for n large enough. Thus, for n large enough, by Lemma 2.8 and inequality (2.12), $\kappa^2(\Sigma, s_+^*) \geq 1/2 - 1/4 = 1/4$. □

Notice that the bound $1/4$ in the lemma above are somewhat arbitrary and an artefact of how we pick our constants in the proof of that Lemma.

2.7.1.3 A rescaled penalized likelihood problem

We mentioned in Section 2.2 that it is possible to present an equivalent formulation of problem (2.4) in terms of a rescaled likelihood problem using the sample size adjusted design matrix $\bar{D}$. We will rely heavily on this formulation which we now make precise.

Recall that in the definition of $\bar{D}$ we effectively blew up the entries belonging to β. The blow-up factor was chosen precisely such that we can now reformulate our problem as a problem in which each parameter effectively has sample size $\binom{n}{2}$. That is, our original penalized likelihood problem can be rewritten as

$$\begin{aligned}\hat{\bar{\theta}} = (\hat{\bar{\beta}}, \hat{\mu}, \hat{\gamma}) = \underset{\bar{\beta},\mu,\gamma}{\arg\min} \frac{1}{\binom{n}{2}} \Bigg(& - \sum_{i=1}^{n} \frac{\sqrt{n}}{\sqrt{2}} \bar{\beta}_i d_i - d_+ \mu - \sum_{i<j} (Z_{ij}^T \gamma) A_{ij} \\ & + \sum_{i<j} \log\left(1 + \exp\left(\frac{\sqrt{n}}{\sqrt{2}} \bar{\beta}_i + \frac{\sqrt{n}}{\sqrt{2}} \bar{\beta}_j + \mu + Z_{ij}^T \gamma\right)\right) \Bigg) \\ & + \bar{\lambda} \|\bar{\beta}\|_1,\end{aligned} \tag{2.13}$$

where $\bar{\lambda} = \frac{\sqrt{n}}{\sqrt{2}}\lambda$ and the argmin is taken over $\bar{\Theta}_{\text{loc}} = \{\bar{\theta} \in \Theta : \|\bar{D}\bar{\theta}\|_\infty \leq r_n\}$. Note that $\bar{\Theta}_{\text{loc}}$ is convex. Given a solution $(\hat{\bar{\beta}}, \hat{\mu}, \hat{\gamma})$ for a given $\bar{\lambda}$ to this modified problem (2.13), we obtain a solution to our original problem (2.4) with penalty $\lambda = \bar{\lambda}\sqrt{2}/\sqrt{n}$, by setting

$$(\hat{\beta}, \hat{\mu}, \hat{\gamma}) = \left(\frac{\sqrt{n}}{\sqrt{2}} \hat{\bar{\beta}}, \hat{\mu}, \hat{\gamma}\right).$$

For a compacter way of writing, introduce the following notation: For any parameter $\theta = (\beta^T, \mu, \gamma^T)^T \in \Theta$, we write

$$\bar{\theta} = \left(\frac{\sqrt{2}}{\sqrt{n}}\beta, \mu, \gamma\right), \quad \text{and} \quad \bar{\beta} = \frac{\sqrt{2}}{\sqrt{n}}\beta.$$

In particular we use $\bar{\theta}_0 = (\bar{\beta}_0^T, \mu_0, \gamma_0^T)^T, \bar{\beta}_0 = \frac{\sqrt{2}}{\sqrt{n}}\beta_0$, to denote the rescaled truth and $\bar{\theta}^* = (\bar{\beta}^{*T}, \mu^*, \gamma^{*T})^T, \bar{\beta}^* = \frac{\sqrt{2}}{\sqrt{n}}\beta^*$ to denote the rescaled best local approximation. Note that for any $\theta \in \Theta$, $D\theta = \bar{D}\bar{\theta}$ and hence the bound r_n is the same in the definitions of Θ_{loc} and $\bar{\Theta}_{\text{loc}}$. Also, since rescaling the set $\mathbb{R}^n_+$ still results in $\mathbb{R}^n_+$, there is no need to introduce a set $\bar{\Theta}$. Note that $\theta \in \Theta_{\text{loc}}$ if and only if $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$.

For any $\bar{\theta} = (\bar{\beta}^T, \mu, \gamma)^T$, denote the negative log-likelihood function corresponding to the rescaled problem (2.13) as

$$\begin{aligned}\bar{\mathcal{L}}(\bar{\theta}) = &-\sum_{i=1}^{n} \frac{\sqrt{n}}{\sqrt{2}}\bar{\beta}_i d_i - d_+\mu - \sum_{i<j}(Z_{ij}^T\gamma)A_{ij} \\ &+ \sum_{i<j} \log\left(1 + \exp\left(\frac{\sqrt{n}}{\sqrt{2}}\bar{\beta}_i + \frac{\sqrt{n}}{\sqrt{2}}\bar{\beta}_j + \mu + Z_{ij}^T\gamma\right)\right).\end{aligned}$$

Then, clearly $\bar{\mathcal{L}}(\bar{\theta}) = \mathcal{L}(\theta)$ and

$$\bar{\mathcal{E}}(\bar{\theta}) \coloneqq \frac{1}{\binom{n}{2}}(\mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})] - \mathbb{E}[\bar{\mathcal{L}}(\bar{\theta}^*)]) = \mathcal{E}(\theta).$$

Thus, $\bar{\theta}^*$ fulfils

$$\bar{\theta}^* = \operatorname*{arg\,min}_{\theta \in \bar{\Theta}_{\text{loc}}} \bar{\mathcal{E}}(\bar{\theta}),$$

i.e. $\bar{\theta}^*$ is the best local rescaled solution.

To give us a more compact way of writing, for any $\theta \in \Theta$ we introduce functions $f_\theta : \mathbb{R}^{n+1+p} \to \mathbb{R}, f_\theta(v) = v^T\theta$ and denote the function space of all such f_θ by $\mathbb{F} \coloneqq \{f_\theta : \theta \in \Theta\}$. We endow $\mathbb{F}$ with two norms as follows. Denote the law of the rows of $\bar{D}$ on $\mathbb{R}^{n+1+p}$, i.e. the probability measure induced by $(\bar{X}_{ij}^T, 1, Z_{ij}^T)^T, i < j$, by $\bar{Q}$. That is, for a measurable set $A = A_1 \times A_2 \subset \mathbb{R}^{n+1} \times \mathbb{R}^p$,

$$\bar{Q}(A) = \frac{1}{\binom{n}{2}} \sum_{i<j} P(\bar{D}_{ij} \in A) = \frac{1}{\binom{n}{2}} \sum_{i<j} \bar{\delta}_{ij}(A_1) \cdot P(Z_{ij} \in A_2),$$

where $\bar{\delta}_{ij}(A_1) = 1$ if $(\bar{X}_{ij}^T, 1)^T \in A_1$ and zero otherwise, is the Dirac-measure. We are interested in the L_2 and L_∞ norm on $\mathbb{F}$ with respect to the measure $\bar{Q}$ on $\mathbb{R}^{n+1} \times \mathbb{R}^p$. Denote the $L_2(\bar{Q})$-norm of $f \in \mathbb{F}$ simply by $\|\,.\,\|_{\bar{Q}}$ and let $\mathbb{E}_Z$ be the

expectation with respect to Z:

$$\|f\|_{\bar{Q}}^2 \coloneqq \|f\|_{L_2(\bar{Q})}^2 = \int_{\mathbb{R}^{n+1}\times\mathbb{R}^p} f(v)^2 \bar{Q}(dv) = \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbb{E}_Z[f((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)^2]$$

and define the $L_\infty(\bar{Q})$-norm as usual as the $\bar{Q}$-a.s. smallest upper bound of f:

$$\|f\|_{\bar{Q},\infty} = \inf\{C \geq 0 : |f(v)| \leq C \text{ for } \bar{Q}\text{-almost every } v \in \mathbb{R}^{n+1+p}\}.$$

In particular, for any $f_\theta \in \mathbb{F}, \theta \in \Theta_{\text{loc}}$: $\|f_\theta\|_\infty \leq \sup_{Z_{ij}} \|D\theta\|_\infty \leq r_n$.

We make the analogous definitions for the unscaled design matrix. Define the probability measure induced by the rows of D on $\mathbb{R}^{n+1+p}$ as Q. It is easy to see that we can switch between these norms as follows. Given a parameter θ and its rescaled version $\bar{\theta}$, then clearly

$$\|f_{\bar{\theta}}\|_{\bar{Q}} = \|f_\theta\|_Q, \quad \|f_{\bar{\theta}}\|_{\bar{Q},\infty} = \|f_\theta\|_{Q,\infty}.$$

Also note that for any $\bar{\theta}$

$$\|f_{\bar{\theta}}\|_{\bar{Q}}^2 = \mathbb{E}_Z \left[\frac{1}{\binom{n}{2}} \sum_{i<j} (\bar{D}_{ij}^T \bar{\theta})^2 \right] = \bar{\theta}^T \Sigma \bar{\theta}. \tag{2.14}$$

We have the following corollary which follows immediately from Proposition 2.3.

Corollary 2.10. *Under Assumption 2.1, for $s^* = o(\sqrt{n})$ and n large enough, for every $\bar{\theta} = \bar{\theta}_1 - \bar{\theta}_2, \bar{\theta}_1, \bar{\theta}_2 \in \bar{\Theta}_{\text{loc}}$ with $\|\bar{\theta}_{S_+^{*c}}\|_1 \leq 3\|\bar{\theta}_{S_+^*}\|_1$, we have*

$$\|\bar{\theta}_{S_+^*}\|_1^2 \leq \frac{2s_+^*}{c_{\min}} \|f_{\bar{\theta}_1} - f_{\bar{\theta}_2}\|_{\bar{Q}}^2.$$

Proof. Follows from Proposition 2.3 and identity (2.14). □

2.7.1.4 Two basic inequalities

A key result in the consistency proofs in classical LASSO settings is the so called *basic inequality* (cf. van de Geer & Bühlmann (2011), Chapter 6). We give two formulations of it, one for the original penalized likelihood problem (2.4) and one, completely analogous result, for the rescaled problem (2.13). To that end, let P_n denote the empirical measure with respect to our observations $(A_{ij}, Z_{ij})_{i,j}$, that is, for any suitable function g, $P_n g \coloneqq \sum_{i<j} g(A_{ij}, Z_{ij})/\binom{n}{2}$. In particular, if we let for each $\theta \in \Theta$,

$$l_\theta(A_{ij}, Z_{ij}) = -A_{ij}(\beta_i + \beta_j + \mu + \gamma^T Z_{ij}) + \log(1 + \exp(\beta_i + \beta_j + \mu + \gamma^T Z_{ij})),$$

then $P_n l_\theta = \mathcal{L}(\theta)/\binom{n}{2}$. Similarly, we define $P = \mathbb{E}P_n$. In particular, $Pl_\theta = \mathbb{E}P_n l_\theta = \mathbb{E}[\mathcal{L}(\theta)/\binom{n}{2}]$, where we suppress the dependence on n in our notation.

We define the *empirical process* as

$$\{v_n(\theta) = (P_n - P)l_\theta : \theta \in \Theta\}.$$

Which can also be written in rescaled form as

$$\bar{v}_n(\bar{\theta}) \coloneqq \frac{1}{\binom{n}{2}}(\bar{\mathcal{L}}(\bar{\theta}) - \mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})]) = v_n(\theta).$$

Lemma 2.11 (Basic Inequality)**.** *For any $\theta = (\beta^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}}$ we have*

$$\mathcal{E}(\hat{\theta}) + \lambda\|\hat{\beta}\|_1 \leq -[v_n(\hat{\theta}) - v_n(\theta)] + \mathcal{E}(\theta) + \lambda\|\beta\|_1.$$

Proof. Plugging in the definitions, the above equation is equivalent to

$$\begin{aligned}\frac{1}{\binom{n}{2}}\left(\mathbb{E}[\mathcal{L}(\hat{\theta})] - \mathbb{E}[\mathcal{L}(\theta_0)]\right) + \lambda\|\hat{\beta}\|_1 \\ \leq -\frac{1}{\binom{n}{2}}\mathcal{L}(\hat{\theta}) + \frac{1}{\binom{n}{2}}\mathbb{E}[\mathcal{L}(\hat{\theta})] + \frac{1}{\binom{n}{2}}\mathcal{L}(\theta) - \frac{1}{\binom{n}{2}}\mathbb{E}[\mathcal{L}(\theta)] \\ + \lambda\|\beta\|_1 + \frac{1}{\binom{n}{2}}\left(\mathbb{E}[\mathcal{L}(\theta)] - \mathbb{E}[\mathcal{L}(\theta_0)]\right).\end{aligned}$$

Rearranging shows that this is true if and only if

$$\frac{1}{\binom{n}{2}}\mathcal{L}(\hat{\theta}) + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{\binom{n}{2}}\mathcal{L}(\theta) + \lambda\|\beta\|_1,$$

which is true by definition of $\hat{\theta}$. □

Remark. For any $0 < t < 1$ and $\theta \in \Theta_{\text{loc}}$, let $\tilde{\theta} = t\hat{\theta} + (1-t)\theta$. Since Γ is convex, $\tilde{\theta} \in \Theta_{\text{loc}}$ and since $\theta \mapsto l_\theta$ and $\|\,.\,\|_1$ are convex functions, we can replace $\hat{\theta}$ by $\tilde{\theta}$ in the basic inequality and still obtain the same result. Plugging in the definitions, we see that the basic inequality is equivalent to the following:

$$\begin{aligned}\mathcal{E}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 &\leq -[v_n(\tilde{\theta}) - v_n(\theta)] + \lambda\|\beta\|_1 + \mathcal{E}(\theta) \\ \iff \frac{1}{\binom{n}{2}}\mathcal{L}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 &\leq \frac{1}{\binom{n}{2}}\mathcal{L}(\theta) + \lambda\|\beta\|_1\end{aligned}$$

and by convexity

$$\begin{aligned}\frac{1}{\binom{n}{2}}\mathcal{L}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 &\leq \frac{1}{\binom{n}{2}}t\mathcal{L}(\hat{\theta}) + \frac{1}{\binom{n}{2}}(1-t)\mathcal{L}(\theta) + t\lambda\|\hat{\beta}\|_1 + (1-t)\lambda\|\beta\|_1 \\ &\leq \frac{1}{\binom{n}{2}}\mathcal{L}(\theta) + \lambda\|\beta\|_1,\end{aligned}$$

where the last inequality follows by definition of $\hat{\theta}$. In particular, for any $M > 0$, choosing

$$t = \frac{M}{M + \|\hat{\theta} - \theta\|_1},$$

gives $\|\tilde{\theta} - \theta\|_1 \leq M$.

Lemma 2.12. *For any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ we have*

$$\bar{\mathcal{E}}(\hat{\bar{\theta}}) + \bar{\lambda}\|\hat{\bar{\beta}}\|_1 \leq -[\bar{v}_n(\hat{\bar{\theta}}) - \bar{v}_n(\bar{\theta})] + \bar{\mathcal{E}}(\bar{\theta}) + \bar{\lambda}\|\bar{\beta}\|_1.$$

Since the proof of Lemma 2.11 only relies on the argmin property of $\hat{\theta}$, the proof of Lemma 2.12 is line by line the same as for Lemma 2.11. We also get the same property for convex combinations of $\hat{\bar{\theta}}$ and $\bar{\theta}$: For any $t \in (0, 1)$ the rescaled basic inequality Lemma 2.12 holds for $\hat{\bar{\theta}}$ replaced by $\tilde{\theta} = t\hat{\bar{\theta}} + (1 - t)\bar{\theta} \in \bar{\Theta}_{\text{loc}}$.

2.7.1.5 Lower quadratic margin for $\mathcal{E}$

In this section we will derive a lower quadratic bound on the excess risk $\mathcal{E}(\theta)$ if the parameter θ is close to the truth θ_0. This is a necessary property for the proof to come and is referred to as the *margin condition* in classical LASSO theory (cf. van de Geer & Bühlmann (2011)). We will conduct our derivations for the original parameter space Θ_{loc}. Since $\mathcal{L}(\theta) = \bar{\mathcal{L}}(\bar{\theta})$ and $\mathcal{E}(\theta) = \bar{\mathcal{E}}(\bar{\theta})$, we will find that the same results hold in the rescaled model.

The proof mainly relies on a second-order Taylor expansion of the function l_θ of introduced in Section 2.7.1.4. Given a fixed θ, we treat l_θ as a function in $\theta^T x$ and define new functions $l_{ij} : \mathbb{R} \to \mathbb{R}, i < j$,

$$l_{ij}(a) = \mathbb{E}[l_\theta(A_{ij}, a)|Z_{ij}] = -p_{ij}a + \log(1 + \exp(a)),$$

where $p_{ij} = P(A_{ij} = 1|Z_{ij})$ and by slight abuse of notation we use $l_\theta(A_{ij}, a) \coloneqq -A_{ij}a + \log(1 + \exp(a))$. Taking derivatives, it is easy to see that

$$f_{\theta_0}((X_{ij}^T, 1, Z_{ij}^T)^T) \in \arg\min_a l_{ij}(a).$$

Note that we are using the actual truth θ_0 in the above equation, not the best local approximation θ^*. Write $f_0 = f_{\theta_0}$.

All l_{ij} are clearly twice continuously differentiable with derivative

$$\frac{\partial^2}{\partial a^2} l_{ij}(a) = \frac{\exp(a)}{(1 + \exp(a))^2} > 0, \forall a \in \mathbb{R}.$$

Using a second-order Taylor expansion around $a_0 = f_0((X_{ij}^T, 1, Z_{ij}^T)^T)$ we get

$$l_{ij}(a) = l_{ij}(a_0) + l'(a_0)(a - a_0) + \frac{l''(\bar{a})}{2}(a - a_0)^2 = l_{ij}(a_0) + \frac{l''(\bar{a})}{2}(a - a_0)^2,$$

with an $\bar{a}$ between a and a_0. Note that $\frac{\exp(a)}{(1+\exp(a))^2}$ is symmetric and monotone decreasing for $a \geq 0$. Then, for any $\eta > 0$ and any a with $|a - a_0| \leq \eta$, we get

$$\begin{aligned} l_{ij}(a) - l_{ij}(a_0) &= \frac{\exp(\bar{a})}{(1+\exp(\bar{a}))^2}\frac{(a-a_0)^2}{2} \\ &= \frac{\exp(|\bar{a}|)}{(1+\exp(|\bar{a}|))^2}\frac{(a-a_0)^2}{2}, \quad \text{by symmetry} \\ &\geq \frac{\exp(|a_0|+\eta)}{(1+\exp(|a_0|+\eta))^2}\frac{(a-a_0)^2}{2}, \quad \text{since } |\bar{a}| \leq |a_0| + \eta \\ &= \frac{\exp(|f_0((X_{ij}^T, 1, Z_{ij}^T)^T)| + \eta)}{(1+\exp(|f_0((X_{ij}^T, 1, Z_{ij}^T)^T)| + \eta))^2}\frac{(a-a_0)^2}{2}. \end{aligned} \tag{2.15}$$

In particular, for any θ with $|f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T) - f_0((X_{ij}^T, 1, Z_{ij}^T)^T)| \leq \eta$, we have

$$\begin{aligned} &l_{ij}(f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T)) - l_{ij}(f_0((X_{ij}^T, 1, Z_{ij}^T)^T)) \\ &\qquad \geq \frac{\exp(|f_0((X_{ij}^T, 1, Z_{ij}^T)^T)| + \eta)}{(1+\exp(|f_0((X_{ij}^T, 1, Z_{ij}^T)^T)| + \eta))^2} \\ &\qquad\qquad \cdot \frac{(f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T) - f_0((X_{ij}^T, 1, Z_{ij}^T)^T))^2}{2} \\ &\qquad \geq \frac{\exp(r_{n,0} + \eta)}{(1+\exp(r_{n,0} + \eta))^2}\frac{(f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T) - f_0((X_{ij}^T, 1, Z_{ij}^T)^T))^2}{2}. \end{aligned}$$

Define

$$\tau = \frac{\exp(r_{n,0} + \eta)}{2(1+\exp(r_{n,0} + \eta))^2}$$

and notice that for K_n defined in (2.5) we have

$$K_n = K_n(\eta) = \frac{1}{\tau}.$$

Define a subset $\mathbb{F}_{\text{local}} \subset \mathbb{F}$ as $\mathbb{F}_{\text{local}} = \{f_\theta \in \mathbb{F} : \|f_\theta - f_0\|_\infty \leq \eta\}$. For all $f_\theta \in \mathbb{F}_{\text{local}}$:

$$\begin{aligned} \mathcal{E}(\theta) &= \frac{1}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[l_\theta(A_{ij}, (X_{ij}^T, Z_{ij}^T)^T) - l_{\theta_0}(A_{ij}, (X_{ij}^T, Z_{ij}^T)^T)] \\ &= \frac{1}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[(l_{ij}(f_\theta((X_{ij}^T, Z_{ij}^T)^T)) - l_{ij}(f_0((X_{ij}^T, Z_{ij}^T)^T)))] \\ &\geq \frac{1}{\binom{n}{2}}\sum_{i<j}\tau\mathbb{E}[(f_\theta((X_{ij}^T, Z_{ij}^T)^T) - f_0((X_{ij}^T, Z_{ij}^T)^T))^2] \\ &\geq \frac{1}{K_n}\|f_\theta - f_0\|_Q^2. \end{aligned}$$

Thus, we have obtained a lower bound for the excess risk given by the quadratic function $G_n(\|f_\theta - f_0\|)$ where $G_n(u) = 1/K_n \cdot u^2$. Since $\mathcal{E}(\theta) = \bar{\mathcal{E}}(\bar{\theta})$ and $\|f_\theta\|_Q^2 =$

$\|f_{\bar{\theta}}\|_{\bar{Q}}^2$ and $\|f_\theta\|_{Q,\infty} = \|f_{\bar{\theta}}\|_{\bar{Q},\infty}$, we obtain the same result for the rescaled problem (2.13): For any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ with $\|f_{\bar{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q},\infty} \leq \eta$, we have

$$\bar{\mathcal{E}}(\bar{\theta}) \geq \frac{1}{K_n}\|f_{\bar{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}}^2.$$

Recall that the convex conjugate of a strictly convex function G on $[0, \infty)$ with $G(0) = 0$ is defined as the function

$$H(v) = \sup_u\{uv - G(u)\}, \quad v > 0$$

and in particular, if $G(u) = cu^2$ for a positive constant c, we have $H(v) = v^2/(4c)$. Hence, the convex conjugate of G_n is

$$H_n(v) = \frac{v^2 K_n}{4}.$$

Keep in mind that by definition for any u, v: $uv \leq G(u) + H(v)$.

2.7.1.6 Consistency on a special set

We will show that the penalized likelihood estimator is consistent in the sense that it converges to the best possible approximation θ^*. We will show that consistency holds on a special set $\mathcal{I}$. It then suffices to show that $P(\mathcal{I}) \to 1$.

The proof follows in spirit van de Geer & Bühlmann (2011), Theorem 6.4. It uses the language of rescaled likelihood problem (2.13). We define some objects that we will need for the proof of consistency. We want to use the quadratic margin condition derived in Section 2.7.1.5. Recall that we defined $\eta \coloneqq 2r_n + 2\|\beta^* - \beta_0\|_\infty + |\mu^* - \mu_0| + 2\kappa$ in Section 2.2. Then, for any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ we have

$$\begin{aligned}
\|f_{\bar{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q},\infty} = \|f_\theta - f_{\theta_0}\|_{Q,\infty} &\leq \|f_\theta - f_{\theta^*}\|_{Q,\infty} + \|f_{\theta^*} - f_{\theta_0}\|_{Q,\infty} \\
&\leq \|f_\theta\|_{Q,\infty} + \|f_{\theta^*}\|_{Q,\infty} \\
&\quad + \inf\{C : \max_{i<j} |\beta_i^* + \beta_j^* + \mu^* - \beta_{0,i} - \beta_{0,j} - \mu_0 + (\gamma^* - \gamma_0)^T Z_{ij}| \\
&\qquad \leq C, \text{ a.s.}\} \\
&\leq 2r_n + 2\|\beta^* - \beta_0\|_\infty + |\mu^* - \mu_0| + 2\kappa = \eta.
\end{aligned}$$

That is, the quadratic margin condition holds for any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$. With that definition of η, we have for K_n defined in (2.5),

$$K_n \leq 2\frac{(1 + \exp(r_{n,0} + 2r_n + 2\|\beta^* - \beta_0\|_\infty + |\mu^* - \mu_0| + 2\kappa))^2}{\exp(r_{n,0} + 2r_n + 2\|\beta^* - \beta_0\|_\infty + |\mu^* - \mu_0| + 2\kappa)}.$$

Define

$$\epsilon^* = \frac{3}{2}\bar{\mathcal{E}}(\bar{\theta}^*) + H_n\left(\frac{4\sqrt{2s_+^*}\bar{\lambda}}{\sqrt{c_{\min}}}\right).$$

Remember that $\bar{\mathcal{E}}(\bar{\theta}^*) = \mathcal{E}(\theta^*)$ corresponds to the approximation error of our model. Let for any $M > 0$

$$Z_M \coloneqq \sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}^*\|_1\le M}} |\bar{v}_n(\bar{\theta}) - \bar{v}_n(\bar{\theta}^*)|,$$

where $\bar{v}_n$ denotes the rescaled empirical process. Recall that for any rescaled $\bar{\theta}$ we have $\bar{v}_n(\bar{\theta}) = v_n(\theta)$. and by construction $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ if and only if $\theta \in \Theta_{\text{loc}}$. Hence, the set over which we are maximizing in the definition of Z_M can be expressed in terms of parameters θ on the original scale as

$$\left\{\theta = (\beta^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}} : \frac{\sqrt{2}}{\sqrt{n}}\|\beta - \beta^*\|_1 + |\mu - \mu^*| + \|\gamma - \gamma^*\|_1 \le M\right\}.$$

Set

$$M^* \coloneqq \epsilon^*/\lambda_0,$$

where λ_0 is a lower bound on $\bar{\lambda}$ that will be made precise in the proof showing that $\mathcal{I}$ has large probability. Define

$$\mathcal{I} \coloneqq \{Z_{M^*} \le \lambda_0 M^*\} = \{Z_{M^*} \le \epsilon^*\}. \tag{2.16}$$

Theorem 2.13. *Let Assumptions 2.1 and 2.2 hold and let $\bar{\lambda} \ge 8\lambda_0$. Then, on the set $\mathcal{I}$, we have*

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{\sqrt{2}}{\sqrt{n}}\|\hat{\beta} - \beta^*\|_1 + |\hat{\mu} - \mu^*| + \|\hat{\gamma} - \gamma^*\|_1\right) \le 6\mathcal{E}(\theta^*) + 4H_n\left(\frac{4\sqrt{2s_+^*}\bar{\lambda}}{\sqrt{c_{\min}}}\right).$$

Proof of Theorem 2.13. We assume that we are on the set $\mathcal{I}$ throughout. Set

$$t = \frac{M^*}{M^* + \|\hat{\bar{\theta}} - \bar{\theta}^*\|_1}$$

and $\tilde{\theta} = (\tilde{\beta}^T, \tilde{\mu}, \tilde{\gamma}^T)^T = t\hat{\bar{\theta}} + (1-t)\bar{\theta}^*$. Then,

$$\|\tilde{\theta} - \bar{\theta}^*\|_1 = t\|\hat{\bar{\theta}} - \bar{\theta}^*\| \le M^*.$$

Since $\hat{\bar{\theta}}, \bar{\theta}^* \in \bar{\Theta}_{\text{loc}}$ and by the convexity of $\bar{\Theta}_{\text{loc}}$, $\tilde{\theta} \in \bar{\Theta}_{\text{loc}}$, and by the remark after

Lemma 2.12, the basic inequality holds for $\tilde{\theta}$:

$$\begin{aligned}\bar{\mathcal{E}}(\bar{\theta}) + \bar{\lambda}\|\tilde{\beta}\|_1 &\leq -(\bar{v}_n(\tilde{\theta}) - \bar{v}_n(\bar{\theta}^*)) + \bar{\mathcal{E}}(\bar{\theta}) + \bar{\lambda}\|\bar{\beta}^*\|_1 \\ &\leq Z_{M^*} + \bar{\lambda}\|\bar{\beta}^*\|_1 + \bar{\mathcal{E}}(\bar{\theta}^*) \\ &\leq \epsilon^* + \bar{\lambda}\|\bar{\beta}^*\|_1 + \bar{\mathcal{E}}(\bar{\theta}^*).\end{aligned}$$

From now on, write $\tilde{\mathcal{E}} = \bar{\mathcal{E}}(\tilde{\theta})$ and $\mathcal{E}^* = \bar{\mathcal{E}}(\bar{\theta}^*)$. Note, that $\|\tilde{\beta}\|_1 = \|\tilde{\beta}_{S^{*c}}\|_1 + \|\tilde{\beta}_{S^*}\|_1$ and thus, by the triangle inequality,

$$\begin{aligned}\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\beta}_{S^{*c}}\|_1 &\leq \epsilon^* + \bar{\lambda}(\|\bar{\beta}^*\|_1 - \|\tilde{\beta}_{S^*}\|_1) + \mathcal{E}^* \\ &\leq \epsilon^* + \bar{\lambda}(\|\bar{\beta}^* - \tilde{\beta}_{S^*}\|_1) + \mathcal{E}^* \\ &\leq \epsilon^* + \bar{\lambda}(\|\bar{\beta}^* - \tilde{\beta}_{S^*}\|_1 + \|(\mu^*, \gamma^{*T})^T - (\tilde{\mu}, \tilde{\gamma}^T)^T\|_1) + \mathcal{E}^* \qquad (2.17) \\ &= \epsilon^* + \bar{\lambda}\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1 + \mathcal{E}^* \\ &\leq 2\epsilon^* + \bar{\lambda}\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1.\end{aligned}$$

Case i) If $\bar{\lambda}\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1 \geq \epsilon^*$, then

$$\bar{\lambda}\|\tilde{\beta}_{S^{*c}}\|_1 \leq \tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\beta}_{S^{*c}}\|_1 \leq 3\bar{\lambda}\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1. \qquad (2.18)$$

Since $\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^{*c}}\|_1 = \|\tilde{\beta}_{S^{*c}}\|_1$, we may thus apply the compatibility condition, Corollary 2.10 (note that $\bar{\beta}^* = \bar{\beta}^*_{S^*}$) to obtain

$$\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1 \leq \frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}} - f_{\bar{\theta}^*}\|_{\bar{Q}},$$

where we have used that $\theta \mapsto f_\theta$ is linear and hence $f_{\tilde{\theta} - \bar{\theta}^*} = f_{\tilde{\theta}} - f_{\bar{\theta}^*}$ Observe that

$$\|\tilde{\theta} - \theta^*\|_1 = \|\tilde{\beta}_{S^{*c}}\|_1 + \|(\tilde{\theta} - \theta^*)_{S_+^*}\|_1. \qquad (2.19)$$

Hence,

$$\begin{aligned}\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}^*\|_1 &= \tilde{\mathcal{E}} + \bar{\lambda}(\|\tilde{\beta}_{S^{*c}}\|_1 + \|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1) \\ &\leq \epsilon^* + 2\bar{\lambda}\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1 + \mathcal{E}^* \\ &\leq \epsilon^* + \mathcal{E}^* + 2\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}} - f_{\bar{\theta}^*}\|_{\bar{Q}}.\end{aligned}$$

Recall that for a convex function G and its convex conjugate H we have $uv \leq G(u) + H(v)$. Since $\tilde{\theta}, \bar{\theta}^* \in \bar{\Theta}_{\text{loc}}$, it holds $\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_\infty \leq \eta$, $\|f_{\bar{\theta}^*} - f_{\bar{\theta}_0}\|_\infty \leq \eta$. Thus,

we obtain

$$
\begin{aligned}
2\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}} - f_{\bar{\theta}^*}\|_{\bar{Q}} &= 4\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\frac{\|f_{\tilde{\theta}} - f_{\bar{\theta}^*}\|_{\bar{Q}}}{2} \\
&\leq 4\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\frac{\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}} + \|f_{\bar{\theta}^*} - f_{\bar{\theta}_0}\|_{\bar{Q}}}{2} \\
&\leq H_n\left(4\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\right) + G_n\left(\frac{\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}} + \|f_{\bar{\theta}^*} - f_{\bar{\theta}_0}\|_{\bar{Q}}}{2}\right) \\
&\leq H_n\left(4\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\right) + \frac{G_n(\|f_{\tilde{\theta}} - f_{\bar{\theta}_0}\|_{\bar{Q}})}{2} + \frac{G_n(\|f_{\bar{\theta}^*} - f_{\bar{\theta}_0}\|_{\bar{Q}})}{2} \\
&\leq H_n\left(4\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\right) + \frac{\tilde{\mathcal{E}}}{2} + \frac{\mathcal{E}^*}{2},
\end{aligned}
$$

where we have used the convexity of G_n in the second-to-last inequality and the margin condition in the last inequality. It follows

$$
\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}^*\|_1 \leq \epsilon^* + \frac{3}{2}\mathcal{E}^* + H_n\left(4\bar{\lambda}\frac{\sqrt{2s_+^*}}{\sqrt{c_{\min}}}\right) + \frac{\tilde{\mathcal{E}}}{2} = 2\epsilon^* + \frac{\tilde{\mathcal{E}}}{2}
$$

and therefore

$$
\frac{\tilde{\mathcal{E}}}{2} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}^*\|_1 \leq 2\epsilon^*. \tag{2.20}
$$

Finally, this gives

$$
\|\tilde{\theta} - \bar{\theta}^*\|_1 \leq \frac{2\epsilon^*}{\bar{\lambda}} = \frac{2\lambda_0 M^*}{\bar{\lambda}} \underbrace{\leq}_{\bar{\lambda}\geq 4\lambda_0} \frac{M^*}{2}.
$$

From this, by using the definition of $\tilde{\theta}$, we obtain

$$
\|\tilde{\theta} - \bar{\theta}^*\|_1 = t\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 = \frac{M^*}{M^* + \|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1}\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 \leq \frac{M^*}{2}.
$$

Rearranging gives

$$
\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 \leq M^*.
$$

Case ii) If $\bar{\lambda}\|(\bar{\theta}^* - \tilde{\theta})_{S_+^*}\|_1 \leq \epsilon^*$, then from (2.17)

$$
\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\beta}_{S^{*c}}\|_1 \leq 3\epsilon^*.
$$

Using once more (2.19), we get

$$
\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}^*\|_1 = \tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\beta}_{S^{*c}}\|_1 + \bar{\lambda}\|(\tilde{\theta} - \bar{\theta}^*)_{S_+^*}\|_1 \leq 4\epsilon^*. \tag{2.21}
$$

Thus,

$$
\|\tilde{\theta} - \bar{\theta}^*\|_1 \leq 4\frac{\epsilon^*}{\bar{\lambda}} = 4\frac{\lambda_0}{\bar{\lambda}}M^* \leq \frac{M^*}{2}
$$

by choice of $\lambda \geq 8\lambda_0$. Again, plugging in the definition of $\tilde{\theta}$, we obtain

$$\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 \leq M^*.$$

Hence, in either case we have $\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 \leq M^*$. That means, we can repeat the above steps with $\hat{\tilde{\theta}}$ instead of $\tilde{\theta}$: Writing $\hat{\mathcal{E}} \coloneqq \bar{\mathcal{E}}(\hat{\tilde{\theta}})$, following the same reasoning as above we arrive once more at (2.17):

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\tilde{\beta}}_{S^{*c}}\|_1 \leq 2\epsilon^* + \bar{\lambda}\|\bar{\beta}^* - \hat{\tilde{\beta}}_{S^*}\|_1 \leq 2\epsilon^* + \bar{\lambda}\|(\hat{\tilde{\theta}} - \bar{\theta}^*)_{S_+^*}\|_1.$$

From this, in **case i)** we obtain (2.18) which allows us to use the compatibility assumption to arrive at (2.20):

$$\frac{\hat{\mathcal{E}}}{2} + \bar{\lambda}\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 \leq 2\epsilon^*,$$

resulting in

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 \leq 4\epsilon^*.$$

In **case ii)** on the other hand, we arrive directly at (2.21), and hence

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\tilde{\theta}} - \bar{\theta}^*\|_1 \leq 3\epsilon^*.$$

Plugging in the definitions of $\hat{\tilde{\theta}}$ and $\bar{\theta}^*$ and using the fact that $\hat{\mathcal{E}} = \bar{\mathcal{E}}(\hat{\tilde{\theta}}) = \mathcal{E}(\hat{\theta})$ proves the claim. □

In the SβM without covariates we have an analogous result. Extend the definitions of $\mathcal{I}, Z_M$ to the SβM canonically by letting $p = 0, \gamma = 0, \kappa = 0, Z_{ij} = 0, i < j$. By Lemma 2.9 the compatibility condition holds for the SβM. The proof of the following corollary follows almost line by line as the proof of Theorem 2.13.

Corollary 2.14. *Assume that in the SβM Assumption 2.2 holds and that $\lambda \geq 8\lambda_0$, with λ_0 as in Theorem 2.13. Then, on the set $\mathcal{I}$ defined in (2.16), we have*

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{\sqrt{2}}{\sqrt{n}}\|\hat{\beta} - \beta^*\|_1 + |\hat{\mu} - \mu^*|\right) \leq 6\mathcal{E}(\theta^*) + 4H_n\left(\frac{4\sqrt{2s_+^*}\bar{\lambda}}{\sqrt{c_{\min}}}\right).$$

Proof. Analogous to the proof of Theorem 2.13. □

2.7.1.7 Controlling the special set $\mathcal{I}$

We show that $\mathcal{I}$ has measure tending to one. We first recall some probability inequalities that we will need.

Concentration inequalities

This is based on Chapter 14 in van de Geer & Bühlmann (2011). Throughout let $Z_1, \ldots, Z_n$ be a sequence of independent random variables in some space $\mathcal{Z}$ and $\mathcal{G}$ be a class of real valued functions on $\mathcal{Z}$.

Definition 2.15. A *Rademacher sequence* is a sequence $\epsilon_1, \ldots, \epsilon_n$ of i.i.d. random variables with $P(\epsilon_i = 1) = P(\epsilon_i = -1) = 1/2$ for all i.

Theorem 2.16 (Symmetrization Theorem as in van der Vaart & Wellner (1996), abridged)**.** *Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence independent of $Z_1, \ldots, Z_n$. Then*

$$\mathbb{E}\left(\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\{g(Z_i) - \mathbb{E}[g(Z_i)]\}\right|\right) \leq 2\mathbb{E}\left(\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^{n}\epsilon_i g(Z_i)\right|\right).$$

Theorem 2.17 (Contraction Theorem as in Ledoux & Talagrand (1991))**.** *Let $z_1, \ldots, z_n$ be non-random elements of $\mathcal{Z}$ and let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{Z}$. Consider Lipschitz functions $g_i : \mathbb{R} \to \mathbb{R}$ with Lipschitz constant $L = 1$, i.e. for all i*

$$|g_i(s) - g_i(s')| \leq |s - s'|, \forall s, s' \in \mathbb{R}.$$

Let $\epsilon_1, \ldots, \epsilon_n$ be a Rademacher sequence. Then for any function $f^* : \mathcal{Z} \to \mathbb{R}$ we have

$$\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{g_i(f(z_i)) - g_i(f^*(z_i))\}\right|\right) \leq 2\mathbb{E}\left(\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\epsilon_i\{f(z_i) - f^*(z_i)\}\right|\right).$$

The last theorem we need is a concentration inequality due to Bousquet (2002). We give a version as presented in van de Geer (2008).

Theorem 2.18 (Bousquet's concentration theorem)**.** *Suppose $Z_1, \ldots, Z_n$ and all $g \in \mathcal{G}$ satisfy the following conditions for some real valued constants η_n and τ_n*

$$\|g\|_\infty \leq \eta_n, \ \forall g \in \mathcal{G}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(g(Z_i)) \leq \tau_n^2, \ \forall g \in \mathcal{G}.$$

Define

$$\boldsymbol{Z} \coloneqq \sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n}g(Z_i) - \mathbb{E}[g(Z_i)]\right|.$$

Then for any $z > 0$

$$P\left(\boldsymbol{Z} \geq \mathbb{E}[\boldsymbol{Z}] + z\sqrt{2(\tau_n^2 + 2\eta_n\mathbb{E}[\boldsymbol{Z}])} + \frac{2z^2\eta_n}{3}\right) \leq \exp(-nz^2).$$

Remark. Looking at the original paper of Bousquet (2002), their result looks quite different at first. To see that the above falls into their framework, set the variables

in Bousquet (2002) as follows

$$f(Z_i) = (g(Z_i) - \mathbb{E}[g(Z_i)])/(2\eta_n), \qquad \tilde{Z}_k = \sup_f |\sum_{i \neq k} f(Z_i)|,$$

$$f_k = \arg\sup_f |\sum_{i \neq k} f(Z_i)|, \qquad \tilde{Z}'_k = |\sum_{i=1}^{n} f_k(Z_i)| - \tilde{Z}_k$$

$$\tilde{Z} = \frac{2\eta_n}{n}\mathbf{Z}.$$

Now apply Theorem 2.1 in Bousquet (2002), choosing for their $(Z, Z_1, \ldots, Z_n)$ the above defined $(\tilde{Z}, \tilde{Z}_1, \ldots, \tilde{Z}_n)$, for their $(Z'_1, \ldots, Z'_n)$ the above defined $(\tilde{Z}'_1, \ldots, \tilde{Z}'_n)$ and setting $u = 1$ and $\sigma^2 = \frac{\tau_n^2}{4\eta_n^2}$ in their theorem: The result is exactly Theorem 2.18 above.

The proof of the next lemma can be found in van de Geer & Bühlmann (2011), Lemma 14.14 (here we use the special case of their lemma for $m = 1$).

Lemma 2.19. *Let $\mathcal{G} = \{g_1, \ldots, g_p\}$ be a set of real valued functions on $\mathcal{Z}$ satisfying for all $i = 1, \ldots, n$ and all $j = 1, \ldots, p$*

$$\mathbb{E}[g_j(Z_i)] = 0, \ |g_j(Z_i)| \leq c_{ij}$$

for some positive constants c_{ij}. Then

$$\mathbb{E}\left[\max_{1 \leq j \leq p} \left|\sum_{i=1}^{n} g_j(Z_i)\right|\right] \leq [2\log(2p)]^{1/2} \max_{1 \leq j \leq p} \left[\sum_{i=1}^{n} c_{ij}^2\right]^{1/2}.$$

The expectation of Z_M

Recall the definition of Z_M

$$Z_M \coloneqq \sup_{\substack{\bar{\theta} \in \bar{\Theta}_{\text{loc}}, \\ \|\bar{\theta} - \bar{\theta}^*\|_1 \leq M}} |\bar{v}_n(\bar{\theta}) - \bar{v}_n(\bar{\theta}^*)|,$$

where $\bar{v}_n$ denotes the rescaled empirical process. Recall that there is a constant $c \in \mathbb{R}$ such that uniformly $|Z_{ij,k}| \leq c, 1 \leq i < j \leq n, k = 1, \ldots, p$.

Lemma 2.20. *For any $M > 0$ we have in the SβM-C*

$$\mathbb{E}[Z_M] \leq 8M(1 \vee c)\sqrt{\frac{2\log(2(n+p+1))}{\binom{n}{2}}}$$

and in the SβM without covariates

$$\mathbb{E}[Z_M] \leq 8M\sqrt{\frac{\log(2(n+1))}{\binom{n}{2}}}.$$

Proof. We only give the proof for the SβM-C. The proof for the case without covariates is exactly the same with the corresponding parts set to zero. Let $\epsilon_{ij}, i < j$, be a Rademacher sequence independent of $A_{ij}, Z_{ij}, i < j$. We first want to use the Symmetrization Theorem 2.16: For the random variables $Z_1, \dots, Z_n$ we choose $T_{ij} = (A_{ij}, \bar{X}_{ij}^T, 1, Z_{ij}^T)^T \in \{0,1\} \times \mathbb{R}^{n+1+p}$. For any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ consider the functions

$$g_{\bar{\theta}}(T_{ij}) = \frac{1}{\binom{n}{2}} \left\{ -A_{ij} \bar{D}_{ij}^T (\bar{\theta} - \bar{\theta}^*) + \log(1 + \exp(\bar{D}_{ij}^T \bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T \bar{\theta}^*)) \right\}$$

and the function set $\mathcal{G} = \mathcal{G}(M) \coloneqq \{g_{\bar{\theta}} : \bar{\theta} \in \bar{\Theta}_{\text{loc}}, \|\bar{\theta} - \bar{\theta}^*\|_1 \leq M\}$. Note, that

$$\bar{v}_n(\bar{\theta}) - \bar{v}_n(\bar{\theta}^*) = \sum_{i<j} \{g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})]\}.$$

Then, the Symmetrization Theorem gives us

$$\begin{aligned} \mathbb{E}[Z_M] &= \mathbb{E}\left[\sup_{g_{\bar{\theta}} \in \mathcal{G}} \left| \sum_{i<j} g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})] \right| \right] \\ &\leq 2\mathbb{E}\left[\sup_{g_{\bar{\theta}} \in \mathcal{G}} \left| \sum_{i<j} \epsilon_{ij} g_{\bar{\theta}}(T_{ij}) \right| \right]. \end{aligned}$$

Next, we want to apply the Contraction Theorem 2.17. Denote $T = (T_{ij})_{i<j}$ and let $\mathbb{E}_T$ be the conditional expectation given T. We need the conditional expectation at this point, because Theorem 2.17 requires non-random arguments in the functions. This does not hinder us, as later we will simply take iterated expectations, cancelling out the conditional expectation, see below. For the functions g_i in Theorem 2.17 we choose

$$g_{ij}(x) = \frac{1}{2}\{-A_{ij}x + \log(1 + \exp(x))\}$$

Note, that $\log(1 + \exp(x))$ has derivative bounded by one and thus is Lipschitz continuous with constant one by the Mean Value Theorem. Thus, all g_{ij} are also Lipschitz continuous with constant 1:

$$|g_{ij}(x) - g_{ij}(x')| \leq \frac{1}{2}\{|A_{ij}(x - x')| + |\log(1 + \exp(x)) - \log(1 + \exp(x'))|\} \leq |x - x'|.$$

For the function class $\mathcal{F}$ in Theorem 2.17 we choose $\mathcal{F} = \mathcal{F}_M \coloneqq \{f_{\bar{\theta}} : \bar{\theta} \in \bar{\Theta}_{\text{loc}}, \|\bar{\theta} -$

$\bar{\theta}^*\|_1 \leq M\}$ and pick $f^* = f_{\bar{\theta}^*}$. Then, by Theorem 2.17

$$\mathbb{E}_T \left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}^*\|_1\leq M}} \left| \frac{1}{\binom{n}{2}} \sum_{i<j} \epsilon_{ij} (g_{ij}(f_{\bar{\theta}}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)) - g_{ij}(f_{\bar{\theta}^*}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T))) \right| \right]$$
$$\leq 2\mathbb{E}_T \left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}^*\|_1\leq M}} \left| \frac{1}{\binom{n}{2}} \sum_{i<j} \epsilon_{ij} (f_{\bar{\theta}}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T) - f_{\bar{\theta}^*}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)) \right| \right].$$

Recall that we can express the functions $f_{\bar{\theta}} = f_{\bar{\beta},\mu,\gamma}$ as

$$f_{\bar{\beta},\mu,\gamma}(\,.\,) = \mu e_{n+1}(\,.\,) + \sum_{i=1}^{n} \bar{\beta}_i e_i(\,.\,) + \sum_{i=1}^{p} \gamma_i e_{n+1+i}(\,.\,),$$

where $e_i(\,.\,)$ is the projection on the ith coordinate. Consider any $\bar{\theta} = (\bar{\theta}_i)_{i=1}^{n+1+p} = (\bar{\beta}^T, \mu, \gamma^T)^T \in \bar{\Theta}_{\text{loc}}$ with $\|\bar{\theta} - \bar{\theta}^*\|_1 \leq M$. We simply write $e_k(X_{ij}, 1, Z_{ij})$ for the projection of the the vector $(X_{ij}^T, 1, Z_{ij}^T)^T \in \mathbb{R}^{n+p+1}$ to its kth component, i.e. instead of $e_k((X_{ij}^T, 1, Z_{ij}^T)^T)$. Then,

$$\left| \frac{1}{\binom{n}{2}} \sum_{i<j} \epsilon_{ij} (f_{\bar{\theta}}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T) - f_{\bar{\theta}^*}((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)) \right|$$
$$= \left| \frac{1}{\binom{n}{2}} \sum_{i<j} \epsilon_{ij} \left(\sum_{k=1}^{n+p+1} (\bar{\theta}_k - \bar{\theta}_k^*) e_k(\bar{X}_{ij}, 1, Z_{ij}) \right) \right|$$
$$\leq \frac{1}{\binom{n}{2}} \sum_{k=1}^{n+p+1} \left\{ |\bar{\theta}_k - \bar{\theta}_k^*| \max_{1\leq l\leq n+p+1} \left| \sum_{i<j} \epsilon_{ij} e_l(\bar{X}_{ij}, 1, Z_{ij}) \right| \right\}$$
$$\leq M \max_{1\leq l\leq n+p+1} \left| \frac{1}{\binom{n}{2}} \sum_{i<j} \epsilon_{ij} e_l(\bar{X}_{ij}, 1, Z_{ij}) \right|.$$

The last expression no longer depends on $\bar{\theta}$. To bind the right hand side in the last expression we use Lemma 2.19: In the language of the lemma, choose $Z_1, \ldots, Z_n$ as $T_{ij} = (\epsilon_{ij}, \bar{X}_{ij}^T, 1, Z_{ij}^T)^T$. We choose for the p in the formulation of the Lemma $n + p + 1$ and pick for our functions

$$g_k(T_{ij}) = \frac{1}{\binom{n}{2}} \epsilon_{ij} e_k(\bar{X}_{ij}, 1, Z_{ij}), k = 1, \ldots, n + p + 1.$$

Note that $\mathbb{E}[g_k(T_{ij})] = 0$. We want to employ Lemma 2.19 which requires us to bound $|g_k(T_{ij})| \leq c_{ij,k}$ for all $i < j$ and $k = 1, \ldots, n+1+p$. For any fixed $1 \leq k \leq n$ we have

$$|g_k(T_{ij})| \leq \begin{cases} \frac{\sqrt{n}}{\sqrt{2}\binom{n}{2}} = \frac{\sqrt{2}}{(n-1)\sqrt{n}}, & i \text{ or } j = k \\ 0, & \text{otherwise.} \end{cases}$$

The first case occurs exactly $(n-1)$ times for each k. Thus, for any $k \leq n$,

$$\sum_{i<j} c_{ij,k}^2 = \left(\frac{\sqrt{2}}{(n-1)\sqrt{n}}\right)^2 (n-1) = \frac{1}{\binom{n}{2}}.$$

If $k = n+1$, $|g_k(T_{ij})| = 1/\binom{n}{2}$ and hence

$$\sum_{i<j} c_{ij,n+1}^2 = \frac{1}{\binom{n}{2}}.$$

Finally, if $k > n+1$, $|g_k(T_{ij})| \leq c/\binom{n}{2}$ and therefore,

$$\sum_{i<j} c_{ij,k}^2 \leq \frac{c^2}{\binom{n}{2}}.$$

In total, this means

$$\max_{1\leq k\leq n+1+p} \sum_{i<j} c_{ij,k}^2 \leq \frac{1 \vee c^2}{\binom{n}{2}}.$$

Therefore, an application of Lemma 2.19 results in

$$\begin{aligned}
&\mathbb{E}\left[\max_{1\leq l\leq n+p+1} \left|\frac{1}{\binom{n}{2}} \sum_{i<j} \epsilon_{ij} e_l(\bar{X}_{ij}, Z_{ij})\right|\right] \\
&\qquad \leq \sqrt{2\log(2(n+1+p))} \max_{1\leq k\leq n+1+p} \left[\sum_{i<j} c_{ij,k}^2\right]^{1/2} \\
&\qquad \leq \sqrt{2\log(2(n+1+p))} \sqrt{\frac{1 \vee c^2}{\binom{n}{2}}} \\
&\qquad = \sqrt{\frac{2\log(2(n+1+p))}{\binom{n}{2}}} (1 \vee c).
\end{aligned}$$

Putting everything together, we obtain

$$\begin{aligned}
\mathbb{E}[Z_M] &\leq 2\mathbb{E}\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}^*\|_1\leq M}}\left|\frac{1}{\binom{n}{2}}\sum_{i<j}\epsilon_{ij}(-A_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},Z_{ij})-f_{\bar{\theta}^*}(\bar{X}_{ij},Z_{ij})))\right|\right] \\
&= 2\mathbb{E}\left[\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}^*\|_1\leq M}}\left|\frac{1}{\binom{n}{2}}\sum_{i<j}\epsilon_{ij}(-A_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},Z_{ij})-f_{\bar{\theta}^*}(\bar{X}_{ij},Z_{ij})))\right|\right]\right] \\
&\leq 8\mathbb{E}\left[\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}^*\|_1\leq M}}\left|\frac{1}{\binom{n}{2}}\sum_{i<j}\epsilon_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},Z_{ij})-f_{\bar{\theta}^*}(\bar{X}_{ij},Z_{ij}))\right|\right]\right] \\
&\leq 8M\mathbb{E}\left[\mathbb{E}_T\left[\max_{1\leq l\leq n+p+1}\left|\frac{1}{\binom{n}{2}}\sum_{i<j}\epsilon_{ij}e_l(\bar{X}_{ij},Z_{ij})\right|\right]\right] \\
&\leq 8M\sqrt{\frac{2\log(2(n+1+p))}{\binom{n}{2}}}(1\vee c).
\end{aligned}$$

This concludes the proof. □

Next, we will show that Z_M does not deviate too far from its expectation. The proof relies on the Concentration Theorem due to Bousquet, Theorem 2.18.

Corollary 2.21. *Pick any confidence level $t>0$. In the SβM-C let*

$$a_n \coloneqq \sqrt{\frac{2\log(2(n+p+1))}{\binom{n}{2}}}(1\vee c).$$

and choose $\lambda_0=\lambda_0(t,n)$ as

$$\lambda_0 = 8a_n + 2\sqrt{\frac{t}{\binom{n}{2}}(11(1\vee(c^2p))+8\sqrt{2}(1\vee c)\sqrt{n}a_n)} + \frac{2\sqrt{2}t(1\vee c)\sqrt{n}}{3\binom{n}{2}}.$$

In the SβM without covariates set

$$a_n = \sqrt{\frac{\log(2(n+1))}{\binom{n}{2}}}, \quad \lambda_0 = 8a_n + 2\sqrt{\frac{t}{\binom{n}{2}}(9+8\sqrt{2n}a_n)} + \frac{2\sqrt{2}t\sqrt{n}}{3\binom{n}{2}}.$$

In either case we have

$$P(Z_M > \lambda_0 M) \leq \exp(-t).$$

Proof. Again, we only give the proof for the case with covariates. The case without covariates is completely analogous by setting the corresponding parts to zero. We want to apply Bousquet's Concentration Theorem 2.18. For the random variables Z_i in the formulation of the theorem we choose once more $T_{ij}=(A_{ij},\bar{X}_{ij},1,Z_{ij}), i<j$,

and as functions we consider

$$g_{\bar\theta}(T_{ij}) = -A_{ij}\bar D_{ij}^T(\bar\theta - \bar\theta^*) + \log(1+\exp(\bar D_{ij}^T\bar\theta)) - \log(1+\exp(\bar D_{ij}^T\bar\theta^*)),$$
$$\mathcal{G} = \mathcal{G}_M \coloneqq \{g_{\bar\theta} : \bar\theta \in \bar\Theta_{\text{loc}}, \|\bar\theta - \bar\theta^*\|_1 \le M\}.$$

Then, by definition we have

$$Z_M = \sup_{g_{\bar\theta}\in\mathcal{G}} \frac{1}{\binom{n}{2}} \left| \sum_{i<j} \{g_{\bar\theta}(T_{ij}) - \mathbb{E}[g_{\bar\theta}(T_{ij})]\} \right|.$$

To apply Theorem 2.18, we need to bound $\|g_{\bar\theta}\|_\infty$. Recall that we denote the distribution of $[\bar X|1|Z]$ by $\bar Q$ and $\|g_{\bar\theta}\|_\infty$ is defined as the $\bar Q$-almost-sure smallest upper bound on the value of $g_{\bar\theta}$. For any $g_{\bar\theta} \in \mathcal{G}$, using the Lipschitz continuity of $\log(1+\exp(x))$:

$$\begin{aligned}
|g_{\bar\theta}(T_{ij})| &\le |\bar D_{ij}^T(\bar\theta - \bar\theta^*)| + |\log(1+\exp(\bar D_{ij}^T\bar\theta)) - \log(1+\exp(\bar D_{ij}^T\bar\theta^*))| \\
&\le 2|\bar D_{ij}^T(\bar\theta - \bar\theta^*)| \\
&\le 2\|\beta - \beta^*\|_1 + |\mu - \mu^*| + c\|\gamma - \gamma^*\|_1.
\end{aligned}$$

Thus,

$$\begin{aligned}
\|g_{\bar\theta}\|_\infty &\le 2\|\beta - \beta^*\|_1 + |\mu - \mu^*| + c\|\gamma - \gamma^*\|_1 \\
&\le 2(1\vee c)\|\theta - \theta^*\|_1 \\
&\le \sqrt{2}(1\vee c)\sqrt{n}M =: \eta_n.
\end{aligned}$$

For the last inequality we used that for any θ with $\|\bar\theta - \bar\theta^*\|_1 \le M$ it follows that $\|\theta - \theta^*\|_1 \le \sqrt{n}/\sqrt{2}M$, which is possibly a very generous upper bound. This does not matter, however, as the term associated with the above bound will be negligible, as we shall see.

The second requirement of Theorem 2.18 is that the average variance of $g_{\bar\theta}(T_{ij})$ has to be uniformly bounded. To that end we calculate

$$\begin{aligned}
&\frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{Var}(g_{\bar\theta}(T_{ij})) \\
&\quad= \frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{Var}(-A_{ij}D_{ij}^T(\theta - \theta^*)) \\
&\quad+ \frac{1}{\binom{n}{2}} \sum_{i<j} \operatorname{Var}(\log(1+\exp(\bar D_{ij}^T\bar\theta)) - \log(1+\exp(\bar D_{ij}^T\bar\theta^*))) \\
&\quad+ \frac{2}{\binom{n}{2}} \sum_{i<j} \operatorname{Cov}(-A_{ij}D_{ij}^T(\theta - \theta^*), \log(1+\exp(\bar D_{ij}^T\bar\theta)) - \log(1+\exp(\bar D_{ij}^T\bar\theta^*))).
\end{aligned}$$

Let us look at these terms in term. For the first term, we obtain

$$\frac{1}{\binom{n}{2}}\sum_{i<j}\mathrm{Var}(-A_{ij}D_{ij}^T(\theta-\theta^*)) \le \frac{1}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[(-A_{ij}D_{ij}^T(\theta-\theta^*))^2] \le \mathbb{E}\left[\frac{1}{\binom{n}{2}}\sum_{i<j}(D_{ij}^T(\theta-\theta^*))^2\right].$$

For the second term we get

$$\begin{aligned}&\frac{1}{\binom{n}{2}}\sum_{i<j}\mathrm{Var}(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*)))\\ &\qquad\le \frac{1}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*)))^2]\\ &\qquad\le \mathbb{E}\left[\frac{1}{\binom{n}{2}}\sum_{i<j}(D_{ij}^T(\theta-\theta^*))^2\right].\end{aligned}$$

The last term decomposes as

$$\begin{aligned}&\frac{2}{\binom{n}{2}}\sum_{i<j}\mathrm{Cov}(-A_{ij}D_{ij}^T(\theta-\theta^*),\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*)))\\ &\quad=\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta^*)\cdot(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*)))]\\ &\quad-\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta^*)]\cdot\mathbb{E}[\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*))]\end{aligned}$$

For the first term in that decomposition we have

$$\begin{aligned}&\frac{2}{\binom{n}{2}}\sum_{i<j}\left|\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta^*)\cdot(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*)))]\right|\\ &\quad\le\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|\cdot|\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*))|]\\ &\quad\le\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|^2]\end{aligned}$$

and for the second term, using the same arguments, we get

$$\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta^*)]\cdot\mathbb{E}[\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*))] \le\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|]^2,$$

meaning that in total

$$\frac{2}{\binom{n}{2}}\sum_{i<j}\left|\mathrm{Cov}(-A_{ij}D_{ij}^T(\theta-\theta^*),\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}^*)))\right|$$
$$\leq \frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|^2]+\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|]^2.$$

In total, we thus get

$$\frac{1}{\binom{n}{2}}\sum_{i<j}\mathrm{Var}(g_{\bar{\theta}}(T_{ij}))\leq 4\cdot\mathbb{E}\left[\frac{1}{\binom{n}{2}}\sum_{i<j}(D_{ij}^T(\theta-\theta^*))^2\right]+\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|]^2. \tag{2.22}$$

Furthermore,

$$\begin{aligned}
&\frac{1}{\binom{n}{2}}\sum_{i<j}(D_{ij}^T(\theta-\theta^*))^2\\
&\quad=\frac{1}{\binom{n}{2}}\sum_{i<j}(\beta_i+\beta_j+\mu-\beta_i^*-\beta_j^*-\mu^*+(\gamma-\gamma^*)^TZ_{ij})^2\\
&\quad\leq\frac{4}{\binom{n}{2}}\sum_{i<j}\left\{(\beta_i-\beta_i^*)^2+(\beta_j-\beta_j^*)^2+(\mu-\mu^*)^2+((\gamma-\gamma^*)^TZ_{ij})^2\right\},
\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality. Recall that for any $x\in\mathbb{R}^p$, $\|x\|_2\leq\|x\|_1\leq\sqrt{p}\|x\|_2$ and note that

$$|(\gamma-\gamma^*)^TZ_{ij}|\leq c\|\gamma-\gamma^*\|_1\leq c\sqrt{p}\|\gamma-\gamma^*\|_2.$$

Then, from the above

$$\begin{aligned}
&\frac{1}{\binom{n}{2}}\sum_{i<j}(D_{ij}^T(\theta-\theta^*))^2\\
&\quad\leq\frac{4}{\binom{n}{2}}\sum_{i<j}\left\{(\beta_i-\beta_i^*)^2+(\beta_j-\beta_j^*)^2+(\mu-\mu^*)^2+c^2p\|\gamma-\gamma^*\|_2^2\right\}\\
&\quad=4\left((\mu-\mu^*)^2+c^2p\|\gamma-\gamma^*\|_2^2+\frac{1}{\binom{n}{2}}(n-1)\|\beta-\beta^*\|_2^2\right)\\
&\quad=4\left((\mu-\mu^*)^2+c^2p\|\gamma-\gamma^*\|_2^2+\left\|\frac{\sqrt{2}}{\sqrt{n}}(\beta-\beta^*)\right\|_2^2\right)\\
&\quad=4\left((\mu-\mu^*)^2+c^2p\|\gamma-\gamma^*\|_2^2+\|\bar{\beta}-\bar{\beta}^*\|_2^2\right)\\
&\quad\leq 4(1\vee(c^2p))\|\bar{\theta}-\bar{\theta}^*\|_2^2\\
&\quad\leq 4(1\vee(c^2p))\|\bar{\theta}-\bar{\theta}^*\|_1^2\\
&\quad\leq 4(1\vee(c^2p))M^2.
\end{aligned} \tag{2.23}$$

Notice that for the second term in (2.22) we have

$$\begin{aligned}\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|]^2 &= \frac{2}{\binom{n}{2}}\sum_{i<j}(\beta_i-\beta_i^*+\beta_j-\beta_j^*+\mu-\mu^*+(\gamma-\gamma^*)^T\mathbb{E}[Z_{ij}])^2\\ &= \frac{2}{\binom{n}{2}}\sum_{i<j}(\beta_i-\beta_i^*+\beta_j-\beta_j^*+\mu-\mu^*)^2\\ &\leq \frac{6}{\binom{n}{2}}\sum_{i<j}\left\{(\beta_i-\beta_i^*)^2+(\beta_j-\beta_j^*)^2+(\mu-\mu^*)^2\right\},\end{aligned}$$

so that we may use the same steps as in (2.23) to conclude that

$$\frac{2}{\binom{n}{2}}\sum_{i<j}\mathbb{E}[|D_{ij}^T(\theta-\theta^*)|]^2 \leq 6M^2 \leq 6(1\vee(c^2p))M^2.$$

Such that in total,

$$\frac{1}{\binom{n}{2}}\sum_{i<j}\mathrm{Var}(g_{\bar{\theta}}(T_{ij})) \leq 22(1\vee(c^2p))M^2 =: \tau_n^2.$$

Applying Bousquet's concentration theorem 2.18 with η_n, τ_n defined above, we obtain for all $z>0$

$$\begin{aligned}\exp\left(-\binom{n}{2}z^2\right) &\geq P\left(Z_M \geq \mathbb{E}[Z_M]+z\sqrt{2(\tau_n^2+2\eta_n\mathbb{E}[Z_M])}+\frac{2z^2\eta_n}{3}\right)\\ &= P\Bigg(Z_M \geq \mathbb{E}[Z_M]\\ &\qquad +z\sqrt{2(22(1\vee(c^2p))M^2+2\sqrt{2}(1\vee c)\sqrt{n}M\mathbb{E}[Z_M])}\\ &\qquad +\frac{2\sqrt{2}z^2(1\vee c)\sqrt{n}M}{3}\Bigg).\end{aligned} \tag{2.24}$$

From Lemma 2.20, we know

$$\mathbb{E}[Z_M] \leq 8M\sqrt{\frac{2\log(2(n+p+1))}{\binom{n}{2}}}(1\vee c) = 8Ma_n.$$

Using this, we obtain from (2.24)

$$\begin{aligned}\exp\left(-\binom{n}{2}z^2\right) &\geq P\Bigg(Z_M \geq 8Ma_n+z\sqrt{2(22(1\vee(c^2p))M^2+16\sqrt{2}(1\vee c)\sqrt{n}M^2a_n)}\\ &\qquad +\frac{2\sqrt{2}z^2(1\vee c)\sqrt{n}M}{3}\Bigg)\\ &= P\left(Z_M \geq M\left(8a_n+2z\sqrt{11(1\vee(c^2p))+8\sqrt{2}(1\vee c)\sqrt{n}a_n}+\frac{2\sqrt{2}z^2(1\vee c)\sqrt{n}}{3}\right)\right).\end{aligned}$$

Now, pick $z = \sqrt{t/\binom{n}{2}}$ to get

$$\exp(-t) \geq$$
$$P\left(Z_M \geq M\left(8a_n + 2\sqrt{\frac{t}{\binom{n}{2}}(11(1 \vee (c^2 p)) + 8\sqrt{2}(1 \vee c)\sqrt{n}a_n)} + \frac{2\sqrt{2}t(1 \vee c)\sqrt{n}}{3\binom{n}{2}}\right)\right),$$

which is the claim. □

2.7.1.8 Proof of Theorems 2.4 and 2.6 and Corollary 2.5

Proof of Theorem 2.4. Follows immediately from Theorem 2.13 and Corollary 2.21. □

Proof of Corollary 2.5. We consider the case in which no approximation error is committed, that is $r_{n,0} \leq r_n$. Let ρ_n be the lower bound on the link probabilities corresponding to r_n. In that case $\theta^* = \theta_0$ and hence $\mathcal{E}(\theta^*) = 0$. By increasing $r_{n,0}$ if needed, we may assume without loss of generality that $r_{n,0} = r_n$ and $\rho_{n,0} = \rho_n$. Also, the definition of η may be simplified. Looking back at the derivation of K_n in (2.15) we see that for $a_0 = f_0((X_{ij}^T, 1, Z_{ij}^T)^T)$ and $a = f_{\hat{\theta}}((X_{ij}^T, 1, Z_{ij}^T)^T)$, we have $|a_0|, |a| \leq r_{n,0}$. Thus, for the intermediate point $\bar{a}$ between a_0 and a we must also have $|\bar{a}| \leq r_{n,0}$ and we may use that upper bound in (2.15) instead of $|\bar{a}| \leq |a_0| + \eta$. K_n then simplifies to

$$K_n = 2\frac{(1 + \exp(r_{n,0}))^2}{\exp(r_{n,0})} = 2\frac{(1 + \exp(-\text{logit}(\rho_{n,0})))^2}{\exp(-\text{logit}(\rho_{n,0}))} \leq \frac{4}{\rho_{n,0}}.$$

Thus, under the conditions of Theorem 2.4, we have with high probability

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{\sqrt{2}}{\sqrt{n}}\|\hat{\beta} - \beta^*\|_1 + |\hat{\mu} - \mu^*| + \|\hat{\gamma} - \gamma^*\|_1\right) \leq C\frac{s_+^* \bar{\lambda}^2}{\rho_{n,0}}.$$

with constant $C = 128/c_{\min}$. □

Proof of Theorem 2.6. Follows immediately from Corollary 2.14 and Corollary 2.21. □

2.7.2 Proof of inference results

2.7.2.1 Inverting population and sample Gram matrices

Recall that by Assumption 2.1, the minimum eigenvalue $\lambda_{\min}$ of $\frac{1}{\binom{n}{2}}\mathbb{E}[Z^T Z]$ stays uniformly bounded away from zero for all n. Consequently the minimum eigenvalue of $\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta]$ is lower bounded by $(1 \wedge \lambda_{\min}) > 0$ which is bounded away from

zero uniformly in n. We show that under Assumption 2.1, with high probability, the minimum eigenvalue of $\frac{1}{\binom{n}{2}} D_\vartheta^T D_\vartheta$ is bounded away from zero. More precisely, recall the definition of $\kappa(A, m)$ for square matrices A and dimensions m. We want to consider the expression $\kappa^2\left(\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta], p+1\right)$ which simplifies to

$$\kappa^2\left(\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta], p+1\right) \coloneqq \min_{v\in\mathbb{R}^{p+1}\setminus\{0\}} \frac{v^T \frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta] v}{\frac{1}{p+1}\|v\|_1^2}$$

and compare it to $\kappa^2\left(\frac{1}{\binom{n}{2}} D_\vartheta^T D_\vartheta, p+1\right)$. By Assumption 2.1 and the argument above, we have

$$\kappa^2\left(\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta], p+1\right) \geq C > 0$$

for a C independent of n. With $\delta = \max_{kl}\left|\left(\frac{1}{\binom{n}{2}} D_\vartheta^T D_\vartheta\right)_{kl} - \left(\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta]\right)_{kl}\right|$, by Lemma 2.8, we have

$$\kappa^2\left(\frac{1}{\binom{n}{2}} D_\vartheta^T D_\vartheta, p+1\right) \geq \kappa^2\left(\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta], p+1\right) - 16\delta(p+1).$$

By looking at the proof of Lemma 2.8, we see that in this particular case we do not even need the factor $16(p+1)$ on the right hand side above, but this does not matter anyway, so we keep it.

Lemma 2.22.

$$\delta = \max_{kl}\left|\left(\frac{1}{\binom{n}{2}} D_\vartheta^T D_\vartheta\right)_{kl} - \left(\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta]\right)_{kl}\right| = O_P\left(\binom{n}{2}^{-1/2}\right).$$

Proof. To make referencing sub-matrices of $1/\binom{n}{2} D_\vartheta^T D_\vartheta$ and its expectation easier, write

$$B \coloneqq \frac{1}{\binom{n}{2}} D_\vartheta^T D_\vartheta = \frac{1}{\binom{n}{2}}\begin{bmatrix} \underbrace{\mathbf{1}^T\mathbf{1}}_{⑤} & \underbrace{\mathbf{1}^T Z}_{⑥} \\ \underbrace{Z^T\mathbf{1}}_{⑧} & \underbrace{Z^T Z}_{⑨} \end{bmatrix},\ A \coloneqq \frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta] = \frac{1}{\binom{n}{2}}\begin{bmatrix} \underbrace{\mathbf{1}^T\mathbf{1}}_{⑤} & \underbrace{\mathbf{0}}_{⑥} \\ \underbrace{\mathbf{0}}_{⑧} & \underbrace{\mathbb{E}[Z^T Z]}_{⑨} \end{bmatrix}$$

where we have chosen our numbering to be consistent with the notation used in the proof of Proposition 2.3. The matrices A and B are equal in block ⑤. For i, j corresponding to the blocks ⑥ and ⑧, $B_{ij} - A_{ij} = B_{ij}$ is the sum of all the entries of some column Z_k of the matrix Z for an appropriate k. That is, there is a $1 \leq k \leq p$ such that

$$B_{ij} - A_{ij} = \frac{1}{\binom{n}{2}} Z_k^T\mathbf{1} = \frac{1}{\binom{n}{2}}\sum_{s<t} Z_{k,st}.$$

Thus, by assumption, $\mathbb{E}[B_{ij} - A_{ij}] = 0$. We know that for each $k, s, t : Z_{k,st} \in [-c, c]$.

Hence, by Hoeffding's inequality, for all $\eta > 0$,

$$P\left(|B_{ij} - A_{ij}| \geq \eta\right) = P\left(\left|\sum_{s<t} Z_{k,st}\right| \geq \binom{n}{2}\eta\right)$$
$$\leq 2\exp\left(-\frac{2\binom{n}{2}^2\eta^2}{\sum_{i<j}(2c)^2}\right) = 2\exp\left(-\binom{n}{2}\frac{\eta^2}{2c^2}\right).$$

For i, j from block ⑨, a typical element has the form

$$B_{ij} - A_{ij} = \frac{1}{\binom{n}{2}}\sum_{s<t}\left\{Z_{k,st}Z_{l,st} - \mathbb{E}[Z_{k,st}Z_{l,st}]\right\},$$

for appropriate k, l. In other words, $B_{ij} - A_{ij}$ is the inner product of two columns of Z, minus their expectation, scaled by $1/\binom{n}{2}$. Since $Z_{k,st}Z_{l,st} \in [-c^2, c^2]$ for all k, l, s, t, we have that for all k, l, s, t: $Z_{k,st}Z_{l,st} - \mathbb{E}[Z_{k,st}Z_{l,st}] \in [-2c^2, 2c^2]$. Thus, by Hoeffding's inequality, for all $\eta > 0$,

$$P\left(|B_{ij} - A_{ij}| \geq \eta\right) = P\left(\left|\sum_{s<t}\{Z_{k,st}Z_{l,st} - \mathbb{E}[Z_{k,st}Z_{l,st}]\}\right| \geq \binom{n}{2}\eta\right)$$
$$\leq 2\exp\left(-\binom{n}{2}\frac{\eta^2}{8c^4}\right).$$

Thus, with $\tilde{c} = c^2 \vee (2c^4)$, we have for any entry in blocks ⑥, ⑧, ⑨ and any $\eta > 0$,

$$P\left(|B_{ij} - A_{ij}| \geq \eta\right) \leq 2\exp\left(-\binom{n}{2}\frac{\eta^2}{2\tilde{c}}\right).$$

The claim will follow from a union bound: Because block ⑥ is the transpose of block ⑧, it is sufficient to control one of them. By symmetry of block ⑨ it suffices to control the upper triangular half, including the diagonal, of block ⑨. Thus, we only need to control the entries $B_{ij} - A_{ij}$ for i, j in the following index set

$\mathcal{A} = \{(i, j) :$

i, j belong to block ⑧ or the upper triangular half or diagonal of block ⑨$\}$.

Keep in mind that block ⑧ has p elements, while the upper triangular part of block

⑨ plus its diagonal has $\binom{p}{2} + p = \binom{p+1}{2}$ elements. Thus, for any $\eta > 0$,

$$\begin{aligned} P\left(\max_{ij} |B_{ij} - A_{ij}| \geq \eta\right) &\leq \sum_{(i,j)\in\mathcal{A}} P\left(|B_{ij} - A_{ij}| \geq \eta\right) \\ &\leq 2p \exp\left(-\binom{n}{2}\frac{\eta^2}{2c^2}\right) + 2\binom{p+1}{2}\exp\left(-\binom{n}{2}\frac{\eta^2}{8c^4}\right) \\ &\leq 2\left(p + \binom{p+1}{2}\right)\exp\left(-\binom{n}{2}\frac{\eta^2}{2\tilde{c}}\right) \\ &= p(p+3)\exp\left(-\binom{n}{2}\frac{\eta^2}{2\tilde{c}}\right). \end{aligned}$$

This proves the claim. □

Thus, for n large enough, we have with high probability $\delta \leq \frac{(1\wedge\lambda_{\min})}{32(p+1)}$. Then, by Lemma 2.8, with high probability and uniformly in n,

$$\kappa^2\left(\frac{1}{\binom{n}{2}}D_\vartheta^T D_\vartheta, p+1\right) \geq \kappa^2\left(\frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta], p+1\right) - 16\delta(p+1) \geq \frac{(1\wedge\lambda_{\min})}{2} > 0.$$

Yet, if $\kappa^2\left(\frac{1}{\binom{n}{2}}D_\vartheta^T D_\vartheta, p+1\right) \geq C > 0$ uniformly in n, then also for any $v \neq 0$, $v^T\frac{1}{\binom{n}{2}}D_\vartheta^T D_\vartheta v \geq C\|v\|_2^2$. But we also know that the minimum eigenvalue of $\frac{1}{\binom{n}{2}}D_\vartheta^T D_\vartheta$ is the largest possible C such that this bound holds (it is actually tight with equality for the eigenvectors corresponding to the minimum eigenvalue). Therefore, with high probability, the minimum eigenvalue of $\frac{1}{\binom{n}{2}}D_\vartheta^T D_\vartheta$ stays uniformly bounded away from zero. Thus, for any $v \in \mathbb{R}^{p+1}\backslash\{0\}$ and any finite n:

$$\frac{1}{\binom{n}{2}}v^T D_\vartheta^T \hat{W}^2 D_\vartheta v \geq \min_{i<j}\{p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))\}\left(v^T\frac{1}{\binom{n}{2}}D_\vartheta^T D_\vartheta v\right) \geq C\rho_n\|v\|_2^2 > 0.$$

Thus, $\text{mineval}\left(\frac{1}{\binom{n}{2}}D_\vartheta^T\hat{W}^2 D_\vartheta\right) \geq C\rho_n \text{mineval}\left(\frac{1}{\binom{n}{2}}D_\vartheta^T D_\vartheta\right) > 0$. That means, for every finite n, $\frac{1}{\binom{n}{2}}D_\vartheta^T\hat{W}^2 D_\vartheta$ is invertible with high probability.

2.7.2.2 Goal and approach

Goal: We want to show that for $k = 1, \ldots, p+1$,

$$\sqrt{\binom{n}{2}}\frac{\hat{\vartheta}_k - \vartheta_{0,k}}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \to \mathcal{N}(0,1).$$

Approach: Recall the definition of the "one-sample-version" of $\mathcal{L}$, i.e. $l_\theta : \{0,1\} \times \mathbb{R}^{n+1+p} \to \mathbb{R}$, for $\theta = (\beta^T, \mu, \gamma^T)^T \in \Theta$,

$$l_\theta(y,x) \coloneqq -y\theta^T x + \log(1+\exp(\theta^T x).$$

Then, the negative log-likelihood is given by

$$\mathcal{L}(\theta) = \sum_{i<j} l_\theta(A_{ij}, (X_{ij}^T, 1, Z_{ij}^T)^T)$$

and

$$\nabla\mathcal{L}(\theta) = \sum_{i<j} \nabla l_\theta(A_{ij}, (X_{ij}^T, 1, Z_{ij}^T)^T), \quad H\mathcal{L}(\theta) = \sum_{i<j} H l_\theta(A_{ij}, (X_{ij}^T, 1, Z_{ij}^T)^T),$$

where H denotes the Hessian with respect to θ. Consider l_θ as a function in $\theta^T x$ and introduce:

$$l(y, a) \coloneqq -ya + \log(1 + \exp(a)), \tag{2.25}$$

with second partial derivative: $\ddot{l}(y, a) = \partial_{a^2} l(y, a) = \frac{\exp(a)}{(1+\exp(a))^2}$. Note, that $\partial_{a^2} l(y, a)$ is Lipschitz continuous (it has bounded derivative $|\partial_{a^3} l(y, a)| \leq 1/(6\sqrt{3})$; Lipschitz continuity then follows by the Mean Value Theorem). Doing a first-order Taylor expansion in a of $\dot{l}(y, a) = \partial_a l(y, a)$ in the point $(A_{ij}, D_{ij}^T\theta_0)$ evaluated at $(A_{ij}, D_{ij}^T\hat{\theta})$, we get

$$\partial_a l(A_{ij}, D_{ij}^T\hat{\theta}) = \partial_a l(A_{ij}, D_{ij}^T\theta_0) + \partial_{a^2} l(A_{ij}, \alpha) D_{ij}^T(\hat{\theta} - \theta_0), \tag{2.26}$$

for an α between $D_{ij}^T\hat{\theta}$ and $D_{ij}^T\theta_0$. By Lipschitz continuity of $\partial_{a^2} l$, we also find

$$\begin{aligned} |\partial_{a^2} l(A_{ij}, \alpha) D_{ij}^T(\hat{\theta} - \theta_0) - \partial_{a^2} l(A_{ij}, D_{ij}^T\hat{\theta}) D_{ij}^T(\hat{\theta} - \theta_0)| \\ \leq |\alpha - D_{ij}^T\hat{\theta}||D_{ij}^T(\hat{\theta} - \theta_0)| \leq |D_{ij}^T(\hat{\theta} - \theta_0)|^2, \end{aligned} \tag{2.27}$$

where the last inequality follows, because α is between $D_{ij}^T\hat{\theta}$ and $D_{ij}^T\theta_0$.

Consider the vector $P_n\nabla l_{\hat{\theta}}$: By (2.26), with α_{ij} between $D_{ij}^T\hat{\theta}$ and $D_{ij}^T\theta_0$,

$$\begin{aligned} P_n\nabla l_{\hat{\theta}} &= \frac{1}{\binom{n}{2}} \sum_{i<j} \left(\partial_{\theta_k} l(A_{ij}, D_{ij}^T\hat{\theta})\right)_{k=1,\ldots,n+1+p}, \quad \text{as } (n+1+p) \times 1\text{-vector} \\ &= \frac{1}{\binom{n}{2}} \sum_{i<j} \dot{l}(A_{ij}, D_{ij}^T\hat{\theta}) D_{ij} \\ &= \frac{1}{\binom{n}{2}} \sum_{i<j} (\dot{l}(A_{ij}, D_{ij}^T\theta_0) + \ddot{l}(A_{ij}, \alpha_{ij}) D_{ij}^T(\hat{\theta} - \theta_0)) D_{ij} \end{aligned}$$

which by (2.27) gives

$$= P_n\nabla l_{\theta_0} + \frac{1}{\binom{n}{2}} \sum_{i<j} D_{ij} \left\{ \ddot{l}(A_{ij}, D_{ij}^T\hat{\theta}) D_{ij}^T(\hat{\theta} - \theta_0) + O(|D_{ij}^T(\hat{\theta} - \theta_0)|^2) \right\}.$$

Since $\ddot{l}(A_{ij}, D_{ij}^T\hat{\theta}) = p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))$ and we thus have $\sum_{i<j}\ddot{l}(A_{ij}, D_{ij}^T\hat{\theta})D_{ij}D_{ij}^T(\hat{\theta}-\theta_0) = D^T\hat{W}^2D(\hat{\theta}-\theta_0)$:

$$= P_n\nabla l_{\theta_0} + P_nHl_{\hat{\theta}}(\hat{\theta}-\theta_0) + O\left(\frac{1}{\binom{n}{2}}\sum_{i<j}D_{ij}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right)$$

$$= P_n\nabla l_{\theta_0} + \frac{1}{\binom{n}{2}}D^T\hat{W}^2D(\hat{\theta}-\theta_0) + O\left(\frac{1}{\binom{n}{2}}\sum_{i<j}D_{ij}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right),$$

where the O notation is to be understood componentwise. Above, we have equality of two $((n+1+p)\times 1)$-vectors. We are only interested in the portion relating to $\vartheta = (\mu, \gamma^T)^T$, that is, in the last $p+1$ entries. Introduce the $((n+1+p)\times(n+1+p))$-matrix

$$M = \begin{pmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \hat{\Theta}_\vartheta\end{pmatrix},$$

where $\mathbf{0}$ are zero-matrices of appropriate dimensions. Multiplying the above with M on both sides gives:

$$MP_n\nabla l_{\hat{\theta}} = MP_n\nabla l_{\theta_0} + M\frac{1}{\binom{n}{2}}D^T\hat{W}^2D(\hat{\theta}-\theta_0) + MO\left(\frac{1}{\binom{n}{2}}\sum_{i<j}D_{ij}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right). \tag{2.28}$$

Let us consider these terms in turn: Multiplication by M means that the first n entries of any of the vectors above are zero. Hence we only need to consider the last $p+1$ entries. The left-hand side of (2.28) is equal to zero by (2.8). The last $p+1$ entries of the first term on the right-hand side are $\hat{\Theta}_\vartheta P_n\nabla_\vartheta l_{\theta_0}$. For the second term on the right hand side, notice that

$$\frac{1}{\binom{n}{2}}D^T\hat{W}^2D = \frac{1}{\binom{n}{2}}\begin{bmatrix}X^T\hat{W}^2X & X^T\hat{W}^2\mathbf{1} & X^T\hat{W}^2Z\\ \mathbf{1}^T\hat{W}^2X & \mathbf{1}^T\hat{W}^2\mathbf{1} & \mathbf{1}^T\hat{W}^2Z\\ Z^T\hat{W}^2X & Z^T\hat{W}^2\mathbf{1} & Z^T\hat{W}^2Z\end{bmatrix}.$$

$\hat{\Theta}_\vartheta = \hat{\Sigma}_\vartheta^{-1}$ and $\hat{\Sigma}_\vartheta$ is the lower-right $(p+1)\times(p+1)$ block of above matrix. Thus,

$$M\frac{1}{\binom{n}{2}}D^T\hat{W}^2D = \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \hat{\Theta}_\vartheta\frac{1}{\binom{n}{2}}D_\vartheta^T\hat{W}^2X & I_{(p+1)\times(p+1)}\end{bmatrix}.$$

Then, for the last $p+1$ entries of $M\frac{1}{\binom{n}{2}}D^T\hat{W}^2D(\hat{\theta}-\theta_0)$,

$$\left(M\frac{1}{\binom{n}{2}}D^T\hat{W}^2D(\hat{\theta}-\theta_0)\right)_{\text{last } p+1 \text{ entries}} = \hat{\Theta}_\vartheta\frac{1}{\binom{n}{2}}D_\vartheta^T\hat{W}^2X(\hat{\beta}-\beta_0) + \begin{pmatrix}\hat{\mu}-\mu_0\\ \hat{\gamma}-\gamma_0\end{pmatrix}.$$

Thus, (2.28) implies

$$\begin{aligned} 0 =& \hat{\Theta}_\vartheta P_n \nabla_\gamma l_{\theta_0} + \hat{\Theta}_\vartheta \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 X(\hat{\beta} - \beta_0) + \begin{pmatrix} \hat{\mu} - \mu_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \\ &+ O\left(\hat{\Theta}_\vartheta \frac{1}{\binom{n}{2}} \sum_{i<j} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} |D_{ij}^T(\hat{\theta} - \theta_0)|^2 \right), \end{aligned}$$

which is equivalent to

$$\begin{aligned} \begin{pmatrix} \hat{\mu} - \mu_0 \\ \hat{\gamma} - \gamma_0 \end{pmatrix} =& - \hat{\Theta}_\vartheta P_n \nabla_\vartheta l_{\theta_0} - \hat{\Theta}_\vartheta \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 X(\hat{\beta} - \beta_0) \\ &+ O\left(\hat{\Theta}_\vartheta \frac{1}{\binom{n}{2}} \sum_{i<j} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} |D_{ij}^T(\hat{\theta} - \theta_0)|^2 \right). \end{aligned} \tag{2.29}$$

Our goal is now to show that for each component $k = 1, \ldots, p+1$,

$$\sqrt{\binom{n}{2}} \frac{\hat{\vartheta}_k - \vartheta_{0,k}}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

as described in the **Goal** section. To that end, by (2.29), we now need to solve the following three problems: Writing $\hat{\Theta}_{\vartheta,k}$ for the kth row of $\hat{\Theta}_\vartheta$,

1. $\sqrt{\binom{n}{2}} \frac{\hat{\Theta}_{\vartheta,k} P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \xrightarrow{d} \mathcal{N}(0,1)$,
2. $\frac{1}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 X(\hat{\beta} - \beta_0) = o_P\left(\binom{n}{2}^{-1/2} \right)$.
3. $O\left(\frac{1}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} \sum_{i<j} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} |D_{ij}^T(\hat{\theta} - \theta_0)|^2 \right) = o_P\left(\binom{n}{2}^{-1/2} \right)$

2.7.2.3 Bounding inverses

The problems (1) - (3) above suggest that it will be essential to bound the norm and the distance of $\hat{\Theta}_\vartheta$ and Θ_ϑ in an appropriate manner. For any invertible matrices $A, B \in \mathbb{R}^{m \times m}$ we have

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}. \tag{2.30}$$

Thus, for any sub-multiplicative matrix norm $\| \cdot \|$, we get

$$\|A^{-1} - B^{-1}\| \leq \|A^{-1}\| \|B^{-1}\| \|B - A\|. \tag{2.31}$$

We are particularly interested in the matrix ∞-norm, defined as

$$\|A\|_\infty := \sup\left\{ \frac{\|Ax\|_\infty}{\|x\|_\infty}, x \neq 0 \right\} = \sup\left\{ \|Ax\|_\infty, \|x\|_\infty = 1 \right\} = \max_{1 \leq i \leq m} \sum_{j=1}^{m} |A_{i,j}|,$$

It is well-known, that any such matrix norm induced by a vector norm is sub-multiplicative ($\|AB\|_\infty \leq \|A\|_\infty \|B\|_\infty$) and consistent with the inducing vector norm ($\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty$ for any $x \in \mathbb{R}^m$). We first want to bound the matrix ∞-norm in terms of the largest eigenvalue.

Lemma 2.23. *For any symmetric, positive semi-definite $(m \times m)$-matrix A with maximal eigenvalue $\lambda > 0$, we have $\|A\|_\infty \leq \sqrt{m}\lambda$.*

Proof.

$$\begin{aligned}
\|A\|_\infty &= \sup\{\|Ax\|_\infty, \|x\|_\infty = 1\} \\
&\leq \sup\{\|Ax\|_2, \|x\|_\infty = 1\}, \quad \|Ax\|_\infty \leq \|Ax\|_2 \\
&= \sup\left\{\frac{\|Ax\|_2}{\|x\|_2}\|x\|_2, \|x\|_\infty = 1\right\} \\
&\leq \sqrt{m}\sup\left\{\frac{\|Ax\|_2}{\|x\|_2}, \|x\|_\infty = 1\right\}, \quad \text{if } \|x\|_\infty = 1, \text{ then } \|x\|_2 \leq \sqrt{m}, \\
&\leq \sqrt{m}\sup\left\{\frac{\|Ax\|_2}{\|x\|_2}, x \neq 0\right\} \\
&= \sqrt{m}\|A\|_2 = \sqrt{m}\lambda,
\end{aligned}$$

where $\|A\|_2$ is the spectral norm of A and we have used that for a symmetric matrix, the spectral norm is equal to the modulus of its largest eigenvalue. □

Also, recall that the inverse of a symmetric matrix A is itself symmetric:

$$I = AA^{-1} = A^T A^{-1} \Rightarrow I = (A^{-1})^T A^T = (A^{-1})^T A \Rightarrow (A^{-1})^T = A^{-1},$$

where the last implication follows from the uniqueness of the inverse. Hence, $\hat{\Theta}_\vartheta$ and Θ_ϑ are symmetric and we may apply Lemma 2.23. Using that $\text{maxeval}(\Sigma_\vartheta^{-1}) = \text{mineval}(\Sigma_\vartheta)^{-1}$, we get

$$\|\Theta_\vartheta\|_\infty \leq \sqrt{p} \cdot \text{maxeval}(\Sigma_\vartheta^{-1}) \leq C\frac{1}{\rho_n},$$

and with high probability

$$\|\hat{\Theta}_\vartheta\|_\infty \leq \sqrt{p} \cdot \text{maxeval}(\hat{\Sigma}_\vartheta^{-1}) \leq C\frac{1}{\rho_n},$$

with some absolute constant C. Finally, by (2.31),

$$\|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty \leq \|\hat{\Theta}_\vartheta\|_\infty \|\Theta_\vartheta\|_\infty \|\hat{\Sigma}_\vartheta - \Sigma_\vartheta\|_\infty \leq \frac{C}{\rho_n^2}\|\hat{\Sigma}_\vartheta - \Sigma_\vartheta\|_\infty.$$

It remains to control $\|\hat{\Sigma}_\vartheta - \Sigma_\vartheta\|_\infty$. We have

$$\begin{aligned}\hat{\Sigma}_\vartheta - \Sigma_\vartheta &= \frac{1}{\binom{n}{2}}\left(D_\vartheta^T \hat{W}^2 D_\vartheta - \mathbb{E}[D_\vartheta^T W_0^2 D_\vartheta]\right) \\ &= \underbrace{\frac{1}{\binom{n}{2}}\left(D_\vartheta^T(\hat{W}^2 - W_0^2)D_\vartheta\right)}_{(I)} + \underbrace{\frac{1}{\binom{n}{2}}\left(D_\vartheta^T W_0^2 D_\vartheta - \mathbb{E}[D_\vartheta^T W_0^2 D_\vartheta]\right)}_{(II)}.\end{aligned}$$

Recall that $\hat{w}_{ij}^2 = p_{ij}(\hat{\theta})(1 - p_{ij}(\hat{\theta})) = \frac{\exp(D_{ij}^T\hat{\theta})}{(1+\exp(D_{ij}^T\hat{\theta}))^2} = \partial_{a^2} l(A_{ij}, D_{ij}^T\hat{\theta})$, with the function l defined in (2.25). Also recall that $\partial_{a^2} l$ is Lipschitz with constant one, by the Mean Value Theorem and the fact that it has derivative $\partial_{a^3} l$ bounded by one. Thus, considering the (k,l)-th element of (I) above, we get:

$$\begin{aligned}\left|\frac{1}{\binom{n}{2}}\left(D_\vartheta^T(\hat{W}^2 - W_0^2)D_\vartheta\right)_{kl}\right| &= \left|\frac{1}{\binom{n}{2}}\sum_{i<j} D_{ij,n+k}D_{ij,n+l}(\hat{w}_{ij}^2 - w_{0,ij}^2)\right| \\ &\leq C\frac{1}{\binom{n}{2}}\sum_{i<j}|\hat{w}_{ij}^2 - w_{0,ij}^2|, \quad \text{by uniform boundedness of } Z_{ij} \\ &\leq C\frac{1}{\binom{n}{2}}\sum_{i<j}|D_{ij}^T(\hat{\theta} - \theta_0)|, \quad \text{by Lipschitz continuity} \\ &\leq \frac{C}{\binom{n}{2}}\sum_{i<j}\left\{|\hat{\beta}_i - \beta_{0,i}| + |\hat{\beta}_j - \beta_{0,j}| + |\hat{\mu} - \mu_0| + |Z_{ij}^T(\hat{\gamma} - \gamma_0)|\right\} \\ &\leq \frac{C}{\binom{n}{2}}\underbrace{\left\{\sum_{i<j}|\hat{\beta}_i - \beta_{0,i}| + |\hat{\beta}_j - \beta_{0,j}|\right\}}_{=(n-1)\|\hat{\beta}-\beta_0\|_1} + C|\hat{\mu} - \mu_0| + C\|\hat{\gamma} - \gamma_0\|_1 \\ &\leq C\left\{\frac{1}{n}\|\hat{\beta} - \beta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1\right\} \\ &= O_P\left(s_+^*\sqrt{\frac{\log(n)}{\binom{n}{2}}}\rho_n^{-1}\right),\end{aligned}$$

where the last equality holds under the conditions of Theorem 2.4. Since the dimension of (I) is $(p+1)\times(p+1)$ and thus remains fixed, any row of (I) has ℓ_1-norm of order $O_P\left(s_+^*\sqrt{\frac{\log(n)}{\binom{n}{2}}}\rho_n^{-1}\right)$ and thus

$$\|(I)\|_\infty = O_P\left(s_+^*\sqrt{\frac{\log(n)}{\binom{n}{2}}}\rho_n^{-1}\right).$$

Taking a look at the (k,l)-th element in (II):

$$\begin{aligned}&\left|\frac{1}{\binom{n}{2}}(D_\vartheta^T W_0^2 D_\vartheta - \mathbb{E}[D_\vartheta^T W_0^2 D_\vartheta])_{kl}\right| \\ &\quad = \left|\frac{1}{\binom{n}{2}}\sum_{i<j}\left\{D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2 - \mathbb{E}[D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2]\right\}\right|.\end{aligned}$$

Note that the random variables $D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2$ are bounded uniformly in i, j, k, l. Thus, by Hoeffding's inequality, for any $t \geq 0$,

$$2\exp\left(-C\binom{n}{2}t^2\right) \geq P\left(\left|\frac{1}{\binom{n}{2}}\sum_{i<j}\left\{D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2 - \mathbb{E}[D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2]\right\}\right| \geq t\right).$$

This means, $\left|\frac{1}{\binom{n}{2}}\left(D_\vartheta^T W_0^2 D_\vartheta - \mathbb{E}[D_\vartheta^T W_0^2 D_\vartheta]\right)_{kl}\right| = O_P\left(\binom{n}{2}^{-1/2}\right)$. Again, since the dimension $p+1$ is fixed, we get by a simple union bound

$$\|(II)\|_\infty = O_P\left(\binom{n}{2}^{-1/2}\right).$$

In total, we thus get

$$\|\hat{\Sigma}_\vartheta - \Sigma_\vartheta\|_\infty = O_P\left(s_+^*\sqrt{\frac{\log(n)}{\binom{n}{2}}}\rho_n^{-1} + \frac{1}{\sqrt{\binom{n}{2}}}\right) = O_P\left(s_+^*\sqrt{\frac{\log(n)}{\binom{n}{2}}}\rho_n^{-1}\right).$$

We can now obtain a rate for $\|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty$:

$$\|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty \leq \frac{C}{\rho_n^2}\|\hat{\Sigma}_\vartheta - \Sigma_\vartheta\|_\infty = O_P\left(s_+^*\sqrt{\frac{\log(n)}{\binom{n}{2}}}\rho_n^{-3}\right).$$

By Assumption 2.3, we have $s_+^*\frac{\sqrt{\log(n)}}{\sqrt{n}\rho_n^2} \to 0, n \to \infty$, which in particular also implies that the above is $o_P(1)$. In particular we have now managed to prove for $k = 1, \ldots, p+1$,

- $\|\hat{\Theta}_{\vartheta,k} - \Theta_{\vartheta,k}\|_1 = o_P(1)$,
- $\hat{\Theta}_{\vartheta,k,k} = \Theta_{\vartheta,k,k} + o_p(1)$.

2.7.2.4 Problem 1

We can now take a look at the problems (1) - (3) outlined above. For problem (1), we want to show:

$$\sqrt{\binom{n}{2}}\frac{\hat{\Theta}_{\vartheta,k}P_n\nabla_\vartheta l_{\theta_0}}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \to \mathcal{N}(0,1).$$

Step 1: Show that

$$\hat{\Theta}_{\vartheta,k}P_n\nabla_\vartheta l_{\theta_0} = \Theta_{\vartheta,k}P_n\nabla_\vartheta l_{\theta_0} + o_P\left(\binom{n}{2}^{-1/2}\right). \tag{2.32}$$

We have

$$|(\hat{\Theta}_{\vartheta,k} - \Theta_{\vartheta,k})P_n\nabla_\vartheta l_{\theta_0}| \leq \|\hat{\Theta}_{\vartheta,k} - \Theta_{\vartheta,k}\|_1 \left\| \frac{1}{\binom{n}{2}} \sum_{i<j} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} (p_{ij}(\theta_0) - A_{ij}) \right\|_\infty$$
$$\leq \|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty \left\| \frac{1}{\binom{n}{2}} \sum_{i<j} D_{\vartheta,ij}(p_{ij}(\theta_0) - A_{ij}) \right\|_\infty .$$

Consider the vector $\sum_{i<j} D_{\vartheta,ij}(p_{ij}(\theta_0) - A_{ij}) \in \mathbb{R}^{p+1}$. The kth component of it has the form $\sum_{i<j}(p_{ij}(\theta_0) - A_{ij})$ for $k = 1$ and $\sum_{i<j} Z_{ij,k-1}(p_{ij}(\theta_0) - A_{ij}), k = 2, \ldots, p+1$. These components are all centred:

$$\mathbb{E}[D_{\vartheta,ij,k}(p_{ij}(\theta_0) - A_{ij})] = \mathbb{E}[D_{\vartheta,ij,k}\mathbb{E}[(p_{ij}(\theta_0) - A_{ij})|Z_{ij}]] = \mathbb{E}[D_{\vartheta,ij,k} \cdot 0] = 0.$$

Also, $|D_{\vartheta,ij,k}(p_{ij}(\theta_0) - A_{ij})| \leq c$, where $c > 1$ is a universal constant bounding $|Z_{ij,k}|$ for all i, j, k. Thus, by Hoeffding's inequality, for any $t > 0$,

$$P\left(\left| \frac{1}{\binom{n}{2}} \sum_{i<j} D_{\vartheta,ij,k}(p_{ij}(\theta_0) - A_{ij}) \right| \geq t \right) \leq 2\exp\left(-2\frac{\binom{n}{2}t^2}{c^2} \right)$$

and thus,

$$\frac{1}{\binom{n}{2}} \sum_{i<j} D_{\vartheta,ij}(p_{ij}(\theta_0) - A_{ij}) = O_P\left(\binom{n}{2}^{-1/2} \right).$$

Since we have $\|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty = o_P(1)$, by Section 2.7.2.3, Step 1 is now concluded.

Step 2: Show that

$$\hat{\Theta}_{\vartheta,k,k} = \Theta_{\vartheta,k,k} + o_P(1).$$

Since $\|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty = o_P(1)$, by Section 2.7.2.3, for all k

$$|\hat{\Theta}_{\vartheta,k,k} - \Theta_{\vartheta,k,k}| \leq \|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty = o_P(1)$$

and Step 2 is concluded.

Step 3: Show that

$$\left| \frac{1}{\Theta_{\vartheta,k,k}} \right| \leq C < \infty,$$

for some universal constant $C > 0$. Then, we may conclude from Step 1 and Step 2 that

$$\sqrt{\binom{n}{2}} \frac{\hat{\Theta}_{\vartheta,k} P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} = \sqrt{\binom{n}{2}} \frac{\Theta_{\vartheta,k} P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{\Theta_{\vartheta,k,k}}} + o_P(1).$$

To prove Step 3, notice that Θ_ϑ is symmetric and hence has only real eigenvalues. Therefore it is unitarily diagonalizable and for any $x \in \mathbb{R}^{p+1}$, we have $x^T\Theta_\vartheta x \geq$

$\text{mineval}(\Theta_\vartheta)\|x\|_2^2$. We also know that

$$\text{mineval}(\Theta_\vartheta) = \frac{1}{\text{maxeval}(\Sigma_\vartheta)}.$$

Under Assumption 2.1 we can deduce an upper bound on the maximum eigenvalue of Σ_ϑ: For any $x \in \mathbb{R}^p$,

$$x^T\Sigma_\vartheta x = x^T \frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T W_0^2 D_\vartheta]x \leq x^T \frac{1}{\binom{n}{2}}\mathbb{E}[D_\vartheta^T D_\vartheta]x \leq (1 \vee \lambda_{\max})\|x\|_2^2,$$

where we used that any entry in W_0^2 is bounded above by one. Since $x^T\Sigma_\vartheta x \leq \text{maxeval}(\Sigma_\vartheta)\cdot\|x\|_2^2$ and since this bound is tight, by Assumption 2.1, $\text{maxeval}(\Sigma_\vartheta) \leq (1 \vee \lambda_{\max}) \leq C < \infty$ for some constant $C > 0$.

In particular, since $\Theta_{\vartheta,k,k} = e_k^T\Theta_\vartheta e_k$, we get

$$\Theta_{\vartheta,k,k} \geq \text{mineval}(\Theta_\vartheta)\|e_k\|_2^2 = \frac{1}{\text{maxeval}(\Sigma_\vartheta)} \geq C > 0,$$

uniformly for all n. Consequently,

$$0 < \frac{1}{\Theta_{\vartheta,k,k}} \leq C < \infty.$$

Step 3 is thus concluded.

Step 4: Finally, show that

$$\sqrt{\binom{n}{2}}\frac{\Theta_{\vartheta,k}P_n\nabla_\vartheta l_{\theta_0}}{\sqrt{\Theta_{\vartheta,k,k}}} \xrightarrow{d} \mathcal{N}(0,1),$$

For brevity, write p_{ij} for the true link probabilities $p_{ij}(\theta_0)$. Keep in mind that $\Theta_{\vartheta,k}$ denotes the kth *row* of Θ_ϑ, while $D_{\vartheta,ij}$ denote $((p+1)\times 1)$-*column* vectors. We want to apply the Lindeberg-Feller Central Limit Theorem (CLT). The random variables we study are the summands in

$$\sqrt{\binom{n}{2}}\Theta_{\vartheta,k}P_n\nabla_\vartheta l_{\theta_0} = \sum_{i<j}\left\{\frac{1}{\sqrt{\binom{n}{2}}}\Theta_{\vartheta,k}D_{\vartheta,ij}(p_{ij} - A_{ij})\right\}.$$

These random variables are centred:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\sqrt{\binom{n}{2}}}\Theta_{\vartheta,k}D_{\vartheta,ij}(p_{ij} - A_{ij})\right] &= \mathbb{E}\left[\frac{1}{\sqrt{\binom{n}{2}}}\Theta_{\vartheta,k}D_{\vartheta,ij}\mathbb{E}[p_{ij} - A_{ij}|Z_{ij}]\right] \\ &= \mathbb{E}\left[\frac{1}{\sqrt{\binom{n}{2}}}\Theta_{\vartheta,k}D_{\vartheta,ij}\cdot 0\right] = 0.\end{aligned}$$

For the Lindeberg-Feller CLT we need to sum up the variances of these random

variables. We claim that

$$\sum_{i<j} \mathrm{Var}\left(\frac{1}{\sqrt{\binom{n}{2}}}\Theta_{\vartheta,k} D_{\vartheta,ij}(p_{ij} - A_{ij})\right) = \Theta_{\vartheta,k,k}.$$

Indeed, consider the vector-valued random variable $\sum_{i<j}\left\{\frac{1}{\sqrt{\binom{n}{2}}} D_{\vartheta,ij}(p_{ij} - A_{ij})\right\} \in \mathbb{R}^{p+1}$. It has covariance matrix

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i<j}\left\{\frac{1}{\sqrt{\binom{n}{2}}} D_{\vartheta,ij}(p_{ij} - A_{ij})\right\}\sum_{i<j}\left\{\frac{1}{\sqrt{\binom{n}{2}}} D_{\vartheta,ij}(p_{ij} - A_{ij})\right\}^T\right] \\
&= \mathbb{E}\left[\sum_{i<j}\frac{1}{\sqrt{\binom{n}{2}}} D_{\vartheta,ij}(p_{ij} - A_{ij})\frac{1}{\sqrt{\binom{n}{2}}} D^T_{\vartheta,ij}(p_{ij} - A_{ij})\right], \text{ independence across } i,j \\
&= \frac{1}{\binom{n}{2}}\sum_{i<j}\left[\mathbb{E}[D_{\vartheta,ij,k} D_{\vartheta,ij,l}(p_{ij} - A_{ij})^2]\right]_{k,l=1,\ldots,p+1}, \quad \text{as } ((p+1)\times(p+1))\text{-matrix} \\
&= \frac{1}{\binom{n}{2}}\mathbb{E}[D^T_\vartheta W_0^2 D_\vartheta] \\
&= \Sigma_\vartheta.
\end{aligned}$$

Thus, by independence across i, j,

$$\begin{aligned}
\sum_{i<j} \mathrm{Var}\left(\frac{1}{\sqrt{\binom{n}{2}}}\Theta_{\vartheta,k} D_{\vartheta,ij}(p_{ij} - A_{ij})\right) &= \mathrm{Var}\left(\Theta_{\vartheta,k}\sum_{i<j}\frac{1}{\sqrt{\binom{n}{2}}} D_{\vartheta,ij}(p_{ij} - A_{ij})\right) \\
&= \Theta_{\vartheta,k}\Sigma_\vartheta\Theta^T_{\vartheta,k} = \Theta_{\vartheta,k,k},
\end{aligned}$$

where for the last equality we have used that Θ_ϑ is the inverse of Σ_ϑ and thus, $\Sigma_\vartheta\Theta^T_{\vartheta,k} = e_k$. Now, we need to show that the Lindeberg condition holds. That is, we want for any $\epsilon > 0$,

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{\Theta_{\vartheta,k,k}}\sum_{i<j}\mathbb{E}\Bigg[&\left\{\frac{1}{\sqrt{\binom{n}{2}}}\Theta_{\vartheta,k} D_{\vartheta,ij}(p_{ij} - A_{ij})\right\}^2 \\
&\mathbb{1}\left(|\Theta_{\vartheta,k} D_{\vartheta,ij}(p_{ij} - A_{ij})| > \epsilon\sqrt{\binom{n}{2}\Theta_{\vartheta,k,k}}\right)\Bigg] = 0.
\end{aligned} \tag{2.33}$$

We have

$$|\Theta_{\vartheta,k} D_{\vartheta,ij}(p_{ij} - A_{ij})| \le p \cdot c \cdot \|\Theta_{\vartheta,k}\|_1 \le C\|\Theta_\vartheta\|_\infty \le C\rho_n^{-1}.$$

We know from Step 3 that $\Theta_{Z,k,k} \ge C > 0$ for some universal C. Then, as long as ρ_n^{-1} goes to infinity at a rate slower than n, which is enforced by Assumption 2.3,

we must have for n large enough

$$|\Theta_{\vartheta,k} D_{\vartheta,ij}(p_{ij} - A_{ij})| < \epsilon \sqrt{\binom{n}{2} \Theta_{\vartheta,k,k}}$$

uniformly in i, j. Thus, the indicator function and therefore each summand in (2.33) is equal to zero for n large enough. Hence, (2.33) holds. Then, by the Lindeberg-Feller CLT,

$$\sqrt{\binom{n}{2}} \frac{\Theta_{\vartheta,k} P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{\Theta_{\vartheta,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Now, by Steps 1-4 and Slutzky's Theorem

$$\begin{aligned}
\sqrt{\binom{n}{2}} \frac{\hat{\Theta}_{\vartheta,k} P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} &= \sqrt{\binom{n}{2}} \frac{(\Theta_{\vartheta,k} + o_P(1)) P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{(\Theta_{\vartheta,k,k} + o_P(1))}} \\
&= \sqrt{\binom{n}{2}} \frac{\Theta_{\vartheta,k} P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{(\Theta_{\vartheta,k,k} + o_P(1))}} + \sqrt{\binom{n}{2}} \frac{o_P(1) P_n \nabla_\vartheta l_{\theta_0}}{\sqrt{(\Theta_{\vartheta,k,k} + o_P(1))}} \\
&\xrightarrow{d} \mathcal{N}(0,1).
\end{aligned}$$

This concludes solving Problem 1.

2.7.2.5 Problem 2

For Problem 2 we must show

$$\frac{1}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 X(\hat{\beta} - \beta_0) = o_P\left(\binom{n}{2}^{-1/2}\right).$$

Since we have $\|\hat{\Theta}_\vartheta - \Theta_\vartheta\|_\infty = o_P(1)$, we do not need to worry about $\hat{\Theta}_{Z,k,k}^{-1/2}$, because $\hat{\Theta}_{Z,k,k} = \Theta_{Z,k,k} + o_P(1)$ and $\Theta_{Z,k,k}^{-1/2} \leq C < \infty$, i.e. $\hat{\Theta}_{Z,k,k}^{-1/2} = O_P(1)$. By Theorem 2.4 we also have a high-probability error bound on $\|\hat{\beta} - \beta_0\|_1$. We have,

$$\left| \hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 X(\hat{\beta} - \beta_0) \right| \leq \left\| \frac{1}{\binom{n}{2}} X^T \hat{W}^2 D_\vartheta \hat{\Theta}_{\vartheta,k}^T \right\|_\infty \|\hat{\beta} - \beta_0\|_1.$$

Notice that in the display above we have the vector ℓ_∞-norm. Also,

$$\left\| \frac{1}{\binom{n}{2}} X^T \hat{W}^2 D_\vartheta \hat{\Theta}_{\vartheta,k}^T \right\|_\infty \leq \|\hat{\Theta}_{\vartheta,k}^T\|_\infty \left\| \frac{1}{\binom{n}{2}} X^T \hat{W}^2 D_\vartheta \right\|_\infty.$$

Here we used the compatibility of the matrix ℓ_∞-norm with the vector ℓ_∞-norm. The first term is the vector norm, the second the matrix norm. We know,

$$\|\hat{\Theta}_{\vartheta,k}^T\|_\infty \leq \|\hat{\Theta}_\vartheta\|_\infty \leq C \rho_n^{-1},$$

where on the left hand side we have the vector norm and in the middle display the matrix norm. Finally, $\frac{1}{\binom{n}{2}} X^T \hat{W}^2 D_\vartheta$ is a $(n \times (p+1))$-matrix. The (k,l)-th element looks like

$$\left| \frac{1}{\binom{n}{2}} \sum_{i=1, i \neq l}^{n} D_{\vartheta, il, k} \hat{w}_{il}^2 \right| \leq \frac{1}{\binom{n}{2}} \cdot (n-1) \cdot C = \frac{C}{n}.$$

Thus, the ℓ_1-norm of any row of $\frac{1}{\binom{n}{2}} X^T \hat{W}^2 D_\vartheta$ is bounded by C/n and thus

$$\left\| \frac{1}{\binom{n}{2}} X^T \hat{W}^2 D_\vartheta \right\|_\infty \leq \frac{C}{n}.$$

Recall that $\|\hat{\beta} - \beta_0\|_1 = O_P\left(s_+^* \frac{\sqrt{\log(n)}}{\sqrt{n}} \rho_n^{-1}\right)$ by Theorem 2.4. Then,

$$\begin{aligned} \left| \hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} X^T \hat{W}^2 D_\vartheta (\hat{\beta} - \beta_0) \right| &\leq \|\hat{\Theta}_{\vartheta,k}^T\|_\infty \left\| \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 X \right\|_\infty \|\hat{\beta} - \beta_0\|_1 \\ &= O_P\left(\frac{s_+^*}{\rho_n^2 \cdot n} \cdot \frac{\sqrt{\log(n)}}{\sqrt{n}} \right). \end{aligned}$$

Multiplying by $\sqrt{\binom{n}{2}} = O(n)$, gives

$$\sqrt{\binom{n}{2}} \left| \hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} D_\vartheta^T \hat{W}^2 X (\hat{\beta} - \beta_0) \right| = O_P\left(\frac{s_+^*}{\rho_n^2} \cdot \frac{\sqrt{\log(n)}}{\sqrt{n}} \right),$$

which is $o_P(1)$ under Assumption 2.3.

2.7.2.6 Problem 3

Finally, we must show

$$O\left(\frac{1}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}} \hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} \sum_{i<j} \begin{pmatrix} 1 \\ Z_{ij} \end{pmatrix} |D_{ij}^T (\hat{\theta} - \theta_0)|^2 \right) = o_P\left(\binom{n}{2}^{-1/2} \right).$$

Again, since $\hat{\Theta}_{\vartheta,k,k} = \Theta_{\vartheta,k,k} + o_P(1)$ and $\Theta_{\vartheta,k,k} \geq C > 0$ uniformly in n, we do not need to worry about the factor $\frac{1}{\sqrt{\hat{\Theta}_{\vartheta,k,k}}}$ and it remains to show

$$O\left(\hat{\Theta}_{\vartheta,k} \frac{1}{\binom{n}{2}} \sum_{i<j} D_{\vartheta,ij} |D_{ij}^\top (\hat{\theta} - \theta_0)|^2 \right) = o_P\left(\binom{n}{2}^{-1/2} \right).$$

We have

$$\begin{aligned}\left|\hat{\Theta}_{\vartheta,k}\frac{1}{\binom{n}{2}}\sum_{i<j}D_{\vartheta,ij}|D_{ij}^{\top}(\hat{\theta}-\theta_0)|^2\right| &\leq \frac{1}{\binom{n}{2}}\sum_{i<j}|\hat{\Theta}_{\vartheta,k}D_{\vartheta,ij}||D_{ij}^{T}(\hat{\theta}-\theta_0)|^2\\ &\leq C\|\hat{\Theta}_{\vartheta,k}\|_1\frac{1}{\binom{n}{2}}\sum_{i<j}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2\\ &\leq C\frac{1}{\rho_n}\frac{1}{\binom{n}{2}}\sum_{i<j}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2,\end{aligned}$$

where for the last inequality we have used $\|\hat{\Theta}_{\vartheta,k}\|_1 \leq \|\hat{\Theta}_{\vartheta}\|_\infty \leq C\frac{1}{\rho_n}$. Remember from (2.23) that

$$\frac{1}{\binom{n}{2}}\sum_{i<j}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2 \leq C\|\hat{\bar{\theta}}-\bar{\theta}_0\|_1^2,$$

where we make use of the fact that $\theta^* = \theta_0$ if there is no approximation error (as assumed by Theorem 2.7) and that $\bar{D}\bar{\theta} = D\theta$. From Theorem 2.4 we know that under the assumptions of Theorem 2.7, $\|\hat{\bar{\theta}}-\bar{\theta}_0\|_1 = O_P\left(s_+^*\sqrt{\frac{\log(n)}{\binom{n}{2}}}\rho_n^{-1}\right)$. Thus,

$$\sqrt{\binom{n}{2}}\left|\hat{\Theta}_{\vartheta,k}\frac{1}{\binom{n}{2}}\sum_{i<j}D_{\vartheta,ij}|D_{ij}^{T}(\hat{\theta}-\theta_0)|^2\right| = O_P\left((s_+^*)^2\frac{\log(n)}{\sqrt{\binom{n}{2}}}\rho_n^{-3}\right).$$

We see that this is $o_P(1)$ by applying Assumption 2.3 twice. Problem 3 is solved.

Proof of Theorem 2.7. Theorem 2.7 now follows from the solved problems (1) - (3). □

Chapter 3

A sparse Erdős-Rényi model with covariates

Organization of this chapter

We introduce the sparse Erdős-Rényi with covariates (ER-C), which is a special case of SβM-C, where the degree heterogeneity parameter has been set to zero. We formally introduce this model in Section 3.1. We prove the asymptotic normality of its MLE in Theorem 3.1 and show that this model can generate networks with almost arbitrary levels of sparsity. This is followed by an extensive set of simulation studies in Section 3.2. We take a brief detour in Section 3.3 and discuss the connectivity patterns in ER-C. All the proofs are relegated to Section 3.4. The content of Sections 3.1 and 3.2 is from Stein & Leng (2020).

3.1 SβM-C without β

We zoom in on a second special case of SβM-C when the heterogeneity parameter β equals zero. This model retains many of the favourable properties of SβM-C and can model networks of almost arbitrary sparsity. In particular, it can avoid the issue of data-selective inference for almost any degree of sparsity.

When $\beta = 0$, the linking probabilities in SβM-C become

$$P(A_{ij} = 1|Z_{ij}) = p_{ij} = \frac{\exp(\mu + Z_{ij}^T\gamma)}{1 + \exp(\mu + Z_{ij}^T\gamma)}. \tag{3.1}$$

We remark that the setup in latter case is different to the usual logistic regression as we allow $\mu \to -\infty$ and thus allow for sparse networks. The model in (3.1) can be seen as an extension of the Erdős-Rényi model by incorporating covariates, with an emphasis on modelling sparse networks. For this reason, we call it the sparse Erdős-Rényi model with covariates (ER-C). To the best of our knowledge, a model of this type has not been studied in the literature before Stein & Leng (2020) and thus the results below can be of independent interest.

We study the properties of the MLE of μ and γ under the sparse network regime. Towards this, following Chen et al. (2020), we encode the sparsity of model (3.1) explicitly by assuming a reparametrization of the global sparsity parameter μ of the form

$$\mu = -\xi \log(n) + \mu^\dagger,$$

where $\xi \in [0, 2)$ effectively takes the role of $\rho_{n,0}$ from the previous chapter and $\mu^\dagger \in [-M, M]$ for a fixed $M < \infty$ independent of n. As before, we assume that the entries $Z_{ij,k}$ are uniformly bounded almost surely and that the homophily parameter γ lies in a compact, convex set $\Gamma \subseteq \mathbb{R}^p$. To appreciate this reformulation, notice that the expected total number of edges of ER-C is of order $O(n^{2-\xi})$. When $\xi = 0$, ER-C becomes a standard logistic regression model with fixed parameters. It can generate almost arbitrarily sparse networks when $\xi > 0$.

We denote the true parameters $\mu_0^\dagger$ and γ_0 respectively. Similar to Section 2.3, we abuse notation slightly and denote a generic parameter as $\theta = (\mu^\dagger, \gamma)$, the true parameter as $\theta_0 = (\mu_0^\dagger, \gamma_0)$ and our estimator (defined below) as $\hat{\theta} = (\hat{\mu}^\dagger, \hat{\gamma})$. We think this abuse of notation is justified as it allows a consistent notation with the previous chapter. We make the following assumptions.

Assumption 3.1. $\theta_0 = (\mu_0^\dagger, \gamma_0^T)^T$ lies in the interior of $[-M, M] \times \Gamma$.

Assumption 3.2. The Z_{ij} are *i.i.d.* realizations of the same random variable. The covariance matrix of Z_{12}, that is, the matrix $\mathbb{E}[Z_{12}Z_{12}^T]$, is strictly positive definite with minimum eigenvalue $\lambda_{\min} > 0$.

Assumption 3.2 is analogous to Assumption 2.1 in the case with non-zero β. We remark that the *i.i.d.* condition is used to simplify parts of the proofs and can be relaxed at the expense of lengthier proofs. We consider the following function which is proportional to the negative log-likelihood of the ER-C up to a summand independent of θ,

$$\mathcal{L}^\dagger(\mu^\dagger, \gamma) = -d_+\mu^\dagger - \sum_{i<j}(\gamma^T Z_{ij})A_{ij} + \sum_{i<j}\log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right). \quad (3.2)$$

In the ER-C, the dimension of the parameter is fixed as $p + 1$. Therefore, it is not necessary to employ a penalized likelihood approach as in the SβM-C and we estimate θ_0 via maximum likelihood:

$$\hat{\theta} = (\hat{\mu}^\dagger, \hat{\gamma}^T)^T = \underset{\theta=(\mu^\dagger,\gamma^T)^T}{\arg\min}\ \mathcal{L}^\dagger(\mu^\dagger, \gamma), \quad (3.3)$$

where the argmin is taken over $[-M, M] \times \Gamma$. The design matrix D now takes the simplified form

$$D = \left[\ \mathbf{1}\ \middle|\ Z\ \right] \in \mathbb{R}^{\binom{n}{2}\times(p+1)}.$$

As before, we enumerate the rows of D as $D_{ij}^T, i < j$, where each D_{ij} is treated as a column vector, i.e. $D = [D_{ij}^T]_{i<j}$. Define the matrix $\Sigma \in \mathbb{R}^{(p+1)\times(p+1)}$ as

$$\Sigma \coloneqq \mathbb{E}\left[(D_{12}D_{12}^T)\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12})\right],$$

which is invertible by Assumption 3.2. We have the following central limit theorem for $\hat{\theta}$, the proof of which can be found in Section 3.4. Denote by $\mathcal{N}(0, B)$ the law of the multivariate normal distribution with zero mean vector and covariance matrix B.

Theorem 3.1. *Under Assumptions 3.1 and 3.2, it holds, as $n \to \infty$,*

$$\sqrt{\frac{\binom{n}{2}}{n^\xi}}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}).$$

Since the expected number of observed edges in the ER-C is of order $n^{2-\xi}$, the factor $\sqrt{\binom{n}{2}/n^\xi}$ in Theorem 3.1 corresponds to the square root of the effective sample size. This means, having the link probabilities go to zero reduces the information we gain about θ_0 and this information loss is made explicit in a rate of convergence slower than what we would obtain in a classical parametric setting. This finding is in line with the results in Chen et al. (2020), Proposition 1 and Theorem 1, in which a similar phenomenon was observed for SβM.

While Theorem 3.1 can be interesting from a theoretical point of view, in practice, the sparsity-rate parameter ξ will not be known, which makes solving (3.3) and finding the MLE $(\hat{\mu}^\dagger, \hat{\gamma})$ impossible. It is possible, though, to circumvent this problem with the following argument.

From Theorem 3.1 we obtain for any $k = 1, \ldots, (p+1)$,

$$\sqrt{\frac{\binom{n}{2}}{n^\xi}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{\Sigma^{-1}_{k,k}}} \xrightarrow{d} \mathcal{N}(0, 1),$$

where $\mathcal{N}(0, 1)$ denotes the law of the univariate standard-normal distribution. We also may make use of the the identity

$$\hat{\mu} = -\xi \log(n) + \hat{\mu}^\dagger, \tag{3.4}$$

where $\hat{\mu}$ is the MLE of the global sparsity parameter *before* reparametrization. In particular, $\hat{\mu}$ can be found without knowledge of ξ. Define the matrix

$$\hat{\Sigma} = \frac{1}{\binom{n}{2}} D^T \operatorname{diag}\left(\frac{\exp(\hat{\mu} + \hat{\gamma}^T Z_{ij})}{(1 + \exp(\hat{\mu} + \hat{\gamma}^T Z_{ij}))^2}, i < j\right) D.$$

In Section 3.4 we show $n^{\xi}\hat{\Sigma} = \Sigma + o_P(1)$, which allows us to show $(n^{\xi}\hat{\Sigma})^{-1}_{k,k} = \Sigma^{-1}_{k,k} + o_P(1)$ for all $k = 1, \ldots, (p+1)$. Then, by Slutzky's Theorem,

$$\sqrt{\binom{n}{2}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{\hat{\Sigma}^{-1}_{k,k}}} = \sqrt{\frac{\binom{n}{2}}{n^{\xi}}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{n^{-\xi}\hat{\Sigma}^{-1}_{k,k}}} = \sqrt{\frac{\binom{n}{2}}{n^{\xi}}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{\Sigma^{-1}_{k,k} + o_P(1)}} \xrightarrow{d} \mathcal{N}(0,1).$$

In other words, the matrix $\hat{\Sigma}$ will be singular in the limit as the link probabilities p_{ij} go to zero. The rate n^{ξ} is precisely the rate with which we need to multiply $\hat{\Sigma}$ to stabilize it and make it converge to the non-singular matrix Σ, whose inverse is the asymptotic covariance matrix in Theorem 3.1. Thus allowing us to derive the component-wise limiting distribution of each $\hat{\theta}_k$ without the knowledge of ξ. In particular, looking at the case $k = 2, \ldots, (p+1)$, we are able to calculate confidence intervals for the components of γ without having to know ξ. In summary, Theorem 3.1 allows the following corollary which is proved in Section 3.4.

Corollary 3.2. *Under Assumptions 3.1 and 3.2, as $n \to \infty$, for $k = 1, \ldots, p$,*

$$\sqrt{\binom{n}{2}} \cdot \frac{\hat{\gamma}_k - \gamma_{0,k}}{\sqrt{\hat{\Sigma}^{-1}_{k+1,k+1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Simulation results corroborating the claims in Corollary 3.2 are shown in Section 3.2.

3.2 Simulation: ER-C

We illustrate the finite sample performance of the MLE (3.3). We focus on inference for γ in the more realistic case of unknown ξ. That is, we use the identity (3.4) to estimate μ_0 rather than $\mu_0^{\dagger}$. Our emphasis is on illustrating that the MLE can be used to perform inference in extremely sparse network settings. To that end we fixed $p = 20$ and a true parameter vector $(\mu_0^{\dagger}, \gamma_0^T)^T$ and varied the sparsity parameter ξ. The exact model setup was as follows. We sampled the covariate values $Z_{ij,k}, k = 1, \ldots, p, i < j$ from a centred Beta $(2,2)$ distribution. We used $\mu_0^{\dagger} = 1$ and $\gamma_0 = (1.5, 1.2, 0.8, 1, \ldots, 1)^T$. For ξ we used the values $\xi = 0.3, 1.0, 1.5$. We sampled networks of sizes $n = 300, 500, 800, 1000$, and for each configuration we drew 1,000 realizations of the ER-C and analysed the performance of the MLE (3.3). Table 3.1 gives the median observed edge densities and median minimum and maximum link probabilities across all 1,000 repetitions. The sparsest case $\xi = 1.5$ is close to the maximum theoretically permissible sparsity and results in extremely sparse networks. For $n =$ 1,000, on average, only 73 out of the almost half million possible edges are observed in this setting.

	n	Median edge density	$\min p_{ij}$	$\max p_{ij}$
$\xi = 0.3$				
	300	0.358	7.5×10^{-3}	0.9698
	500	0.330	5.1×10^{-3}	0.9712
	800	0.304	3.8×10^{-3}	0.9724
	1,000	0.292	3.2×10^{-3}	0.9730
$\xi = 1.0$				
	300	1.48×10^{-2}	1.4×10^{-4}	0.3724
	500	9.03×10^{-3}	6.6×10^{-5}	0.3058
	800	5.69×10^{-3}	3.5×10^{-5}	0.2515
	1,000	4.56×10^{-3}	2.6×10^{-5}	0.2211
$\xi = 1.5$				
	300	8.70×10^{-4}	8.1×10^{-6}	0.0331
	500	4.17×10^{-4}	3.0×10^{-6}	0.0193
	800	2.03×10^{-4}	1.2×10^{-6}	0.0117
	1,000	1.46×10^{-4}	8.0×10^{-7}	0.0089

Table 3.1: Network density in the ER-C for different n and ξ. The columns $\min p_{ij}$ and $\max p_{ij}$ give the median minimum and median maximum link probability between two nodes i and j across all 1,000 repetitions.

The asymptotic normality for each component of $\hat{\gamma}$ allows us to construct confidence intervals at the 95%-level as prescribed by Corollary 3.2. We assess the performance of our MLE by calculating the empirical coverage for each component. There is no significant difference in the empirical coverage or the average length of the confidence intervals between the various components of γ, which is why we only present them for $\gamma_{0,1}$ in Table 3.2. As we can see, coverage is very close to the nominal confidence level of 95% and the length of the confidence intervals decreases with increasing network size. As expected, confidence intervals are larger for sparse networks. For $\xi = 1.5$ we observe very wide confidence intervals, which is due to the very low effective sample size.

	Coverage	CI	Coverage	CI	Coverage	CI
n	$\xi = 0.3$		$\xi = 1.0$		$\xi = 1.5$	
300	0.941	0.193	0.956	0.711	0.944	2.892
500	0.955	0.118	0.938	0.541	0.967	2.505
800	0.943	0.075	0.950	0.424	0.951	2.235
1000	0.949	0.061	0.935	0.379	0.948	2.107

Table 3.2: Empirical coverage under nominal 95% coverage and median lengths of confidence intervals for γ_1. The results are similar for the other components of γ.

3.3 A quick aside: Connectivity threshold in the ER-C

We mostly focus on inference of the model parameters driving network formation. What we do not touch upon is what the generated networks "look like" other than being sparse. This is for good reason, since the study of the connectivity properties of random networks is a very broad and complex field in its own right. Nonetheless, in the special case of ER-C, it is possible to draw upon some of the deep results

derived for the connectivity behaviour in the Erdős-Rényi model, using a coupling argument. This allows us to derive similar results for ER-C almost "for free". The proof is quite elegant and simple and presented in Section 3.4.3.

Recall that a undirected, simple graph $G = (V, E)$ is called *connected*, if for any pair $i, j \in V$ there exists a path in E from u to j. We call G *disconnected* otherwise. We make the following claim.

Theorem 3.3. *As n grows, the realizations of ER-C are connected with probability approaching one if $\xi \in [0, 1)$. Also, realizations of ER-C will be disconnected with probability approaching one if $\xi \geq 1$.*

To prove Theorem 3.3, we make use of an analogous result in the classical Erdős-Rényi model. Denote by $\mathrm{ER}(n, p)$ the law of the Erdős-Rényi model with n nodes and link probability p. Define $\lambda = \lambda(n) = np$. The connectivity of the $\mathrm{ER}(n, p)$ depends almost entirely on the value of λ, see for example van der Hofstad (2016), Chapters 4 and 5. In particular, the following holds.

Theorem 3.4 (Connectivity Threshold, Theorem 5.8 in van der Hofstad (2016))**.** *For $\lambda - \log(n) \to \infty$, the Erdős-Rényi random graph is with high probability connected, while for $\lambda - \log(n) \to -\infty$ the Erdős-Rényi random graph is with high probability disconnected.*

Our strategy for proving Theorem 3.3 will be to couple the law of ER-C to the law of $\mathrm{ER}(n, p)$ for appropriately chosen p, which allows us to make use of Theorem 3.4. We then show that under this coupling, a realization from ER-C will contain at least the same edges as a realization from $\mathrm{ER}(n, p)$. This phenomenon is referred to as *monotonicity in the edge probabilities* in the literature (cf. van der Hofstad (2016), Section 4.1.1). Thus, if $\mathrm{ER}(n, p)$ is connected with high probability, the same must hold for ER-C. Reversing the roles of ER-C and $\mathrm{ER}(n, p)$ allows us to prove the statement about disconnected networks.

To give a visual interpretation of Theorem 3.3, we sample from the ER-C on $n = 300$ nodes, with $p = 5, \mu^\dagger = 1, \gamma_0 = (1.5, 1.2, 0.8, 1, 1)^T$ and varying $\xi = 0.8, 1, 1.2$. Figure 3.1 shows a realization of each of these three models. For $\xi = 0.8$ we observe a single connected component. In the threshold case for $\xi = 1$ there still is a giant component present, but we also observe many isolated nodes, while for $\xi = 1.2$ the graph breaks apart into many small pieces. The observed edge densities are: 4.34% for $\xi = 0.8$, 1.50% for $\xi = 1$ and 0.59% for $\xi = 1.2$. Hence, although the graph is connected for $\xi = 0.8$, it is still very sparse. Take note of how quickly this phase transition happens: In an Euclidean sense $\xi = 0.8$ and $\xi = 1.2$ are not too far apart. However, the networks they produce differ vastly in their geometric properties.

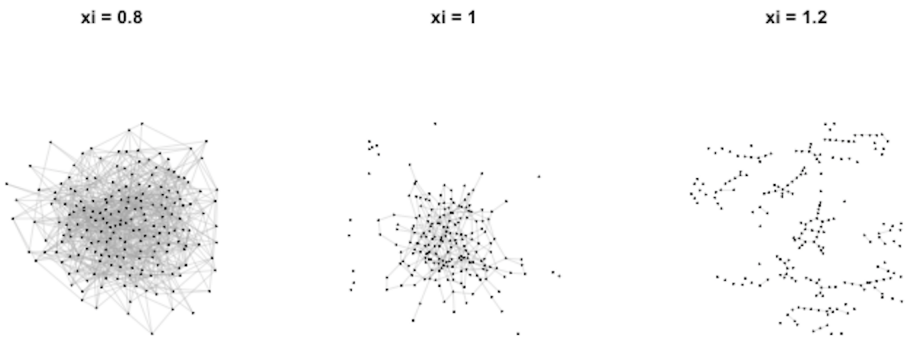


Figure 3.1: ER-C on $n = 300$ nodes, with $p = 5, \mu^\dagger = 1, \gamma_0 = (1.5, 1.2, 0.8, 1, 1)^T$ and varying $\xi = 0.8, 1, 1.2$. For $\xi = 0.8$ we observe a single connected component. In the threshold case for $\xi = 1$ there still is a giant component present, but we also observe many isolated nodes, while in for $\xi = 1.2$ the graph breaks apart into many small pieces.

3.4 Proofs of Chapter 3

We first prove the consistency of the MLE $\hat{\theta} = (\hat{\mu}^\dagger, \hat{\gamma}^T)^T$ and then its asymptotic normality.

3.4.1 Consistency of $(\hat{\mu}^\dagger, \hat{\gamma})$

We want to find a limit for an appropriately scaled version of $\mathcal{L}^\dagger$. To that end, we first prove a concentration result of d_+ around its expectation. Consider

$$\begin{aligned}
\mathbb{E}[d_+] = \mathbb{E}[\mathbb{E}[d_+|Z]] &= \sum_{i<j} \mathbb{E}\left[\frac{n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{ij})}{1+n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{ij})}\right] \\
&= n^{-\xi}\exp(\mu_0^\dagger)\sum_{i<j}\mathbb{E}\left[\frac{\exp(\gamma_0^T Z_{ij})}{1+n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{ij})}\right] \\
&= n^{-\xi}\exp(\mu_0^\dagger)\binom{n}{2}\mathbb{E}\left[\frac{\exp(\gamma_0^T Z_{12})}{1+n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12})}\right], \quad \text{since } Z_{ij} \text{ are i.i.d.} \\
&= \frac{n^{2-\xi}}{2}\exp(\mu_0^\dagger)\mathbb{E}\left[\frac{\exp(\gamma_0^T Z_{12})}{1+n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12})}\right] + o(n^{2-\xi}).
\end{aligned}$$

By the Law of Total Variance, we may write the variance of d_+ as

$$\mathrm{Var}(d_+) = \mathbb{E}[\mathrm{Var}(d_+|Z)] + \mathrm{Var}(\mathbb{E}[d_+|Z]).$$

We have,

$$\begin{aligned}
\mathrm{Var}(\mathbb{E}[d_+|Z]) &= \mathrm{Var}\left(\sum_{i<j} p_{ij}\right) \\
&= \sum_{i<j} n^{-2\xi}\mathrm{Var}\left(\frac{\exp(\mu_0^\dagger + \gamma_0^T Z_{ij})}{1+n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{ij})}\right) = O\left(n^{2-2\xi}\right).
\end{aligned}$$

Also, by independence of the A_{ij} given Z,

$$\mathrm{Var}(d_+|Z) = \sum_{i<j}\mathrm{Var}(A_{ij}|Z) = \sum_{i<j} p_{ij}(1-p_{ij}) = O(n^{2-\xi}).$$

Therefore,

$$\mathrm{Var}(d_+) = O\left(n^{2-2\xi}\right) + O(n^{2-\xi}) = O(n^{2-\xi}).$$

By Chebychev's inequality, for any $t > 0$,

$$P(|d_+ - \mathbb{E}[d_+]| \geq t) \leq \frac{\mathrm{Var}(d_+)}{t^2}.$$

Letting $\epsilon > 0$ and picking $t = n^{2-\xi}\epsilon$, we obtain

$$P(n^{-2+\xi}|d_+ - \mathbb{E}[d_+]| \geq \epsilon) \leq \frac{O(n^{2-\xi})}{n^{4-2\xi}} = \frac{O(1)}{n^{2-\xi}} \to 0, \quad n \to \infty,$$

since $\xi \in [0, 2)$. This implies

$$d_+ = \mathbb{E}[d_+] + o_P(n^{2-\xi}) = \frac{n^{2-\xi}}{2} \exp(\mu_0^\dagger) \mathbb{E}\left[\frac{\exp(\gamma_0^T Z_{12})}{1 + n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12})}\right] + o_P(n^{2-\xi}).$$

In particular, this implies

$$2n^{-2+\xi} d_+ \xrightarrow{P} \exp(\mu_0^\dagger)\mathbb{E}\left[\exp(\gamma_0^T Z_{12})\right], \quad n \to \infty. \tag{3.5}$$

Next, we deal with the second term in $\mathcal{L}^\dagger$:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i<j}(\gamma^T Z_{ij})A_{ij}\right] &= \sum_{i<j}\mathbb{E}\left[(\gamma^T Z_{ij})\mathbb{E}[A_{ij}|Z_{ij}]\right] = \sum_{i<j}\mathbb{E}\left[(\gamma^T Z_{ij})p_{ij}\right] \\
&= \sum_{i<j} n^{-\xi}\exp(\mu_0^\dagger)\mathbb{E}\left[(\gamma^T Z_{ij})\frac{\exp(\gamma_0^T Z_{ij})}{1 + n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{ij})}\right] \\
&= n^{-\xi}\exp(\mu_0^\dagger)\binom{n}{2}\mathbb{E}\left[(\gamma^T Z_{12})\frac{\exp(\gamma_0^T Z_{12})}{1 + n^{-\xi}\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12})}\right] \\
&=: n^{-\xi}\exp(\mu_0^\dagger)\binom{n}{2}\bar{\alpha}_n,
\end{aligned}$$

where we used that the Z_{ij} are i.i.d. in the penultimate equality and where we suppress the dependence of $\bar{\alpha}_n$ on γ in our notation. Pay special attention to the distinction between the generic γ and the true parameter γ_0 here. The last equality in the previous display can be written as

$$\mathbb{E}\left[\sum_{i<j}(\gamma^T Z_{ij})A_{ij}\right] = \frac{n^{2-\xi}}{2}\exp(\mu_0^\dagger)\bar{\alpha}_n + o(n^{2-\xi}).$$

We use the Law of Total Variance once more to bound $\mathrm{Var}(\sum_{i<j}(\gamma^T Z_{ij})A_{ij})$. For any i, j,

$$\mathrm{Var}((\gamma^T Z_{ij})A_{ij}) = \mathbb{E}[\mathrm{Var}((\gamma^T Z_{ij})A_{ij}|Z)] + \mathrm{Var}(\mathbb{E}[(\gamma^T Z_{ij})A_{ij}|Z]).$$

We have,

$$\mathrm{Var}(\mathbb{E}[(\gamma^T Z_{ij})A_{ij}|Z]) = \mathrm{Var}((\gamma^T Z_{ij})p_{ij}) \leq \mathbb{E}[\big((\gamma^T Z_{ij})p_{ij}\big)^2] \leq Cn^{-2\xi}$$

and

$$\mathrm{Var}((\gamma^T Z_{ij})A_{ij}|Z) = (\gamma^T Z_{ij})^2 p_{ij}(1 - p_{ij}) \leq Cn^{-\xi},$$

where in both instances we may choose some constant $C > 0$ independent of i, j and n. Thus,

$$\operatorname{Var}\left(\sum_{i<j}(\gamma^T Z_{ij})A_{ij}\right) \leq \sum_{i<j} C(n^{-2\xi} + n^{-\xi}) = O(n^{2-\xi}).$$

Using Chebyshev's inequality, we obtain for any $t > 0$,

$$P\left(\left|\sum_{i<j}(\gamma^T Z_{ij})A_{ij} - \mathbb{E}\left[\sum_{i<j}(\gamma^T Z_{ij})A_{ij}\right]\right| \geq t\right) \leq \frac{\operatorname{Var}\left(\sum_{i<j}(\gamma^T Z_{ij})A_{ij}\right)}{t^2}.$$

Letting $\epsilon > 0$ and picking $t = n^{2-\xi}\epsilon$, we obtain

$$P\left(n^{-2+\xi}\left|\sum_{i<j}(\gamma^T Z_{ij})A_{ij} - \mathbb{E}\left[\sum_{i<j}(\gamma^T Z_{ij})A_{ij}\right]\right| \geq \epsilon\right) \leq \frac{O(n^{2-\xi})}{n^{2-\xi}\cdot n^{2-\xi}} \to 0.$$

This implies

$$\sum_{i<j}(\gamma^T Z_{ij})A_{ij} = \mathbb{E}\left[\sum_{i<j}(\gamma^T Z_{ij})A_{ij}\right] + o_P(n^{2-\xi}) = \frac{n^{2-\xi}}{2}\exp(\mu_0^\dagger)\bar{\alpha}_n + o_P(n^{2-\xi}).$$

Since $\bar{\alpha}_n \to \mathbb{E}[(\gamma^T Z_{12})\exp(\gamma_0^T Z_{12})]$ almost surely, we end up with

$$2n^{-2+\xi}\sum_{i<j}(\gamma^T Z_{ij})A_{ij} \xrightarrow{P} \exp(\mu_0^\dagger)\mathbb{E}[(\gamma^T Z_{12})\exp(\gamma_0^T Z_{12})], \quad n \to \infty. \tag{3.6}$$

It remains to analyse $\sum_{i<j}\log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right)$, i.e. the last term in $\mathcal{L}^\dagger$. Since $\log(1+x) \leq x$ for $x > -1$:

$$\begin{aligned}
\sum_{i<j}\log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right) &\leq n^{-\xi}\exp(\mu^\dagger)\sum_{i<j}\exp(Z_{ij}^T\gamma) \\
&= n^{-\xi}\exp(\mu^\dagger)\binom{n}{2}\underbrace{\frac{1}{\binom{n}{2}}\sum_{i<j}\exp(Z_{ij}^T\gamma)}_{=:\alpha_n} \\
&= \frac{n^{2-\xi}}{2}\exp(\mu^\dagger)\alpha_n + o(n^{2-\xi}).
\end{aligned}$$

On the other hand, we also have $x/(1+x) \leq \log(1+x)$ for all $x > -1$. Also recall

that $|\gamma^T Z_{ij}| \leq \kappa$ almost surely. Thus,

$$\begin{aligned}
\sum_{i<j} \log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right) &\geq n^{-\xi}\exp(\mu^\dagger)\sum_{i<j}\frac{\exp(\gamma^T Z_{ij})}{1 + n^{-\xi}\exp(\mu^\dagger)\exp(\gamma^T Z_{ij})} \\
&\geq n^{-\xi}\exp(\mu^\dagger)\frac{1}{1 + n^{-\xi}\exp(\mu^\dagger + \kappa)}\sum_{i<j}\exp(\gamma^T Z_{ij}) \\
&= n^{-\xi}\exp(\mu^\dagger)\frac{1}{1 + n^{-\xi}\exp(\mu^\dagger + \kappa)}\binom{n}{2}\alpha_n \\
&= \frac{n^{2-\xi}}{2}\exp(\mu^\dagger)\frac{1}{1 + n^{-\xi}\exp(\mu^\dagger + \kappa)}\alpha_n + o(n^{2-\xi}).
\end{aligned}$$

Since the Z_{ij} are i.i.d. and since $\gamma^T Z_{ij}$ is uniformly bounded,

$$\alpha_n \xrightarrow{a.s.} \mathbb{E}[\exp(\gamma^T Z_{12})].$$

We have found an upper and a lower bound on $\sum_{i<j}\log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right)$. Multiplying both sides with $2n^{-2+\xi}$ and taking the limit $n \to \infty$, we see that the lower as well as the upper bound converge to $\exp(\mu^\dagger)\mathbb{E}[\exp(\gamma^T Z_{12})]$. But then, this already must be the limit for $2n^{-2+\xi}\sum_{i<j}\log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right)$:

$$2n^{-2+\xi}\sum_{i<j}\log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right) \xrightarrow{P} \exp(\mu^\dagger)\mathbb{E}[\exp(\gamma^T Z_{12})], \tag{3.7}$$

as $n \to \infty$. Putting equations (3.5), (3.6) and (3.7) together, we obtain that for any $(\mu^\dagger, \gamma) \in [-M, M] \times \Gamma$:

$$\begin{aligned}
2n^{-2+\xi}\mathcal{L}^\dagger(\mu^\dagger, \gamma) \xrightarrow{P} &- \mu^\dagger\exp(\mu_0^\dagger)\mathbb{E}[\exp(\gamma_0^T Z_{12})] \\
&- \exp(\mu_0^\dagger)\mathbb{E}[(\gamma^T Z_{12})\exp(\gamma_0^T Z_{12})] \\
&+ \exp(\mu^\dagger)\mathbb{E}[\exp(\gamma^T Z_{12})],
\end{aligned} \tag{3.8}$$

as $n \to \infty$. We define this limiting function as $M : \mathbb{R}^{p+1} \to \mathbb{R}$,

$$\begin{aligned}
M(\mu^\dagger, \gamma) \coloneqq &- \mu^\dagger\exp(\mu_0^\dagger)\mathbb{E}[\exp(\gamma_0^T Z_{12})] - \exp(\mu_0^\dagger)\mathbb{E}[\gamma^T Z_{12}\cdot\exp(\gamma_0^T Z_{12})] \\
&+ \exp(\mu^\dagger)\mathbb{E}[\exp(\gamma^T Z_{12})].
\end{aligned}$$

We want to employ Theorem 5.7 in van der Vaart (1998).

Theorem 3.5 (Theorem 5.7 in van der Vaart (1998))**.** *Let (Θ, d) be a metric space. Let M_n be random functions and let M be a fixed function of $\theta \in \Theta$, such that for every $\epsilon > 0$,*

1. $\sup_{\theta\in\Theta}|M_n(\theta) - M(\theta)| \xrightarrow{P} 0$
2. $\sup_{\theta: d(\theta,\theta_0)\geq\epsilon} M(\theta) > M(\theta_0)$.

Then any sequence of estimators $\hat{\theta}_n$ with $M_n(\hat{\theta}_n) \geq M_n(\theta_0) + o_p(1)$ converges in

probability to θ_0.

To apply Theorem 3.5, we must show that the convergence in (3.8) is uniform in probability. That is, we must show that

$$\sup_{\theta\in[-M,M]\times\Gamma} |2n^{-2+\xi}\mathcal{L}^\dagger(\theta) - M(\theta)| = o_P(1). \tag{3.9}$$

To shorten notation, introduce $M_n(\theta) \coloneqq 2n^{-2+\xi}\mathcal{L}^\dagger(\theta)$. Since we already have pointwise convergence in probability of M_n to M, it will suffice to show that for any $\epsilon > 0$,

$$\lim_{\delta\downarrow 0}\limsup_{n\to\infty} P\left(\sup_{\|\theta_1-\theta_2\|_2\leq\delta} |M_n(\theta_1) - M_n(\theta_2)| \geq \epsilon\right) = 0. \tag{3.10}$$

Property (3.9) then follows from the pointwise convergence, the continuity of M and the compactness of the parameter space $[-M, M]\times\Gamma$. To ease notation further, define for any $\delta \geq 0$,

$$\Delta_\delta^n \coloneqq \sup_{\|\theta_1-\theta_2\|_2\leq\delta} |M_n(\theta_1) - M_n(\theta_2)|.$$

Let $\epsilon, \eta > 0$. We have to show that there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} P(\Delta_\delta^n \geq \epsilon) \leq \eta. \tag{3.11}$$

Consider the following representation of $\mathcal{L}^\dagger(\theta)$:

$$\begin{aligned}
\mathcal{L}^\dagger(\theta) &= -d_+\mu^\dagger - \sum_{i<j}(\gamma^T Z_{ij})A_{ij} + \sum_{i<j}\log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right) \\
&= \sum_{i<j} -(\mu^\dagger + \gamma^T Z_{ij})A_{ij} + \log\left(1 + n^{-\xi}\exp(\mu^\dagger + \gamma^T Z_{ij})\right) \\
&= \sum_{i<j} \underbrace{-D_{ij}^T\theta A_{ij} + \log\left(1 + n^{-\xi}\exp(D_{ij}^T\theta)\right)}_{=:l_{ij}(\theta)}.
\end{aligned}$$

For any $\delta > 0$ and any θ_1, θ_2 with $\|\theta_1 - \theta_2\|_2 < \delta$ and any $i < j$, we obtain:

$$\begin{aligned}
\mathbb{E}|l_{ij}(\theta_1) - l_{ij}(\theta_2)| = \mathbb{E}\Big| &- D_{ij}^T(\theta_1 - \theta_2)A_{ij} + \log\left(1 + n^{-\xi}\exp(D_{ij}^T\theta_1)\right) \\
&- \log\left(1 + n^{-\xi}\exp(D_{ij}^T\theta_2)\right)\Big|.
\end{aligned}$$

Hence, by the Mean Value Theorem with α between $D_{ij}\theta_1$ and $D_{ij}\theta_2$:

$$\begin{aligned}
\mathbb{E}|l_{ij}(\theta_1) - l_{ij}(\theta_2)| &\leq \mathbb{E}\left[|D_{ij}^T(\theta_1 - \theta_2)| \cdot A_{ij}\right] + \frac{n^{-\xi}\exp(\alpha)}{1 + n^{-\xi}\exp(\alpha)}\mathbb{E}\left|D_{ij}^T(\theta_1 - \theta_2)\right| \\
&\leq C\|\theta_1 - \theta_2\|_2\mathbb{E}[p_{ij}] + Cn^{-\xi}\|\theta_1 - \theta_2\|_2 \\
&\leq C\|\theta_1 - \theta_2\|_2\left(\mathbb{E}\left[n^{-\xi}\frac{\exp(D_{ij}^T\theta_0)}{1 + n^{-\xi}\exp(D_{ij}^T\theta_0)}\right] + n^{-\xi}\right) \\
&\leq Cn^{-\xi}\|\theta_1 - \theta_2\|_2 \\
&\leq Cn^{-\xi}\delta,
\end{aligned}$$

where $C > 0$ denotes some generic constant that may change between displays. By the compactness of our parameter space and the resulting uniform boundedness of $|D_{ij}(\theta_1 - \theta_2)|$, we may in particular choose this C independent of n, i and j. Then, almost surely,

$$\mathbb{E}|\mathcal{L}^\dagger(\theta_1) - \mathcal{L}^\dagger(\theta_2)| \leq C\binom{n}{2}n^{-\xi}\delta$$

and thus, almost surely,

$$\mathbb{E}\Delta_\delta^n \leq Cn^{-2+\xi}n^{-\xi}\binom{n}{2}\delta \leq C\delta.$$

Thus, we can choose a $\delta > 0$ independent of n, such that $\mathbb{E}\Delta_\delta^n \leq \epsilon\eta$. But then an application of Markov's inequality yields for all n large enough

$$P(\Delta_\delta^n \geq \epsilon) \leq \eta.$$

It follows (3.11), which implies (3.10), which yields (3.9).

The second condition of Theorem 3.5 requires that the true parameter be a well-separated extrema of M. That is, we must show: For any fixed $\epsilon > 0$,

$$\sup_{\theta : d(\theta,\theta_0) \geq \epsilon} M(\theta) > M(\theta_0). \tag{3.12}$$

Consider the first partial derivatives of M:

$$\begin{aligned}
\partial_{\mu^\dagger}M(\mu^\dagger, \gamma) &= -\exp(\mu_0^\dagger)\mathbb{E}[\exp(\gamma_0^T Z_{12})] + \exp(\mu^\dagger)\mathbb{E}[\exp(\gamma^T Z_{12})], \\
\partial_{\gamma_k}M(\mu^\dagger, \gamma) &= -\exp(\mu_0^\dagger)\mathbb{E}[Z_{12,k}\exp(\gamma_0^T Z_{12})] + \exp(\mu^\dagger)\mathbb{E}[Z_{12,k}\exp(\gamma^T Z_{12})].
\end{aligned}$$

Clearly, by Assumption 3.1 the true parameter is a critical point of M, i.e. the first partial derivatives of M evaluated at $\theta_0 = (\mu_0^\dagger, \gamma_0^T)^T$ are zero:

$$\nabla M(\theta_0) = 0.$$

Consider the Hessian $HM(\mu^\dagger, \gamma)$ of M at the point $(\mu^\dagger, \gamma)$:

$$\begin{aligned}
\frac{\partial^2}{\partial(\mu^\dagger)^2} M(\mu^\dagger, \gamma) &= \exp(\mu^\dagger)\mathbb{E}[\exp(\gamma^T Z_{12})], \\
\frac{\partial^2}{\partial\mu^\dagger\gamma_k} M(\mu^\dagger, \gamma) &= \exp(\mu^\dagger)\mathbb{E}[Z_{12,k}\exp(\gamma^T Z_{12})], \\
\frac{\partial^2}{\partial\gamma_k^2} M(\mu^\dagger, \gamma) &= \exp(\mu^\dagger)\mathbb{E}[Z_{12,k}^2\exp(\gamma^T Z_{12})], \\
\frac{\partial^2}{\partial\gamma_k\gamma_l} M(\mu^\dagger, \gamma) &= \exp(\mu^\dagger)\mathbb{E}[Z_{12,k}Z_{12,l}\exp(\gamma^T Z_{12})].
\end{aligned}$$

We thus see that $HM(\mu^\dagger, \gamma)$ allows a matrix representation as

$$HM(\mu^\dagger, \gamma) = \exp(\mu^\dagger)\mathbb{E}\left[\exp(\gamma^T Z_{12})\begin{bmatrix} 1 & Z_{12}^T \\ Z_{12} & Z_{12}Z_{12}^T \end{bmatrix}\right] \in \mathbb{R}^{(p+1)\times(p+1)}.$$

By the compactness of our parameter space and the boundedness of Z_{12}, we now obtain for any $v \in \mathbb{R}^{p+1}$:

$$\begin{aligned}
v^T HM(\mu^\dagger, \gamma)v &= \exp(\mu^\dagger)\mathbb{E}\left[\exp(\gamma^T Z_{12})v^T D_{12}D_{12}^T v\right] \\
&\geq C\mathbb{E}\left[v^T D_{12}D_{12}^T v\right] \\
&= Cv^T\mathbb{E}\left[\begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & Z_{12}Z_{12}^T \end{bmatrix}\right] v \\
&\geq C\|v\|_2^2,
\end{aligned}$$

where for the last inequality we have used that the matrix is strictly positive definite by Assumption 3.2. That means, $HM(\mu^\dagger, \gamma)$ is strictly positive definite on the entire parameter space $[-M, M] \times \Gamma$. Hence, M is strictly convex and its minimum θ_0 already must be a global minimum. Now, since our parameter space is compact, M is continuous and θ_0 is a global maximum, it is easy to see that (3.12) must hold.

Finally, since (3.9) and (3.12) hold, we have consistency by Theorem 3.5:

$$\hat{\theta} \xrightarrow{P} \theta_0.$$

3.4.2 Asymptotic normality

The proof of asymptotic normality in spirit follows to some extent the proof of Theorem 2.7. By Assumption 3.1, the MLE $\hat{\theta}$ fulfils the first-order estimating equations:

$$0 = \nabla\mathcal{L}^\dagger(\hat{\theta}),$$

which, when looking at the individual components, means that for $k = 1, \dots, p$

$$0 = \partial_{\mu^\dagger} \mathcal{L}^\dagger(\hat{\theta}) = -d_+ + n^{-\xi} \exp(\hat{\mu}^\dagger) \sum_{i<j} \frac{\exp(\hat{\gamma}^T Z_{ij})}{1 + n^{-\xi} \exp(\hat{\mu}^\dagger + \hat{\gamma}^T Z_{ij})},$$
$$0 = \partial_{\gamma_k} \mathcal{L}^\dagger(\hat{\theta}) = \sum_{i<j} Z_{ij,k} A_{ij} + n^{-\xi} \exp(\hat{\mu}^\dagger) \sum_{i<j} \frac{Z_{ij,k} \exp(\hat{\gamma}^T Z_{ij})}{1 + n^{-\xi} \exp(\hat{\mu}^\dagger + \hat{\gamma}^T Z_{ij})}.$$

We want to make use of a Taylor expansion. Define the functions $l_n(y, a) : \{0, 1\} \times \mathbb{R} \to \mathbb{R}$,

$$l_n(y, a) = -ya + \log(1 + n^{-\xi} \exp(a)).$$

In particular,

$$\mathcal{L}^\dagger(\mu^\dagger, \gamma) = \sum_{i<j} l_n(A_{ij}, (\mu^\dagger, \gamma^T)^T D_{ij}).$$

The l_n have the following partial derivatives with respect to a:

$$\dot{l}_n(y, a) \coloneqq \partial_a l_n(y, a) = -y + n^{-\xi} \frac{\exp(a)}{1 + n^{-\xi} \exp(a)},$$
$$\ddot{l}_n(y, a) \coloneqq \partial_{a^2}^2 l_n(y, a) = n^{-\xi} \frac{\exp(a)}{(1 + n^{-\xi} \exp(a))^2},$$
$$\partial_{a^3}^3 l_n(y, a) = n^{-\xi} \frac{\exp(a)}{(1 + n^{-\xi} \exp(a))^2} \cdot \frac{1 - n^{-\xi} \exp(a)}{1 + n^{-\xi} \exp(a)}.$$

Note that $|\partial_{a^3}^3 l_n(y, a)| \leq C n^{-\xi}$ and hence $\ddot{l}_n(y, a)$ is Lipschitz continuous in a with constant $C n^{-\xi}$ by the Mean Value Theorem. Doing a first-order Taylor expansion in a of $\dot{l}_n(y, a) = \partial_a l_n(y, a)$ in the point $a_0 = (A_{ij}, D_{ij}^T \theta_0)$ evaluated at $a = (A_{ij}, D_{ij}^T \hat{\theta})$, we get

$$\partial_a l_n(A_{ij}, D_{ij} \hat{\theta}) = \partial_a l_n(A_{ij}, D_{ij}^T \theta_0) + \partial_{a^2} l_n(A_{ij}, \alpha) D_{ij}^T (\hat{\theta} - \theta_0), \tag{3.13}$$

for an α between $D_{ij}^T \hat{\theta}$ and $D_{ij}^T \theta_0$.

Consider the vector $1/\binom{n}{2} \nabla \mathcal{L}^\dagger(\hat{\theta})$: By equation (3.13), with α_{ij} between $D_{ij}^T \hat{\theta}$ and $D_{ij}^T \theta_0$,

$$\begin{aligned}
0 = \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\hat{\theta}) &= \frac{1}{\binom{n}{2}} \sum_{i<j} \left(\partial_{\theta_k} l_n(A_{ij}, D_{ij}^T \hat{\theta}) \right)_{k=1,\dots,p+1}, \quad \text{as a } (p+1) \times 1\text{-vector} \\
&= \frac{1}{\binom{n}{2}} \sum_{i<j} \dot{l}_n(A_{ij}, D_{ij}^T \hat{\theta}) D_{ij}, \quad \text{by the Chain Rule} \\
&= \frac{1}{\binom{n}{2}} \sum_{i<j} (\dot{l}_n(A_{ij}, D_{ij}^T \theta_0) + \ddot{l}_n(A_{ij}, \alpha_{ij}) D_{ij}^T (\hat{\theta} - \theta_0)) D_{ij}, \quad \text{by (3.13)} \\
&= \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\theta_0) + \frac{1}{\binom{n}{2}} \sum_{i<j} \ddot{l}_n(A_{ij}, \alpha_{ij}) D_{ij}^T (\hat{\theta} - \theta_0) \cdot D_{ij}.
\end{aligned}$$

Proving Theorem 3.1 now breaks down into three problems.

3.4.2.1 Problem 1

We first show that under appropriate scaling $\frac{1}{\binom{n}{2}}\nabla\mathcal{L}^\dagger(\theta_0)$ is asymptotically normal. We may write the kth component of $\nabla\mathcal{L}^\dagger(\theta_0)$ more compactly as

$$\nabla\mathcal{L}^\dagger(\theta_0)_k = \sum_{i<j} D_{ij,k}(p_{ij} - A_{ij}),$$

where $D_{ij,k}$ is the kth component of the (i,j)-th row of D, i.e. $D_{ij,k} = 1$, if $k = 1$ and $D_{ij,k} = Z_{ij,k-1}$, if $k = 2, \dots, p+1$ and

$$p_{ij} = \mathbb{E}[A_{ij}|Z_{ij}] = n^{-\xi} \cdot \frac{\exp(\mu_0^\dagger + \gamma_0^T Z_{ij})}{1 + n^{-\xi}\exp(\mu_0^\dagger + \gamma_0^T Z_{ij})}.$$

Notice that all components of $\nabla\mathcal{L}^\dagger(\theta_0)$ are centred. Indeed,

$$\mathbb{E}[\nabla\mathcal{L}^\dagger(\theta_0)_k] = \sum_{i<j}\mathbb{E}[D_{ij,k}(p_{ij} - A_{ij})] = \sum_{i<j}\mathbb{E}[D_{ij,k}\mathbb{E}[(p_{ij} - A_{ij})|Z_{ij}]] = 0.$$

We want to apply the Lindeberg-Feller Central Limit Theorem to the term

$$\sqrt{\binom{n}{2}}n^{\xi/2} \cdot \frac{1}{\binom{n}{2}}\nabla\mathcal{L}^\dagger(\theta_0) = \sum_{i<j} D_{ij}(p_{ij} - A_{ij}) \cdot \sqrt{\frac{n^\xi}{\binom{n}{2}}}.$$

To that end, define the triangular array $Y_{n,ij} = D_{ij}(p_{ij} - A_{ij}) \cdot \sqrt{\frac{n^\xi}{\binom{n}{2}}}, 1 \le i < j \le n, n \in \mathbb{N}$. Since the $Y_{n,ij}$ are centred, their covariance matrix is given by

$$\begin{aligned}\mathrm{Cov}(Y_{n,ij}) = \mathbb{E}[Y_{n,ij}Y_{n,ij}^T] &= \mathbb{E}\left[D_{ij}D_{ij}^T(p_{ij} - A_{ij})^2 \cdot \frac{n^\xi}{\binom{n}{2}}\right] \\ &= \mathbb{E}\left[D_{ij}D_{ij}^T p_{ij}(1-p_{ij}) \cdot \frac{n^\xi}{\binom{n}{2}}\right],\end{aligned}$$

where for the last equality we have used that $\mathbb{E}[(p_{ij} - A_{ij})^2|Z_{ij}] = p_{ij}(1-p_{ij})$. In analogy to the case with non-zero β, we write $W_0^2 = \mathrm{diag}(p_{ij}(1-p_{ij}), i<j) \in \mathbb{R}^{\binom{n}{2}\times\binom{n}{2}}$. Then, we get for the sum of covariance matrices

$$\sum_{i<j}\mathrm{Cov}(Y_{n,ij}) = \sum_{i<j}\mathbb{E}\left[D_{ij}D_{ij}^T p_{ij}(1-p_{ij}) \cdot \frac{n^\xi}{\binom{n}{2}}\right] = \frac{n^\xi}{\binom{n}{2}}\mathbb{E}[D^T W_0^2 D] =: \Sigma^{(n)}.$$

For any pair $i < j$, we have $p_{ij}(1-p_{ij}) = n^{-\xi}\exp(\mu_0^\dagger)\frac{\exp(\gamma_0^T Z_{ij})}{(1+n^{-\xi}\exp(\mu_0^\dagger+\gamma_0^T Z_{ij}))^2}$. Hence, $n^\xi p_{ij}(1-p_{ij}) \to \exp(\mu_0^\dagger + \gamma_0^T Z_{ij})$, as $n \to \infty$. Consider the (k,l)-th entry of $\Sigma^{(n)}$:

$$\begin{aligned}
\Sigma_{k,l}^{(n)} &= \frac{1}{\binom{n}{2}} \sum_{i<j} \mathbb{E}\left[(D_{ij}D_{ij}^T)_{k,l} \exp(\mu_0^\dagger) \frac{\exp(\gamma_0^T Z_{ij})}{(1+n^{-\xi}\exp(\mu_0^\dagger + \gamma_0^T Z_{ij}))^2} \right] \\
&= \mathbb{E}\left[(D_{12}D_{12}^T)_{k,l} \exp(\mu_0^\dagger) \frac{\exp(\gamma_0^T Z_{12})}{(1+n^{-\xi}\exp(\mu_0^\dagger + \gamma_0^T Z_{12}))^2} \right], \quad Z_{ij} \text{ i.i.d.} \\
&\overset{n\to\infty}{\longrightarrow} \mathbb{E}\left[(D_{12}D_{12}^T)_{k,l} \exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12}) \right] =: \Sigma_{kl},
\end{aligned}$$

by dominated convergence. Hence, with $\Sigma = (\Sigma_{kl})_{k,l} \in \mathbb{R}^{(p+1)\times p+1}$, as $n \to \infty$,

$$\sum_{i<j} \mathrm{Cov}(Y_{n,ij}) \to \Sigma,$$

where convergence is to be understood componentwise. We claim that Σ is strictly positive definite. Indeed, since $\mu_0^\dagger + \gamma_0^T Z_{12}$ lies in some compact set, there is a constant $C > 0$ such that $\exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12}) > C > 0$ almost surely. Then, for any vector $v = (v_1, v_R^T)^T \in \mathbb{R}^{p+1}, v_1 \in \mathbb{R}, v_R \in \mathbb{R}^p$,

$$v^T \Sigma v = \mathbb{E}[(D_{12}^T v)^2 \exp(\mu_0^\dagger)\exp(\gamma_0^T Z_{12})] > C v^T \mathbb{E}[D_{12}D_{12}^T] v.$$

Yet, by Assumption 3.2,

$$\begin{aligned}
v^T \mathbb{E}[D_{12}D_{12}^T] v &= v^T \mathbb{E}\begin{bmatrix} 1 & Z_{12}^T \\ Z_{12} & Z_{12}Z_{12}^T \end{bmatrix} v = v^T \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \mathbb{E}[Z_{12}Z_{12}^T] \end{bmatrix} v \\
&= v_1^2 + v_R^T \mathbb{E}[Z_{12}Z_{12}^T] v_R \geq (1 \wedge \lambda_{\min}) \|v\|_2^2.
\end{aligned}$$

Thus, for any $v \neq 0$,

$$v^T \Sigma v \geq C\|v\|_2^2 > 0$$

and therefore Σ is positive definite.

Furthermore, we clearly have $\mathbb{E}[\|Y_{n,ij}\|_2^2] < C < \infty$ for any i, j, n. Finally, let $\epsilon > 0$. Since $\|D_{ij}(p_{ij} - A_{ij})\|_2$ is uniformly bounded for all $i < j$, we we may find an $n_0 \in \mathbb{N}$ such that for all $n > n_0$ we have $\|Y_{n,ij}\|_2 < \epsilon$ for all $i < j$. This gives us that, as $n \to \infty$,

$$\sum_{i<j} \mathbb{E}[\|Y_{n,ij}\|_2^2 \mathbb{1}(\|Y_{n,ij}\|_2 > \epsilon)] \to 0.$$

Then, by the vector-valued Lindeberg-Feller Central Limit Theorem, we obtain

$$\sqrt{\binom{n}{2}} n^{\xi/2} \cdot \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\theta_0) = \sum_{i<j} Y_{n,ij} \xrightarrow{d} \mathcal{N}(0, \Sigma). \tag{3.14}$$

3.4.2.2 Problem 2

Next, we must find a bound on the speed of convergence of $\hat{\theta} - \theta_0$. Recall that we obtained the equality

$$0 = \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^{\dagger}(\theta_0) + \frac{1}{\binom{n}{2}} \sum_{i<j} \ddot{l}_n(A_{ij}, \alpha_{ij}) D_{ij} D_{ij}^T (\hat{\theta} - \theta_0). \tag{3.15}$$

Consider the matrix

$$\Sigma_\alpha \coloneqq \frac{1}{\binom{n}{2}} \sum_{i<j} \ddot{l}_n(A_{ij}, \alpha_{ij}) D_{ij} D_{ij}^T = \frac{1}{\binom{n}{2}} D^T \cdot \operatorname{diag}(\ddot{l}_n(A_{ij}, \alpha_{ij}), i < j) \cdot D.$$

Since α_{ij} lies between $D_{ij}^T \hat{\theta}$ and $D_{ij}^T \theta_0$ and both of these points lie in some compact set, we have for some universal constant $C > 0$, independent of i, j,

$$\ddot{l}_n(A_{ij}, \alpha_{ij}) \geq C n^{-\xi}.$$

Thus, for any $v \in \mathbb{R}^{p+1}$,

$$v^T \Sigma_\alpha v \geq C n^{-\xi} v^T \left(\frac{1}{\binom{n}{2}} D^T D \right) v.$$

Completely analogously to the case with non-zero β, we can show that $\frac{1}{\binom{n}{2}} D^T D$ is positive definite with high probability by using Lemma 6 in Kock & Tang (2019) (cf. Section 2.7.2.1). Therefore, with high probability, $\operatorname{mineval}(\Sigma_\alpha) \geq C n^{-\xi} > 0$. Thus,

$$\operatorname{maxeval}(\Sigma_\alpha^{-1}) = \frac{1}{\operatorname{mineval}(\Sigma_\alpha)} \leq C n^{\xi}.$$

From (3.15) we now obtain

$$\Sigma_\alpha (\hat{\theta} - \theta_0) = -\frac{1}{\binom{n}{2}} \nabla \mathcal{L}^{\dagger}(\theta_0)$$

which is equivalent to

$$\hat{\theta} - \theta_0 = -\Sigma_\alpha^{-1} \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^{\dagger}(\theta_0),$$

which after rescaling gives

$$\begin{aligned} \sqrt{\frac{\binom{n}{2}}{n^\xi}} (\hat{\theta} - \theta_0) &= -\sqrt{\frac{\binom{n}{2}}{n^\xi}} \Sigma_\alpha^{-1} \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^{\dagger}(\theta_0) \\ &= -n^{-\xi} \Sigma_\alpha^{-1} \cdot \sqrt{\binom{n}{2}} n^{\xi/2} \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^{\dagger}(\theta_0). \end{aligned}$$

From the previous section we know $\sqrt{\binom{n}{2}} n^{\xi/2} \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^{\dagger}(\theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. Also, the maximum eigenvalue of $n^{-\xi} \Sigma_\alpha^{-1}$ is uniformly bounded by some universal constant

$C < \infty$, making the right-hand side above $O_P(1)$. This means

$$\hat{\theta} - \theta_0 = O_P\left(\sqrt{\frac{n^\xi}{\binom{n}{2}}}\right).$$

3.4.2.3 Problem 3

Finally, we derive the desired central limit theorem for our estimator. We claim that $n^\xi \Sigma_\alpha = \Sigma + o_P(1)$. To prove this, first consider the functions

$$f_n(x) = \frac{\exp(x)}{(1 + n^{-\xi}\exp(x))^2}.$$

For every x, we have pointwise convergence $f_n(x) \to f(x) \coloneqq \exp(x)$ as $n \to \infty$. Since $\hat{\theta}$ and θ_0 lie in some compact set and since Z_{ij} is uniformly bounded, the values α_{ij} in (3.15) and $\mu_0^\dagger + \gamma_0^T Z_{ij}, i < j$ all lie in some compact interval $I \subset \mathbb{R}$ independent of i, j and n. Also notice that $f_n(x) \leq f_{n+1}(x)$ for all $n \in \mathbb{N}$ and $x \in I$. Recall that by Dini's Theorem a sequence of monotonically increasing, continuous, real-valued functions that converges pointwise to some continuous limit function on a compact topological space, must already converge uniformly. Hence, f_n converges uniformly to f on I:

$$\lim_{n\to\infty} \sup_{x\in I} |f_n(x) - f(x)| = 0.$$

Furthermore, since I is compact and hence bounded, f has bounded derivative on I and thus is Lipschitz continuous on I with some finite constant C by the Mean Value Theorem:

$$|f(x) - f(y)| \leq C|x - y|, \quad \text{for all } x, y \in I.$$

Consider the (k,l)-th entry of $n^\xi \Sigma_\alpha - \Sigma$:

$$\begin{aligned}
&|(n^\xi \Sigma_\alpha - \Sigma)_{kl}| \\
&\quad = \left| \frac{1}{\binom{n}{2}} \sum_{i<j} D_{ij,k} D_{ij,l} \frac{\exp(\alpha_{ij})}{(1 + n^{-\xi}\exp(\alpha_{ij}))^2} - \mathbb{E}[D_{12,k} D_{12,l} \exp(\mu_0^\dagger + \gamma_0^T Z_{12})] \right| \\
&\quad \leq \underbrace{\left| \frac{1}{\binom{n}{2}} \sum_{i<j} D_{ij,k} D_{ij,l} \left\{ \frac{\exp(\alpha_{ij})}{(1 + n^{-\xi}\exp(\alpha_{ij}))^2} - \exp(\mu_0^\dagger + \gamma_0^T Z_{ij}) \right\} \right|}_{(I)} \\
&\qquad + \underbrace{\left| \frac{1}{\binom{n}{2}} \sum_{i<j} D_{ij,k} D_{ij,l} \exp(\mu_0^\dagger + \gamma_0^T Z_{ij}) - \mathbb{E}[D_{12,k} D_{12,l} \exp(\mu_0^\dagger + \gamma_0^T Z_{12})] \right|}_{(II)}
\end{aligned}$$

By the strong law of large numbers, (II) goes to zero almost surely. Let us consider (I).

$$\begin{aligned}
(I) &\leq \frac{1}{\binom{n}{2}} \sum_{i<j} |D_{ij,k} D_{ij,l}| \left| \frac{\exp(\alpha_{ij})}{(1+n^{-\xi}\exp(\alpha_{ij}))^2} - \exp(\mu_0^\dagger + \gamma_0^T Z_{ij}) \right| \\
&\leq C \cdot \max_{i<j} \left| \frac{\exp(\alpha_{ij})}{(1+n^{-\xi}\exp(\alpha_{ij}))^2} - \exp(\mu_0^\dagger + \gamma_0^T Z_{ij}) \right| \\
&= C \cdot \max_{i<j} |f_n(\alpha_{ij}) - f(\mu_0^\dagger + \gamma_0^T Z_{ij})| \\
&\leq C \cdot \left\{ \max_{i<j} |f_n(\alpha_{ij}) - f(\alpha_{ij})| + \max_{i<j} |f(\alpha_{ij}) - f(\mu_0^\dagger + \gamma_0^T Z_{ij})| \right\} \\
&\leq C \cdot \left\{ \sup_{x \in I} |f_n(x) - f(x)| + \max_{i<j} |\alpha_{ij} - \mu_0^\dagger + \gamma_0^T Z_{ij}| \right\},
\end{aligned}$$

where we have used the Lipschitz continuity of f on I for the last inequality. By the uniform convergence of f_n to f on I, we know that the first term in the last line goes to zero. For the second term, recall that α_{ij} is a point between $D_{ij}^T\hat{\theta}$ and $D_{ij}^T\theta_0 = \mu_0^\dagger + \gamma_0^T Z_{ij}$. Hence,

$$\max_{i<j} |\alpha_{ij} - \mu_0^\dagger + \gamma_0^T Z_{ij}| \leq \max_{i<j} |(\hat{\mu}^\dagger - \mu_0^\dagger) + (\hat{\gamma} - \gamma_0)^T Z_{ij}| \leq C\|\hat{\theta} - \theta_0\|_1 \xrightarrow{P} 0,$$

by the consistency of $\hat{\theta}$. Thus, $(I) \xrightarrow{P} 0$ as $n \to \infty$.

In conclusion, $|(n^\xi \Sigma_\alpha - \Sigma)_{kl}| \xrightarrow{P} 0$ and therefore,

$$n^\xi \Sigma_\alpha = \Sigma + o_P(1),$$

where $o_P(1)$ is to be understood as a matrix in which each component is $o_P(1)$. Now, we get from (3.15),

$$0 = \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\theta_0) + \Sigma_\alpha(\hat{\theta} - \theta_0)$$

which after multiplying with n^ξ is equivalent to

$$0 = n^\xi \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\theta_0) + (\Sigma + o_P(1))(\hat{\theta} - \theta_0).$$

Rearranging gives

$$\Sigma(\hat{\theta} - \theta_0) = -n^\xi \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\theta_0) + o_P(1)(\hat{\theta} - \theta_0).$$

Now, remember that Σ is positive definite and thus invertible. Hence,

$$(\hat{\theta} - \theta_0) = -\Sigma^{-1} n^\xi \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\theta_0) + \Sigma^{-1} o_P(1)(\hat{\theta} - \theta_0).$$

Observe that Σ^{-1} has bounded maximum eigenvalue due to Assumption 3.2 and thus $\Sigma^{-1} o_P(1) = o_P(1)$:

$$(\hat{\theta} - \theta_0) = -\Sigma^{-1} n^\xi \frac{1}{\binom{n}{2}} \nabla \mathcal{L}^\dagger(\theta_0) + o_P(1)(\hat{\theta} - \theta_0).$$

Finally, multiply by $\sqrt{\frac{\binom{n}{2}}{n^\xi}}$ and remember that $\hat{\theta} - \theta_0 = O_P\left(\sqrt{\frac{n^\xi}{\binom{n}{2}}}\right)$

$$\sqrt{\frac{\binom{n}{2}}{n^\xi}}(\hat{\theta} - \theta_0) = -\Sigma^{-1}\sqrt{\binom{n}{2}}n^{\xi/2}\frac{1}{\binom{n}{2}}\nabla\mathcal{L}^\dagger(\theta_0) + o_P(1).$$

With this, due to (3.14), we have proven

$$\sqrt{\frac{\binom{n}{2}}{n^\xi}}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}). \tag{3.16}$$

Proof of Theorem 3.1. By the solved problems 1 - 3 above. □

It remains to prove Corollary 3.2.

Proof of Corollary 3.2. Notice that from (3.16) we get: For any $k = 1, \ldots, (p+1)$,

$$\sqrt{\frac{\binom{n}{2}}{n^\xi}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{\Sigma^{-1}_{k,k}}} \xrightarrow{d} \mathcal{N}(0, 1). \tag{3.17}$$

By the exact same arguments that we have used to show that $n^\xi\Sigma_\alpha = \Sigma + o_P(1)$, we can also show that

$$n^\xi\hat{\Sigma} = \Sigma + o_P(1),$$

where $\hat{\Sigma}$ is the same matrix as Σ_α with α_{ij} replaced by $\hat{\mu}_0^\dagger + \hat{\gamma}^T Z_{ij}$:

$$\hat{\Sigma} = \frac{1}{\binom{n}{2}}D^T\operatorname{diag}\left(\frac{n^{-\xi}\exp(\hat{\mu}^\dagger + \hat{\gamma}^T Z_{ij})}{(1 + n^{-\xi}\exp(\hat{\mu}^\dagger + \hat{\gamma}^T Z_{ij}))^2}, i < j\right)D.$$

By the same arguments as before, we can show that the minimum eigenvalue of $n^\xi\hat{\Sigma}$ is bounded away from zero, uniformly in n. This implies that the maximum eigenvalue of $(n^\xi\hat{\Sigma})^{-1}$ is bounded by some finite constant C. We already know that the same property holds for Σ and Σ^{-1}. Therefore, we have for the matrix ∞-norm (recall (2.30) and Lemma 2.23):

$$\|(n^\xi\hat{\Sigma})^{-1} - \Sigma^{-1}\|_\infty \leq \|(n^\xi\hat{\Sigma})^{-1}\|_\infty\|\Sigma^{-1}\|_\infty\|n^\xi\hat{\Sigma} - \Sigma\|_\infty \leq C\|n^\xi\hat{\Sigma} - \Sigma\|_\infty = o_P(1).$$

Thus, in particular, for the diagonal elements:

$$(n^\xi\hat{\Sigma})^{-1}_{k,k} = n^{-\xi}\hat{\Sigma}^{-1}_{k,k} = \Sigma^{-1}_{k,k} + o_P(1).$$

But then, from (3.17) and by Slutzky's Theorem,

$$\sqrt{\binom{n}{2}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{\hat{\Sigma}^{-1}_{k,k}}} = \sqrt{\frac{\binom{n}{2}}{n^\xi}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{n^{-\xi}\hat{\Sigma}^{-1}_{k,k}}} = \sqrt{\frac{\binom{n}{2}}{n^\xi}} \cdot \frac{\hat{\theta}_k - \theta_{0,k}}{\sqrt{\Sigma^{-1}_{k,k} + o_P(1)}} \xrightarrow{d} \mathcal{N}(0, 1).$$

□

3.4.3 Proof of Theorem 3.3

Proof of Theorem 3.3. Denote the law of ER-C with parameters μ, γ on n nodes by ER-C(n), where we suppress the dependence on μ, γ in our notation. Denote the law of the ER-C, *given the realizations* of $Z = (Z_{ij})_{i<j}$ as ER-C(n, Z).

Let $\xi \in [0, 1)$. Since $\mu^\dagger + Z_{ij}^T\gamma$ is bounded almost surely, we can find a finite constant $c > 0$, such that almost surely, for all i, j,

$$\frac{\exp(\mu^\dagger + Z_{ij}^T\gamma)}{1 + n^{-\xi}\exp(\mu^\dagger + Z_{ij}^T\gamma)} \geq c.$$

In particular, since the Z_{ij} are uniformly bounded, we may choose a universal c independent of Z. Define $p_{\min} \coloneqq n^{-\xi}c$. Then, almost surely, for all i, j,

$$p_{ij} \geq p_{\min}.$$

We now construct a coupling between ER-C (n, Z) and ER$(n, p_{\min})$ as follows: Start out with n nodes, numbered $1, \ldots, n$, without any connections between them.

1. For each pair of nodes $i < j$ draw independent, uniform random variables $U_{ij} \sim \mathcal{U}([0, 1])$.
2. Place a link between i, j if and only if $U_{ij} \leq p_{\min}$. Since $P(U_{ij} \leq p_{\min}) = p_{\min}$ the resulting graph has distribution ER$(n, p_{\min})$.
3. On another copy of the set $\{1, \ldots, n\}$, place a link between i and j if and only if $U_{ij} \leq p_{ij}$, with p_{ij} from (3.1), using the same realizations of the U_{ij}. The resulting graph has distribution ER-C(n, Z).

By construction, the realization of ER-C (n, Z) will contain at least the same edges as the realization of ER$(n, p_{\min})$, possibly more.

Define $\lambda_n = np_{\min} = n^{1-\xi}c$. Since $\xi \in [0, 1)$, $\lambda_n - \log(n) \to \infty$ as $n \to \infty$. Thus, by Theorem 3.4, a realization of ER$(n, p_{\min})$ will be connected with high probability. But due to the coupling, the realization of ER-C (n, Z) contains at least the same edges as that realization of ER$(n, p_{\min})$, plus potentially some other edges. Thus, it must be connected with high probability, too. Since $p_{\min}$ is independent of the realization of Z, this must hold for any realization of ER-C (n).

To prove the converse, let $\xi \geq 1$. By the boundedness of $\mu^\dagger + Z_{ij}^T\gamma$, we can find a finite constant $C > 0$, such that almost surely, uniformly in i, j,

$$\frac{\exp(\mu^\dagger + Z_{ij}^T\gamma)}{1 + n^{-\xi}\exp(\mu^\dagger + Z_{ij}^T\gamma)} \leq C.$$

Again, since the Z_{ij} are uniformly bounded, we may choose a universal C independent of the realization of Z. Define $p_{\max} \coloneqq n^{-\xi}C$. Then, almost surely, for all i, j,

$$p_{ij} \leq p_{\max}.$$

Construct a coupling between $\mathrm{ER}(n, p_{\max})$ and ER-C (n, Z) using the same procedure as above, replacing $p_{\min}$ with $p_{\max}$. Define $\lambda_n = np_{\max} = n^{1-\xi}C$. Since $\xi \geq 1$, $\lambda_n - \log(n) \to -\infty$ and thus, by Theorem 3.4, a realization of $\mathrm{ER}(n, p_{\max})$ will be disconnected with high probability. Due to the coupling, a realization of ER-C (n, Z) will contain at most the same edges as that realization of $\mathrm{ER}(n, p_{\max})$, possibly fewer. Thus, it must be disconnected with high probability, too. Since $p_{\max}$ was chosen independent of the realization of Z, this must hold for any realization of ER-C(n). □

Chapter 4

A sparse random graph model for sparse directed networks

Organization of this chapter

We extend the SβM-C to directed networks, introducing what we call the parameter-Sparse Random Graph Model (SRGM) in Section 4.1. We focus particularly on the interplay of the rates of convergence for the network sparsity, ρ_n, the parameter sparsity, s_0 and the penalty parameter, λ. We define an ℓ_1-penalized estimator and emphasize how different regimes for s_0, ρ_n and λ allow us to prove different properties of our estimator.

Our main results are presented in Section 4.2. Model selection consistency is the most refined of our results and is presented in Section 4.2.1. We state our parameter estimation consistency result in Section 4.2.2 and the central limit theorem for the covariate parameter in Section 4.2.3. This is followed by a set of simulation studies in Section 4.3. Large parts of the proofs of consistency and asymptotic normality are similar to those for SβM-C, which is why those proofs have been relegated to Appendix A. The proofs relating model selection consistency are presented in Section 4.4. The content of this chapter is from Stein & Leng (2021).

4.1 Estimation

We saw the flexibility of the parameter-Sparse Random Graph Model (SRGM) and its ability to provide reliable inference when we fitted it to Lazega's lawyer data in Section 1.2.2. It is now time to put this model on a solid theoretical foundation. Recall the definition of SRGM in equation (1.2): We study a directed network model in which for each ordered pair of nodes (i, j) we observe a covariate vector $Z_{ij} \in \mathbb{R}^p$. In particular, we may have $Z_{ij} \neq Z_{ji}$. The probability of observing a directed edge *from* node i *to* node j, given the covariate vector Z_{ij}, is given by

$$P(A_{ij} = 1|Z_{ij}) = p_{ij} = \frac{\exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}{1 + \exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}, \tag{4.1}$$

where $\gamma \in \mathbb{R}^p$ are covariate weights and μ is a global sparsity parameter, for which we allow $\mu \to -\infty$, as $n \to \infty$. The degree heterogeneity of the network is characterized by two parameters $\alpha, \beta \in \mathbb{R}^n$: Each node i has an *outgoingness parameter* α_i that determines how likely a node is to send out directed links to other nodes and an *incomingness parameter* β_i that determines how likely the node is to receive directed links from other nodes. For identifiability we impose $\min\{\alpha_i : i = 1, \dots, n\} = \min\{\beta_j : j = 1, \dots n\} = 0$. Notice the subtle change in our identifiability condition: Were we to translate the identifiability condition used in SβM-C directly, one might assume that we should impose $\min\{\alpha_i, \beta_j : i, j = 1, \dots, n\} = 0$. Our condition is the slightest bit stricter. This stricter assumption is necessary for showing model selection consistency and really does not change the flavour of the model.

For brevity, let $\vartheta = (\alpha^T, \beta^T)^T$ denote the degree heterogeneity parameters and $\xi = (\mu, \gamma^T)^T \in \mathbb{R}^{p+1}$ the global parameters. Write $\theta = (\vartheta^T, \xi^T)^T$ with its true value denoted as $\theta_0 = (\vartheta_0^T, \xi_0^T)^T$. We let $\Theta = \mathbb{R}_+^n \times \mathbb{R}_+^n \times \mathbb{R} \times \Gamma$ denote the *global parameter space*. As in the case of SβM-C and as is commonly assumed in general for LASSO type problems (van de Geer & Bühlmann 2011, Chapter 6), we assume the parameters that will be left unpenalized to be active: $\mu_0 \neq 0, \gamma_{0,i} \neq 0, i = 1, \dots, p$.

We write $S_0 = S(\vartheta_0)$ and denote its cardinality as $s_0 = |S_0|$. We write $S_{0,+} \coloneqq S_0 \cup \{2n + 1, 2n + 2, \dots 2n + 1 + p\}$ with cardinality $s_{0,+} = |S_{0,+}| = s_0 + p + 1$ to refer to all active indices including those of μ and γ. Let $S_\alpha = \{i : \alpha_{0,i} > 0\}, S_\beta = \{j : \beta_{0,j} > 0\}$ and $s_\alpha = |S_\alpha|, s_\beta = |S_\beta|$. When we want to make the dependence of the link probabilities given Z_{ij} on different values of θ explicit, we write $p_{ij}(\theta) = \frac{\exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})}{1 + \exp(\beta_i + \beta_j + \mu + \gamma^T Z_{ij})}$.

Without the covariates and by absorbing μ into α_i and β_j as $\mu/2 + \alpha_i$ and $\mu/2 + \beta_j$, respectively, this model is the p_0-model introduced in Holland & Leinhardt (1981). It was also studied in Yan, Leng & Zhu (2016). When the covariates (but not the μ) are added, it becomes the model in Yan et al. (2019). The results and derivations therein only hold for dense networks, however. By adding the global sparsity parameter μ and imposing the identifiability assumption $\min_i\{\alpha_i\} = \min_j\{\beta_j\} = 0$, we are able to perform estimation and inference in sparse networks.

Given the adjacency matrix A and the covariates $\{Z_{ij}\}_{i \neq j}$, it is easily seen that the negative log-likelihood of the SRGM in (4.1) at $\theta = (\alpha^T, \beta^T, \mu, \gamma^T)^T$ is

$$\begin{aligned} \mathcal{L}(\theta) = &- \sum_{i=1}^n \alpha_i b_i - \sum_{i=1}^n \beta_i d_i - d_+ \mu - \sum_{\substack{i,j=1 \\ i \neq j}}^n (\gamma^T Z_{ij}) A_{ij} \\ &+ \sum_{\substack{i,j=1 \\ i \neq j}}^n \log(1 + \exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij})), \end{aligned} \tag{4.2}$$

where b_i is the out-degree of node i and d_i its in-degree (see Definition 1.5) and $d_+ = \sum_{i=1}^n d_i = \sum_{i=1}^n b_i$. It is easy to see by taking derivatives that

$$\theta_0 = \arg\min_{\theta \in \Theta} \mathbb{E}[\mathcal{L}(\theta)],$$

where the expectation is taken with respect to the A_{ij} and Z_{ij}. We define $N = n(n-1)$.

We make a lot of the same definitions as in SβM-C. For the sake of completeness, we do restate them in the language of model (4.1). To estimate θ_0 and identify the support of $\vartheta = (\alpha^T, \beta^T)^T$, a natural idea is to resort to the method of penalized likelihood by solving

$$\arg\min_{\theta \in \Theta} \frac{1}{N} \mathcal{L}(\alpha, \beta, \mu, \gamma) + \lambda(\|\alpha\|_1 + \|\beta\|_1), \tag{4.3}$$

where λ is a tuning parameter and we have used the same amount of penalty on α and β because $\sum_{i=1}^n d_i = \sum_{i=1}^n b_i$. The objective function in (4.3) is similar to the penalized logistic regression with an ℓ_1-penalty and thus can be easily solved similarly to the case of SβM-C, using for example the R package `glmnet` (Friedman et al. 2010). This makes our estimation approach extremely scalable.

Since our focus is on sparse networks, we need $p_{ij} \to 0$ as $n \to \infty$ at least for some i and j. It is natural to impose restrictions on how fast this decay can be. Therefore, we once more assume the existence of a non-random sequence $\rho_{n,0} \in (0, 1/2]$ with $\rho_{n,0} \to 0$ as $n \to \infty$, such that for all i, j, almost surely,

$$1 - \rho_{n,0} \geq p_{ij} \geq \rho_{n,0},$$

or equivalently,

$$|\alpha_{0,i} + \beta_{0,j} + \mu_0 + \gamma_0^T Z_{ij}| \leq -\text{logit}(\rho_{n,0}) =: r_{n,0} \geq 0, \quad \forall i, j,$$

where the positivity follows from $\rho_n \leq 1/2$. The previous inequality can also be expressed in terms of the design matrix D associated with the corresponding logistic regression problem – defined in (4.5) below – and is equivalent to $\|D\theta_0\|_\infty \leq r_{n,0}$. This motivates the following tweak to the estimation procedure in (4.3): Given a sufficiently large constant r_n we define the local parameter space

$$\Theta_{\text{loc}} = \Theta_{\text{loc}}(r_n) \coloneqq \{\theta \in \Theta : \|D\theta\|_\infty \leq r_n\}$$

and propose to perform estimation via

$$\hat{\theta} = (\hat{\alpha}^T, \hat{\beta}^T, \hat{\mu}, \hat{\gamma}^T)^T = \underset{\theta=(\alpha^T,\beta^T,\mu,\gamma^T)^T \in \Theta_{\text{loc}}}{\arg\min} \frac{1}{N}\mathcal{L}(\alpha, \beta, \mu, \gamma) + \lambda(\|\alpha\|_1 + \|\beta\|_1), \quad (4.4)$$

which is more amenable for theoretical analysis. Notice that Θ_{loc} is convex.

We now give an explicit form of the associated design matrix D. We have to consider the presence/ absence of $N = n(n-1)$ directed edges and our model has $2n + 1 + p$ parameters. Thus, D has dimension $N \times (2n + 1 + p)$. Define the *out-matrix* $X^{\text{out}} \in \mathbb{R}^{N\times n}$ with rows $X^{\text{out}}_{ij} \in \mathbb{R}^{1\times n}, i \neq j$, such that for each component $k = 1, \ldots, n$, $X^{\text{out}}_{ij,k} = 1$ if $k = i$ and zero otherwise. Likewise, define the *in-matrix* $X^{\text{in}} \in \mathbb{R}^{N\times n}$ with rows $X^{\text{in}}_{ij} \in \mathbb{R}^{1\times n}, i \neq j$, such that for each component $k = 1, \ldots, n$, $X^{\text{in}}_{ij,k} = 1$ if $k = j$ and zero otherwise. Let $Z = (Z^T_{ij})_{i\neq j} \in \mathbb{R}^{N\times p}$ be the matrix of the covariate vectors written below each other. Then, D consists of four blocks, written next to each other:

$$D = \left[\begin{array}{c|c|c|c} X^{\text{out}} & X^{\text{in}} & \mathbf{1} & Z \end{array}\right] \in \mathbb{R}^{N\times(2n+p+1)}, \quad (4.5)$$

where $\mathbf{1} \in \mathbb{R}^N$ is a vector of all ones. We use the shorthand $X = [X^{\text{out}} \mid X^{\text{in}}] \in \mathbb{R}^{N\times 2n}$.

As in the case of SβM-C, the design matrix D reveals an important property of our model (4.1). While the columns of the the global parameters μ and γ have effective sample size of order $N \sim n^2$, the local parameters α and β only have n non-zero entries in their respective columns. Thus, the effective sample size for α and β is of order n smaller than the one for the global parameters μ and γ. This will also be reflected in the different rates of convergence we obtain in Theorem 4.4 below.

Rescaled parameters

While in SβM-C the rescaled parameters (c.f. Section 2.7.1.3) were mostly a mathematical device needed for proving consistency of the estimator (2.4) (Theorem 2.4), the analogous notion in SRGM is integral for understanding the model selection consistency result, Theorem 4.1. Therefore, we will introduce them formally in the main text of this chapter.

Were we to naively ignore the differing sample sizes, our proofs would fail. In particular, the *compatibility condition* (cf. Section 4.2.2), crucial for proofs for LASSO-type problems, would not hold. We therefore need to adjust for the differing sample sizes. To that end, we introduce the matrix

$$T = \begin{bmatrix} \sqrt{n-1}I_{2n} & 0 \\ 0 & \sqrt{N}I_{p+1} \end{bmatrix},$$

where I_m is the $m \times m$ identity matrix and define the *sample size adjusted Gram matrix* Σ as

$$\Sigma = T^{-1}\mathbb{E}[D^T D]T^{-1}. \tag{4.6}$$

It will be convenient to cast problem (4.4) in terms of rescaled parameters $\bar{\theta}$ which adjust for the discrepancy in effective sample sizes. This new formulation is equivalent to the one in (4.4), but gives us a unified framework for treating convergence properties of our estimators. We will rely heavily on that rescaled version in our proofs. Precisely, define the *sample size adjusted design matrix* $\bar{D}$ as

$$\bar{D} = \left[\begin{array}{c|c|c} \bar{X} & \mathbf{1} & Z \end{array} \right] \in \mathbb{R}^{N \times (2n+p+1)},$$

where

$$\bar{X} = \left[\begin{array}{c|c} \bar{X}^{\text{out}} & \bar{X}^{\text{in}} \end{array} \right] = \left[\begin{array}{c|c} \sqrt{n}X^{\text{out}} & \sqrt{n}X^{\text{in}} \end{array} \right],$$

is blowing up the entries in D belonging to ϑ. For any parameter $\theta = (\vartheta^T, \mu, \gamma^T)^T \in \Theta$, we introduce the notation

$$\bar{\theta} = (\bar{\vartheta}, \mu, \gamma) = \left(\frac{1}{\sqrt{n}}\vartheta, \mu, \gamma \right). \tag{4.7}$$

In particular we use the notation $\bar{\theta}_0 = (\bar{\vartheta}_0^T, \mu_0, \gamma_0^T)^T$, to denote the rescaled true parameter. The blow-up factor $\sqrt{n}$ was chosen such that we can now reformulate our problem as a problem in which each parameter effectively has sample size N in the sense that

$$\Sigma = \frac{1}{N}\mathbb{E}[\bar{D}^T \bar{D}].$$

We find that the negative log-likelihood corresponding to the rescaled parameters (4.7) is

$$\begin{aligned} \bar{\mathcal{L}}(\bar{\theta}) = &- \sum_{i=1}^{n} \sqrt{n}\bar{\alpha}_i b_i - \sum_{i=1}^{n} \sqrt{n}\bar{\beta}_i d_i - d_+ \mu - \sum_{i \neq j} (Z_{ij}^T \gamma) A_{ij} \\ &+ \sum_{i \neq j} \log\left(1 + \exp\left(\sqrt{n}\bar{\alpha}_i + \sqrt{n}\bar{\beta}_j + \mu + Z_{ij}^T \gamma\right)\right) \end{aligned}$$

Our original penalized likelihood problem can be rewritten as

$$\hat{\bar{\theta}} = (\hat{\bar{\vartheta}}^T, \hat{\mu}, \hat{\gamma})^T = \underset{\bar{\vartheta}, \mu, \gamma}{\arg\min} \frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}) + \bar{\lambda}\|\bar{\vartheta}\|_1, \tag{4.8}$$

where $\bar{\lambda} = \sqrt{n}\lambda$ and the argmin is taken over $\bar{\Theta}_{\text{loc}} = \{\bar{\theta} : \theta \in \Theta, \|\bar{D}\bar{\theta}\|_\infty \leq r_n\}$. Note that $\bar{\Theta}_{\text{loc}}$ is convex. Given a solution $\hat{\bar{\theta}}$ for penalty $\bar{\lambda}$ to this modified problem (4.8), we can obtain a solution to our original problem (4.4) with penalty $\lambda = \bar{\lambda}/\sqrt{n}$ via

$$(\hat{\vartheta}, \hat{\mu}, \hat{\gamma}) = \left(\sqrt{n}\hat{\bar{\vartheta}}, \hat{\mu}, \hat{\gamma} \right).$$

Since for any $\theta \in \Theta$, $D\theta = \bar{D}\bar{\theta}$, the bound r_n is the same in the definitions of Θ_{loc} and $\bar{\Theta}_{\text{loc}}$. Note that $\theta \in \Theta_{\text{loc}}$ if and only if $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$. Clearly $\bar{\mathcal{L}}(\bar{\theta}) = \mathcal{L}(\theta)$ and $\mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})] = \mathbb{E}[\mathcal{L}(\theta)]$. Thus, $\bar{\theta}_0$ satisfies that $\bar{\theta}_0 = \arg\min_{\theta \in \bar{\Theta}} \mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})]$.

4.2 Theory

We outline the main assumptions first. For the covariates, we focus on the case when Z_{ij} is finite dimensional following a random design and γ_0 is a fixed vector.

Assumption 4.1. The Z_{ij} are independent with $\mathbb{E}[Z_{ij}] = 0$ and $|Z_{ij}|$ is uniformly bounded. The covariate parameter γ_0 lies in some compact, convex set $\Gamma \subset \mathbb{R}^p$ and p remains fixed. Further assume that there are constants $0 < c_{\min} < C$, independent of n, such that for all $n \in \mathbb{N}$, the minimum eigenvalue $\lambda_{\min} = \lambda_{\min}(n)$ and the maximum eigenvalue $\lambda_{\max} = \lambda_{\max}(n)$ of $\frac{1}{N}\mathbb{E}[Z^T Z]$ fulfil $c_{\min} \leq \lambda_{\min} \leq \lambda_{\max} \leq C < \infty$. Without loss of generality assume $c_{\min} < 1/2$.

As a result of Assumption 4.1, there exist constants $\kappa, c > 0$ such that $|Z_{ij}^T \gamma| \leq \kappa$ for all $1 \leq i \neq j \leq n$ and $|Z_{ij,k}| \leq c$ for all $1 \leq i \neq j \leq n, k = 1, \ldots, p$.

Assumption 4.2. $\theta_0 \in \Theta_{\text{loc}}$ or equivalently $r_{n,0} \leq r_n$.

Assumption 4.1 is quite standard. Note that Z_{ij}'s are not necessarily i.i.d., possibly having correlated entries and that Z_{ij} can be asymmetric in that $Z_{ij} \neq Z_{ji}$. We note that we could have made a fixed-design assumption but the random-design assumption is somewhat more interesting (cf. the analogous discussion for SβM-C in Section 2.1). Assumption 4.2 is rather harmless as it simply states that the parameter we are estimating is actually contained in the space over which we are optimizing. Therefore, without loss of generality we assume $r_n = r_{n,0}$ and thus $\rho_n = \rho_{n,0}$. Indeed, this can always be achieved by simply increasing r_n as needed. Results that take a potential model-misspecification when $r_n < r_{n,0}$ into account and quantify the resulting bias can be derived similarly to the results in Chapter 2 (Section 2.2, Theorem 2.4), but are omitted for reasons of space. Indeed, in practice, it will not be necessary to choose r_n explicitly, as discussed in Section 4.3. The existence of $r_{n,0}$ and $\rho_{n,0}$ are technical artefacts that encode the permissible sparsity of the networks we study and enter our rates of convergence.

Assumption B1. $\sqrt{n} s_+^2 \bar{\lambda} \rho_n^{-2} \to 0, n \to \infty$.

For all of our theorems striking the right balance between parameter sparsity s_+, network sparsity ρ_n and penalty $\bar{\lambda}$ is crucial. The restrictiveness of these balancing assumptions will depend on the complexity of the results being proven and we number them separately from the general assumptions as "Assumption Bi", $i = 1, 2, 3$,

to make their special standing explicit in our notation. Our main result on model selection consistency, Theorem 4.1, is the most refined of our theorems and hence Assumption B1 is the strongest such balancing assumption. In particular, the weaker balancing assumptions required to establish parameter estimation consistency (Theorem 4.4; Assumption B2), and asymptotic normality of $\hat{\gamma}$ (Theorem 4.5; Assumption B3), follow from Assumptions B1 and 4.3.

In particular, with probability tending to one, our estimator $\hat{\theta}$ in (4.4) will simultaneously recover the correct support S_0, estimate the true parameter θ_0 at the classical LASSO rate of convergence up to a factor ρ_n^{-1} and produce asymptotically normal estimators for γ_0.

4.2.1 Model selection consistency

Our main result for this section, Theorem 4.1, states that under the appropriate conditions our estimator $\hat{\theta}$ will exclude all the truly inactive parameters and correctly include all those truly active parameters whose value exceeds a certain threshold (that tends to zero with increasing n). The latter "ϑ-min"-condition is typical for model selection in high-dimensional logistic regression type problems (Ravikumar et al. 2010, Chen et al. 2020).

Recall that we use S_0 to refer to the active set of indices associated with $\vartheta_0 = (\alpha_0^T, \beta_0^T)^T$, whereas $S_{0,+} = S_0 \cup \{2n+1, \ldots, 2n+1+p\}$. In the following derivations it will be crucial to distinguish the two correctly. We use $S_{0,+}^c$ to denote the complement of $S_{0,+}$ in $[2n+1+p]$, that is $S_{0,+}^c = [2n+1+p] \backslash S_{0,+}$. Let S_0^c refer to the complement of S_0 in $[2n]$ *only*: $S_0^c = [2n] \backslash S_0$. While this may seem like a potential notational pitfall, this allows for much cleaner notation in our proofs.

We first state the main theorem of this section before giving more details on its derivation. Recall that $\bar{\lambda}$ is the penalty parameter in the rescaled version (4.8) of our problem (4.4). Also notice that $\hat{S} \coloneqq \{i : \hat{\bar{\vartheta}}_i > 0\} = \{i : \hat{\vartheta}_i > 0\}$. That is, the estimator (4.4) and (4.8) will always select the same active set of parameters.

Assumption 4.3. $-\frac{N\bar{\lambda}^2}{18} + \log(n) \to -\infty, n \to \infty$.

Assumption 4.3 requires $\bar{\lambda} > 3\sqrt{2} \cdot \sqrt{\log(n)/N}$, which is the typical rate for the penalty we would expect from classical LASSO literature (van de Geer & Bühlmann 2011). Theorem 4.1 is proved in Section 4.4.

Theorem 4.1. *Under Assumptions 4.1, 4.2, B1 and 4.3, and for n sufficiently large, with probability approaching one, the estimator $\hat{\theta}$ from* (4.4)*:*

1. excludes all the truly inactive parameters: $\hat{S} \cap S^c = \emptyset$ and,

2. *with penalty of order* $\bar{\lambda} \asymp \sqrt{\frac{\log(n)}{N}}$*, it includes all those truly active parameters whose value is larger than* $C \cdot \rho_n^{-1} \frac{\sqrt{\log(n)}}{\sqrt{n}}$*:*

$$\left\{ i : \vartheta_{0,i} > C \cdot \rho_n^{-1} \frac{\sqrt{\log(n)}}{\sqrt{n}} \right\} \subseteq \hat{S},$$

where the form of C *and the exact probability are given in the proof.*

Remark. 1. Assumption 4.3 requires the same regime for $\bar{\lambda}$ as specified by Theorems 4.4 and 4.5 which ensure consistent parameter estimation and asymptotic normality of $\hat{\gamma}$. Hence, consistent parameter estimation, inference on γ and support recovery are all possible simultaneously.

2. If we choose $\bar{\lambda} \asymp \sqrt{\frac{\log(n)}{N}}$ and s_+ is of lower order, such as growing logarithmically or constant, then, up to log-terms, Assumption B1 means the permissible network sparsity ρ_n may go to zero at most as fast as $n^{-1/4}$.

Our tool of choice for proving Theorem 4.1 is a *primal-dual witness construction*, similar to the one in Ravikumar et al. (2010). The idea is to construct a tuple $(\bar{\theta}^\dagger, \bar{z}^\dagger)$, such that $\bar{\theta}^\dagger$ solves (4.8), while identifying the correct support S_0 for ϑ_0 and $\bar{z}^\dagger$ is a solution to the Karush-Kuhn-Tucker (KKT) conditions (4.9) as outlined below. In the construction of $(\bar{\theta}^\dagger, \bar{z}^\dagger)$, we make use of knowledge of the true active set S_0, which makes it infeasible to use in practice. However, by Lemma 4.2 below, if the construction succeeds – we make precise what we mean by that below – any solution to (4.8) must have the same support as $\bar{\theta}^\dagger$. In summary, if the construction succeeds, our estimator $\hat{\bar{\theta}}$ must identify the correct support S_0, too. The bulk of the work in proving Theorem 4.1 is to show that the construction of $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ will be successful with high probability for large n.

It is important to point out that due to the mixture of deterministic and random columns in D and the differing sample sizes between ϑ and ξ, the standard assumptions in Ravikumar et al. (2010) imposed on the Hessian of $\mathcal{L}$ cannot simply be imposed in our model. Rather, a careful argument is needed to prove that analogous properties hold for sufficiently large n with high probability. See Section 4.4.1 for details.

Our starting point for proving Theorem 4.1 are the KKT conditions (Bertsekas 1995, Chapter 5): Equation (4.8) is a convex optimization problem. Hence, by subdifferential calculus, a vector $\bar{\theta}$ is a minimizer of (4.8) if and only if zero is contained in the subdifferential of $\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}) + \bar{\lambda}\|\bar{\vartheta}\|_1$ at $\bar{\theta}$. That is, if and only if there is a vector $\bar{z} \in \mathbb{R}^{2n+1+p}$ such that

$$0 = \frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}) + \bar{\lambda}\bar{z}, \tag{4.9}$$

and

$$\bar{z}_i = 1, \text{ if } \bar{\vartheta}_i > 0, i = 1, \ldots, 2n, \tag{4.10a}$$

$$\bar{z}_i \in [-1, 1], \text{ if } \bar{\vartheta}_i = 0, i = 1, \ldots, 2n, \tag{4.10b}$$

$$\bar{z}_i = 0, i = 2n + 1, \ldots, 2n + 1 + p. \tag{4.10c}$$

We call such a pair $(\bar{\theta}, \bar{z}) \in \mathbb{R}^{2n+1+p} \times \mathbb{R}^{2n+1+p}$ *primal-dual optimal* for the rescaled problem (4.8). Note that in the first $2n$ components of $\nabla\bar{\mathcal{L}}$ we are taking the derivative with respect to $\bar{\vartheta}$ instead of ϑ. This means we need to pay attention to additional $\sqrt{n}$-factors. For such a pair to identify the correct support S_0, it is sufficient for

$$\bar{\theta}_i > 0, \text{ for all } i \in S_0, \text{ and} \tag{4.11a}$$

$$\|\bar{z}_{S_{0,+}^c}\|_\infty < 1 \tag{4.11b}$$

to hold. Where (4.11a) ensures that all truly active indices are included and (4.11b) ensures that all truly inactive indices are excluded (due to (4.10a)). We call (4.11b) the *strict feasibility condition* as in Ravikumar et al. (2010).

We will proceed to construct a pair $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ that satisfies condition (4.9), (4.10a) - (4.10c) and (4.11a) - (4.11b) with high probability and for sufficiently large n. We say the construction *succeeds*, if $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ fulfils (4.9) - (4.11b), which in particular implies that $\bar{\theta}^\dagger$ identifies the correct support S_0 and also is a solution to (4.8).

By the following lemma, if the construction succeeds, any solution to (4.8) must have the same support as $\bar{\theta}^\dagger$. Thus, if the construction succeeds, our estimator $\hat{\bar{\theta}}$ must identify the correct support S_0, too.

Lemma 4.2. *Suppose the construction* $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ *fulfils equations* (4.9) *and* (4.10a) - (4.10c) *and* (4.11b). *Let* $S^\dagger = \{i : \bar{\vartheta}_i^\dagger > 0\}$. *Then,*

$$\hat{S} = S^\dagger.$$

In particular, if $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ *additionally fulfils* (4.11a), *then* $S^\dagger = S_0$, *and thus,* $\hat{S} = S_0$.

Lemma 4.2 is proved in Section 4.4.1. We now give a detailed description of the primal-dual witness construction.

Primal-dual witness construction.

1. Solve the restricted penalized likelihood problem

$$\bar{\theta}^\dagger = (\bar{\vartheta}^{\dagger,T}, \mu^\dagger, \gamma^{\dagger,T})^T = \arg\min \frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}) + \bar{\lambda}\|\bar{\vartheta}\|_1, \tag{4.12}$$

where the argmin is taken over all $\bar{\theta} = (\bar{\vartheta}^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}}$ with support $S_{0,+}$,

i.e. $\bar{\theta}^\dagger_{S_{0,+}} = \bar{\theta}^\dagger$ or equivalently $\bar{\theta}^\dagger_{S^c_{0,+}} = 0$. Thus, by construction, $\bar{\theta}^\dagger$ correctly excludes all inactive indices.

2. Since (4.12) is a convex problem, zero must be contained in its subdifferential at $\bar{\theta}^\dagger$. Thus, we set $\bar{z}^\dagger_i = 1$, if $\bar{\vartheta}^\dagger_i > 0$ such that (4.10a) holds and $\bar{z}^\dagger_i = 0, i = 2n+1, \ldots, 2n+1+p$, such that (4.10c) holds. By subdifferential calculus we find $\bar{z}^\dagger_i \in [-1, 1]$, for those $i \in S$ with $\bar{\vartheta}^\dagger_i = 0$ (in case there are any), such that (4.9) holds for those components in S.
3. Plug $\bar{\theta}^\dagger$ and $\bar{z}^\dagger$ into (4.9) and solve for the remaining components of $\bar{z}^\dagger$, such that (4.9) holds for $(\bar{\theta}^\dagger, \bar{z}^\dagger)$.

The challenge will be proving that (4.11a) and (4.11b) also hold, which together ensure that (4.10b) holds, too. This will be shown in Section 4.4.

4.2.2 Consistency

After having seen in Theorem 4.1 that our estimator $\hat{\theta}$ will recover the true set of active indices S_0 with high probability for sufficiently large n, in this section we will show that under similar assumptions it will also be consistent in terms of excess risk and ℓ_1-error. This section is very similar to Section 2.2 for SβM-C.

We follow once more the empirical risk literature (cf. Greenshtein & Ritov (2004), Koltchinskii (2011)) and analyse the performance of our estimator in terms of *excess risk*. To that end, define the excess risk for a parameter θ and its sample-size adjusted version $\bar{\theta}$ as

$$\mathcal{E}(\theta) \coloneqq \frac{1}{N}\mathbb{E}[\mathcal{L}(\theta) - \mathcal{L}(\theta_0)] \quad \text{and} \quad \bar{\mathcal{E}}(\bar{\theta}) = \frac{1}{N}\mathbb{E}[\bar{\mathcal{L}}(\bar{\theta}) - \bar{\mathcal{L}}(\bar{\theta}_0)]$$

respectively. By construction, $\theta_0 = \arg\min_{\theta\in\Theta} \mathcal{E}(\theta) = \arg\min_{\theta\in\Theta_{\text{loc}}(r_{n,0})} \mathcal{E}(\theta)$, where the second equality follows from Assumption 4.2. Also, $\bar{\mathcal{E}}(\bar{\theta}) = \mathcal{E}(\theta)$.

A compatibility condition. As in SβM-C, the *compatibility condition* is crucial for proving the consistency of our estimator (4.4). As in Chapter 2, the classical compatibility condition as for example defined for generalized linear models in van de Geer et al. (2014) does not hold. The reason for this is that ϑ and $(\mu, \gamma^T)^T$ have different effective sample sizes. Using similar techniques as for SβM-C, we can now show that the sample size adjusted Gram matrix fulfils the compatibility condition, see Appendix A.1.1 for the proof.

Proposition 4.3 (Compatibility condition)**.** *Under Assumption 4.1, for $s_0 = o(\sqrt{n})$ and n large enough, for every $\theta \in \mathbb{R}^{2n+1+p}$ with $\|\theta_{S^c_{0,+}}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1$, we have*

$$\|\theta_{S_{0,+}}\|_1^2 \leq \frac{2s_{0,+}}{c_{\min}}\theta^T\Sigma\theta.$$

Parameter estimation consistency is the most lenient of our theorems in terms of restrictions that we have to impose on the parameter sparsity s_0 and the network sparsity ρ_n. We may replace the stricter Assumption B1 by the following.

Assumption B2. $\sqrt{n} s_0 \rho_n^{-1} \bar{\lambda} \to 0, n \to \infty$.

Theorem 4.4 below suggests a choice of $\bar{\lambda} \asymp \sqrt{\log(n)/N}$. Under these conditions, Assumption B2 becomes $s_0 \rho_n^{-1} \sqrt{\log(n)/n} \to 0$, which is the same as Assumption 2.2 needed for estimation consistency in SβM-C. Up to an additional factor ρ_n^{-1}, which is the price we have to pay for allowing vanishing link probabilities, the permissible sparsity for ϑ_0 is thus the permissible sparsity in classical LASSO theory for an effective sample size of order n. This makes sense, considering the discussion of the differing effective sample sizes in Section 4.2. Also, this choice of $\bar{\lambda}$ together with Assumption B2 imply $s_0 = o(\sqrt{n})$, as is required by Proposition 4.3 and which thus is not a restriction.

Theorem 4.4. *Let Assumptions 4.1, 4.2 and B2 hold. Fix a confidence level t and let*

$$a_n \coloneqq \sqrt{\frac{2\log(2(2n+p+1))}{N}}(1 \vee c).$$

Choose $\lambda_0 = \lambda_0(t, n)$ as

$$\lambda_0 = 8a_n + 2\sqrt{\frac{t}{N}(11(1 \vee (c^2 p)) + 16(1 \vee c)\sqrt{n} a_n)} + \frac{4t(1 \vee c)\sqrt{n}}{3N}.$$

Let $\bar{\lambda} = \sqrt{n}\lambda \geq 8\lambda_0$. Then, with probability at least $1 - \exp(-t)$ we have

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{1}{\sqrt{n}}\|\hat{\vartheta} - \vartheta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1\right) \leq C\frac{s_{0,+}\bar{\lambda}^2}{\rho_{n,0}},$$

with constant $C = 128/c_{\min}$.

Remark. The proof for consistency follows almost exactly, sometimes even line by line, as for the SβM-C in Theorem 2.4. Therefore, we only state the theorem here and defer its proof, which we include for completeness, to Appendix A.1. Indeed, our proof strategy is the same as in SβM-C: We first assert that the basic inequality holds for our estimator (4.4) and its rescaled version (4.8) (Section A.1.2) and derive a lower quadratic margin condition (Section A.1.4). We prove consistency on a special set $\mathcal{I}$ (Section A.1.5) and conclude that $\mathcal{I}$ has probability approaching one (A.1.6). Putting these pieces together results in Theorem 4.4.

Remark. The conditions in Theorem 4.4 imply $\lambda_0 \asymp \sqrt{\log(n)/N}$, suggesting that we may choose $\bar{\lambda}$ of the same order. Thus, up to the additional factor ρ_n^{-1}, we obtain the classical LASSO rates of convergence for a parameter of effective sample size N for μ and γ and those for a parameter of effective sample size n for α and β.

If s_0 is a lower order term, such as growing logarithmically or constant, then up to log-factors Assumption B2 requires that ρ_n tend to zero at rate at most as fast as $1/\sqrt{n}$, which allows for sparser networks than what we had for model selection consistency in Theorem 4.1.

4.2.3 Inference

We derive the limiting distribution of $\hat{\gamma}$ and as a by-product of our proofs we also obtain an analogous limiting result for $\hat{\mu}$. The employed methodology and results are analogous to the ones used for SβM-C in Section 2.4. Therefore, we give the general idea and the main result here and defer the derivations to the appendix.

Our strategy for proving Theorem 4.5 below will be inverting the KKT conditions similar to van de Geer et al. (2014) and to what we did in Section 2.4 for SβM-C. This relies on a Taylor expansion followed by the inversion of the Hessian of the negative log-likelihood $\mathcal{L}$ with respect to $\xi = (\mu, \gamma^T)^T$. See Appendix A.2 for details. The difficulty is that the Hessian will be singular in the limit, because we allow our link probabilities to go to zero.

In detail, denote by $H(\hat{\theta}) \coloneqq H_{\xi\times\xi}(\theta)|_{\theta=\hat{\theta}} \in \mathbb{R}^{(p+1)\times(p+1)}$ the Hessian of $\frac{1}{N}\mathcal{L}(\theta)$ with respect to ξ only, evaluated at $\hat{\theta}$. Let $D_\xi = [\mathbf{1}|Z]$ be the part of D corresponding to ξ with rows $D^T_{\xi,ij} = (1, Z^T_{ij}), i \neq j$. Also, let $\hat{W}^2 = \operatorname{diag}\Big(p_{ij}(\hat{\theta})(1 - p_{ij}(\hat{\theta})), i \neq j\Big)$. Then,

$$H(\hat{\theta}) = \frac{1}{N} D_\xi^T \hat{W}^2 D_\xi.$$

Let $W_0^2 = \operatorname{diag}(p_{ij}(\theta_0)(1-p_{ij}(\theta_0)), i \neq j)$ and consider the corresponding population version:

$$\mathbb{E}[H(\theta_0)] = \frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi].$$

To be consistent with commonly used notation, call $\hat{\Sigma}_\xi = H(\hat{\theta})$ and $\Sigma_\xi = \mathbb{E}[H(\theta_0)]$ and $\hat{\Theta}_\xi \coloneqq \hat{\Sigma}_\xi^{-1}, \Theta_\xi \coloneqq \Sigma_\xi^{-1}$.

For the proof of asymptotic normality we need to invert $\hat{\Sigma}_\xi$ and Σ_ξ and show that these inverses are close to each other in an appropriate sense. It is commonly assumed in LASSO theory (cf. van de Geer et al. (2014)) that the minimum eigenvalues of these matrices stay bounded away from zero, uniformly in n. In our case, however, such an assumption is invalid. As we have argued, it is a necessary condition for modelling sparse networks to allow $p_{ij} \to 0$, since otherwise each node degree will scale linearly in n, putting us in the dense graph regime. Alas, for the general setting described here, any lower bound on the diagonal entries of W_0 and $\hat{W}$ and thus also any lower bound on the minimum eigenvalue of Σ_ξ and $\hat{\Sigma}_\xi$ will tend to zero with growing n. A careful argument is needed and we have to impose

a slightly stricter balancing assumption than the Assumption B2 we used for our consistency result.

Assumption B3. $\sqrt{n}s_0\rho_n^{-2}\bar{\lambda} \to 0, n \to \infty$.

Theorem 4.5. *Under Assumptions 4.1, 4.2 and B3, with λ fulfilling the conditions of Theorem 4.4, we have for any $k = 1, \ldots, p$, as $n \to \infty$,*

$$\sqrt{N}\frac{\hat{\gamma}_k - \gamma_{0,k}}{\sqrt{\hat{\Theta}_{\vartheta,k+1,k+1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

We also have for our estimator of the global sparsity parameter, $\hat{\mu}$, as $n \to \infty$,

$$\sqrt{N}\frac{\hat{\mu} - \mu_0}{\sqrt{\hat{\Theta}_{\vartheta,1,1}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Contrary to what is commonly seen in the penalized likelihood literature (Zhang & Zhang 2014, van de Geer et al. 2014), no debiasing of $\hat{\gamma}$ and $\hat{\mu}$ is needed, as was also the case for SβM-C. The reason for this is that columns of D pertaining to those parameters which are indeed biased, that is to ϑ, and those pertaining to $\xi = (\mu, \gamma^T)^T$ become asymptotically orthogonal, meaning that the bias in $\hat{\xi}$ vanishes fast enough for the derivation of Theorem 4.5 to be possible. Notice that for a lower order s_0, Assumption B3 essentially allows for the same level of network sparsity as Assumption B1, up to lower order factors. Also, under the stated conditions on λ, Assumption B3 is analogous to Assumption 2.3 in SβM-C.

4.3 Simulation: SRGM

We demonstrate the effectiveness of our estimator (4.4) in consistently performing simultaneous parameter estimation and model selection. To this end, we tested it on networks of varying sizes. Specifically, we let n vary from 150 to 800 in steps of 50 and chose s_0 close to $\sqrt{n}/2$ and $s_\alpha = s_\beta = s_0/2$. The values for s_0 are given in Table 4.1. We selected a heterogeneous configuration for the assignment of non-zero α and β values. That is, we included dedicated "spreader" nodes, with large α and zero β value, as well as "attractor" nodes, with large β and zero α, as well as some nodes with active α and β. In detail, we let

$$\alpha = (2, 1.5, 1, 0.8, \ldots, 0.8, 0, \ldots, 0),$$
$$\beta = (0, \ldots, 0, 2, 1.5, 1, 0.8, \ldots, 0.8, 0, \ldots, 0),$$

where the number of entries with value 0.8 was chosen to match the chosen sparsity level (zero for the first three values of n) and the number of leading zeros in β was

n	150	200	250	300	350	400	450	500	550	600	650	700	750	800
s_0	6	6	6	8	8	10	10	10	10	12	12	12	12	14

Table 4.1: The sparsity level s_0 for each value of n.

n	Median edge density	$\min p_{ij}$	$\max p_{ij}$
150	0.140	0.059	0.888
200	0.130	0.055	0.886
250	0.123	0.052	0.879
300	0.119	0.050	0.873
350	0.115	0.048	0.868
400	0.112	0.047	0.867
450	0.109	0.046	0.864
500	0.107	0.045	0.861
550	0.105	0.044	0.859
600	0.104	0.043	0.858
650	0.102	0.043	0.854
700	0.100	0.042	0.856
750	0.099	0.041	0.851
800	0.098	0.041	0.853

Table 4.2: Network summary statistics for directed network model for various values of n.

chosen such that there were exactly two nodes with both active α and β. We let the networks get progressively sparser by setting $\mu = -1.2 \cdot \log(\log(n))$. We used $p = 2$ and sampled the covariates $Z_{ij,k}, k = 1, 2, i \neq j$, from centred Beta$(2, 2)$ distributions, that is $Z_{ij,k} \sim \text{Beta}(2, 2) - 1/2$. We weighted the covariates with $\gamma = (1, 0.8)^T$. For each value of n we drew $M = 500$ realizations of this model. The observed median edge density, as well as median minimum and maximum link probabilities p_{ij} are recorded in Table 4.2.

Our estimator requires us to choose a tuning parameter λ and as in the SβM-C we explored the use of the Bayesian Information Criterion (BIC) as well as a heuristic based on our developed theory for model selection. We will see that, while the two model selection procedures perform similarly in terms of parameter estimation and inference for γ, with BIC achieving slightly better results, the heuristic based on our developed theory is superior to BIC in terms of model selection consistency.

Recall the definition of BIC from Section 2.5: We denote the solution of (4.4) when using penalty λ by $\hat{\theta}(\lambda) = (\hat{\alpha}(\lambda)^T, \hat{\beta}(\lambda)^T, \hat{\mu}(\lambda), \hat{\gamma}(\lambda)^T)^T$ and write $s(\lambda) = |\{i : \hat{\vartheta}_i(\lambda) > 0\}|$ for its sparsity. The value of the BIC at λ is given by

$$\text{BIC} = 2\mathcal{L}(\hat{\theta}(\lambda)) + s(\lambda) \log(N)$$

and the penalty λ was chosen to minimize BIC.

Our heuristic is motivated by the theory developed in the previous sections. We have two restrictions on the choice of the penalty parameter λ, namely the ones specified in Theorems 4.1 and 4.4. While both Theorems demand λ to be of the

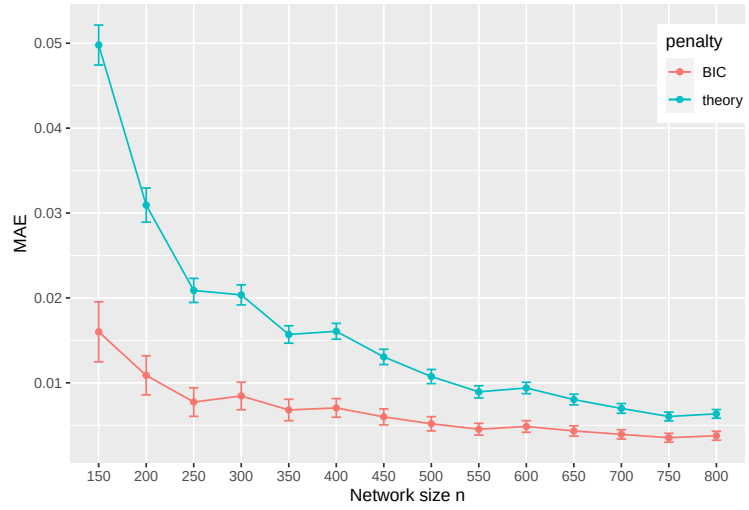
same order, upon inspection we see that the conditions imposed by Theorem 4.4 demand λ to be larger in terms of the leading constant. Thus, we may use essentially the same heuristic as in Section 2.5 for SβM-C. In detail, recall that Theorem 4.4 suggests that based on a confidence level t picked by us, we should first define λ_0 as given in the theorem. We pick $t = 3$ and set c to the maximum observed covariate value and, as in the case of SβM-C, we drop the factor eight in the relation between λ_0 and $\bar{\lambda}$, as it is a technical artefact. Decreasing the penalty in this manner is in line with empirical findings that suggest that in high-dimensional settings the penalty values prescribed by mathematical theory in practice tend to over-penalize the parameter values (Yu et al. 2019).

We drew $M = 500$ realizations for each value of n and recorded the mean absolute error for estimation of $(\alpha^T, \beta^T)^T$, the absolute error for estimation of μ and the ℓ_1-error for estimation of γ. We also constructed confidence intervals as prescribed by Theorem 4.5 and recorded the empirical coverage at the nominal 95% level. Finally, we studied how well BIC and our heuristic did in terms of identifying the correct model.

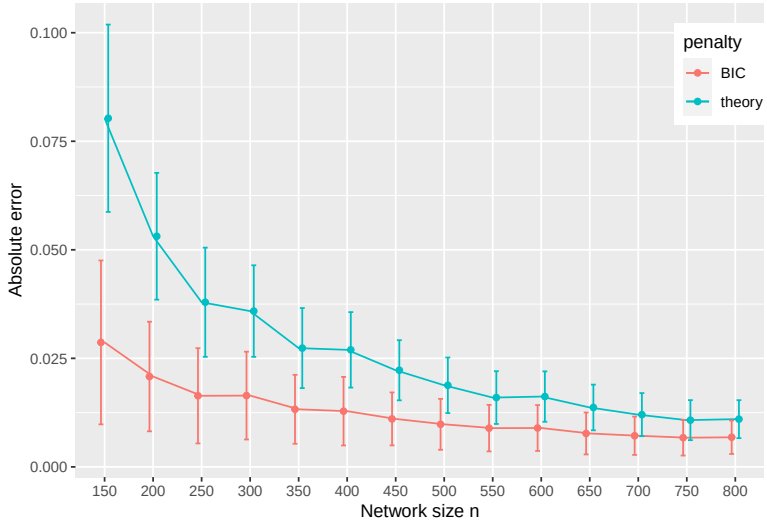
Consistency. We display the error statistics for estimation of $\vartheta_0 = (\alpha_0^T, \beta_0^T)^T$, μ_0 and γ_0 in Figures 4.1a, 4.1b and 4.1c respectively. We see that the error decreases with increasing network size for both model selection procedures. Especially for small n, BIC outperforms the heuristic for ϑ_0 and μ_0, while they both give essentially the same results for estimation of γ_0. The better performance of BIC is less prominent as n increases. BIC selects the penalty in a purely data driven manner, which allows it to adapt to differing degrees of sparsity in the network, while for the heuristic the penalty value only depends on n and p. This additional flexibility is what allows BIC to achieve lower error values.

Asymptotic normality. We construct confidence intervals at the nominal 95% level for our estimators of $\gamma_{0,1}$ and $\gamma_{0,2}$ as prescribed by Theorem 4.5. Table 4.3 shows the results for $\gamma_{0,1}$ across the values of n. The results for $\gamma_{0,2}$ are similar and are omitted for reasons of space. The coverage is very close to the 95%-level across all network sizes, independent of which model selection criterion we use. This is to be expected, considering that there was hardly any difference for the estimation of γ between our two model selection criteria. This empirically illustrates the validity of the asymptotic results derived in Theorem 4.5. As expected, the median length of the confidence interval decreases with increasing network size.

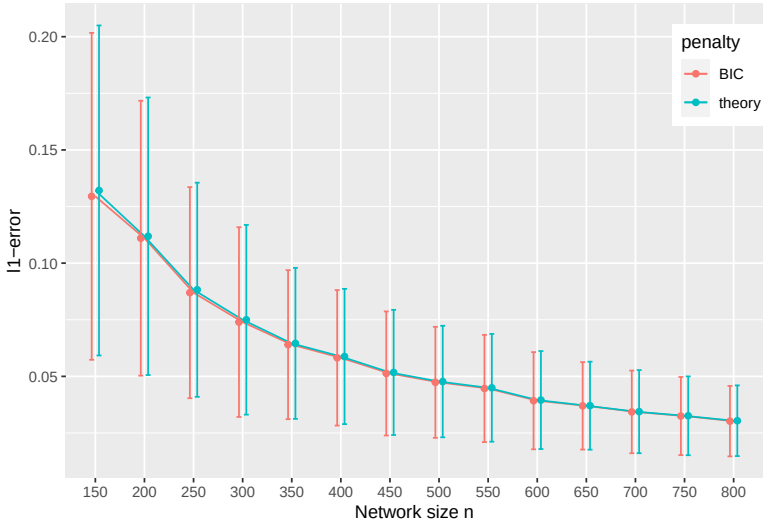
Model selection. Figure 4.2a shows the empirical probability of selecting the correct model for the various network sizes for BIC and the heuristic. We see very clearly that, as n grows, our heuristic outperforms BIC, achieving correct model selection almost all the time. Nonetheless, it is worth pointing out that even though



(a) Mean absolute error for $\vartheta_0 = (\alpha_0^T, \beta_0^T)^T$.



(b) Absolute error for μ_0



(c) ℓ_1-error for γ_0.

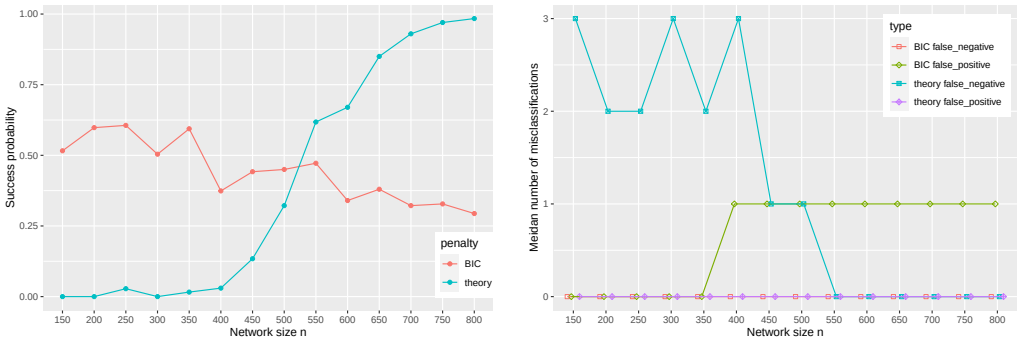
Figure 4.1: Mean absolute error for $\vartheta_0 = (\alpha_0^T, \beta_0^T)$, absolute error for μ_0 and ℓ_1-error for γ_0 for varying n. The results for BIC are presented in red, the ones for our heuristic in green. The dots are the mean errors and the error bars are of length one standard deviation.

BIC may not select the exact correct model, the number of misclassifications it does on average is not very large, as shown in Figure 4.2b. Figure 4.2b also shows that the heuristic, by virtue of selecting a larger penalty than BIC, will on average incur

n	Coverage	CI	Coverage	CI
	Pre-determined λ		BIC	
150	0.960	0.342	0.962	0.344
200	0.952	0.265	0.962	0.266
250	0.952	0.217	0.954	0.218
300	0.944	0.183	0.948	0.184
350	0.946	0.160	0.952	0.160
400	0.950	0.141	0.964	0.141
450	0.962	0.127	0.960	0.127
500	0.940	0.115	0.944	0.115
550	0.952	0.106	0.950	0.106
600	0.954	0.097	0.956	0.097
650	0.960	0.091	0.964	0.091
700	0.946	0.085	0.950	0.085
750	0.944	0.079	0.946	0.079
800	0.946	0.075	0.952	0.075

Table 4.3: Empirical coverage for estimation of $\gamma_{0,1}$ under nominal 95% coverage and median lengths of confidence intervals.

more false negatives for small n. On the other hand, as n grows, BIC will incur false positives, resulting in the decreasing probability of selecting the correct model. Running the risk of selecting a slightly misspecified model with BIC may seem an acceptable price to pay in many applications, when considering that, unlike the heuristic, BIC is capable of a data driven selection of the penalty parameter and results in slightly better parameter estimation on average. In the end, it will depend on the preference of the statistician and the application at hand if they value exact model recovery over improved parameter estimation.



(a) Probability of correct model selection.

(b) Median number of misclassifications.

Figure 4.2: (a): The empirical probability of selecting the correct subset of active indices. (b): The median number of misclassifications for each model selection procedure, split up into false positives and false negatives.

4.4 Proofs of Chapter 4

4.4.1 Proof of Lemmas 4.2, 4.6, 4.7

To make the representation cleaner, for the remainder of Section 4.4 we will simply write S for S_0 and S_+ for $S_{0,+}$. Recall that we use S_+^c to denote the complement of S_+ in $[2n+1+p]$, that is $S_+^c = [2n+1+p]\backslash S_+$. We also use S^c to refer to the complement of S in $[2n]$ *only*: $S^c = [2n]\backslash S$.

Proof of Lemma 4.2. Since $\bar{\theta}^\dagger$ and $\hat{\bar{\theta}}$ both solve (4.8), we must have

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda}\|\bar{\vartheta}^\dagger\|_1 = \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) + \bar{\lambda}\|\hat{\bar{\vartheta}}\|_1.$$

Denote by $\bar{z}_\vartheta^\dagger$ the first $2n$ components of $\bar{z}^\dagger$. Then, by (4.10a) and (4.10b), $\langle \bar{z}_\vartheta^\dagger, \bar{\vartheta}^\dagger\rangle = \|\bar{\vartheta}^\dagger\|_1$. Thus,

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda}\langle \bar{z}_\vartheta^\dagger, \bar{\vartheta}^\dagger\rangle = \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) + \bar{\lambda}\|\hat{\bar{\vartheta}}\|_1.$$

Hence, using that the last $p+1$ components of $\bar{z}^\dagger$ are zero,

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda}\langle \bar{z}^\dagger, \bar{\theta}^\dagger - \hat{\bar{\theta}}\rangle = \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) + \bar{\lambda}\left(\|\hat{\bar{\vartheta}}\|_1 - \langle \bar{z}^\dagger, \hat{\bar{\theta}}\rangle\right).$$

But by (4.9), $\bar{\lambda}\bar{z}^\dagger = -1/N \cdot \nabla\bar{\mathcal{L}}(\bar{\theta}^\dagger)$ and therefore

$$\frac{1}{N}\bar{\mathcal{L}}(\bar{\theta}^\dagger) - \langle 1/N \cdot \nabla\bar{\mathcal{L}}(\bar{\theta}^\dagger), \bar{\theta}^\dagger - \hat{\bar{\theta}}\rangle - \frac{1}{N}\bar{\mathcal{L}}(\hat{\bar{\theta}}) = \bar{\lambda}\left(\|\hat{\bar{\vartheta}}\|_1 - \langle \bar{z}^\dagger, \hat{\bar{\theta}}\rangle\right).$$

By the convexity of $\bar{\mathcal{L}}$, the left-hand side in the above display is negative. Therefore,

$$\|\hat{\bar{\vartheta}}\|_1 \leq \langle \bar{z}^\dagger, \hat{\bar{\theta}}\rangle = \langle \bar{z}_\vartheta^\dagger, \hat{\bar{\vartheta}}\rangle \leq \|\bar{z}_\vartheta^\dagger\|_\infty \|\hat{\bar{\vartheta}}\|_1 \leq \|\hat{\bar{\vartheta}}\|_1.$$

Hence, $\langle \bar{z}_\vartheta^\dagger, \hat{\bar{\vartheta}}\rangle = \|\hat{\bar{\vartheta}}\|_1$. But since $\|\bar{z}_{S^{\dagger c}}^\dagger\|_\infty < 1$ by (4.11b), this can only hold if $\hat{\bar{\vartheta}}_{S^{\dagger c}} = 0$. The claim follows. □

For the proof of Theorem 4.1 we need conditions akin to those used in Ravikumar et al. (2010). The first condition is the so-called *dependency condition* which demands that the population Hessian of $\bar{\mathcal{L}}$ with respect to the variables contained in the active set S is invertible. For our specific case, the Hessian of $1/N \cdot \bar{\mathcal{L}}$ with respect to $\bar{\vartheta}$ only is

$$Q \coloneqq \frac{1}{n-1} X^T W_0^2 X = H_{\bar{\vartheta}\times\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}) \in \mathbb{R}^{2n\times 2n}, \tag{4.13}$$

where $W_0 = \operatorname{diag}(\sqrt{p_{ij}(\theta_0)(1-p_{ij}(\theta_0))}, i \neq j)$.

Lemma 4.6 (Dependency condition)**.** *For any n,*

$$\operatorname{mineval}(Q_{S,S}) \geq \frac{1}{2}\rho_n \cdot \left(1 - \frac{\max\{s_\alpha, s_\beta\}}{n-1}\right) > 0.$$

Proof of Lemma 4.6. Notice that

$$\frac{1}{n-1}X^TX = \begin{bmatrix} I_n & B \\ B & I_n \end{bmatrix} \in \mathbb{R}^{2n\times 2n},$$

where I_n is the $(n \times n)$ identity matrix and B is a matrix with zeros on the diagonal and $1/(n-1)$ everywhere else. Consider the sub-matrix with only those rows and columns belonging to S:

$$P \coloneqq \frac{1}{n-1}(X^TX)_{S\times S} = \begin{bmatrix} I_{s_\alpha} & B_{S_\alpha,S_\beta} \\ B_{S_\beta,S_\alpha} & I_{s_\beta} \end{bmatrix} \in \mathbb{R}^{s\times s}.$$

This matrix P is strictly diagonally dominant. Indeed,

$$\sum_{j\in S, j\neq i} P_{ij} = \frac{s_\beta}{n-1} < 1 = P_{ii}, \quad i \in S_\alpha$$
$$\sum_{j\in S, j\neq i} P_{ij} = \frac{s_\alpha}{n-1} < 1 = P_{ii}, \quad i \in S_\beta.$$

Thus, P is strictly positive definite. More, by the Gershgorin Circle Theorem, all the eigenvalues of P must lie in one of the discs $D(P_{ii}, R_i)$, where $R_i = \sum_{j\in S, j\neq i} P_{ij}$ and $D(P_{ii}, R_i)$ is the disc with radius R_i centred at P_{ii}. In particular,

$$\text{mineval}(P) \geq 1 - \frac{\max\{s_\alpha, s_\beta\}}{n-1}.$$

But now, for any $v \in \mathbb{R}^s$,

$$v^TQ_{S,S}v \geq \frac{1}{2}\rho_n \cdot v^TPv \geq \frac{1}{2}\rho_n \left(1 - \frac{\max\{s_\alpha, s_\beta\}}{n-1}\right)\|v\|_2^2$$

and the claim follows. □

The next condition we need is the so-called *incoherence condition.*

Lemma 4.7 (Incoherence condition)**.** *For any n,*

$$\|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty \leq \frac{1}{2}\rho_n^{-1} \cdot \frac{\max\{s_\alpha, s_\beta\}}{n - \max\{s_\alpha, s_\beta\}}.$$

By Lemma 4.6 the left-hand side in Lemma 4.7 is well-defined. Under Assumption B1, the right-hand side in Lemma 4.7 tends to zero as $n \to \infty$.

Proof of Lemma 4.7. We make use of the following bound of a the infinity norm of the inverse of a diagonally dominant matrix (see for example Varah (1975))

$$\|Q_{S,S}^{-1}\|_\infty \leq \max_{i\in S}\left\{\frac{1}{|q_{ii}| - R_i}\right\},$$

where q_{ii} is the ith diagonal entry of $Q_{S,S}$ and R_i is the sum of the off-diagonal elements of the ith row of $Q_{S,S}$. That is, for $i \in S_\alpha$,

$$q_{ii} - R_i = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} p_{ij}(1-p_{ij}) - \frac{1}{n-1} \sum_{j \in S_\beta, j \neq i} p_{ij}(1-p_i j) \geq \frac{1}{2(n-1)} \rho_n (n - s_\beta),$$

and analogously for $i \in S_\beta$,

$$q_{ii} - R_i \geq \frac{1}{2(n-1)} \rho_n (n - s_\alpha).$$

Thus,

$$q_{ii} - R_i \geq \frac{1}{2(n-1)} \rho_n (n - \max\{s_\alpha, s_\beta\})$$

and therefore,

$$\|Q_{S,S}^{-1}\|_\infty \leq 2\rho_n^{-1} \cdot \frac{n-1}{n - \max\{s_\alpha, s_\beta\}}. \tag{4.14}$$

Furthermore, any row of $Q_{S^c,S}$ has either s_α or s_β non-zero entries, each of the form $1/(n-1) \cdot p_{ij}(1-p_{ij}) \leq 1/(4(n-1))$. Hence,

$$\|Q_{S^c,S}\|_\infty \leq \frac{\max\{s_\alpha, s_\beta\}}{4(n-1)}.$$

The claim follows by the sub-multiplicativity of the matrix infinity norm. □

4.4.2 General strategy

The proof of Theorem 4.1 hinges on the construction of $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ succeeding with high probability. The challenge in proving this is proving that $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ fulfils conditions (4.11a) and (4.11b). Our proof relies on the following derivations. From (4.9) we obtain

$$0 = \frac{1}{N} \nabla \bar{\mathcal{L}}(\bar{\theta}^\dagger) + \bar{\lambda} \bar{z}^\dagger - \frac{1}{N} \nabla \bar{\mathcal{L}}(\bar{\theta}_0) + \frac{1}{N} \nabla \bar{\mathcal{L}}(\bar{\theta}_0).$$

Doing a Taylor expansion along the same lines as (A.16) and (A.17), we obtain

$$\frac{1}{N} \nabla \bar{\mathcal{L}}(\bar{\theta}^\dagger) - \frac{1}{N} \nabla \bar{\mathcal{L}}(\bar{\theta}_0) = \frac{1}{N} \bar{D}^T W_0^2 \bar{D} (\bar{\theta}^\dagger - \bar{\theta}_0) + O\left(\frac{1}{N} \sum_{i \neq j} \bar{D}_{ij} |\bar{D}_{ij}^T (\bar{\theta}^\dagger - \bar{\theta}_0)|^2 \right),$$

where we have used the fact that we are taking derivatives with respect to $\bar{\theta}$ and used $\bar{D}_{ij}\bar{\theta}_0$ in (A.17), to obtain W_0^2 instead of $\hat{W}^2$ above. Combining the last two equations, we obtain

$$\frac{1}{N} \bar{D}^T W_0^2 \bar{D} (\bar{\theta}^\dagger - \bar{\theta}_0) = -\bar{\lambda} \bar{z}^\dagger - \frac{1}{N} \nabla \bar{\mathcal{L}}(\bar{\theta}_0) + O\left(\frac{1}{N} \sum_{i \neq j} \bar{D}_{ij} |\bar{D}_{ij}^T (\bar{\theta}^\dagger - \bar{\theta}_0)|^2 \right).$$

Taking only the first $2n$ entries of that equation we obtain

$$\frac{1}{N}\bar{X}^T W_0^2 \bar{X}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0) = -\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0) + \frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n} + \bar{R} \tag{4.15}$$

where we use $\bar{z}^\dagger_{1:2n}$ to refer to the first $2n$ components of $\bar{z}^\dagger_{1:2n}$, use our shorthand notation $\xi = (\mu, \gamma^T)^T$ and let

$$\bar{R} = O\left(\frac{1}{N}\sum_{i\neq j}\bar{X}_{ij}|\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2\right).$$

The left-hand side in (4.15) is equal to

$$Q(\bar{\vartheta}^\dagger - \bar{\vartheta}_0) = Q_{-,S}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S + Q_{-,S^c}\underbrace{(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_{S^c}}_{=0}.$$

Plugging this into (4.15) and splitting up by rows, we get

$$\begin{aligned} Q_{S,S}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S =& -\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}(\bar{X}_{-,S})^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0) \\ & - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S, \end{aligned} \tag{4.16a}$$

$$\begin{aligned} Q_{S^c,S}(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S =& -\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_{S^c} + \frac{1}{N}(\bar{X}_{-,S^c})^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0) \\ & - \bar{\lambda}\bar{z}^\dagger_{1:2n,S^c} + \bar{R}_{S^c}, \end{aligned} \tag{4.16b}$$

where it is important to remember that $S^c = [2n]\backslash S$. We solve (4.16a) for $(\bar{\vartheta}^\dagger - \bar{\vartheta}_0)_S$ and plug the result into (4.16b). Finally we rearrange for $-\bar{\lambda}\bar{z}^\dagger_{1:2n,S^c}$,

$$\begin{aligned} -\bar{\lambda}\bar{z}^\dagger_{1:2n,S^c} =& Q_{S^c,S}Q_{S,S}^{-1}\Big\{ -\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}(\bar{X}_{-,S})^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0) \\ & - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S\Big\} \\ & + \frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_{S^c} - \frac{1}{N}(\bar{X}_{-,S^c})^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0) - \bar{R}_{S^c}. \end{aligned}$$

Now, divide by $\bar{\lambda}$ and take the ∞-norm on both sides. Rearrange corresponding terms and use (4.10a).

$$\begin{aligned} \|\bar{z}^\dagger_{1:2n,S^c}\|_\infty \leq & \frac{1}{\bar{\lambda}}\left\{\|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty + 1\right\}\left\|\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty && (I) \\ & + \frac{1}{\bar{\lambda}}\left\{\|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty + 1\right\}\left\|\bar{R}\right\|_\infty && (II) \\ & + \frac{1}{\bar{\lambda}}\left\{\|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty + 1\right\}\left\|\frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0)\right\|_\infty && (III) \\ & + \left\|Q_{S^c,S}Q_{S,S}^{-1}\right\|_\infty && (IV). \end{aligned}$$

By appropriately bounding the terms $(I) - (IV)$ on the right-hand side, we will proceed to show that for sufficiently large n, with high probability, $\|\bar{z}^\dagger_{1:2n,S^c}\|_\infty < 1$,

which is equivalent to (4.11b). Notice that we already may control term (IV) as well as the terms $\|Q_{S^c,S}Q_{S,S}^{-1}\|_\infty + 1$ by the incoherence condition, Lemma 4.7.

4.4.3 Controlling term (I)

Notice that the ith component of $\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)$ is of the form

$$\frac{1}{N}\sqrt{n}\sum_{j=1,j\neq i}(A_{ij}-p_{ij}) = \frac{1}{\sqrt{n}}\cdot\frac{1}{n-1}\sum_{j=1,j\neq i}(A_{ij}-p_{ij}).$$

In particular, each summand is a centred, bounded random variable. By Hoeffding's inequality, we have for every $t>0$,

$$P\left(\left|\frac{1}{n-1}\sum_{j=1,j\neq i}(A_{ij}-p_{ij})\right|\geq t\right)\leq 2\exp\left(-\frac{n-1}{2}t^2\right).$$

Thus, for any $\epsilon>0$, picking $t=\epsilon\sqrt{n}\bar{\lambda}$, gives

$$P\left(\frac{1}{\bar{\lambda}}\left|\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)_i\right|\geq\epsilon\right)\leq 2\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2\right).$$

Taking a union bound over all $2n$ components of $\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)$, leads to

$$P\left(\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty\geq\epsilon\right)\leq 4n\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2\right)=4\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2+\log(n)\right). \tag{4.17}$$

In the next section, when controlling term (II), we will also need a similar bound on the components of $\frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0)$ corresponding to $\xi=(\mu,\gamma^T)^T$, which is why we derive the respective bounds now. Using analogous arguments to the above, we obtain

$$P\left(\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\nabla_{\xi}\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty\geq\epsilon\right)\leq 2(p+1)\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2(1\vee c^2)}\epsilon^2\right). \tag{4.18}$$

Combining (4.17) and (4.18), we obtain a bound on the infinity norm of the full gradient,

$$\begin{aligned}&P\left(\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\nabla\bar{\mathcal{L}}(\bar{\theta}_0)\right\|_\infty\geq\epsilon\right)\\&\quad\leq 4\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2+\log(n)\right)+2(p+1)\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2(1\vee c^2)}\epsilon^2\right),\end{aligned} \tag{4.19}$$

which tends to zero, as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2+\log(n)\to\infty$, as n tends to infinity.

4.4.4 Controlling term (II)

Controlling term (II) is by far the most involved step in controlling $\|\bar{z}^{\dagger}_{1:2n,S^c}\|_{\infty}$. We start by controlling the ℓ_2-error between our construction $\bar{\theta}^{\dagger}$ and the truth $\bar{\theta}_0$.

Lemma 4.8. *Under Assumptions 4.1, 4.2, B1, 4.3, for n large enough, for any $\epsilon > 0$, with probability at least*

$$1 - 4 \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n)\right) - 2(p+1) \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2(1 \vee c^2)}\epsilon^2\right) - p(p+3)\exp\left(-N\frac{c^2_{\min}}{2048 s_+^2 \tilde{c}}\right),$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n) \to -\infty$, as $n \to \infty$, we have

$$\|\bar{\theta}^{\dagger} - \bar{\theta}_0\|_1 \leq (1+\epsilon)\frac{9}{c_{\min}}\rho_n^{-1} s_+ \bar{\lambda}.$$

Proof. Keep in mind that $\bar{\theta}^{\dagger} - \bar{\theta}_0 = \bar{\theta}^{\dagger}_{S_+} - \bar{\theta}_{0,S_+}$. Define a function $G : \mathbb{R}^{s+1+p} \to \mathbb{R}$,

$$G(u) = \frac{1}{N}\left\{\bar{\mathcal{L}}(\bar{\theta}_{0,S_+} + u) - \bar{\mathcal{L}}(\bar{\theta}_{0,S_+})\right\} + \bar{\lambda}\left(\|\bar{\theta}_{0,S} + u_S\|_1 - \|\bar{\theta}_{0,S}\|_1\right),$$

where for the addition $\bar{\theta}_{0,S_+} + u$ to be well-defined, we use the canonical embedding of $\mathbb{R}^{s+1+p} \hookrightarrow \mathbb{R}^{2n+1+p}$, by setting the components not contained in S to zero. In the following we will make use of that embedding without explicitly mentioning it if there is no chance of confusion. Also, pay close attention to the distinction between S_+ and S in above display. Clearly, $G(0) = 0$ and G is minimized at $\bar{u}^{\dagger} = \bar{\theta}^{\dagger}_{S_+} - \bar{\theta}_{0,S_+}$, which implies that $G(\bar{u}^{\dagger}) \leq 0$. Also, G is convex.

Now suppose we manage to find some $B \in \mathbb{R}, B > 0$, such that for all $u \in \mathbb{R}^{s+1+p}$ with $\|u\|_1 = B$ we have $G(u) > 0$. We claim that in that case it must hold $\|\bar{u}^{\dagger}\|_1 \leq B$. Indeed, if $\|\bar{u}^{\dagger}\|_1 > B$, then there exists a $t \in (0,1)$ such that for $\tilde{u} = t\bar{u}^{\dagger}$ we have $\|\tilde{u}\|_1 = B$. But then, by convexity of G, $G(\tilde{u}) \leq tG(\bar{u}^{\dagger}) + (1-t)G(0) = tG(\bar{u}^{\dagger}) \leq 0$. A contradiction.

Thus, we need to find an appropriate B. Let $B > 0$, the correct form to be determined later. Now, pick any $u \in \mathbb{R}^{s+1+p}$ with $\|u\|_1 = B$. We do a first-order Taylor expansion of $\bar{\mathcal{L}}$, with respect to the components in S_+, in the point $\bar{\theta}_{0,S_+}$, evaluated at $\bar{\theta}_{0,S_+} + u$. This yields

$$G(u) = \frac{1}{N}\left\{\nabla_{S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+})^T(\bar{\theta}_{0,S_+} + u - \bar{\theta}_{0,S_+}) + \frac{1}{2}\cdot u^T H_{S_+,S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+} + u\alpha)u\right\} + \bar{\lambda}\left(\|\bar{\theta}_{0,S} + u_S\|_1 - \|\bar{\theta}_{0,S}\|_1\right),$$

for some $\alpha \in [0,1]$. Now, using (4.19), we know that with high-probability,

$$\left|\frac{1}{N}\nabla_{S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+})^T u\right| \leq \left\|\frac{1}{N}\nabla_{S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+})\right\|_\infty \|u\|_1 \leq \epsilon\bar{\lambda}B \tag{4.20}$$

with the ϵ from (4.19). Furthermore, by using the triangle inequality, we obtain

$$\bar{\lambda}\left(\|\bar{\theta}_{0,S} + u_S\|_1 - \|\bar{\theta}_{0,S}\|_1\right) \geq -\bar{\lambda}\|u\|_1 = -\bar{\lambda}B \tag{4.21}$$

Clearly, the canonical embedding of u into $\mathbb{R}^{2n+1+p}$ fulfils the condition of the empirical compatibility condition, Proposition A.2. Also, keep in mind that Assumptions B1 and 4.3 together imply $n^{-1/2}\rho_n^{-1}s_+ \to 0$, which in particular implies $s_+ = o(\sqrt{n})$. Thus, Proposition A.2 is applicable and with high probability as prescribed in Proposition A.2, we have

$$\begin{aligned}
\frac{1}{2}\cdot\frac{1}{N}\cdot u^T H_{S_+,S_+}\bar{\mathcal{L}}(\bar{\theta}_{0,S_+} + u\alpha)u &\geq \frac{1}{4}\rho_n u^T\left\{\frac{1}{N}\bar{D}^T\bar{D}\right\}_{S_+,S_+} u \\
&= \frac{1}{4}\rho_n u^T\hat{\Sigma}u \\
&\geq \frac{1}{8}\rho_n\frac{c_{\min}}{s_+}\|u\|_1^2 \\
&= \frac{1}{8}\rho_n\frac{c_{\min}}{s_+}B^2
\end{aligned} \tag{4.22}$$

Combining (4.20), (4.21), (4.22), we find

$$G(u) \geq -\epsilon\bar{\lambda}B - \bar{\lambda}B + \frac{1}{8}\rho_n\frac{c_{\min}}{s_+}B^2.$$

The right-hand side of this equation is strictly larger zero, whenever

$$B > (1+\epsilon)\frac{8}{c_{\min}}\rho_n^{-1}s_+\bar{\lambda}.$$

Thus, the claim follows from picking

$$B = (1+\epsilon)\frac{9}{c_{\min}}\rho_n^{-1}s_+\bar{\lambda}.$$

□

Lemma 4.9. *Under Assumptions 4.1, 4.2, B1, 4.3, for n large enough, for any $\epsilon > 0$, with probability at least*

$$\begin{aligned}
&1 - 4\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n)\right) - 2(p+1)\cdot\exp\left(-\frac{N\bar{\lambda}^2}{2(1\vee c^2)}\epsilon^2\right) \\
&\quad - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_+^2\tilde{c}}\right),
\end{aligned}$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n) \to -\infty$*, as* $n \to \infty$*, we have*

$$\frac{1}{\bar{\lambda}}\|\bar{R}\|_\infty \leq \frac{324(1 \vee (c^2 p))(1+\epsilon)^2}{c_{\min}^2} \cdot \sqrt{n}\rho_n^{-2} s_+^2 \bar{\lambda}.$$

Proof. Consider the ith component of $\bar{R}$, for $i \in S_\alpha$. Similar to (2.23), using the Cauchy-Schwarz Inequality, we obtain,

$$\begin{aligned}
\bar{R}_i &= \frac{1}{N}\sum_{j=1, j\neq i}^{n} \bar{X}_{ij}|\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2 \\
&= \frac{1}{\sqrt{n}} \cdot \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} |\bar{D}_{ij}^T(\bar{\theta}^\dagger - \bar{\theta}_0)|^2 \\
&= \frac{1}{\sqrt{n}} \cdot \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \left(\alpha_i^\dagger - \alpha_{0,i} + \beta_j^\dagger - \beta_{0,j}^\dagger + \mu^\dagger - \mu_0 + Z_{ij}^T(\gamma^\dagger - \gamma_0)\right)^2 \\
&\leq \frac{4}{\sqrt{n}} \cdot \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \left((\alpha_i^\dagger - \alpha_{0,i})^2 + (\beta_j^\dagger - \beta_{0,j})^2 + (\mu^\dagger - \mu_0)^2 + c^2 p\|\gamma^\dagger - \gamma_0\|_2^2\right) \\
&= \frac{4}{\sqrt{n}}\left\{(\alpha_i^\dagger - \alpha_{0,i})^2 + (\mu^\dagger - \mu_0)^2 + c^2 p\|\gamma^\dagger - \gamma_0\|_2^2\right\} + \frac{4}{\sqrt{n}} \cdot \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} (\beta_j^\dagger - \beta_{0,j})^2 \\
&\leq \frac{4}{\sqrt{n}}(1 \vee (c^2 p))\left\{(\alpha_i^\dagger - \alpha_{0,i})^2 + (\mu^\dagger - \mu_0)^2 + \|\gamma^\dagger - \gamma_0\|_2^2\right\} + \frac{4\sqrt{n}}{n-1}\|\bar{\beta}^\dagger - \bar{\beta}_0\|_2^2.
\end{aligned}$$

We have

$$(\alpha_i^\dagger - \alpha_{0,i})^2 = n(\bar{\alpha}_i^\dagger - \bar{\alpha}_{0,i})^2 \leq n\|\bar{\alpha}^\dagger - \bar{\alpha}_0\|_2^2.$$

Thus, by Lemma 4.8, with at least the prescribed probability and for all $i \in S_\alpha$,

$$\begin{aligned}
\frac{R_i}{\bar{\lambda}} &\leq \frac{4(1 \vee (c^2 p))}{\bar{\lambda}}\sqrt{n}\|\bar{\theta}^\dagger - \bar{\theta}_0\|_2^2 \leq \frac{4(1 \vee (c^2 p))}{\bar{\lambda}}\sqrt{n}\|\bar{\theta}^\dagger - \bar{\theta}_0\|_1^2 \\
&\leq \frac{324(1 \vee (c^2 p))(1+\epsilon)^2}{c_{\min}^2} \cdot \sqrt{n}\rho_n^{-2} s_+^2 \bar{\lambda}.
\end{aligned}$$

The same bound is found for all $i \in S_\beta$ using the exact same steps. Since the right-hand side above does not depend on i the claim follows. □

4.4.5 Controlling term (III)

Lemma 4.10. *Under Assumptions 4.1, 4.2, B1, 4.3 for* n *large enough, for any* $\epsilon > 0$*, with probability at least*

$$\begin{aligned}
&1 - 4 \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n)\right) - 2(p+1) \cdot \exp\left(-\frac{N\bar{\lambda}^2}{2(1 \vee c^2)}\epsilon^2\right) \\
&\quad - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_+^2 \tilde{c}}\right),
\end{aligned}$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{2}\epsilon^2 + \log(n) \to -\infty$, *as* $n \to \infty$, *we have*

$$\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0)\right\|_\infty \leq \frac{9(1\vee c)(1+\epsilon)(p+1)}{4c_{\min}} \cdot \frac{1}{\sqrt{n}}\rho_n^{-1} s_+.$$

Proof. We have

$$\begin{aligned}\left\|\frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0)\right\|_\infty &\leq \left\|\frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right]\right\|_\infty \|(\xi^\dagger - \xi_0)\|_\infty \\ &\leq \left\|\frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right]\right\|_\infty \|\bar{\theta}^\dagger - \bar{\theta}_0\|_1.\end{aligned}$$

Consider the ith row of the matrix $\frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right]$,

$$\left\|\left(\frac{1}{N}(\bar{X}_{-,i})^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right]\right)^T\right\|_1 \leq \frac{1}{N}\sqrt{n}(n-1)\cdot\frac{1}{4}(1\vee c)(p+1) = \frac{p+1}{4}(1\vee c)\frac{1}{\sqrt{n}},$$

where we have used that the ith column of $\bar{X}$ has exactly $(n-1)$ non-zero entries, each with value $\sqrt{n}$, each entry of W_0^2 is upper bounded by $1/4$ and any row of $\left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right]$ has $p+1$ entries, each upper bounded by $1\vee c$. Thus, by Lemma 4.8, with the prescribed probability,

$$\frac{1}{\bar{\lambda}}\left\|\frac{1}{N}\bar{X}^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array}\right](\xi^\dagger - \xi_0)\right\|_\infty \leq \frac{9(1\vee c)(1+\epsilon)(p+1)}{4c_{\min}} \cdot \frac{1}{\sqrt{n}}\rho_n^{-1} s_+.$$

□

4.4.6 Condition (4.11b)

Lemma 4.11. *Under Assumptions 4.1, 4.2, B1, 4.3, for n large enough, with probability at least*

$$1 - 4\exp\left(-\frac{N\bar{\lambda}^2}{18} + \log(n)\right) - 2(p+1)\exp\left(-\frac{N\bar{\lambda}^2}{18(1\vee c^2)}\right) - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_+^2 \tilde{c}}\right),$$

which tends to one as long as $-\frac{N\bar{\lambda}^2}{18} + \log(n) \to -\infty$, *as* $n \to \infty$, *we have*

$$\|\bar{z}^\dagger_{1:2n,S^c}\|_\infty < 1.$$

Proof. By equation (4.17), Lemmas 4.7, 4.9, 4.10, with the probability given in those

Lemmas, for any $\epsilon > 0$,

$$\begin{aligned}\|\bar{z}^{\dagger}_{1:2n,S^c}\|_\infty \leq & \left\{\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty + 1\right\}\epsilon \\ & + \left\{\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty + 1\right\}\frac{324(1\vee(c^2p))(1+\epsilon)^2}{c^2_{\min}}\cdot\sqrt{n}\rho_n^{-2}s_+^2\bar{\lambda} \\ & + \left\{\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty + 1\right\}\frac{9(1\vee c)(1+\epsilon)(p+1)}{4c_{\min}}\cdot\frac{1}{\sqrt{n}}\rho_n^{-1}s_+ \\ & + \frac{1}{2}\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty.\end{aligned}$$

By Lemma 4.7, for n sufficiently large, we have $\|Q_{S^c,S}Q^{-1}_{S,S}\|_\infty < 1/2$. Thus, by equation (4.17), Lemmas 4.9 and 4.10, for n sufficiently large, with the prescribed probability,

$$\begin{aligned}\|\bar{z}^{\dagger}_{1:2n,S^c}\|_\infty \leq & \frac{3}{2}\epsilon + \frac{1}{4} \\ & + \frac{486(1\vee(c^2p))(1+\epsilon)^2}{c^2_{\min}}\cdot\sqrt{n}\rho_n^{-2}s_+^2\bar{\lambda} \\ & + \frac{27(1\vee c)(1+\epsilon)(p+1)}{8c_{\min}}\cdot\frac{1}{\sqrt{n}}\rho_n^{-1}s_+.\end{aligned}$$

Pick $\epsilon = 1/3$, to obtain

$$\|\bar{z}^{\dagger}_{1:2n,S^c}\|_\infty \leq \frac{3}{4} + \frac{864(1\vee(c^2p))}{c^2_{\min}}\cdot\sqrt{n}\rho_n^{-2}s_+^2\bar{\lambda} + \frac{9(1\vee c)(p+1)}{2c_{\min}}\cdot\frac{1}{\sqrt{n}}\rho_n^{-1}s_+.$$

The second and third term go to zero, as $n \to \infty$, by Assumption B1. Indeed, the second term is Assumption B1 exactly. For the third term, by Assumption B1, $\sqrt{n}s_+^2\bar{\lambda}\rho_n^{-2} = n^{-1/2}\rho_n^{-1}s_+ \cdot n\rho_n s_+\bar{\lambda} \to 0$ as, $n \to \infty$. On the other hand, by Assumption 4.3, $n\rho_n s_+\bar{\lambda} \geq C\rho_n^{-1}s_+\log(n) \to \infty$. Therefore it must hold true that $n^{-1/2}\rho_n^{-1}s_+ \to 0$. The claim follows. □

4.4.7 Proof of Theorem 4.1

Proof of Theorem 4.1. By Lemma 4.11, we know that with probability at least as large as

$$1 - 4\exp\left(-\frac{N\bar{\lambda}^2}{18} + \log(n)\right) - 2(p+1)\exp\left(-\frac{N\bar{\lambda}^2}{18(1\vee c^2)}\right) - p(p+3)\exp\left(-N\frac{c^2_{\min}}{2048s_+^2\tilde{c}}\right),$$

property (4.11b) holds for the construction $(\bar{\theta}^\dagger, \bar{z}^\dagger)$. Thus, by construction and (4.11b), $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ fulfils (4.10b). Furthermore, $(\bar{\theta}^\dagger, \bar{z}^\dagger)$ fulfils (4.9), (4.10a) and (4.10c) by construction. Thus, by Lemma 4.2, $\hat{S} = S^\dagger$ and in particular $\hat{S} \cap S^c = \emptyset$.

Recall that by equation (4.16a),

$$\bar{\vartheta}^\dagger_S = \bar{\vartheta}_{0,S} + Q^{-1}_{S,S}\left\{-\frac{1}{N}\left(\nabla_{\bar{\vartheta}}\bar{\mathcal{L}}(\bar{\theta}_0)\right)_S + \frac{1}{N}\bar{X}_S^T W_0^2\left[\begin{array}{c|c}\mathbf{1} & Z\end{array}\right](\xi^\dagger - \xi_0) - \bar{\lambda}\bar{z}^\dagger_{1:2n,S} + \bar{R}_S\right\}. \tag{4.23}$$

Thus, $S^\dagger$ contains all those indices i with

$$\left\| Q_{S,S}^{-1} \left\{ -\frac{1}{N} \left(\nabla_{\bar{\vartheta}} \bar{\mathcal{L}}(\bar{\theta}_0) \right)_S + \frac{1}{N} \bar{X}_S^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array} \right] (\xi^\dagger - \xi_0) - \bar{\lambda} \bar{z}_{1:2n,S}^\dagger + \bar{R}_S \right\} \right\|_\infty < \bar{\vartheta}_{0,i}.$$

Consider

$$\begin{aligned}
&\left\| Q_{S,S}^{-1} \left\{ -\frac{1}{N} \left(\nabla_{\bar{\vartheta}} \bar{\mathcal{L}}(\bar{\theta}_0) \right)_S + \frac{1}{N} \bar{X}_S^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array} \right] (\xi^\dagger - \xi_0) - \bar{\lambda} \bar{z}_{1:2n,S}^\dagger + \bar{R}_S \right\} \right\|_\infty \\
&\quad \leq \|Q_{S,S}^{-1}\|_\infty \left\| -\frac{1}{N} \left(\nabla_{\bar{\vartheta}} \bar{\mathcal{L}}(\bar{\theta}_0) \right)_S + \frac{1}{N} \bar{X}_S^T W_0^2 \left[\begin{array}{c|c} \mathbf{1} & Z \end{array} \right] (\xi^\dagger - \xi_0) - \bar{\lambda} \bar{z}_{1:2n,S}^\dagger + \bar{R}_S \right\|_\infty \\
&\quad \leq 2\rho_n^{-1} \cdot \frac{n-1}{n - \max\{s_\alpha, s_\beta\}} \\
&\qquad \left\{ \epsilon\bar{\lambda} + \bar{\lambda} \right. \\
&\qquad + \frac{9(1 \vee c)(1+\epsilon)(p+1)}{4c_{\min}} \cdot \frac{1}{\sqrt{n}} \rho_n^{-1} s_+ \bar{\lambda} \\
&\qquad \left. + \frac{324(1 \vee (c^2 p))(1+\epsilon)^2}{c_{\min}^2} \cdot \sqrt{n} \rho_n^{-2} s_+^2 \bar{\lambda}^2 \right\}
\end{aligned}$$

where we used (4.14), (4.19) and Lemmas 4.9 and 4.10.

By assumption, $\bar{\lambda} \leq C \cdot \sqrt{\log(n)/N}$ for some $C > 0$, thus the first two terms in the bracket may be upper bound by $C \cdot \sqrt{\log(n)/N}$, for a possibly different C. The third term is $o(1) \cdot 1/n$ by Assumption B1 and the last term is $o(1) \cdot \sqrt{\log(n)}/n$ by Assumption B1. Since $(n-1)/(n - \max\{s_\alpha, s_\beta\}) = O(1)$, the entire right-hand side is less or equal

$$C\rho_n^{-1} \frac{\sqrt{\log(n)}}{n}.$$

Multiply (4.23) by $\sqrt{n}$ to transition to the unscaled parameters $\vartheta_S^\dagger$ and the claim follows. In particular, for n large enough, with at least the prescribed probability the construction fulfils (4.11a) and thus $\hat{S} = S^\dagger = S$ by Lemma 4.2. □

Chapter 5

Conclusion

In Chapter 1 we reviewed recent advances in the field of random network models. We identified several features commonly observed in real-world networks which we would like to capture in our model. Most notably *sparsity*, which refers to the phenomenon that the number of observed edges scales sub-quadratically in the number of nodes, *degree heterogeneity*, which means that the degree distribution of the observed network is heavy-tailed and *homophily*, which means that similar nodes are more likely to connect to one another.

We saw that SβM-C and SRGM emerge as a natural extension to a very active branch of research on extensions of the original β-model (Chatterjee et al. 2011). While previous models in this line of research were able to capture degree heterogeneity and covariates (Graham 2017, Yan et al. 2019) or degree heterogeneity and sparsity (Chen et al. 2020), SβM-C and SRGM are the first models that capture all three of these characteristics explicitly in a single model.

The β-model and many of its extensions, such as Yan, Leng & Zhu (2016), Yan et al. (2019), pursue some type of maximum likelihood approach to parameter estimation. This entails that they are over-parametrized and require the observed network to be dense in order to be able to perform consistent parameter estimation. Therefore, a popular exercise in the literature is to focus on a sub-network induced by nodes with better connectivity, leaving out a substantial portion of the nodes. We have highlighted in Section 1.2 the fallacy of the resulting data-selective inference for analysis. Its associated non-random sampling often brings biased estimates as we have demonstrated in two fundamental network models, under the idealistic scenario when the assumed model does produce the realized data. As a result of data-selective inference, it is not clear whether any findings are genuine or artefact of biased sampling – Statisticians are well aware of the pitfalls of what systematic sampling bias can bring to data analysis. Having biased initial parameter estimates calls into question the validity of any downstream statistical inference, including

consistency, model selection, hypothesis testing, and so on, creating a demand for models that can account for sparsity.

SβM-C builds on earlier work by Chen et al. (2020) by including covariates into the likelihood of the sparse β-model. SRGM extends and refines the results from SβM-C to directed networks. The key assumption in these models is that the degree heterogeneity parameter, which assigns individual parameters to each node, is sparse. As a consequence, the number of parameters these models need to estimate is much lower than the number of parameters in the aforementioned models. Thus, they avoid over-parametrization and can be fitted to sparse networks without the need of removing low-degree nodes. They may be a stepping stone towards avoiding data-selective inference in the future.

In Chapter 2 we have shown that SβM-C is well suited to model sparse networks, thanks to the sparsity assumption on the nodal parameter that can effectively reduce the dimensionality of the model. We have presented theory for the penalized likelihood estimator based on an ℓ_1-penalty on the nodal parameter, including ℓ_1-consistency and a central limit theorem for the homophily parameter. The key to our success was proving that once we account for differing sample sizes of the model parameters, the analogue to the compatibility condition – which is crucial for consistency results in classical LASSO theory – holds. Built on LASSO theory, our theoretical contributions go beyond existing theory for LASSO as we have argued.

Additionally to the general theory for SβM-C developed in Section 2.2, we treated two special cases of SβM-C: in Section 2.3 we showed that the consistency results from Section 2.2 can be applied to the model studied in Chen et al. (2020). We compared our findings for SβM when employing an ℓ_1-penalty for regularization with the results obtained by Chen et al. (2020), who used an ℓ_0-penalty. We dedicated Chapter 3 to the second special case in which we set the degree heterogeneity parameter $\beta = 0$. We named this model the sparse Erdős-Rényi model with covariates (ER-C). Since the number of parameters remains fixed in this model, we were able to use a standard MLE approach for parameter estimation and showed that ER-C can model networks of almost arbitrary sparsity.

In Chapter 4 we introduced direction to the edges in SβM-C by studying the parameter-Sparse Random Graph Model (SRGM). Without covariates and without the global sparsity parameter, this model is the p_0-model introduced in Holland & Leinhardt (1981). Once we were able to argue that a compatibility condition also holds in this model, we were able to carry over the consistency and asymptotic normality proofs from SβM-C without much effort. As a novel aspect to the developed theory, we derived conditions under which our ℓ_1-penalized estimator will identify the correct support of the degree heterogeneity parameter ϑ. We gave a nuanced

account of how the network sparsity, the parameter sparsity and the penalty used need to be appropriately balanced for our penalized likelihood estimator for SRGM to simultaneously consistently identify the support of ϑ, consistently estimate the whole parameter θ, and provide valid inference for the covariates parameter γ.

The computation of the estimators in SβM-C and SRGM capitalizes on the tremendous progress made in the algorithmic development of LASSO-type estimators and thus is straightforward to implement and extremely scalable

There are a host of important issues for future research. Firstly, in SRGM, we have assumed that, given the covariates, directed links are formed independently between node-pairs. This may be a limitation because empirically, reciprocity – a measure of the likelihood of vertices in a directed network to be mutually linked – may be present. For example, in the lawyer friendship network we studied in Section 1.2, lawyer j may be more likely to call lawyer i a friend if the converse is true. To address this layer of sophistication, the next natural step is to add a reciprocity parameter to the model. One approach to do this is to add an extra term $\delta \sum_{i<j} A_{ij}A_{ji}$ in the model, where $\delta \in \mathbb{R}$ is an unknown parameter, as adopted in the p_1 model (Holland & Leinhardt 1981). For the models in Yan, Leng & Zhu (2016) and Yan et al. (2019), including this extra reciprocity parameter turns out very challenging for their theoretical framework, because their analytical tool needed to approximate the inverse of a Fisher information matrix accurately is no longer applicable. On the other hand, by assuming parameter sparsity in α and β, SRGM will deal with a much smaller number of parameters and thus some of the theoretical arguments presented in this thesis may go through. Secondly, we assume that the dimension of the covariates is fixed but it need not be the case in practice. Recent data deluge brings more and more data sets that have more variables than observations. How to generalize SβM-C and SRGM to include growing dimensional covariates is worth further investigation. Thirdly, in some applications inference on α and β might be of interest. Since $\hat{\alpha}$ and $\hat{\beta}$ are biased due to the shrinkage incurred by our ℓ_1-penalty, this will require a debiasing step. We believe that it is possible to derive inference results with suitable balancing assumptions, in a manner similar to what was done in van de Geer et al. (2014). Finally, it will be interesting to incorporate a low rank component in SβM-C in order to capture transitivity, the phenomenon that nodes with common neighbours are more likely to connect, as is done in Ma, Ma & Yuan (2020).

Part II

A guided analytics tool for feature selection in steel manufacturing

Chapter 6

Data science in industry

Organization of this chapter

We present the 'initial Guided Analytics for parameter Testing and control-band Extraction (iGATE)' framework. iGATE is designed to provide a standardized, expert knowledge driven approach to feature selection. It is a middle ground between autonomous and manual feature selection.

In Section 6.1 we introduce the general concept of *guided analytics* and provide a high-level overview of the iGATE methodology. This is followed by an in depth treatment of iGATE for the case of continuous response variables in Section 6.2. In Section 6.3 we show how the same methodology can be extended to categorical response variables. We follow up with an application of iGATE to top-gas efficiency in Section 6.4. Concluding remarks are presented in Section 6.5. The content of this chapter is taken from Stein et al. (2021).

6.1 Guided analytics

Over the past decade or so, so-called data science has become an increasingly important topic in all aspects of business and industry. This reflects the increasing availability and power of computing resource and associated big data technologies over the same period. Data from manufacturing and business processes is being increasingly recognized as holding enormous business development potential. The vehicle for realization of this potential is systematic data analysis and this has evolved from the traditional niche domain of the statistician to an organized *Advanced Analytics* business function commanding an increasingly prominent position on the senior management agenda (Jensen 2020). Advanced Analytics present many opportunities, including optimization of manufacturing, maximization of equipment effectiveness and enhanced logistics for customer service.

In steel manufacturing in particular, even small improvements to stability, yield or quality make big differences to costs and profitability, making application of Advanced Analytics a lucrative endeavour. In this chapter we present a novel, expert knowledge driven approach to feature selection in industrial applications. This ap-

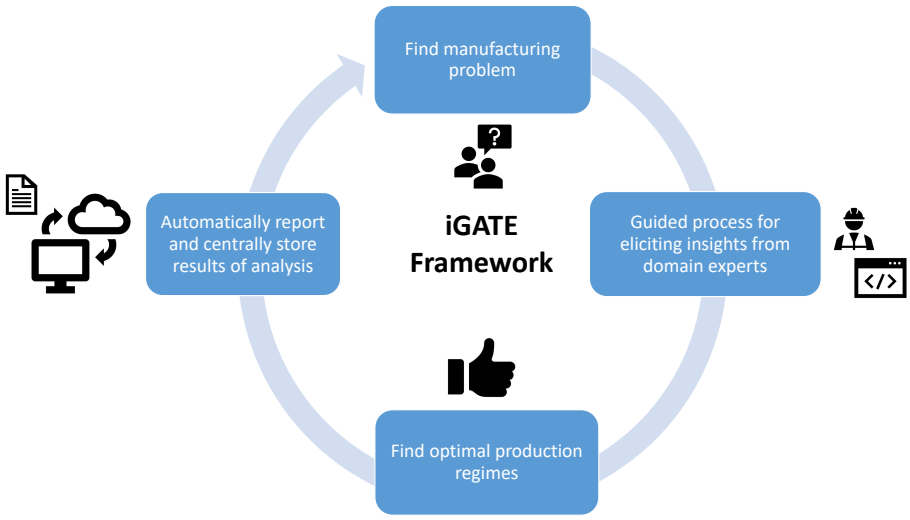


Figure 6.1: Overview of the iGATE framework: Once a manufacturing problem has been found, iGATE will automatically find potentially influential process parameters using hypothesis tests. These are then systematically reviewed by a domain expert for their suitability for the problem under study. The expert is guided through this process by the application and encouraged to comment on their decision to include or discard parameters. For the retained parameters, iGATE finds optimal production regimes and validates these findings. At the end, a report of the analysis is automatically produced and centrally stored. This helps long-term knowledge retention within a company and can inform future data science projects.

proach is called the 'initial Guided Analytics for parameter Testing and controlband Extraction (iGATE)' framework. It has been published in Stein et al. (2021) under the *Creative Commons Attribution 4.0 International License* (2013) and this chapter is based on said paper. The technical component of iGATE has been made publicly available in form of the `igate` package (Stein 2019) for the R programming language and is available on the Comprehensive R Archive Network (CRAN)[1].

iGATE is a *guided analytics* tool. The premise of guided analytics is that the successful application of data science requires the combination of statistical or machine learning techniques with domain knowledge of the physical process being modelled (Anderson-Cook et al. 2019). Failure to include expert knowledge from the outset may result in models that conflict with expert knowledge and ultimately physical reality (Liu et al. 2020). A guided analytics application is meant to facilitate the knowledge transfer between domain experts and statisticians by guiding them through an analysis, prompting them to interactively use their respective expertise to influence its outcome. The long-term value to the company lies in domain experts and statisticians commenting on their decisions, enabling the company to capture and centrally store their expert judgement in a structured way. While each individual method used in iGATE already existed beforehand, Stein et al. (2021) combined them in a novel way to create the iGATE framework. The iGATE framework is schematically visualized in Figure 6.1.

[1] `https://cran.r-project.org`

Steel manufacturing, like many other process industries, can be considered particularly knowledge intensive. That is, a high degree of human judgement is involved in decision making processes, making it ideal for the application of guided analytics. This is due in part to typically high degrees of legacy in information systems, but also because of practical limitations to implementation of robust sensor technologies. At Tata Steel in Europe, iGATE has been made available as a web application providing on-demand analytics functionality and streamlining overall analytics usage. The tool has proven especially useful as an early step in data science projects as it allows for initial dimensionality reduction and elicitation of expert knowledge.

In Section 6.4 we present an application of iGATE to blast furnace data from Tata Steel. Blast furnace steelmaking accounted for roughly 70% of steel produced worldwide in 2015 (Geerdes et al. 2015) and therefore, improving efficiency of blast furnaces is a lucrative area of application for Advanced Analytics technologies. The blast furnace is particularly interesting for guided analytics, as accurate measurement of parameters is a major bottleneck (Agarwal et al. 2010) and thus data tends to be inherently messy and values need to be placed into context by domain experts. For a long time, the blast furnace has been considered a "black-box" (Omori 1987).

6.2 The iGATE methodology

On a high level, the assumptions on the data for iGATE resemble a standard regression setup. We assume that the quality of a manufacturing output can be measured by some univariate response variable y (the *target*; either continuous or categorical). Alongside y, we observe an array of covariates X that may or may not have an influence on y. We call each covariate a *suspected source of variation* (SSV) and wish to determine which SSVs significantly influence the product quality y. Different from a regression problem, iGATE is not concerned with modelling the actual relationship $y \approx f(X)$, for some suitable function f. The emphasis is rather on fostering discussion between domain experts and statisticians so that they may come up with a (small) subset of potentially influential variables that is worth further analysis.

While there are plenty of statistical techniques available for quantifying statistical significance in regression models, in real-world industrial contexts there may be factors other than statistical accuracy at play, which we have to consider when deciding which techniques to employ. In the following we highlight some of these factors.

Firstly, there is the aforementioned need for domain expertise. Inclusion of expert feedback is especially important for identifying *target leakage* in the feature selection step of an analysis. Target leakage refers to using illegitimate variables to

predict the target variable (Kaufman et al. 2011). In the context of manufacturing, it usually occurs, when using variables as predictors that may be highly correlated with the target, but that cannot be physically controlled during the manufacturing process. Any sensible tool would identify such variables as a strong predictors for the target. In terms of actionable business insights, however, these findings would be useless. Target leakage is considered one of the most insidious problems of automated machine learning (Larsen & Becker 2018).[2]

However, eliciting expert knowledge usually requires much iteration between domain experts and statisticians and is an inefficient path to follow. In industrial contexts, such as modern factories, often hundreds or even thousands of process parameters are captured automatically by sensors and asking a domain expert to review them all for their suitability for an analytics project is unrealistic.

Property 1. *iGATE systematically compares good and bad products by applying statistical hypothesis tests to find a* small *subset of potentially influential variables for review by the domain expert. They may then decide which SSVs to keep for further study.*

Secondly, we have to recognize that any machine learning project will only be of value to a company, if the knowledge gained from it can be transformed into actionable business insights. In the context of manufacturing, this means that the insights gained must be moved to the shop floor – where the actual wealth is created – by educating the workforce about the findings and by providing them with actionable knowledge (Brimacombe 1999). Furthermore, the findings need to be understandable to domain experts and decision-makers alike who may not have had prior statistical training. Especially "black-box" models, that give predictions without explanatory context, rarely enjoy the confidence of decision-makers.

Property 2. *iGATE uses intuitive concepts easily grasped by personnel who possibly have had no prior statistical training, combating concerns about so-called "black-box" models. iGATE also provides an initial estimation of favourable controlbands that – under regular manufacturing conditions – will result in good product quality. These controlbands, once validated, can immediately be translated into actionable instructions for process operators.*

Thirdly, eliciting expert knowledge on an ad-hoc basis for an analysis at hand tends to be suboptimal with regard to long-term organizational knowledge capture. Unless results are stored in such a way that they are easily accessible and interpretable by other data science teams in the future, any knowledge gained might be

[2]This illustrates that while autonomous feature engineering and by extension autonomous machine learning might be the ultimate goal in manufacturing, they are often not yet feasible and can only work for very specific and well defined tasks.

lost when the employees involved in an analysis leave or move to a different position within the company. It is also imperative to accurately document the assumptions underlying the analysis as assumptions may become outdated and invalid over time.

Property 3. *iGATE captures and centrally stores comments made by domain experts for future reference in the form of standardized reports, aiding long-term knowledge capture.*

Finally, depending on the industrial context, procuring samples may be very expensive and data on them may be messy; either because it is missing altogether or because it has been recorded incorrectly. Especially in the latter case the insights from a domain expert can be of great help.

Property 4. *iGATE uses techniques that have reasonable statistical power for small sample sizes. It uses non-parametric hypothesis tests to avoid making distributional assumptions. It has been robustified against messy data in the sense made precise in the next section.*

The current implementation of iGATE can be considered a skeleton pipeline for analytics projects to which new steps and methods can be added as user confidence in the use of guided analytics tools increases. After running iGATE it is possible to employ more powerful, but possibly less transparent machine learning algorithms to find correlations in the features selected by iGATE.

In summary, iGATE's main features are:

1. It is a fast framework for initial feature selection and expert knowledge elicitation that is applicable beyond the steel manufacturing application presented here.
2. It works with messy data with potentially many missing or misrecorded values.
3. Results are easy to interpret and explain.
4. It contains a standardized way of documenting the analysis, its underlying assumptions and results, aiding knowledge capture.

6.2.1 iGATE overview

The main idea of iGATE is to compare the best products with the worst products and determine those production parameters in which they differ significantly, which are then concluded to be potentially influential for the product quality. This allows us to automatically exclude many parameters that are irrelevant to the problem under investigation. iGATE iteratively applies the Tukey-Duckworth test as proposed in Tukey (1959), which performs well even for small sample sizes. We explain how it works below. This statistical hypothesis test was chosen for its ease of interpretation, making it possible to effectively explain any findings to people without

statistical training. As a non-parametric hypothesis test it is also robust against different underlying distributions.

Having identified a manufacturing problem to be investigated, a data set is assembled for a typical period of operation, i.e. excluding known disturbances such as maintenance or equipment failures. This data set includes the *target variable* as well as a number of features we consider potentially influential for the value of the target (*suspected sources of variation*; SSVs). We explain the general concept for continuous targets in this section and show how it can be generalized to categorical targets in the next section. The iGATE procedure consists of the following steps (detailed explanations follow below):

1. Select the eight best and eight worst products.
2. Perform the Tukey-Duckworth test for each SSV.
3. Optional: For each SSV perform unpaired Wilcoxon rank sum test.
4. Extract upper/ lower control bands for kept parameters.
5. Perform sanity check via regression plot; based on whether a trend is discernible and expert judgement, decide which SSV to keep.
6. Validate choice of SSVs and control bands.
7. Report findings in standardized format.

6.2.2 Products selection

When running iGATE with default settings, a box-plot approach is used for outlier detection and all observations with a target value that lies beyond 1.5 times the interquartile range of the 25th and the 75th quantile are removed before the analysis. This is justified, because we want to understand the behaviour of the target under regular production conditions. If one is interested in the insights outliers provide into the behaviour of the target, outlier removal can be switched off.

From the remaining dataset we select the eight products that produced the best quality in terms of our target variable ("Best of the Best", BOB) and the eight products that produced the worst quality ("Worst of the Worst", WOW). If many samples are readily available, the number of BOB/ WOW can be specified manually by the user. To avoid selecting observations with missing values, rather than selecting the same eight BOB and WOW for all SSV, we select them dynamically: For the SSV we are currently investigating, we first remove those observations that contain missing values for that SSV and then select our BOB and WOW from the remaining data. We conduct the analysis with the 16 selected observations and determine whether or not the current SSV is potentially influential for the target variable. For the next SSV we repeat the process, starting again with the full data set. The

user may choose to perform outlier removal for each SSV before selecting the BOB and WOW. If ties occur when selecting BOB and WOW, we select from the tied observations at random.

6.2.3 The Tukey-Duckworth hypothesis test

The Tukey-Duckworth test used in Step 2 is a distribution free hypothesis test pioneered in Tukey (1959). Its null hypothesis is that the two samples come from the same distribution and it works as follows. After selecting the BOB and WOW, we are left with 16 observations of the SSV under consideration. Denote this vector as $X = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_8$ being the values of the SSV of the BOB and $X_9, \ldots, X_{16}$ the values of the WOW respectively. Define the vector of labels $v = (v_1, \ldots, v_n)$, where $v_i = \text{BOB}$ if X_i is a value corresponding to a BOB and $v_i = \text{WOW}$ otherwise. Consider the order statistics $X_{(i)}$, where $X_{(i)}$ the i-th smallest entry of X. The *rank of* X_i is

$$R_i = \sum_{j=1}^{n} \mathbb{1}(X_j \leq X_i), \tag{6.1}$$

where $\mathbb{1}(X_j \leq X_i)$ denotes the indicator function for the event $\{X_j \leq X_i\}$. That is, R_i is the position of X_i in the ordered vector $\bar{X} = (X_{(1)}, \ldots, X_{(n)})$. Consider the label vector ordered according to the ranks R_i,

$$\bar{v} = (v_{(1)}, \ldots, v_{(n)}),$$

where $v_{(i)}$ is the label of $X_{(i)}$. The count summary statistic s is defined as

$$s = \begin{cases} 0, & \text{if } v_{(1)} = v_{(n)}, \\ s_l + s_u, & \text{otherwise,} \end{cases}$$

where s_l and s_u the *lower* and *upper counts* given by

$$s_l = \sum_{j=1}^{n} \mathbb{1}(v_{(1)} = \cdots = v_{(j)}), \quad s_u = \sum_{j=1}^{n} \mathbb{1}(v_{(n)} = \cdots = v_{v_{(n-(j-1))}}),$$

i.e. s_l counts how many of the entries of $\bar{X}$ at the *lower end* have the same label as $X_{(1)}$ and s_u counts how many entries of $\bar{X}$ at the *upper end* have the same label as $X_{(n)}$. If ties occur we take the average over all the possible values of s for each of the ties.

If the distribution of the BOB and WOW differ significantly in the current SSV, the BOB will cluster on one end of $\bar{X}$ and the WOW on the other and s will be large. It will be small otherwise. See Figure 6.2 for an example. If $s \geq 6$, we keep the SSV

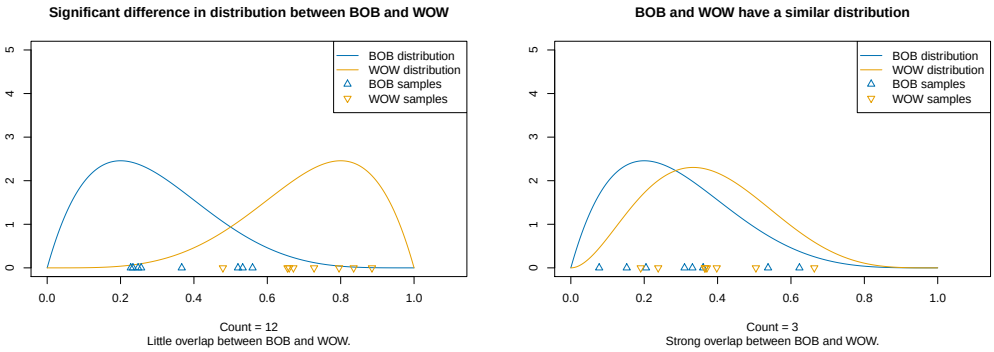


Figure 6.2: Count summary statistic example. If the distribution of a specific SSV differs significantly between good and bad products, good products will cluster at one end of the sorted vector $\bar{X}$ and bad products at the other (left plot; BOB are sampled from a Beta(2,5) distribution, WOW from a Beta(5,2) distribution). If the distribution of that SSV is similar for good and bad products, no such overlap will be observable (right plot; BOB sampled from a Beta(2,5) distribution, WOW from a Beta(3,5) distribution).

as potentially influential. If $s < 6$, we discard it. The critical value 6 corresponds to a p-value of roughly 0.05 and is independent of the sample size 16 (Tukey 1959). The Tukey-Duckworth test is used as a preliminary step to greatly reduce the number of parameters under consideration and was chosen for its easy interpretability.

6.2.4 The Wilcoxon rank sum test

Optionally, we can choose to perform the two-sample Wilcoxon rank sum test described in Wilcoxon (1945) instead of the Tukey-Duckworth test. It serves as a possibly more widely known alternative to the Tukey-Duckworth test, that might, however, be harder to explain to non-statisticians. Given the rank vector $R = (R_1, \ldots, R_n)$ from equation (6.1), we calculate the summary statistics

$$W_{\text{BOB}} = \sum_{i=1}^{8} R_i, \quad W_{\text{WOW}} = \sum_{i=9}^{16} R_i.$$

That is, we sum up all the ranks of the BOB and all the ranks of the WOW. Simple algebra shows that $W_{\text{BOB}} + W_{\text{WOW}} = \sum_{i=1}^{16} i = 136$ and the values of W_{BOB} and W_{WOW} must lie between $\sum_{i=1}^{8} i = 36$ and $\sum_{i=9}^{16} = 100$. Under the null-hypothesis that the samples from the BOB and the WOW come from the same distribution, W_{BOB} and W_{WOW} will take similar values. If they come from significantly different distributions, they will produce significantly different values and we keep the SSV under study as potentially influential. This hypothesis test is frequently also referred to as the *Mann-Whitney U Test*, which uses slightly different – but equivalent – test statistics and was introduced independently from Wilcoxon (1945) in Mann & Whitney (1947).

This is a multiple testing problem and we therefore adjust the p-values using the Bonferroni-Holm procedure presented in (Holm 1979). The main function of these

steps is to facilitate dimensionality reduction in the data to generate a manageable population for expert consideration. These two tests are preferred over the t-Test, because they are distribution free and, while the t-Test may be optimal for normally distributed data, for non-normal data it can get arbitrarily weak.

6.2.5 Controlband extraction

For those SSVs retained after conducting the above hypothesis tests, control bands are extracted in Step 4 as follows. For each SSV retained, we have $s > 0$. If $s_l = k > 0$, then $v_{(1)} = \cdots = v_{(k)}$, but $v_{(k)} \neq v_{(k+1)}$. The control band for the group with label $v_{(1)}$ is then given as $I_l = [X_{(1)}, X_{(k)}]$ and we conjecture, that if the SSV under consideration is kept within I_l during production, we are more likely to obtain a good target value if $v_{(1)} = \text{BOB}$, and a bad target value if $v_{(1)} = \text{WOW}$. Similarly, if $s_u = k > 0$, we define the control band corresponding to the group of $v_{(n)}$ as $I_u = [X_{(n-k+1)}, X_{(n)}]$.[3]

6.2.6 Sanity check via regression plots

As sanity check of the results obtained by the hypothesis test a linear regression plot of for each retained SSV against the target is produced in Step 5. The domain experts can now review these plots, the extracted controlbands, the count summary statistic and the adjusted p-value together with their domain expertise to make a final decision on which parameters to keep and which to discard. If a trend can be seen in the plots and the order of magnitude of the extracted control bands align with their expertise, an SSV will be kept, otherwise discarded. Note that manual inspection of regression plots for all SSVs is often not feasible for processes with hundreds of parameters. In iGATE the user will only have to check regression plots for those SSVs that passed the hypothesis tests. At this point, control bands may also be adjusted manually based on their expert knowledge.

6.2.7 Validation of controlbands

For the validation step, the production period from which the validation data is selected is dependent on the business situation, but should be from a period of operation consistent with that from which the initial population was drawn, i.e. similar product types, similar level of equipment status etc. The validation step extracts from the validation sample all the records for which any of the retained SSVs lies within these bands. We expect that if the SSV lies within the good band, then

[3] By construction $v_{(1)} \neq v_{(n)}$.

the target should also correspond to good performance, and vice versa for bad performance. The application gives feedback on the extent to which this criterion is satisfied, such as how many observations fall within the good/ bad band of each individual SSV. See Section 6.4 for an example.

6.2.8 Automatic report generation

In the last step, a report of the conducted analysis and its findings is automatically created. We provide a visual template of the report generated by iGATE in Figure 6.3. For reasons of space, we limit ourselves to presenting the conceptual idea of what the report looks like. It starts with the *Overview* section containing the metadata of the analysis, such as when it was conducted, which data set was used with which target variable etc. It follows an outline of the analysis, describing the techniques used and showing a box-plot of the target variable (in case of a continuous target). In the *Results* section the retained SSVs together with their count summary statistics, their adjusted p-values and extracted control-bands are shown. This section also contains any comments made by domain experts about the SSVs. This is followed by the *Validation* section. If validation of the results has been conducted, the results of it are presented here. The appendix of the report contains a list of all the SSVs that were analysed as well as all the regression plots. Thus, if at a later stage the results of the analysis are reviewed by a different data scientist, it is clear to them how data decisions were taken in the original analysis.

6.3 Extending iGATE to categorical target variables

Using iGATE with categorical target variables is analogous to the continuous target case. The main difference is that when selecting the eight best and eight worst observations, this selection is unlikely to be unique. In this case, eight observations are selected from the best and from the worst category at random. However, especially in the case of few categories with many observations in each category, this can be problematic. The variance of each SSV within each category may be very large and we might obtain a different result each time we run iGATE. To robustify against this, we implemented a multiple sampling approach. In this ensemble method we run iGATE with categorical target 50 times and only return those SSVs that come up as influential in at least 50% of the runs. This prevents a scenario in which we obtain a different outcome every time the analysis is conducted. The rest of the analysis follows the same steps as in the case with continuous target. The only difference is that in Step 5, we do not produce regression plots. Instead, a normalized frequency

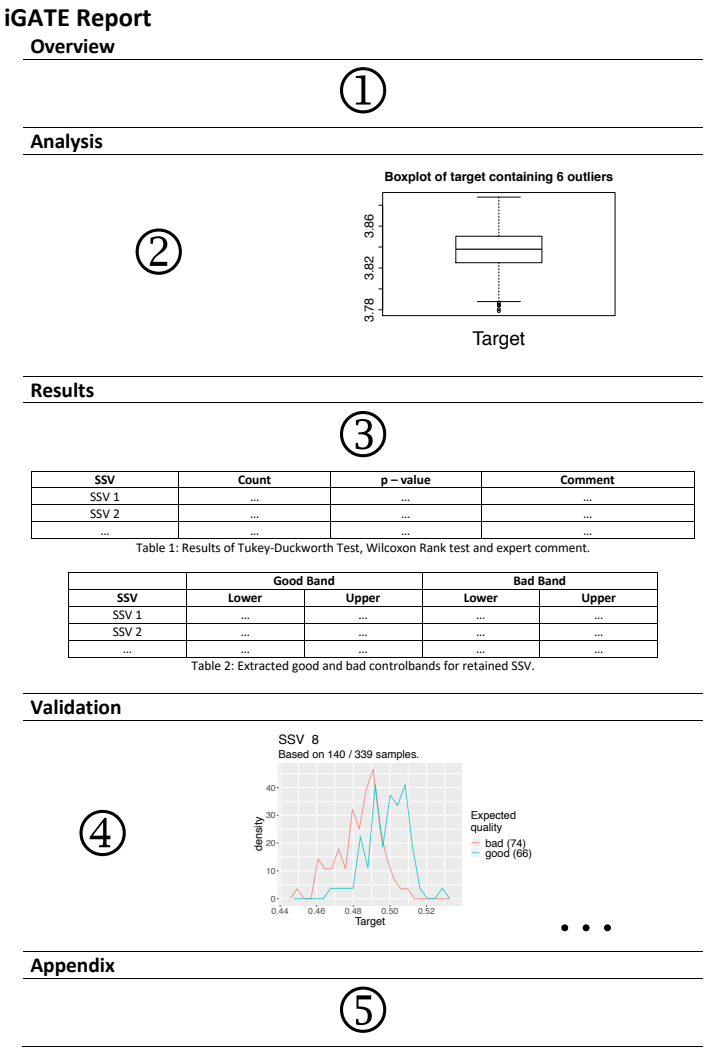


Figure 6.3: A schematic representation of the report automatically generated by iGATE. 1) An overview of the conducted analysis is presented, including the date of the analysis, the name of the data set used and of the target variable. 2) A detailed description of the methods used, such as which hypothesis tests were used and what plots were created. In case of a continuous target variable it also contains a box-plot of the target. 3) A table with the SSVs selected by iGATE, the obtained count statistics, p–values and expert comments is presented, followed by another table containing the extracted controlbands for the retained SSV. 4) Summary statistics about the validation results, such as how many samples of the validation set fall within the extracted controlbands and the distribution of the target variable amongst these samples. 5) The appendix contains a possibly long list of all the SSV that were studied and the produced regression/ frequency plots for future reference.

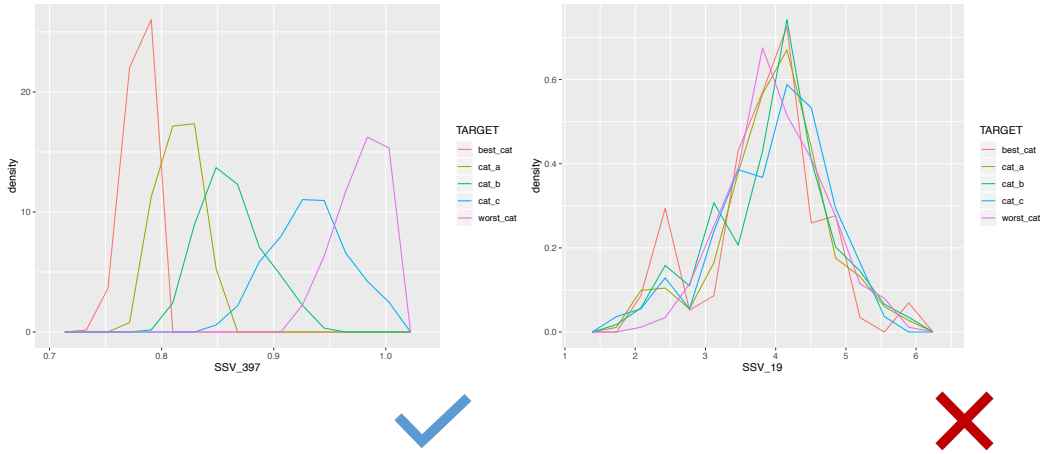


Figure 6.4: Frequency polygon example: In this case the user might decide to keep the SSV on the left and discard the one on the right.

plot for each retained SSV, split up by the various categories, is created[4]. If there is a clear separation between the curve for the best category and the curve for the worst category, the SSV is kept as potentially influential. Otherwise it is discarded. See Figure 6.4 for an example.

6.4 An application to blast furnace top-gas efficiency

We applied iGATE to blast furnace data provided by Tata Steel. For reasons of confidentiality we suppress the real variable names and simply refer to them generically by "SSV i" with $i = 1, \ldots, 218$. Our target was *top-gas efficiency* (η_{CO}). The efficiency of a blast furnace is the amount of reductant (i.e. coke and other injectants containing carbon) used per tonne of hot metal produced. As a proxy for the efficiency of the furnace, the chemical decomposition of the *top-gas* (the gas escaping at the top of the furnace) can be studied. More precisely, the top-gas efficiency η_{CO} measures how efficiently the oxygen from the burden in the blast furnace is removed.[5] It is calculated as

$$\eta_{\mathrm{CO}} = \frac{\mathrm{CO}_2}{\mathrm{CO} + \mathrm{CO}_2}.$$

That is, an increase in η_{CO} means, more CO_2 is produced rather than CO, meaning, the oxygen is removed using less coke, making the furnace more efficient. Typical values for η_{CO} are in the range of $45\% - 50\%$ (Geerdes et al. 2015). It was known

[4] A normalized frequency plot shows the same data as a normalized histogram, only that instead of bars, we are connecting the bins via lines. They are more suitable than histograms when comparing a distribution across the levels of a categorical variable, as we do here.

[5] We want to remove as much oxygen (O_2) as possible, using as little carbon (C) as possible.

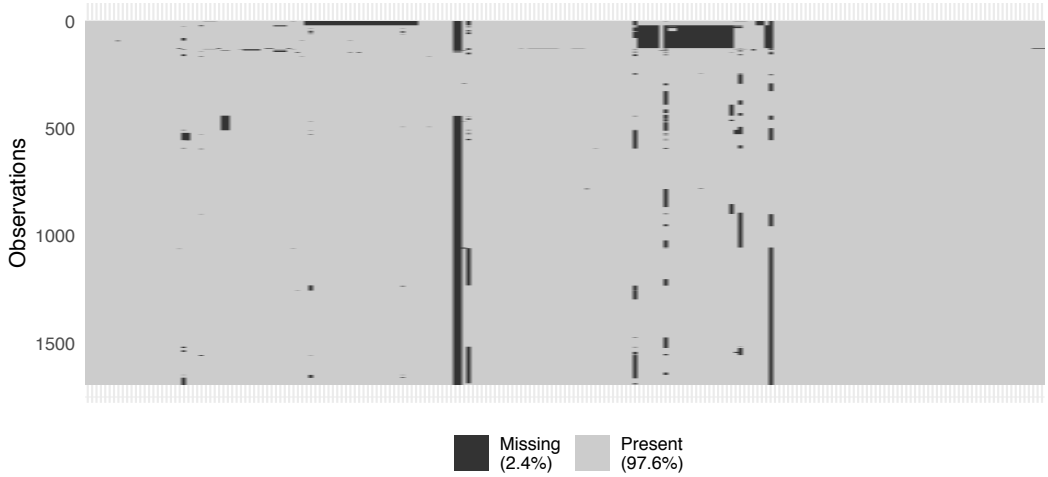


Figure 6.5: Positions of missing values in the blast furnace data. Each tick along the x-axis represents one feature, while the y-axis enumerates the rows in the data frame. Grey cells correspond to correctly recorded data. Black cells correspond to missing data. A total of 2.4% of data is missing.

beforehand that η_{CO} is a definite indicator for process stability and thus, better understanding of η_{CO} would lead to better process control. Also, since it is negatively correlated with the amount of coke used (the higher η_{CO}, the less coke is needed to fuel the furnace) improvements to η_{CO} will have a direct, quantifiable business impact.

The data spanned around five years of daily mean η_{CO} values. It contained 1692 observations and 218 SSVs. In total, 2.4% of all entries were missing or wrongly recorded, cf. Figure 6.5. The missing data required no additional pre-processing as iGATE handles missing values automatically. Of this data set, we randomly selected 80% as training data (1353 observations), while we retained 20% for validation (339 observations).

Running iGATE on the training data returned 88 potentially influential features for η_{CO}, meaning it was able to reduce the number of variables under study by around 60%. Upon presenting these variables to domain experts, several variables corresponding to target leakage were identified and removed from further analysis. For example, iGATE identified the coke-rate (the amount of coke burned per tonne of hot metal produced) as a significant predictor for the values of η_{CO}. While coke-rate is highly correlated with η_{CO}, it is a quality measure in and off itself and cannot be controlled directly. iGATE also found the chemical decomposition of the coke (to be precise: the concentration of a specific chemical element) to be influential. High concentrations of it were found to result in worse performance. A domain expert explained that this SSV can be interpreted as an indicator for the type of coke that is being burned and that it was known that certain types of coke performed better than others. This suggested that separate analyses for different coke types might be sensible. While it is statistical best practice to account for different group effects such as this one, it would have been difficult to find the right groups and parameters

to adjust for without this expert's insight. This is especially true in a case like this, where the group membership of the samples is only encoded implicitly in their chemical decomposition. The experts also confirmed some of the selected SSVs. For example, a certain temperature setting was found to produce bad results when it was too high. This was interpreted to mean that if the temperature is too high, more fuel is used, producing lower values of η_{CO}. Having an expert confirm such findings and recording their comment on it can be equally valuable for the long-term knowledge capture within the company as it creates a knowledge pool that future data science projects can build on. This once more illustrates that expert feedback is essential for successful analytics projects as fully autonomous approaches would not have been able to provide the necessary context to these findings. After incorporating the expert feedback, we retained 16 potentially influential variables for further analysis.

iGATE does not yet employ any statistical regression modelling, hence there is no explicit loss function that can be used for validation purposes. Instead, in the validation step we try to gauge how well the control bands extracted by iGATE capture the differences between good and bad products. To that end, we took the 339 validation samples and for each retained SSV extracted those observations that fell into any of the good or bad control bands. The results are displayed in Figure 6.6. It shows frequency plots of η_{CO} for those observations that fell into either of the control bands. The plots have been normalized to have density one to account for differing group sizes. We see that in most cases those observations that we would expect to yield a good, i.e. high, value of η_{CO} based on the extracted control bands indeed have higher η_{CO} values than those we would expect to yield bad η_{CO} values. This is particularly prominent in the plots **A, D, E, F, O**. There are several plots in which the distribution of η_{CO} overlaps strongly between those observations we would expect to have good quality and those we would expect to have bad quality, e.g. plots **B** or **J**. This means that these variables on their own might not be impacting η_{CO} significantly after all and further analysis is needed.

6.5 Conclusion

iGATE is a guided analytics framework that presents a middle ground between autonomous and manual feature selection. It is fast, easy to explain to people without statistical training and the controlbands extracted by it can be translated into actionable instructions for process operators. The automated reporting feature is an integral part of iGATE that promotes knowledge capture within a company. We recognize that there are statistically more powerful tools available for assessing the influence of covariates on a target variable, but chose the tools used in iGATE for

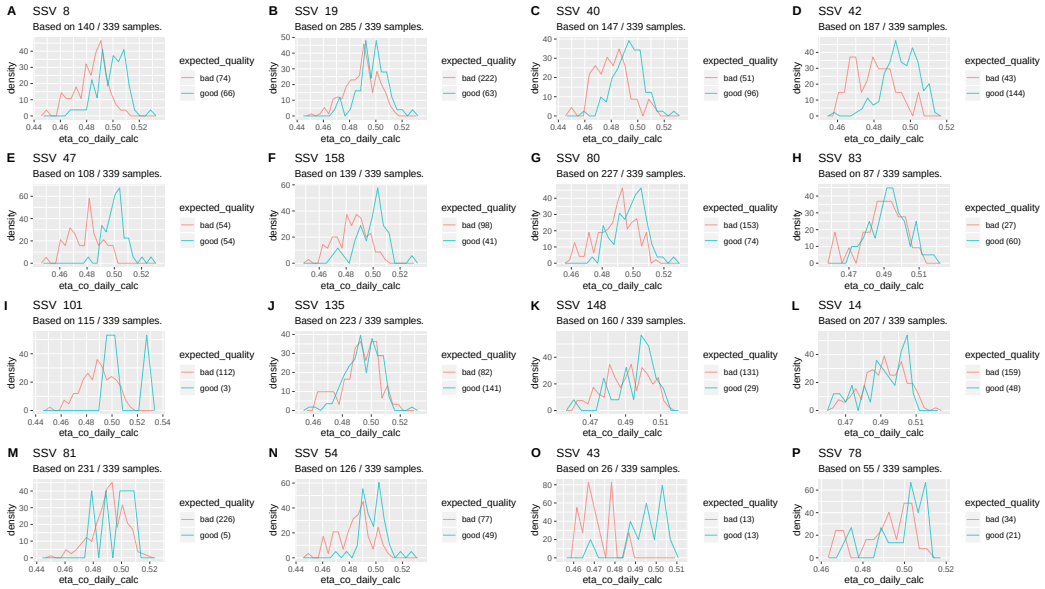


Figure 6.6: Validation results. For each of the 16 retained SSVs we checked which of the 339 observations retained for validation had values for that SSV within the extracted good or bad control bands. We plotted the η_{CO} value of these variables as a frequency plot with normalized density curve. In most cases we observe that the mode of the η_{CO} values of the observations we expect to have good η_{CO} values is to the right of the mode of the observations we expect to have bad η_{CO} values. This indicates that there indeed is a difference in distribution for that SSV between good and bad η_{CO} values. This is particularly prominent for plots **A, D, E, F, O**. In plots **B** or **J**, for example, no such difference is apparent, suggesting that these SSV by themselves might not be significantly influential to the target variable after all.

their easy interpretability and robustness against messy data. Much of the value of the traditional manual approach of domain experts and statisticians exchanging information lies in its interactivity and mutual guidance, an element which was retained in iGATE but significantly streamlined. While the methods used in iGATE already existed beforehand, novelty was added by combining them in this manner and extending them to categorical target variables.

The emphasis on explainable results seems justified to us as there commonly are concerns about basing business decisions with far-reaching consequences on results obtained from "black-box" models. We consider iGATE as a stepping stone in fostering user confidence in the use of guided analytics tools.

Appendix

Appendix A

Proofs for Chapter 4

A.1 Proof of Theorem 4.4

A.1.1 Proof of the compatibility condition, Proposition 4.3

We first prove a *sample compatibility condition* before providing a proof for the population compatibility condition in Proposition 4.3. That is, we first want to find a suitable relation between the quantities $\|\hat{\theta} - \theta_0\|_1$ and $(\hat{\theta} - \theta_0)\hat{\Sigma}(\hat{\theta} - \theta_0)$, where $\hat{\Sigma} = T^{-1}D^T DT^{-1}$ is the sample version of the sample size adjusted Gram matrix Σ.

Recall from Section 2.7.1.2 that the compatibility condition is equivalent to the condition that

$$\kappa^2(A, s_0) := \min_{\substack{\theta \in \mathbb{R}^{2n+1+p} \setminus \{0\} \\ \|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1}} \frac{\theta^T A \theta}{\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2}$$

stays bounded away from zero. We first show that the compatibility condition holds for the matrix

$$\Sigma_A := \begin{bmatrix} I_{2n} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbb{E}[Z^T Z/N] \end{bmatrix} \in \mathbb{R}^{(2n+1+p)\times(2n+1+p)},$$

where I_{2n} is the $(2n) \times (2n)$ identity matrix.

Recall that by Assumption 4.1 there is a finite constant $c_{\min} > 0$ independent of n, such that $\lambda_{\min} > c_{\min} > 0$ for all n, where $\lambda_{\min} = \lambda_{\min}(n)$ of is the the minimum eigenvalue of $\frac{1}{N}\mathbb{E}[Z^T Z]$. Then, clearly, for any $\theta = (\vartheta^T, \mu, \gamma^T)^T$,

$$\theta^T \Sigma_A \theta = \|\vartheta\|_2^2 + \mu^2 + \gamma^T \frac{1}{N}\mathbb{E}[Z^T Z]\gamma \geq \|\vartheta\|_2^2 + \mu^2 + c_{\min}\|\gamma\|_2^2 \geq (1 \wedge c_{\min})\|\theta\|_2^2.$$

Thus, Σ_A is strictly positive definite. Furthermore, by Cauchy-Schwarz' inequality,

for any $\theta \in \mathbb{R}^{2n+1+p}$ with $\|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1$,

$$\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2 \leq \|\theta_{S_{0,+}}\|_2^2 \leq \|\theta\|_2^2.$$

Thus,

$$\kappa^2(\Sigma_A, s_0) = \min_{\substack{\theta \in \mathbb{R}^{2n+1+p} \setminus \{0\} \\ \|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1}} \frac{\theta^T \Sigma_A \theta}{\frac{1}{s_{0,+}}\|\theta_{S_{0,+}}\|_1^2} \geq \frac{(1 \wedge c_{\min})\|\theta\|_2^2}{\|\theta\|_2^2} > 0.$$

We conclude that the compatibility condition holds for Σ_A. Now we need to show that with high probability $\kappa(\hat{\Sigma}, s_0) \geq \kappa(\Sigma_A, s_0)$, which would imply that the compatibility condition holds with high probability for $\hat{\Sigma}$. To that end, we employ once more Lemma 2.8.

Introduce the set

$$\mathcal{J} = \left\{ \max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \leq \frac{c_{\min}}{32 s_{0,+}} \right\}.$$

On the set $\mathcal{J}$, by Lemma 2.8, we have $\kappa^2(\hat{\Sigma}, s_0) \geq \kappa(\Sigma_A, s_0) - \frac{c_{\min}}{2} \geq \frac{c_{\min}}{2} > 0$ and thus the compatibility condition holds for $\hat{\Sigma}$ on $\mathcal{J}$.

Lemma A.1. *If $s_0 = o(\sqrt{n})$, for n large enough, with $\delta = \frac{c_{\min}}{32 s_{0,+}}$ and $\tilde{c} = c^2 \vee (2c^4)$, where $c > 0$ is the universal constant such that $|Z_{k,ij}| \leq c$ for all k, i, j, we have*

$$P(\mathcal{J}) \geq 1 - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048 s_{0,+}^2 \tilde{c}}\right).$$

Proof. To make referencing of sections of $\hat{\Sigma}$ easier, we number its blocks as follows

$$\hat{\Sigma} = T^{-1} \begin{bmatrix} \underbrace{X^T X}_{①} & \underbrace{X^T \mathbf{1}}_{②} & \underbrace{X^T Z}_{③} \\ \underbrace{\mathbf{1}^T X}_{④} & \underbrace{\mathbf{1}^T \mathbf{1}}_{⑤} & \underbrace{\mathbf{1}^T Z}_{⑥} \\ \underbrace{Z^T X}_{⑦} & \underbrace{Z^T \mathbf{1}}_{⑧} & \underbrace{Z^T Z}_{⑨} \end{bmatrix} T^{-1}.$$

For block ①, i.e. $i, j = 1, \ldots, 2n$, notice that $(X^{\text{out}})^T X^{\text{out}} = (X^{\text{in}})^T X^{\text{in}} = (n-1) I_n$ and $(X^{\text{out}})^T X^{\text{in}}$ is a matrix with zero on the diagonal and ones everywhere else. Therefore, we have either $\hat{\Sigma}_{ij} = \Sigma_{A,ij}$ or

$$|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| = \frac{1}{n-1} < \frac{c_{\min}}{32 s_{0,+}},$$

for n large enough, since $s_{0,+} = o(\sqrt{n})$. Blocks ② and ④ are a $2n$ dimensional column and row vector respectively in which each entry is equal to $n-1$. Thus, for

i, j corresponding to these blocks,

$$|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| = \frac{n-1}{\sqrt{(n-1)N}} = \frac{1}{\sqrt{n}} \leq \frac{c_{\min}}{32 s_{0,+}},$$

for n large enough, since $s_{0,+} = o(\sqrt{n})$. For i, j corresponding to blocks ③ and ⑦, we have

$$|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \leq \frac{c}{\sqrt{n}} < \frac{c_{\min}}{32 s_{0,+}},$$

for n large enough. Block ⑤ is a single real number and equal for $\hat{\Sigma}$ and Σ_A.

The only cases left to consider are those entries corresponding to blocks ⑥, ⑧ and ⑨. For the blocks ⑥ and ⑧, that is for $i = 2n+1, j = 2n+2, \ldots, 2n+1+p$ and $i = 2n+2, \ldots, 2n+1+p, j = 2n+1$, $\hat{\Sigma}_{ij} - \Sigma_{A,ij} = \hat{\Sigma}_{ij}$ is the scaled sum of all the entries of some column Z_k of the matrix Z for an appropriate k. That is, there is a $1 \leq k \leq p$ such that

$$\hat{\Sigma}_{ij} - \Sigma_{A,ij} = \frac{1}{N} Z_k^T \mathbf{1} = \frac{1}{N} \sum_{s \neq t} Z_{k,st}.$$

By model assumption, $\mathbb{E}[\hat{\Sigma}_{ij} - \Sigma_{A,ij}] = 0$. We know that for each $k, s, t : Z_{k,st} \in [-c, c]$. Hence, by Hoeffding's inequality, for all $\delta > 0$,

$$\begin{aligned} P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \delta\right) &= P\left(\left|\sum_{s \neq t} Z_{k,st}\right| \geq N\delta\right) \\ &\leq 2\exp\left(-\frac{2N^2\delta^2}{\sum_{i \neq j}(2c)^2}\right) = 2\exp\left(-N\frac{\delta^2}{2c^2}\right). \end{aligned}$$

For block ⑨, that is for $i, j = 2n+2, \ldots, 2n+1+p$, a typical element has the form

$$\hat{\Sigma}_{ij} - \Sigma_{A,ij} = \frac{1}{N} \sum_{s \neq t} \left\{ Z_{k,st} Z_{l,st} - \mathbb{E}[Z_{k,st} Z_{l,st}] \right\},$$

for appropriate k, l. In other words, $\hat{\Sigma}_{ij} - \Sigma_{A,ij}$ is the inner product of two columns of Z, minus their expectation, scaled by $1/N$. Since $Z_{k,st} Z_{l,st} \in [-c^2, c^2]$ for all k, l, s, t, we have that for all k, l, s, t: $Z_{k,st} Z_{l,st} - \mathbb{E}[Z_{k,st} Z_{l,st}] \in [-2c^2, 2c^2]$. Thus, by Hoeffding's inequality, for all $\delta > 0$,

$$\begin{aligned} P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \delta\right) &= P\left(\left|\sum_{s \neq t} \{Z_{k,st} Z_{l,st} - \mathbb{E}[Z_{k,st} Z_{l,st}]\}\right| \geq N\delta\right) \\ &\leq 2\exp\left(-N\frac{\delta^2}{8c^4}\right). \end{aligned}$$

Thus, with $\tilde{c} = c^2 \vee (2c^4)$, we have for any entry in blocks ⑥, ⑧, ⑨, that for any

$\delta > 0$,

$$P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \delta\right) \leq 2\exp\left(-N\frac{\delta^2}{2\tilde{c}}\right).$$

Choosing $\delta = \frac{c_{\min}}{32s_{0,+}}$, by the exposition above we know that all entries in blocks ① - ⑤ and ⑦ are bounded by δ for $n \gg 0$. Also, because block ⑥ is the transpose of block ⑧, it is sufficient to control one of them. By symmetry of block ⑨ it suffices to control the upper triangular half, including the diagonal, of block ⑨. Thus, we only need to control the entries $\hat{\Sigma}_{ij} - \Sigma_{A,ij}$ for i, j in the following index set

$$\begin{aligned}\mathcal{A} &= \{i, j : i, j \text{ belong to ⑧ or the upper triangular half or diagonal of ⑨}\} \\ &= \{(i,j) \in \{n+2, \ldots, n+1+p\} \times \{n+1\}\} \cup \{i \leq j : i, j = n+2, \ldots, n+1+p\}.\end{aligned}$$

Keep in mind that block ⑧ has p elements, while the upper triangular part of block ⑨ plus its diagonal has $\binom{p}{2} + p = \binom{p+1}{2}$ elements. Thus, for $n \gg 0$,

$$\begin{aligned}P(\mathcal{J}^c) &= P\left(\max_{ij} |\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \frac{c_{\min}}{32s_{0,+}}\right) \\ &\leq \sum_{i,j \in \mathcal{A}} P\left(|\hat{\Sigma}_{ij} - \Sigma_{A,ij}| \geq \frac{c_{\min}}{32s_{0,+}}\right) \\ &\leq 2p\exp\left(-N\frac{\delta^2}{2c^2}\right) + 2\binom{p+1}{2}\exp\left(-N\frac{\delta^2}{8c^4}\right) \\ &\leq 2\left(p + \binom{p+1}{2}\right)\exp\left(-N\frac{\delta^2}{2\tilde{c}}\right) \\ &= p(p+3)\exp\left(-N\frac{\delta^2}{2\tilde{c}}\right).\end{aligned}$$

This proves the claim. □

For our results on model selection consistency in Section 4.2.1 we require a sample version of the population compatibility condition (Proposition 4.3). Therefore, we prove that the compatibility condition holds for the sample version of Σ, that is for $\hat{\Sigma} = T^{-1}D^T DT^{-1}$ first in Proposition A.2 from which Proposition 4.3 readily follows. The tools we employ are similar to what we saw in Section 2.7.1.2.

Proposition A.2. *Under Assumption 4.1, for $s_0 = o(\sqrt{n})$ and n large enough, with $\tilde{c} = c^2 \vee (2c^4)$, where $c > 0$ is the universal constant such that $|Z_{k,ij}| \leq c$ for all k, i, j: With probability at least*

$$1 - p(p+3)\exp\left(-N\frac{c_{\min}^2}{2048s_{0,+}^2\tilde{c}}\right)$$

we have for every $\theta \in \mathbb{R}^{2n+1+p}$ with $\|\theta_{S_{0,+}^c}\|_1 \leq 3\|\theta_{S_{0,+}}\|_1$, that

$$\|\theta_{S_{0,+}}\|_1^2 \leq \frac{2s_{0,+}}{c_{\min}}\theta^T\hat{\Sigma}\theta.$$

Proof. This follows from Lemma A.1. □

Proof of Proposition 4.3. To prove that the compatibility condition holds for the population sample size adjusted Gram matrix Σ we may follow the same steps as in the proof of Proposition 2.3: Number the blocks of Σ as ① - ⑨ as we did for $\hat{\Sigma}$. Σ and Σ_A are equal on blocks ③, ⑤, ⑥, ⑦, ⑧ and ⑨. For blocks ①, ② and ④ we use the exact same arguments as in the proof of Proposition 2.3 to find that for n sufficiently large, almost surely,

$$\max_{ij} |\Sigma_{ij} - \Sigma_{A,ij}| \leq \frac{c_{\min}}{32 s_{0,+}}.$$

The claim follows from Lemma 2.8. □

A.1.2 A basic Inequality

Recall that P_n denotes the empirical measure with respect to our observations (A_{ij}, Z_{ij}), that is, for any suitable function g,

$$P_n g \coloneqq \frac{1}{N} \sum_{i \neq j} g(A_{ij}, Z_{ij}).$$

In particular, if we let for each $\theta \in \Theta$, $l_\theta(A_{ij}, Z_{ij}) = -A_{ij}(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij}) + \log(1 + \exp(\alpha_i + \beta_j + \mu + \gamma^T Z_{ij}))$, then $P_n l_\theta = \mathcal{L}(\theta)/N$. Similarly, we define the theoretical risk as $P = \mathbb{E}P_n$. In particular, $P l_\theta = \mathbb{E}P_n l_\theta = \mathbb{E}[\mathcal{L}(\theta)]/N$, where we suppress the dependence of the theoretical risk on n in our notation. Note that we have for the excess risk

$$\mathcal{E}(\theta) = P(l_\theta - l_{\theta_0}).$$

We derive a basic inequality for model (4.1) as we did in Lemma 2.11. We define the *empirical process* as

$$\{v_n(\theta) = (P_n - P) l_\theta : \theta \in \Theta\}.$$

Lemma A.3 (Basic Inequality)**.** *For any $\theta = (\beta^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}}$ we have*

$$\mathcal{E}(\hat{\theta}) + \lambda \|\hat{\beta}\|_1 \leq -[v_n(\hat{\theta}) - v_n(\theta)] + \mathcal{E}(\theta) + \lambda \|\beta\|_1.$$

Proof. By plugging in the definitions and rearranging, we see that the above equation is equivalent to

$$\frac{1}{N}\mathcal{L}(\hat{\theta}) + \lambda \|\hat{\beta}\|_1 \leq \frac{1}{N}\mathcal{L}(\theta) + \lambda \|\beta\|_1,$$

which is true by definition of $\hat{\theta}$. □

An analogous result follows line by line for the rescaled parameter $\hat{\bar{\theta}}$. Writing

$$\bar{v}_n(\bar{\theta}) \coloneqq \frac{1}{N}(\bar{\mathcal{L}}(\bar{\theta}) - \mathbb{E}[\bar{\mathcal{L}}(\bar{\theta})]) = v_n(\theta).$$

for the rescaled empirical process, we have the following.

Lemma A.4. *For any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ we have*

$$\bar{\mathcal{E}}(\hat{\bar{\theta}}) + \bar{\lambda}\|\hat{\bar{\vartheta}}\|_1 \leq -[\bar{v}_n(\hat{\bar{\theta}}) - \bar{v}_n(\bar{\theta})] + \bar{\mathcal{E}}(\bar{\theta}) + \bar{\lambda}\|\bar{\vartheta}\|_1.$$

Remark. For any $0 < t < 1$ and $\theta \in \Theta_{\text{loc}}$, let $\tilde{\theta} = t\hat{\theta} + (1-t)\theta$. Since Γ is convex, $\tilde{\theta} \in \Theta_{\text{loc}}$ and since $\theta \to l_\theta$ and $\|\,.\,\|_1$ are convex functions, we can replace $\hat{\theta}$ by $\tilde{\theta}$ in the basic inequality and still obtain the same result. Plugging in the definitions, we see that the basic inequality is equivalent to the following:

$$\begin{aligned} &\mathcal{E}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 \leq -[v_n(\tilde{\theta}) - v_n(\theta)] + \lambda\|\beta\|_1 + \mathcal{E}(\theta) \\ \iff &\frac{1}{N}\mathcal{L}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 \leq \frac{1}{N}\mathcal{L}(\theta) + \lambda\|\beta\|_1 \end{aligned}$$

and by convexity

$$\frac{1}{N}\mathcal{L}(\tilde{\theta}) + \lambda\|\tilde{\beta}\|_1 \leq \frac{1}{N}t\mathcal{L}(\hat{\theta}) + \frac{1}{N}(1-t)\mathcal{L}(\theta) + t\lambda\|\hat{\beta}\|_1 + (1-t)\lambda\|\beta\|_1 \leq \frac{1}{N}\mathcal{L}(\theta) + \lambda\|\beta\|_1,$$

where the last inequality follows by definition of $\hat{\theta}$. In particular, for any $M > 0$, choosing

$$t = \frac{M}{M + \|\hat{\theta} - \theta\|_1},$$

gives $\|\tilde{\theta} - \theta\|_1 \leq M$. The completely analogous result holds for $\bar{\theta}$.

A.1.3 Two norms and one function space

To give us a more compact way of writing, for any $\bar{\theta} \in \Theta$ we introduce functions $f_{\bar{\theta}} : \mathbb{R}^{2n+1+p} \to \mathbb{R}, f_{\bar{\theta}}(v) = v^T\bar{\theta}$ and denote the function space of all such $f_{\bar{\theta}}$ by $\bar{\mathbb{F}} \coloneqq \{f_{\bar{\theta}} : \bar{\theta} \in \Theta\}$. We endow $\bar{\mathbb{F}}$ with two norms as follows:
Denote the law of the rows of $\bar{D}$ on $\mathbb{R}^{2n+1+p}$, i.e. the probability measure induced by $(\bar{X}_{ij}^T, 1, Z_{ij}^T)^T, i \neq j$, by $\bar{Q}$. That is, for a measurable set $A = A_1 \times A_2 \subset \mathbb{R}^{2n+1} \times \mathbb{R}^p$,

$$\bar{Q}(A) = \frac{1}{N}\sum_{i \neq j} P(\bar{D}_{ij} \in A) = \frac{1}{N}\sum_{i \neq j} \delta_{ij}(A_1) \cdot P(Z_{ij} \in A_2),$$

where $\delta_{ij}(A_1) = 1$ if $(\bar{X}_{ij}^T, 1)^T \in A_1$ and zero otherwise, is the Dirac-measure. We are interested in the L_2 and L_∞ norm on $\bar{\mathbb{F}}$ with respect to the measure $\bar{Q}$ on $\mathbb{R}^{2n+1} \times \mathbb{R}^p$. Denote the $L_2(\bar{Q})$-norm of $f \in \bar{\mathbb{F}}$ simply by $\|\,.\,\|_{\bar{Q}}$ and let $\mathbb{E}_Z$ be the

expectation with respect to Z:

$$\|f\|_{\bar{Q}}^2 \coloneqq \|f\|_{L_2(\bar{Q})}^2 = \int_{\mathbb{R}^{2n+1}\times\mathbb{R}^p} f(v)^2 \bar{Q}(dv) = \frac{1}{N}\sum_{i\neq j} \mathbb{E}_Z[f((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)^2]$$

and define the $L_\infty(\bar{Q})$-norm as usual as the $\bar{Q}$-a.s. smallest upper bound of f:

$$\|f\|_{\bar{Q},\infty} = \inf\{C \geq 0 : |f(v)| \leq C \text{ for } \bar{Q}\text{-almost every } v \in \mathbb{R}^{2n+1+p}\}.$$

In particular, for any $f_{\bar{\theta}} \in \bar{\mathbb{F}}, \bar{\theta} \in \bar{\Theta}_{\text{loc}}$: $\|f_{\bar{\theta}}\|_\infty \leq \sup_{Z_{ij}} \|\bar{D}\bar{\theta}\|_\infty \leq r_n$.

We make the analogous definitions for the unscaled design matrix. Let Q denote the probability measure induced by the rows of D. Since $\bar{D}\bar{\theta} = D\theta$, for any θ with rescaled version $\bar{\theta}$, we have

$$\|f_{\bar{\theta}}\|_{L_2(\bar{Q})} = \|f_\theta\|_{L_2(Q)}, \quad \|f_{\bar{\theta}}\|_{\bar{Q},\infty} = \|f_\theta\|_{Q,\infty}.$$

We want to apply the compatibility condition to vectors of the form $\theta = \theta_1 - \theta_2, \theta_1, \theta_2 \in \Theta_{\text{loc}}$.

We have the following relation between the $L_2(Q)$-norm and the sample size adjusted Gram matrix Σ: For any θ we have

$$\|f_\theta\|_Q^2 = \mathbb{E}_Z\left[\frac{1}{N}\sum_{i\neq j}(D_{ij}^T\theta)^2\right] = \bar{\theta}^T\Sigma\bar{\theta}. \tag{A.1}$$

We have the following corollary which follows immediately from Proposition 4.3 (see e.g. van de Geer & Bühlmann (2011), Section 6.12 for a general treatment).

Corollary A.5. *Under Assumption 4.1, for $s_0 = o(\sqrt{n})$ and n large enough, for every $\bar{\theta} = \bar{\theta}_1 - \bar{\theta}_2, \bar{\theta}_1, \bar{\theta}_2 \in \bar{\Theta}_{\text{loc}}$ with $\|\bar{\theta}_{S_{0,+}^c}\|_1 \leq 3\|\bar{\theta}_{S_{0,+}}\|_1$, we have*

$$\|\bar{\theta}_{S_{0,+}}\|_1^2 \leq \frac{2s_{0,+}}{c_{\min}}\|f_{\theta_1} - f_{\theta_2}\|_Q^2.$$

Proof. By Proposition 4.3,

$$\|\bar{\theta}_{S_{0,+}}\|_1^2 \leq \frac{2s_{0,+}}{c_{\min}}\bar{\theta}\Sigma\bar{\theta}.$$

The claim follows from (A.1) and the fact that $\theta \mapsto f_\theta$ is linear. □

A.1.4 Lower quadratic margin for $\mathcal{E}$

We derive a lower quadratic bound on the excess risk $\mathcal{E}(\theta)$ if the parameter θ is close to the truth θ_0. This is referred to as the *margin condition* in classical LASSO theory (cf. van de Geer & Bühlmann (2011)). The proof relies on a second-order

Taylor expansion of the function l_θ of introduced in Section 4.2.2. Given a fixed θ, we treat l_θ as a function in $\theta^T x$ and define new functions $l_{ij} : \mathbb{R} \to \mathbb{R}, i \neq j$,

$$l_{ij}(a) = \mathbb{E}[l_\theta(A_{ij}, a)|Z_{ij}] = -p_{ij}a + \log(1 + \exp(a)),$$

where $p_{ij} = P(A_{ij} = 1|Z_{ij})$ and by slight abuse of notation we use $l_\theta(A_{ij}, a) \coloneqq -A_{ij}a + \log(1 + \exp(a))$. Taking the derivative, it is easy to see that

$$f_{\theta_0}((X_{ij}^T, 1, Z_{ij}^T)^T) \in \arg\min_a l_{ij}(a).$$

Write $f_0 = f_{\theta_0}$. All l_{ij} are clearly twice continuously differentiable with derivative

$$\frac{\partial^2}{\partial a^2} l_{ij}(a) = \frac{\exp(a)}{(1 + \exp(a))^2} > 0, \forall a \in \mathbb{R}.$$

Using a second-order Taylor expansion around $a_0 = f_0((X_{ij}^T, 1, Z_{ij}^T)^T)$ we get

$$l_{ij}(a) = l_{ij}(a_0) + l'(a_0)(a - a_0) + \frac{l''(\bar{a})}{2}(a - a_0)^2 = l_{ij}(a_0) + \frac{l''(\bar{a})}{2}(a - a_0)^2,$$

with an $\bar{a}$ between a and a_0. Note that $|a_0| \leq r_n$. Then, for any a with $|a| \leq r_n$, we must have that for any intermediate point $\bar{a}$ between a_0 and a it also holds true that $|\bar{a}| \leq r_n$. Also note that $\frac{\exp(a)}{(1+\exp(a))^2}$ is symmetric and monotone decreasing for $a \geq 0$. Thus, for any a with $|a| \leq r_n$,

$$\begin{aligned} l_{ij}(a) - l_{ij}(a_0) &= \frac{\exp(\bar{a})}{(1 + \exp(\bar{a}))^2} \frac{(a - a_0)^2}{2} \\ &= \frac{\exp(|\bar{a}|)}{(1 + \exp(|\bar{a}|))^2} \frac{(a - a_0)^2}{2}, \quad \text{by symmetry} \\ &\geq \frac{\exp(r_n)}{(1 + \exp(r_n))^2} \frac{(a - a_0)^2}{2}. \end{aligned} \tag{A.2}$$

In particular, if we pick any $\theta \in \Theta_{\text{loc}}$ and let $a = f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T)$, we have

$$\begin{aligned} &l_{ij}(f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T)) - l_{ij}(f_0((X_{ij}^T, 1, Z_{ij}^T)^T)) \\ &\qquad \geq \frac{\exp(r_n)}{(1 + \exp(r_n))^2} \frac{(f_\theta((X_{ij}^T, 1, Z_{ij}^T)^T) - f_0((X_{ij}^T, 1, Z_{ij}^T)^T))^2}{2}. \end{aligned}$$

Let

$$K_n = \frac{2(1 + \exp(r_n))^2}{\exp(r_n)}. \tag{A.3}$$

Define a subset $\mathbb{F}_{\text{local}} \subset \mathbb{F}$ as $\mathbb{F}_{\text{local}} = \{f_\theta : \theta \in \Theta_{\text{loc}}\}$. Now, for all $f_\theta \in \mathbb{F}_{\text{local}}$:

$$\begin{aligned}
\mathcal{E}(\theta) &= \frac{1}{N} \sum_{i \neq j} \mathbb{E}[l_\theta(A_{ij}, D_{ij}) - l_{\theta_0}(A_{ij}, D_{ij})] \\
&= \frac{1}{N} \sum_{i \neq j} \mathbb{E}[(l_{ij}(f_\theta(D_{ij}) - l_{ij}(f_0(D_{ij})))] \\
&\geq \frac{1}{K_n} \cdot \frac{1}{N} (\theta - \theta_0)^T \mathbb{E}_Z[D^T D](\theta - \theta_0) \\
&= \frac{1}{K_n} \cdot \|f_\theta - f_0\|_Q^2.
\end{aligned}$$

Thus, we have obtained a lower bound for the excess risk given by the quadratic function $G_n(\|f_\theta - f_0\|)$ where $G_n(u) = 1/K_n \cdot u^2$. Recall that the convex conjugate of a strictly convex function G on $[0, \infty)$ with $G(0) = 0$ is defined as the function

$$H(v) = \sup_u \{uv - G(u)\}, \quad v > 0,$$

and in particular, if $G(u) = cu^2$ for a positive constant c, we have $H(v) = v^2/(4c)$. Hence, the convex conjugate of G_n is

$$H_n(v) = \frac{v^2 K_n}{4}.$$

Keep in mind that by definition for any u, v: $uv \leq G(u) + H(v)$.

A.1.5 Consistency on a special set

We show that the penalized likelihood estimator is consistent on a specific set $\mathcal{I}$. It will then suffice to show that $P(\mathcal{I}) \to 1$. The proof follows in spirit Theorem 6.4 in van de Geer & Bühlmann (2011) and is analogous to the derivations in Section 2.7.1.6 for SβM-C.

We define some objects that we will need for the proof of consistency. We want to use the quadratic margin condition derived in Section A.1.4. Recall that the quadratic margin condition holds for any $\theta \in \Theta_{\text{loc}}$. Define

$$\epsilon^* = H_n \left(\frac{4\sqrt{2}\sqrt{s_{0,+}}\bar{\lambda}}{\sqrt{c_{\min}}} \right).$$

Recall the definition of $\bar{\theta}$ in equation (4.7) and let for any $M > 0$

$$Z_M \coloneqq \sup_{\substack{\theta \in \Theta_{\text{loc}}, \\ \|\bar{\theta} - \bar{\theta}_0\|_1 \leq M}} |v_n(\theta) - v_n(\theta_0)|,$$

where v_n denotes the empirical process. The set over which we are maximizing in the definition of Z_M can be expressed in terms of parameters θ on the original scale

as

$$\left\{\theta = (\vartheta^T, \mu, \gamma^T)^T \in \Theta_{\text{loc}} : \frac{1}{\sqrt{n}}\|\vartheta - \vartheta_0\|_1 + |\mu - \mu_0| + \|\gamma - \gamma_0\|_1 \leq M\right\}.$$

Set

$$M^* \coloneqq \epsilon^*/\lambda_0,$$

where λ_0 is a lower bound on $\bar{\lambda}$ that will be made precise in the proof showing that $\mathcal{I}$ has large probability. Define

$$\mathcal{I} \coloneqq \{Z_{M^*} \leq \lambda_0 M^*\} = \{Z_{M^*} \leq \epsilon^*\}. \tag{A.4}$$

Theorem A.6. *Assume that Assumptions 4.1 and B2 hold and that $\bar{\lambda} \geq 8\lambda_0$. Then, on the set $\mathcal{I}$, we have*

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{1}{\sqrt{n}}\|\hat{\vartheta} - \vartheta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1\right) \leq 4\epsilon^* = 4H_n\left(\frac{4\sqrt{2}\sqrt{s_{0,+}}\bar{\lambda}}{\sqrt{c_{\min}}}\right).$$

Proof of Theorem A.6. We assume that we are on the set $\mathcal{I}$ throughout. Set

$$t = \frac{M^*}{M^* + \|\hat{\bar{\theta}} - \bar{\theta}_0\|_1}$$

and $\tilde{\theta} = (\tilde{\vartheta}^T, \tilde{\mu}, \tilde{\gamma}^T)^T = t\hat{\bar{\theta}} + (1-t)\bar{\theta}_0$. Then,

$$\|\tilde{\theta} - \bar{\theta}_0\|_1 = t\|\hat{\bar{\theta}} - \bar{\theta}_0\| \leq M^*.$$

Since $\hat{\bar{\theta}}, \bar{\theta}_0 \in \bar{\Theta}_{\text{loc}}$ and by the convexity of $\bar{\Theta}_{\text{loc}}$, $\tilde{\theta} \in \bar{\Theta}_{\text{loc}}$, and by the remark after Lemma A.4, the basic inequality holds for $\tilde{\theta}$. Also, recall that $\bar{\mathcal{E}}(\bar{\theta}_0) = 0$:

$$\begin{aligned}
\bar{\mathcal{E}}(\tilde{\theta}) + \bar{\lambda}\|\tilde{\vartheta}\|_1 &\leq -(\bar{v}_n(\tilde{\theta}) - \bar{v}_n(\bar{\theta}_0)) + \bar{\mathcal{E}}(\bar{\theta}_0) + \bar{\lambda}\|\bar{\vartheta}_0\|_1 \\
&\leq Z_{M^*} + \bar{\lambda}\|\bar{\vartheta}_0\|_1 \\
&\leq \epsilon^* + \bar{\lambda}\|\bar{\vartheta}_0\|_1.
\end{aligned}$$

From now on write $\tilde{\mathcal{E}} = \bar{\mathcal{E}}(\tilde{\theta})$. Note, that $\|\tilde{\vartheta}\|_1 = \|\tilde{\vartheta}_{S_0^c}\|_1 + \|\tilde{\vartheta}_{S_0}\|_1$ and thus, by the triangle inequality,

$$\begin{aligned}
\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\vartheta}_{S_0^c}\|_1 &\leq \epsilon^* + \bar{\lambda}(\|\bar{\vartheta}_0\|_1 - \|\tilde{\vartheta}_{S_0}\|_1) \\
&\leq \epsilon^* + \bar{\lambda}(\|\bar{\vartheta}_0 - \tilde{\vartheta}_{S_0}\|_1) \\
&\leq \epsilon^* + \bar{\lambda}(\|\bar{\vartheta}_0 - \tilde{\vartheta}_{S_0}\|_1 + \|(\mu_0, \gamma_0^T)^T - (\tilde{\mu}, \tilde{\gamma}^T)^T\|_1) \\
&= \epsilon^* + \bar{\lambda}\|(\tilde{\theta} - \bar{\theta}_0)_{S_{0,+}}\|_1.
\end{aligned} \tag{A.5}$$

Case i) If $\bar{\lambda}\|(\tilde{\theta}-\bar{\theta}_0)_{S_{0,+}}\|_1 \geq \epsilon^*$, then

$$\bar{\lambda}\|\tilde{\vartheta}_{S_0^c}\|_1 \leq \tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\vartheta}_{S_0^c}\|_1 \leq 2\bar{\lambda}\|(\tilde{\theta}-\bar{\theta}_0)_{S_{0,+}}\|_1. \tag{A.6}$$

Since $\|(\tilde{\theta}-\bar{\theta}_0)_{S_{0,+}^c}\|_1 = \|\tilde{\vartheta}_{S_0^c}\|_1$, we may thus apply the compatibility condition, Corollary A.5 (note that $\bar{\vartheta}_0 = \bar{\vartheta}_{0,S_0}$) to obtain

$$\|(\tilde{\theta}-\bar{\theta}_0)_{S_{0,+}}\|_1 \leq \sqrt{2}\cdot\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}}-f_{\bar{\theta}_0}\|_{\bar{Q}},$$

where we have used that $\theta \mapsto f_\theta$ is linear and hence $f_{\tilde{\theta}-\bar{\theta}_0} = f_{\tilde{\theta}} - f_{\bar{\theta}_0}$. Observe that

$$\|\tilde{\theta}-\theta_0\|_1 = \|\tilde{\vartheta}_{S_0^c}\|_1 + \|(\tilde{\theta}-\theta_0)_{S_{0,+}}\|_1. \tag{A.7}$$

Hence,

$$\begin{aligned}
\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta}-\bar{\theta}_0\|_1 &= \tilde{\mathcal{E}} + \bar{\lambda}(\|\tilde{\vartheta}_{S_0^c}\|_1 + \|(\tilde{\theta}-\bar{\theta}_0)_{S_{0,+}}\|_1) \\
&\leq \epsilon^* + 2\bar{\lambda}\|(\tilde{\theta}-\bar{\theta}_0)_{S_{0,+}}\|_1 \\
&\leq \epsilon^* + 2\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}}-f_{\bar{\theta}_0}\|_{\bar{Q}}.
\end{aligned}$$

Recall that for a convex function G and its convex conjugate H we have $uv \leq G(u)+H(v)$. Thus, we obtain

$$\begin{aligned}
2\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\|f_{\tilde{\theta}}-f_{\bar{\theta}_0}\|_{\bar{Q}} &= 4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\frac{\|f_{\tilde{\theta}}-f_{\bar{\theta}_0}\|_{\bar{Q}}}{2} \\
&\leq H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + G_n\left(\frac{\|f_{\tilde{\theta}}-f_{\bar{\theta}_0}\|_{\bar{Q}}}{2}\right) \\
&\overset{G_n \text{ convex}}{\leq} H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + \frac{G_n(\|f_{\tilde{\theta}}-f_{\bar{\theta}_0}\|_{\bar{Q}})}{2} \\
&\overset{\text{margin condition}}{\leq} H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + \frac{\tilde{\mathcal{E}}}{2}.
\end{aligned}$$

It follows

$$\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta}-\bar{\theta}_0\|_1 \leq \epsilon^* + H_n\left(4\sqrt{2}\bar{\lambda}\frac{\sqrt{s_{0,+}}}{\sqrt{c_{\min}}}\right) + \frac{\tilde{\mathcal{E}}}{2} = 2\epsilon^* + \frac{\tilde{\mathcal{E}}}{2}$$

and therefore

$$\frac{\tilde{\mathcal{E}}}{2} + \bar{\lambda}\|\tilde{\theta}-\bar{\theta}_0\|_1 \leq 2\epsilon^*. \tag{A.8}$$

Finally, this gives

$$\|\tilde{\theta}-\bar{\theta}_0\|_1 \leq \frac{2\epsilon^*}{\bar{\lambda}} = \frac{2\lambda_0 M^*}{\bar{\lambda}} \underbrace{\leq}_{\bar{\lambda}\geq 4\lambda_0} \frac{M^*}{2}.$$

From this, by using the definition of $\tilde{\theta}$, we obtain

$$\|\tilde{\theta} - \bar{\theta}_0\|_1 = t\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 = \frac{M^*}{M^* + \|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1}\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 \leq \frac{M^*}{2}.$$

Rearranging gives

$$\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 \leq M^*.$$

Case ii) If $\bar{\lambda}\|(\bar{\theta}_0 - \tilde{\theta})_{S_{0,+}}\|_1 \leq \epsilon^*$, then from (A.5)

$$\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\vartheta}_{S_0^c}\|_1 \leq 2\epsilon^*.$$

Using once more (A.7), we get

$$\tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\theta} - \bar{\theta}_0\|_1 = \tilde{\mathcal{E}} + \bar{\lambda}\|\tilde{\vartheta}_{S_0^c}\|_1 + \bar{\lambda}\|(\tilde{\theta} - \bar{\theta}_0)_{S_{0,+}}\|_1 \leq 3\epsilon^*. \tag{A.9}$$

Thus,

$$\|\tilde{\theta} - \bar{\theta}_0\|_1 \leq 3\frac{\epsilon^*}{\bar{\lambda}} = 3\frac{\lambda_0}{\bar{\lambda}}M^* \leq \frac{M^*}{2}$$

by choice of $\lambda \geq 6\lambda_0$. Again, plugging in the definition of $\tilde{\theta}$, we obtain

$$\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 \leq M^*.$$

Hence, in either case we have $\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 \leq M^*$. That means, we can repeat the above steps with $\hat{\tilde{\theta}}$ instead of $\tilde{\theta}$: Writing $\hat{\mathcal{E}} \coloneqq \bar{\mathcal{E}}(\hat{\tilde{\theta}})$, following the same reasoning as above we arrive once more at (A.5):

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\tilde{\vartheta}}_{S_0^c}\|_1 \leq \epsilon^* + \bar{\lambda}\|\bar{\vartheta}^* - \hat{\tilde{\vartheta}}_{S_0}\|_1 \leq 2\epsilon^* + \bar{\lambda}\|(\hat{\tilde{\theta}} - \bar{\theta}_0)_{S_{0,+}}\|_1.$$

From this, in **case i)** we obtain (A.6) which allows us to use the compatibility condition to arrive at (A.8):

$$\frac{\hat{\mathcal{E}}}{2} + \bar{\lambda}\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 \leq 2\epsilon^*,$$

resulting in

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 \leq 4\epsilon^*.$$

In **case ii)** on the other hand, we arrive directly at (A.9), and hence

$$\hat{\mathcal{E}} + \bar{\lambda}\|\hat{\tilde{\theta}} - \bar{\theta}_0\|_1 \leq 3\epsilon^*.$$

Plugging in the definitions of $\hat{\tilde{\theta}}$ and $\bar{\theta}_0$ and using the fact that $\hat{\mathcal{E}} = \bar{\mathcal{E}}(\hat{\tilde{\theta}}) = \mathcal{E}(\hat{\theta})$ proves the claim. □

A.1.6 Controlling the special set $\mathcal{I}$

In this section we seek to control The expectation of Z_M. Recall the definition

$$Z_M \coloneqq \sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}} |\bar{v}_n(\bar{\theta}) - \bar{v}_n(\bar{\theta}_0)|,$$

where $\bar{v}_n$ denotes the rescaled empirical process. Recall, that there is a constant $c \in \mathbb{R}$ such that uniformly $|Z_{ij,k}| \leq c, 1 \leq i \neq j \leq n, k = 1, \ldots, p$.

Lemma A.7. *For any $M > 0$ we have in model* (4.1)

$$\mathbb{E}[Z_M] \leq 8M(1 \vee c)\sqrt{\frac{2\log(2(2n+p+1))}{N}}.$$

Proof. Let $\epsilon_{ij}, i \neq j$, be a Rademacher sequence independent of $A_{ij}, Z_{ij}, i \neq j$. We first want to use the Symmetrization Theorem 2.16: For the random variables $Z_1, \ldots, Z_n$ we choose $T_{ij} = (A_{ij}, \bar{X}_{ij}^T, 1, Z_{ij}^T)^T \in \{0,1\} \times \mathbb{R}^{2n+1+p}$. For any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ we consider the functions

$$g_{\bar{\theta}}(T_{ij}) = \frac{1}{N}\left\{-A_{ij}\bar{D}_{ij}^T(\bar{\theta}-\bar{\theta}_0) + \log(1+\exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0))\right\}$$

and the function set $\mathcal{G} = \mathcal{G}(M) \coloneqq \{g_{\bar{\theta}} : \bar{\theta} \in \bar{\Theta}_{\text{loc}}, \|\bar{\theta}-\bar{\theta}_0\|_1 \leq M\}$. Note, that

$$\bar{v}_n(\bar{\theta}) - \bar{v}_n(\bar{\theta}_0) = \sum_{i\neq j}\{g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})]\}.$$

Then, by the Symmetrization Theorem,

$$\begin{aligned}\mathbb{E}[Z_M] &= \mathbb{E}\left[\sup_{g_{\bar{\theta}}\in\mathcal{G}}\left|\sum_{i\neq j} g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})]\right|\right] \\ &\leq 2\mathbb{E}\left[\sup_{g_{\bar{\theta}}\in\mathcal{G}}\left|\sum_{i\neq j}\epsilon_{ij}g_{\bar{\theta}}(T_{ij})\right|\right].\end{aligned}$$

Next, we want to apply the Contraction Theorem 2.17. Denote $T = (T_{ij})_{i\neq j}$ and let $\mathbb{E}_T$ be the conditional expectation given T. We need the conditional expectation at this point, because Theorem 2.17 requires non-random arguments in the functions. This does not hinder us, as later we will simply take iterated expectations, cancelling out the conditional expectation, see below. For the functions g_i in Theorem 2.17 we choose

$$g_{ij}(x) = \frac{1}{2}\{-A_{ij}x + \log(1+\exp(x))\}$$

Note, that $\log(1+\exp(x))$ has derivative bounded by one and thus is Lipschitz continuous with constant one by the Mean Value Theorem. Thus, all g_{ij} are also

Lipschitz continuous with constant 1:

$$|g_{ij}(x) - g_{ij}(x')| \leq \frac{1}{2}\{|A_{ij}(x-x')| + |\log(1+\exp(x)) - \log(1+\exp(x'))|\} \leq |x-x'|.$$

For the function class $\mathcal{F}$ in Theorem 2.17 we choose $\mathcal{F} = \mathcal{F}_M \coloneqq \{f_{\bar{\theta}} : \bar{\theta} \in \bar{\Theta}_{\text{loc}}, \|\bar{\theta} - \bar{\theta}_0\|_1 \leq M\}$ and pick $f^* = f_{\bar{\theta}_0}$. Then, by Theorem 2.17

$$\begin{aligned}
&\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(g_{ij}(f_{\bar{\theta}}((\bar{X}_{ij}^T,1,Z_{ij}^T)^T)) - g_{ij}(f_{\bar{\theta}_0}((\bar{X}_{ij}^T,1,Z_{ij}^T)^T)))\right|\right] \\
&\quad\leq 2\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(f_{\bar{\theta}}((\bar{X}_{ij}^T,1,Z_{ij}^T)^T) - f_{\bar{\theta}_0}((\bar{X}_{ij}^T,1,Z_{ij}^T)^T))\right|\right].
\end{aligned}$$

Recall that we can express the functions $f_{\bar{\theta}} = f_{\bar{\alpha},\bar{\beta},\mu,\gamma}$ as

$$f_{\bar{\alpha},\bar{\beta},\mu,\gamma}(\,.\,) = \sum_{i=1}^{n}\bar{\alpha}_i e_i(\,.\,) + \sum_{i=n+1}^{2n}\bar{\beta}_{i-n}e_i(\,.\,) + \mu e_{2n+1}(\,.\,) + \sum_{i=1}^{p}\gamma_i e_{2n+1+i}(\,.\,),$$

where $e_i(\,.\,)$ is the projection on the i-th coordinate. Consider any $\bar{\theta} \in \bar{\Theta}_{\text{loc}}$ with $\|\bar{\theta} - \bar{\theta}_0\|_1 \leq M$. For the sake of a compact representation we use our shorthand notation $\bar{\theta} = (\bar{\theta}_i)_{i=1}^{2n+1+p}$ where the components θ_i are defined in the canonical way and we also simply write $e_k(\bar{X}_{ij}, 1, Z_{ij})$ for the projection of the the vector $(\bar{X}_{ij}^T, 1, Z_{ij}^T)^T \in \mathbb{R}^{2n+p+1}$ to its k-th component, i.e. instead of $e_k((\bar{X}_{ij}^T, 1, Z_{ij}^T)^T)$. Then,

$$\begin{aligned}
&\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(f_{\bar{\theta}}((\bar{X}_{ij}^T,1,Z_{ij}^T)^T) - f_{\bar{\theta}_0}((\bar{X}_{ij}^T,1,Z_{ij}^T)^T))\right| \\
&= \left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}\left(\sum_{k=1}^{2n+p+1}(\bar{\theta}_k - \bar{\theta}_{0,k})e_k(\bar{X}_{ij},1,Z_{ij})\right)\right| \\
&\leq \frac{1}{N}\sum_{k=1}^{2n+p+1}\left\{|\bar{\theta}_k - \bar{\theta}_{0,k}|\max_{1\leq l\leq 2n+p+1}\left|\sum_{i\neq j}\epsilon_{ij}e_l(\bar{X}_{ij},1,Z_{ij})\right|\right\} \\
&\leq M\max_{1\leq l\leq 2n+p+1}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}e_l(\bar{X}_{ij},1,Z_{ij})\right|.
\end{aligned}$$

Note, that the last expression no longer depends on $\bar{\theta}$. To bind the right hand side in the last expression we use Lemma 2.19: In the language of the Lemma, choose $Z_1, \ldots, Z_n$ as $T_{ij} = (\epsilon_{ij}, \bar{X}_{ij}^T, 1, Z_{ij}^T)^T$. We choose for the p in the formulation of the Lemma $2n + p + 1$ and pick for our functions

$$g_k(T_{ij}) = \frac{1}{N}\epsilon_{ij}e_k(\bar{X}_{ij}, 1, Z_{ij}), k = 1, \ldots, 2n + p + 1.$$

Then, $\mathbb{E}[g_k(T_{ij})] = 0$. We want to employ Lemma 2.19 which requires us to bound $|g_k(T_{ij})| \leq c_{ij,k}$ for all $i \neq j$ and $k = 1, \ldots, n+1+p$. For any fixed $1 \leq k \leq n$ we have

$$|g_k(T_{ij})| \leq \begin{cases} \frac{\sqrt{n}}{N} = \frac{1}{(n-1)\sqrt{n}}, & i \text{ or } j = k \\ 0, & \text{otherwise.} \end{cases}$$

The first case occurs exactly $(n-1)$ times for each k. Thus, for any $k \leq 2n$,

$$\sum_{i \neq j} c_{ij,k}^2 = \left(\frac{1}{(n-1)\sqrt{n}}\right)^2 (n-1) = \frac{1}{N}.$$

If $k = 2n+1$, $|g_k(T_{ij})| = 1/N$ and hence

$$\sum_{i \neq j} c_{ij,2n+1}^2 = \frac{1}{N}.$$

Finally, if $k > 2n+1$, $|g_k(T_{ij})| \leq c/N$ and therefore,

$$\sum_{i \neq j} c_{ij,k}^2 \leq \frac{c^2}{N}.$$

In total, this means

$$\max_{1 \leq k \leq 2n+1+p} \sum_{i \neq j} c_{ij,k}^2 \leq \frac{1 \vee c^2}{N}.$$

Therefore, an application of Lemma 2.19 results in

$$\begin{aligned} &\mathbb{E}\left[\max_{1 \leq l \leq 2n+p+1} \left| \frac{1}{N} \sum_{i \neq j} \epsilon_{ij} e_l(\bar{X}_{ij}, Z_{ij}) \right| \right] \\ &\qquad \leq \sqrt{2\log(2(2n+1+p))} \max_{1 \leq k \leq 2n+1+p} \left[\sum_{i \neq j} c_{ij,k}^2 \right]^{1/2} \\ &\qquad \leq \sqrt{\frac{2\log(2(2n+1+p))}{N}} (1 \vee c). \end{aligned}$$

Putting everything together, we obtain

$$\begin{aligned}
\mathbb{E}[Z_M] &\leq 2\mathbb{E}\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(-A_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},1,Z_{ij})-f_{\bar{\theta}_0}(\bar{X}_{ij},1,Z_{ij})))\right|\right] \\
&= 2\mathbb{E}\left[\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(-A_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},1,Z_{ij})-f_{\bar{\theta}_0}(\bar{X}_{ij},1,Z_{ij})))\right|\right]\right] \\
&\leq 8\mathbb{E}\left[\mathbb{E}_T\left[\sup_{\substack{\bar{\theta}\in\bar{\Theta}_{\text{loc}},\\ \|\bar{\theta}-\bar{\theta}_0\|_1\leq M}}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}(f_{\bar{\theta}}(\bar{X}_{ij},1,Z_{ij})-f_{\bar{\theta}_0}(\bar{X}_{ij},1,Z_{ij}))\right|\right]\right] \\
&\leq 8M\mathbb{E}\left[\mathbb{E}_T\left[\max_{1\leq l\leq 2n+p+1}\left|\frac{1}{N}\sum_{i\neq j}\epsilon_{ij}e_l(\bar{X}_{ij},1,Z_{ij})\right|\right]\right] \\
&\leq 8M\sqrt{\frac{2\log(2(2n+1+p))}{N}}(1\vee c).
\end{aligned}$$

This concludes the proof. □

We now show that Z_M does not deviate too far from its expectation. The proof relies on the concentration theorem due to Bousquet, Theorem 2.18.

Corollary A.8. *Pick any confidence level $t>0$. Let*

$$a_n \coloneqq \sqrt{\frac{2\log(2(2n+p+1))}{N}}(1\vee c)$$

and choose $\lambda_0=\lambda_0(t,n)$ as

$$\lambda_0 = 8a_n + 2\sqrt{\frac{t}{N}(11(1\vee(c^2p))+16(1\vee c)\sqrt{n}a_n)} + \frac{4t(1\vee c)\sqrt{n}}{3N}$$

Then, we have the inequality

$$P\left(Z_M \geq M\lambda_0\right) \leq \exp(-t).$$

Proof. We want to apply Bousquet's Concentration Theorem 2.18. For the random variables Z_i in the formulation of the theorem we choose once more $T_{ij}=(A_{ij},\bar{X}_{ij},1,Z_{ij}), i\neq j$, and as functions we consider

$$\begin{aligned}
g_{\bar{\theta}}(T_{ij}) &= -A_{ij}\bar{D}_{ij}^T(\bar{\theta}-\bar{\theta}_0)+\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)), \\
\mathcal{G}=\mathcal{G}_M &\coloneqq \{g_{\bar{\theta}}:\bar{\theta}\in\bar{\Theta}_{\text{loc}},\|\bar{\theta}-\bar{\theta}_0\|_1\leq M\}.
\end{aligned}$$

Then, we have

$$Z_M = \sup_{g_{\bar{\theta}} \in \mathcal{G}} \frac{1}{N} \left| \sum_{i \neq j} \{ g_{\bar{\theta}}(T_{ij}) - \mathbb{E}[g_{\bar{\theta}}(T_{ij})] \} \right|.$$

To apply Theorem 2.18, we need to bound $\|g_{\bar{\theta}}\|_\infty$. Recall that we denote the distribution of $[\bar{X}|1|Z]$ by $\bar{Q}$ and $\|g_{\bar{\theta}}\|_\infty$ is defined as the $\bar{Q}$-a.s. smallest upper bound on the value of $g_{\bar{\theta}}$. We have for any $g_{\bar{\theta}} \in \mathcal{G}$, using the Lipschitz continuity of $\log(1+\exp(x))$:

$$\begin{aligned}
|g_{\bar{\theta}}(T_{ij})| &\leq |\bar{D}_{ij}^T(\bar{\theta} - \bar{\theta}_0)| + |\log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0))| \\
&\leq 2|\bar{D}_{ij}^T(\bar{\theta} - \bar{\theta}_0)| \\
&\leq 2\|\vartheta - \vartheta_0\|_1 + |\mu - \mu_0| + c\|\gamma - \gamma_0\|_1.
\end{aligned}$$

Thus,

$$\begin{aligned}
\|g_{\bar{\theta}}\|_\infty &\leq 2\|\vartheta - \vartheta_0\|_1 + |\mu - \mu_0| + c\|\gamma - \gamma_0\|_1 \\
&\leq 2(1 \vee c)\|\theta - \theta_0\|_1 \\
&\leq 2(1 \vee c)\sqrt{n}M =: \eta_n.
\end{aligned}$$

For the last inequality we used that for any θ with $\|\bar{\theta} - \bar{\theta}_0\|_1 \leq M$ it follows that $\|\theta - \theta_0\|_1 \leq \sqrt{n}M$, which is possibly a very generous upper bound. This does not matter, however, as the term associated with the above bound will be negligible, as we shall see.

The second requirement of Theorem 2.18 is that the average variance of $g_{\bar{\theta}}(T_{ij})$ has to be uniformly bounded. To that end we calculate

$$\begin{aligned}
&\frac{1}{N}\sum_{i \neq j} \mathrm{Var}(g_{\bar{\theta}}(T_{ij})) \\
&\quad = \frac{1}{N}\sum_{i \neq j} \mathrm{Var}(-A_{ij}D_{ij}^T(\theta - \theta_0)) \\
&\quad + \frac{1}{N}\sum_{i \neq j} \mathrm{Var}(\log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0))) \\
&\quad + \frac{2}{N}\sum_{i \neq j} \mathrm{Cov}(-A_{ij}D_{ij}^T(\theta - \theta_0), \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta})) - \log(1 + \exp(\bar{D}_{ij}^T\bar{\theta}_0))).
\end{aligned}$$

Let us look at these terms in term. For the first term, we obtain

$$\begin{aligned}
\frac{1}{N}\sum_{i \neq j} \mathrm{Var}(-A_{ij}D_{ij}^T(\theta - \theta_0)) &\leq \frac{1}{N}\sum_{i \neq j} \mathbb{E}[(-A_{ij}D_{ij}^T(\theta - \theta_0))^2] \\
&\leq \mathbb{E}\left[\frac{1}{N}\sum_{i \neq j} (D_{ij}^T(\theta - \theta_0))^2 \right].
\end{aligned}$$

For the second term, we get by Lipschitz continuity,

$$\begin{aligned}
&\frac{1}{N}\sum_{i\neq j}\mathrm{Var}(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))\\
&\qquad\leq \frac{1}{N}\sum_{i\neq j}\mathbb{E}[(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))^2]\\
&\qquad\leq \mathbb{E}\left[\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2\right].
\end{aligned}$$

The last term decomposes as

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\mathrm{Cov}(-A_{ij}D_{ij}^T(\theta-\theta_0),\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))\\
&\quad=\frac{2}{N}\sum_{i\neq j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)\cdot(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))]\\
&\quad\quad-\frac{2}{N}\sum_{i\neq j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)]\cdot\mathbb{E}[\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0))]
\end{aligned}$$

For the first term in that decomposition we have

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\left|\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)\cdot(\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))]\right|\\
&\quad\leq\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|\cdot|\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0))|]\\
&\quad\leq\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|^2]
\end{aligned}$$

and for the second term using the same arguments, we get

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\mathbb{E}[-A_{ij}D_{ij}^T(\theta-\theta_0)]\cdot\mathbb{E}[\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0))]\\
&\qquad\qquad\leq\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2.
\end{aligned}$$

Meaning that in total

$$\begin{aligned}
&\frac{2}{N}\sum_{i\neq j}\left|\mathrm{Cov}(-A_{ij}D_{ij}^T(\theta-\theta_0),\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}))-\log(1+\exp(\bar{D}_{ij}^T\bar{\theta}_0)))\right|\\
&\qquad\leq\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|^2]+\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2.
\end{aligned}$$

In total, we thus get

$$\frac{1}{N}\sum_{i\neq j}\mathrm{Var}(g_{\bar{\theta}}(T_{ij})) \leq 4\cdot\mathbb{E}\left[\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2\right] + \frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2. \tag{A.10}$$

Furthermore, by Cauchy-Schwarz' inequality,

$$\begin{aligned}
&\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2 \\
&\quad= \frac{1}{N}\sum_{i\neq j}(\alpha_i+\beta_j+\mu-\alpha_{0,i}-\beta_{0,j}-\mu_0+(\gamma-\gamma_0)^T Z_{ij})^2 \\
&\quad\leq \frac{4}{N}\sum_{i\neq j}\left\{(\alpha_i-\alpha_{0,i})^2+(\beta_j-\beta_{0,j})^2+(\mu-\mu_0)^2+((\gamma-\gamma_0)^T Z_{ij})^2\right\}.
\end{aligned}$$

Recall that for any $x\in\mathbb{R}^p$, $\|x\|_2\leq\|x\|_1\leq\sqrt{p}\|x\|_2$ and note that

$$|(\gamma-\gamma_0)^T Z_{ij}| \leq c\|\gamma-\gamma_0\|_1 \leq c\sqrt{p}\|\gamma-\gamma_0\|_2.$$

Then, from the above

$$\begin{aligned}
&\frac{1}{N}\sum_{i\neq j}(D_{ij}^T(\theta-\theta_0))^2 \\
&\quad\leq \frac{4}{N}\sum_{i\neq j}\left\{(\alpha_i-\alpha_{0,i})^2+(\beta_j-\beta_{0,j})^2+(\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2\right\} \\
&\quad= 4\left((\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2+\frac{1}{N}(n-1)\|\vartheta-\vartheta_0\|_2^2\right) \\
&\quad= 4\left((\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2+\left\|\frac{1}{\sqrt{n}}(\vartheta-\vartheta_0)\right\|_2^2\right) \\
&\quad= 4\left((\mu-\mu_0)^2+c^2p\|\gamma-\gamma_0\|_2^2+\|\bar{\vartheta}-\bar{\vartheta}_0\|_2^2\right) \\
&\quad\leq 4(1\vee(c^2p))\|\bar{\theta}-\bar{\theta}_0\|_2^2 \\
&\quad\leq 4(1\vee(c^2p))\|\bar{\theta}-\bar{\theta}_0\|_1^2 \\
&\quad\leq 4(1\vee(c^2p))M^2.
\end{aligned} \tag{A.11}$$

For the second summand on the right-hand side in (A.10), we have

$$\begin{aligned}
\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2 &= \frac{2}{N}\sum_{i\neq j}(\alpha_i+\beta_j+\mu-\alpha_{0,i}-\beta_{0,j}-\mu_0+(\gamma-\gamma_0)^T\mathbb{E}[Z_{ij}])^2 \\
&= \frac{2}{N}\sum_{i\neq j}(\alpha_i+\beta_j+\mu-\alpha_{0,i}-\beta_{0,j}-\mu_0)^2.
\end{aligned}$$

So that we may use the same steps as in (A.11) to conclude that

$$\frac{2}{N}\sum_{i\neq j}\mathbb{E}[|D_{ij}^T(\theta-\theta_0)|]^2 \leq 6(1\vee(c^2p))M^2.$$

Such that in total,

$$\frac{1}{N}\sum_{i\neq j}\mathrm{Var}(g_{\bar{\theta}}(T_{ij})) \leq 22(1\vee(c^2p))M^2 \coloneqq \tau_n^2.$$

Applying Bousquet's Concentration Theorem 2.18 with η_n, τ_n defined above, we obtain for all $z > 0$

$$\begin{aligned}\exp\left(-Nz^2\right) &\geq P\left(Z_M \geq \mathbb{E}[Z_M] + z\sqrt{2(\tau_n^2 + 2\eta_n\mathbb{E}[Z_M])} + \frac{2z^2\eta_n}{3}\right)\\ &= P\left(Z_M \geq \mathbb{E}[Z_M] + z\sqrt{2(22(1\vee(c^2p))M^2 + 4(1\vee c)\sqrt{n}M\mathbb{E}[Z_M])} + \frac{4z^2(1\vee c)\sqrt{n}M}{3}\right).\end{aligned} \tag{A.12}$$

From Lemma A.7, we know

$$\mathbb{E}[Z_M] \leq 8M\sqrt{\frac{2\log(2(2n+p+1))}{N}}(1\vee c) = 8Ma_n.$$

Using this, we obtain from (A.12)

$$\begin{aligned}&\exp\left(-Nz^2\right)\\ &\quad\geq P\left(Z_M \geq 8Ma_n + z\sqrt{2(22(1\vee(c^2p))M^2 + 32(1\vee c)\sqrt{n}M^2a_n)} + \frac{4z^2(1\vee c)\sqrt{n}M}{3}\right)\\ &\quad= P\left(Z_M \geq M\left(8a_n + 2z\sqrt{11(1\vee(c^2p)) + 16(1\vee c)\sqrt{n}a_n} + \frac{4z^2(1\vee c)\sqrt{n}}{3}\right)\right).\end{aligned}$$

Now, pick $z = \sqrt{t/N}$ to get

$$\begin{aligned}&\exp(-t)\\ &\quad\geq P\left(Z_M \geq M\left(8a_n + 2\sqrt{\frac{t}{N}(11(1\vee(c^2p)) + 16(1\vee c)\sqrt{n}a_n)} + \frac{4t(1\vee c)\sqrt{n}}{3N}\right)\right).\end{aligned}$$

which is the claim. □

A.1.7 Putting it all together

Proof of Theorem 4.4. Theorem 4.4 follows from Theorem A.6 and Corollary A.8. Recall the definition of K_n in (A.3), which simplifies to

$$K_n = 2\frac{(1+\exp(r_{n,0}))^2}{\exp(r_{n,0})} = 2\frac{(1+\exp\left(-\mathrm{logit}(\rho_{n,0})\right))^2}{\exp\left(-\mathrm{logit}(\rho_{n,0})\right)} \leq \frac{4}{\rho_{n,0}}.$$

Thus, under the conditions of Theorem 4.4, we have with high probability by Theorem A.6 and Corollary A.8,

$$\mathcal{E}(\hat{\theta}) + \bar{\lambda}\left(\frac{1}{\sqrt{n}}\|\hat{\vartheta} - \vartheta_0\|_1 + |\hat{\mu} - \mu_0| + \|\hat{\gamma} - \gamma_0\|_1\right) \leq C\frac{s_{0,+}\bar{\lambda}^2}{\rho_{n,0}}.$$

with constant $C = 128/c_{\min}$. □

A.2 Proof of Theorem 4.5

Recall our discussion of the KKT conditions in Section 4.2.1. By the same arguments we find that 0 has to be contained in the subdifferential of $\frac{1}{N}\mathcal{L}(\theta)+\lambda\|\beta\|_1$ at $\hat{\theta}$, where this time we consider the KKT conditions with respect to the original parameters θ. That is, there exists a $\hat{z} \in \mathbb{R}^{2n+1+p}$ such that

$$0 = \frac{1}{N}\nabla\, \mathcal{L}(\theta)|_{\theta=\hat{\theta}} + \lambda\hat{z}, \tag{A.13}$$

where $\nabla\, \mathcal{L}(\theta)|_{\theta=\hat{\theta}}$ is the gradient of $\mathcal{L}(\theta)$ evaluated at $\hat{\theta}$ and for $i = 1, \ldots, 2n$, $\hat{z}_i = 1$ if $\hat{\vartheta}_i > 0$ and $\hat{z}_i \in [-1, 1]$ if $\hat{\vartheta}_i = 0$, and for $i = 2n+1, \ldots, 2n+1+p$, $\hat{z}_i = 0$.

Denoting $\nabla_\xi\, \mathcal{L}(\theta)|_{\theta=\hat{\theta}} \in \mathbb{R}^{p+1}$ the gradient of $\mathcal{L}$ with respect to the unpenalized parameters $\xi = (\mu, \gamma^T)^T$ only, evaluated at $\hat{\theta}$, we have

$$0 = \nabla_\xi\, \mathcal{L}(\theta)|_{\theta=\hat{\theta}}\,. \tag{A.14}$$

Recall the form of the Hessian $H(\hat{\theta}) \coloneqq H_{\xi\times\xi}(\theta)|_{\theta=\hat{\theta}}$ of $\frac{1}{N}\mathcal{L}(\theta)$ with respect to ξ only, evaluated at $\hat{\theta}$:

$$H(\hat{\theta}) = \frac{1}{N} D_\xi^T \hat{W}^2 D_\xi,$$

where $D_\xi = [\mathbf{1}|Z]$ is the part of the design matrix D corresponding to ξ with rows $D_{\xi,ij}^T = (1, Z_{ij}^T), i \neq j$, and

$$\hat{W} = \operatorname{diag}\left(\sqrt{p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))}, i \neq j\right).$$

Also recall the corresponding population version

$$\mathbb{E}[H(\theta_0)] = \frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi],$$

where $W_0 = \operatorname{diag}(\sqrt{p_{ij}(\theta_0)(1-p_{ij}(\theta_0))}, i \neq j)$. Finally, recall that to be consistent with commonly used notation, we write $\hat{\Sigma}_\xi = H(\hat{\theta}) = \frac{1}{N} D_\xi^T \hat{W}^2 D_\xi$ and $\Sigma_\xi = \mathbb{E}[H(\theta_0)] = \frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi]$ and $\hat{\Theta}_\xi \coloneqq \hat{\Sigma}_\xi^{-1}, \Theta_\xi \coloneqq \Sigma_\xi^{-1}$.

A.2.1 Inverting population and sample Gram matrices

Note that the function $f(x) = x(1-x)$ is monotonically increasing in x for $x \leq 1/2$ and monotonically decreasing in x for $x \geq 1/2$. Thus, by considering the cases $p_{ij} \leq 1/2$ and $p_{ij} \geq 1/2$ separately and using that $\rho_n \leq 1/2$, we may employ the following lower bound for all $i \neq j$: $p_{ij}(\theta_0)(1-p_{ij}(\theta_0)) \geq 1/2\rho_n$. Also, recall that by Assumption 4.1, for the minimum eigenvalue $\lambda_{\min}$ of $\mathbb{E}[Z^T Z/N]$ we have $\lambda_{\min} = \lambda_{\min}(n) \geq c_{\min} > 0$. Then, for any n and $v \in \mathbb{R}^{p+1}\backslash\{0\}$ with components

$v = (v_1, v_R^T)^T, v_R \in \mathbb{R}^p$, we have

$$\begin{aligned} v^T \Sigma_\xi v &\geq \frac{1}{2}\rho_n v^T \frac{1}{N}\mathbb{E}[D_\xi^T D_\xi] v = \frac{1}{2}\rho_n v^T \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \frac{1}{N}\mathbb{E}[Z^T Z] \end{pmatrix} v \\ &= \frac{1}{2}\rho_n \left(v_1^2 + v_R^T \frac{1}{N}\mathbb{E}[Z^T Z] v_R \right) \\ &\geq \frac{1}{2}\rho_n (v_1^2 + \lambda_{\min}\|v_R\|_2^2) \geq \frac{1}{2}\rho_n (1 \wedge c_{\min})\|v\|_2^2 > 0. \end{aligned}$$

Hence, for finite n all eigenvalues of Σ_ξ are strictly positive and consequently this matrix is invertible. We want to show that the same holds with high probability for the sample matrix $\hat{\Sigma}_\xi$. Using the tools deployed in the proofs of Lemma 2.8 and A.1 we can show that with high probability the minimum eigenvalue of $D_\xi^T D_\xi / N$ is also strictly larger than zero and thus, $D_\xi^T D_\xi / N$ is invertible with high probability. From this the desired properties of $\hat{\Sigma}_\xi$ follow.

More precisely, recall the definition of $\kappa(A, m)$ for square matrices A and dimensions m. We want to consider the expression $\kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right)$ which simplifies to

$$\kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) \coloneqq \min_{v \in \mathbb{R}^{p+1}\setminus\{0\}} \frac{v^T \frac{1}{N}\mathbb{E}[D_\xi^T D_\xi] v}{\frac{1}{p+1}\|v\|_1^2}$$

and compare it to $\kappa^2\left(\frac{1}{N} D_\xi^T D_\xi, p+1\right)$. By Assumption 4.1 and the argument above, we have

$$\kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) \geq C > 0$$

for a universal constant C independent of n. By Lemma 2.8, with

$$\delta = \max_{kl} \left| \left(\frac{1}{N} D_\xi^T D_\xi\right)_{kl} - \left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi]\right)_{kl} \right|,$$

we have

$$\kappa^2\left(\frac{1}{N} D_\xi^T D_\xi, p+1\right) \geq \kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) - 16\delta(p+1).$$

By looking at the proof of Lemma 2.8, we see that in this particular case we do not even need the factor $16(p+1)$ on the right hand side above, but this does not matter anyway, so we keep it. By the exact same arguments we have used in the proof of Lemma A.1 for the blocks ⑤, ⑥, ⑧ and ⑨, we now get

$$\delta = O_P\left(N^{-1/2}\right).$$

Thus, for n large enough, we have with high probability $\delta \leq \frac{\lambda_{\min}}{32}$. Then, by Lemma

2.8, with high probability and uniformly in n,

$$\kappa^2\left(\frac{1}{N}D_\xi^T D_\xi, p+1\right) \geq \kappa^2\left(\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi], p+1\right) - 16\delta(p+1) \geq \frac{\lambda_{\min}(p+1)}{2} \geq C > 0.$$

If $\kappa^2\left(\frac{1}{N}D_\xi^T D_\xi, p+1\right) \geq C > 0$ uniformly in n, then for any $v \neq 0, v^T\frac{1}{N}D_\xi^T D_\xi v \geq C\|v\|_2^2$. Thus, for any $v \in \mathbb{R}^{p+1}\backslash\{0\}$ and any finite n:

$$\frac{1}{N}v^T D_\xi^T \hat{W}^2 D_\xi v \geq \min_{i\neq j}\{p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))\}\left(v^T\frac{1}{N}D_\xi^T D_\xi v\right) \geq C\rho_n\|v\|_2^2 > 0.$$

Thus, for every finite n, $\frac{1}{N}D_\xi^T\hat{W}^2 D_\xi$ is positive definite and invertible with high probability.

A.2.2 Goal and approach

Goal: We want to show that for $k = 1, \ldots, p+1$,

$$\sqrt{N}\frac{\hat{\xi}_k - \xi_{0,k}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \to \mathcal{N}(0,1).$$

Approach: Recall the definition of the "one-sample-version" of $\mathcal{L}$: For $\theta \in \Theta$,

$$l_\theta : \{0,1\} \times \mathbb{R}^{2n+1+p} \to \mathbb{R}$$
$$l_\theta(y,x) \coloneqq -y\theta^T x + \log(1+\exp(\theta^T x).$$

Then,

$$\mathcal{L}(\theta) = \sum_{i\neq j} l_\theta(A_{ij}, D_{ij}^T)$$

and

$$\nabla\mathcal{L}(\theta) = \sum_{i\neq j}\nabla l_\theta(A_{ij}, D_{ij}^T), \quad H\mathcal{L}(\theta) = \sum_{i\neq j} H l_\theta(A_{ij}, D_{ij}^T),$$

where H denotes the Hessian with respect to θ. Consider l_θ as a function in $\theta^T x$ and introduce:

$$l(y,a) \coloneqq -ya + \log(1+\exp(a)), \tag{A.15}$$

with second derivative: $\ddot{l}(y,a) = \partial_{a^2} l(y,a) = \frac{\exp(a)}{(1+\exp(a))^2}$. Note, that $\partial_{a^2} l(y,a)$ is Lipschitz continuous (it has bounded derivative $|\partial_{a^3} l(y,a)| \leq 1/(6\sqrt{3})$; Lipschitz continuity then follows by the Mean Value Theorem). Doing a first-order Taylor expansion in a of $\dot{l}(y,a) = \partial_a l(y,a)$ in the point $(A_{ij}, D_{ij}^T\theta_0)$ evaluated at $(A_{ij}, D_{ij}^T\hat{\theta})$, we get

$$\partial_a l(A_{ij}, D_{ij}\hat{\theta}) = \partial_a l(A_{ij}, D_{ij}^T\theta_0) + \partial_{a^2} l(A_{ij}, \alpha)D_{ij}^T(\hat{\theta} - \theta_0), \tag{A.16}$$

for an α between $D_{ij}^T\hat{\theta}$ and $D_{ij}^T\theta_0$. By Lipschitz continuity of $\partial_{a^2} l$, we also find

$$\begin{aligned}|\partial_{a^2} l(A_{ij},\alpha)D_{ij}^T(\hat{\theta}-\theta_0) - \partial_{a^2} l(A_{ij},D_{ij}^T\hat{\theta})D_{ij}^T(\hat{\theta}-\theta_0)| \\ \leq |\alpha - D_{ij}^T\hat{\theta}||D_{ij}^T(\hat{\theta}-\theta_0)| \leq |D_{ij}^T(\hat{\theta}-\theta_0)|^2,\end{aligned} \tag{A.17}$$

where the last inequality follows, because α is between $D_{ij}^T\hat{\theta}$ and $D_{ij}^T\theta_0$.

Consider the vector $P_n\nabla l_{\hat{\theta}}$: By equation (A.16), with α_{ij} between $D_{ij}^T\hat{\theta}$ and $D_{ij}^T\theta_0$,

$$\begin{aligned}P_n\nabla l_{\hat{\theta}} &= \frac{1}{N}\sum_{i\neq j}\left(\partial_{\theta_k} l(A_{ij},D_{ij}^T\hat{\theta})\right)_{k=1,\ldots,2n+1+p}, \quad \text{as a } (2n+1+p)\times 1\text{-vector} \\ &= \frac{1}{N}\sum_{i\neq j}\dot{l}(A_{ij},D_{ij}^T\hat{\theta})D_{ij} \\ &= \frac{1}{N}\sum_{i\neq j}(\dot{l}(A_{ij},D_{ij}^T\theta_0) + \ddot{l}(A_{ij},\alpha_{ij})D_{ij}^T(\hat{\theta}-\theta_0))D_{ij}\end{aligned}$$

which by (A.17) gives

$$= P_n\nabla l_{\theta_0} + \frac{1}{N}\sum_{i\neq j} D_{ij}\left\{\ddot{l}(A_{ij},D_{ij}^T\hat{\theta})D_{ij}^T(\hat{\theta}-\theta_0) + O(|D_{ij}^T(\hat{\theta}-\theta_0)|^2)\right\}.$$

Noticing that $\ddot{l}(A_{ij},D_{ij}^T\hat{\theta}) = p_{ij}(\hat{\theta})(1-p_{ij}(\hat{\theta}))$ and thus $\sum_{i\neq j}\ddot{l}(A_{ij},D_{ij}^T\hat{\theta})D_{ij}D_{ij}^T(\hat{\theta}-\theta_0) = D^T\hat{W}^2D(\hat{\theta}-\theta_0)$:

$$\begin{aligned}&= P_n\nabla l_{\theta_0} + P_nHl_{\hat{\theta}}(\hat{\theta}-\theta_0) + O\left(\frac{1}{N}\sum_{i\neq j} D_{ij}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right) \\ &= P_n\nabla l_{\theta_0} + \frac{1}{N}D^T\hat{W}^2D(\hat{\theta}-\theta_0) + O\left(\frac{1}{N}\sum_{i\neq j} D_{ij}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right),\end{aligned}$$

where the O notation is to be understood componentwise. Above, we have equality of two $((2n+1+p)\times 1)$-vectors. We are only interested in the portion relating to $\xi = (\mu,\gamma^T)^T$, i.e. in the last $p+1$ entries. Introduce the $((2n+1+p)\times(2n+1+p))$-matrix

$$A = \begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\Theta}_\xi\end{pmatrix},$$

where $\mathbf{0}$ are zero-matrices of appropriate dimensions. Multiplying the above with A from the left on both sides gives:

$$AP_n\nabla l_{\hat{\theta}} = AP_n\nabla l_{\theta_0} + A\frac{1}{N}D^T\hat{W}^2D(\hat{\theta}-\theta_0) + AO\left(\frac{1}{N}\sum_{i\neq j} D_{ij}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right). \tag{A.18}$$

Let us consider these terms in turn: Multiplication by A means that the first n entries of any of the vectors above are zero. Hence we only need to consider the last $p+1$ entries. The left-hand side of (A.18) is equal to zero by (A.14). The last $p+1$ entries of the first term on the right-hand side are $\hat{\Theta}_\xi P_n\nabla_\xi l_{\theta_0}$. For the second term

on the right hand side, notice that

$$\frac{1}{N}D^T\hat{W}^2D = \frac{1}{N}\begin{bmatrix} X^T\hat{W}^2X & X^T\hat{W}^2\mathbf{1} & X^T\hat{W}^2Z \\ \mathbf{1}^T\hat{W}^2X & \mathbf{1}^T\hat{W}^2\mathbf{1} & \mathbf{1}^T\hat{W}^2Z \\ Z^T\hat{W}^2X & Z^T\hat{W}^2\mathbf{1} & Z^T\hat{W}^2Z \end{bmatrix}.$$

Since $\hat{\Theta}_\xi = \hat{\Sigma}_\xi^{-1}$ and $\hat{\Sigma}_\xi^{-1}$ is the lower-right $(p+1)\times(p+1)$ block of above matrix,

$$A\frac{1}{N}D^T\hat{W}^2D = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \hat{\Theta}_\xi\frac{1}{N}D_\xi^T\hat{W}^2X & I_{(p+1)\times(p+1)} \end{bmatrix}.$$

Then, for the last $p+1$ entries of $A\frac{1}{N}D^T\hat{W}^2D(\hat{\theta}-\theta_0)$

$$\left(A\frac{1}{N}D^T\hat{W}^2D(\hat{\theta}-\theta_0)\right)_{\text{last } p+1 \text{ entries}} = \hat{\Theta}_\xi\frac{1}{N}D_\xi^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0) + \begin{pmatrix}\hat{\mu}-\mu_0\\ \hat{\gamma}-\gamma_0\end{pmatrix}.$$

Thus, (A.18) implies

$$0 = \hat{\Theta}_\xi P_n\nabla_\gamma l_{\theta_0} + \hat{\Theta}_\xi\frac{1}{N}D_\xi^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0) + \begin{pmatrix}\hat{\mu}-\mu_0\\ \hat{\gamma}-\gamma_0\end{pmatrix} + O\left(\hat{\Theta}_\xi\frac{1}{N}\sum_{i\neq j}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right),$$

which is equivalent to

$$\begin{pmatrix}\hat{\mu}-\mu_0\\ \hat{\gamma}-\gamma_0\end{pmatrix} = -\hat{\Theta}_\xi P_n\nabla_\xi l_{\theta_0} - \hat{\Theta}_\xi\frac{1}{N}D_\xi^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0) + O\left(\hat{\Theta}_\xi\frac{1}{N}\sum_{i\neq j}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right). \tag{A.19}$$

Our goal is now to show that for each component $k = 1,\ldots,p+1$,

$$\sqrt{N}\frac{\hat{\xi}_k - \xi_{0,k}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

as described in the **Goal** section. To that end, by equation (A.19), we now need to solve the following three problems: Writing $\hat{\Theta}_{\xi,k}$ for the k-th row of $\hat{\Theta}_\xi$,

1. $\sqrt{N}\frac{\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1)$,
2. $\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\hat{\Theta}_{\xi,k}\frac{1}{N}D_\xi^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0) = o_P\left(N^{-1/2}\right)$,
3. $O\left(\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right) = o_P\left(N^{-1/2}\right)$.

A.2.3 Bounding inverses

The problems (1) - (3) above suggest that it will be essential to bound the norm and the distance of $\hat{\Theta}_\xi$ and Θ_ξ in an appropriate manner. Recall that for any invertible

matrices $A, B \in \mathbb{R}^{m\times m}$ and any sub-multiplicative matrix norm $\|\,.\,\|$, we have

$$\|A^{-1} - B^{-1}\| \leq \|A^{-1}\|\|B^{-1}\|\|B - A\|. \tag{A.20}$$

where we are particularly interested in the matrix ∞-norm. Also recall that since the matrix ∞-norm is induced by a vector norm, it is sub-multiplicative and consistent with the inducing vector norm. Also recall Lemma 2.23 to bound the matrix ∞-norm in terms of the largest eigenvalue and that the inverse of a symmetric matrix A is itself symmetric.

Hence, $\hat{\Theta}_\xi$ and Θ_ξ are symmetric and we may apply Lemma 2.23. Using that $\lambda_{\max}(\Sigma_\xi^{-1}) = \frac{1}{\lambda_{\min}(\Sigma_\xi)}$, we get

$$\|\Theta_\xi\|_\infty \leq \sqrt{p}\lambda_{\max}(\Sigma_\xi^{-1}) \leq C\frac{1}{\rho_n},$$

and with high probability

$$\|\hat{\Theta}_\xi\|_\infty \leq \sqrt{p}\lambda_{\max}(\hat{\Sigma}_\xi^{-1}) \leq C\frac{1}{\rho_n},$$

with some absolute constant C. Finally, by (A.20),

$$\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty \leq \|\hat{\Theta}_\xi\|_\infty\|\Theta_\xi\|_\infty\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty \leq \frac{C}{\rho_n^2}\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty.$$

It remains to control $\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty$. We have

$$\begin{aligned}\hat{\Sigma}_\xi - \Sigma_\xi &= \frac{1}{N}\left(D_\xi^T\hat{W}^2D_\xi - \mathbb{E}[D_\xi^TW_0^2D_\xi]\right)\\ &= \underbrace{\frac{1}{N}\left(D_\xi^T(\hat{W}^2 - W_0^2)D_\xi\right)}_{(I)} + \underbrace{\frac{1}{N}\left(D_\xi^TW_0^2D_\xi - \mathbb{E}[D_\xi^TW_0^2D_\xi]\right)}_{(II)}.\end{aligned}$$

Recall that $\hat{w}_{ij}^2 = p_{ij}(\hat{\theta})(1 - p_{ij}(\hat{\theta})) = \frac{\exp(D_{ij}^T\hat{\theta})}{(1+\exp(D_{ij}^T\hat{\theta}))^2} = \partial_{a^2}l(A_{ij}, D_{ij}^T\hat{\theta})$, with the function l defined in (A.15). Also recall that $\partial_{a^2}l$ is Lipschitz with constant one, by the Mean Value Theorem and the fact that it has derivative $\partial_{a^3}l$ bounded by one.

Thus, considering the (k,l)-th element of (I) above, we get:

$$\begin{aligned}
&\left|\frac{1}{N}\left(D_\xi^T(\hat{W}^2-W_0^2)D_\xi\right)_{kl}\right| \\
&\qquad= \left|\frac{1}{N}\sum_{i\neq j} D_{ij,n+k}D_{ij,n+l}(\hat{w}_{ij}^2 - w_{0,ij}^2)\right| \\
&\qquad\leq C\frac{1}{N}\sum_{i\neq j}|\hat{w}_{ij}^2 - w_{0,ij}^2|, \quad \text{by uniform boundedness of } Z_{ij} \\
&\qquad\leq C\frac{1}{N}\sum_{i\neq j}|D_{ij}^T(\hat{\theta}-\theta_0)|, \quad \text{by Lipschitz continuity} \\
&\qquad\leq \frac{C}{N}\sum_{i\neq j}\left\{|\hat{\alpha}_i-\alpha_{0,i}| + |\hat{\beta}_j-\beta_{0,j}| + |\hat{\mu}-\mu_0| + |Z_{ij}^T(\hat{\gamma}-\gamma_0)|\right\} \\
&\qquad\leq \frac{C}{N}\underbrace{\left\{\sum_{i\neq j}|\hat{\alpha}_i-\alpha_{0,i}| + |\hat{\beta}_j-\beta_{0,j}|\right\}}_{=(n-1)\|\hat{\vartheta}-\vartheta_0\|_1} + C|\hat{\mu}-\mu_0| + C\|\hat{\gamma}-\gamma_0\|_1 \\
&\qquad\leq C\left\{\frac{1}{n}\|\hat{\vartheta}-\vartheta_0\|_1 + |\hat{\mu}-\mu_0| + \|\hat{\gamma}-\gamma_0\|_1\right\} \\
&\qquad= O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right),
\end{aligned}$$

where the last equality holds under the conditions of Theorem 4.4. Since the dimension of (I) is $(p+1)\times(p+1)$ and thus remains fixed, any row of (I) has ℓ_1-norm of order $O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right)$ and thus

$$\|(I)\|_\infty = O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right).$$

Taking a look at the (k,l)-th element in (II):

$$\begin{aligned}
&\left|\frac{1}{N}\left(D_\xi^T W_0^2 D_\xi - \mathbb{E}[D_\xi^T W_0^2 D_\xi]\right)_{kl}\right| \\
&\qquad= \left|\frac{1}{N}\sum_{i\neq j}\left\{D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2 - \mathbb{E}[D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2]\right\}\right|.
\end{aligned}$$

The random variables $D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2$ are bounded uniformly in i,j,k,l. Thus, by Hoeffding's inequality, for any $t\geq 0$,

$$P\left(\left|\frac{1}{N}\sum_{i\neq j}\left\{D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2 - \mathbb{E}[D_{ij,n+k}D_{ij,n+l}w_{0,ij}^2]\right\}\right| \geq t\right) \leq 2\exp\left(-CNt^2\right).$$

This means, $\left|\frac{1}{N}\left(D_\xi^T W_0^2 D_\xi - \mathbb{E}[D_\xi^T W_0^2 D_\xi]\right)_{kl}\right| = O_P\left(N^{-1/2}\right)$. Again, since the di-

mension $p+1$ is fixed, we get by a simple union bound

$$\|(II)\|_\infty = O_P\left(N^{-1/2}\right).$$

In total, we thus get

$$\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty = O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-1} + \frac{1}{\sqrt{N}}\right) = O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right).$$

We can now obtain a rate for $\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty$.

$$\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty \leq \frac{C}{\rho_n^2}\|\hat{\Sigma}_\xi - \Sigma_\xi\|_\infty = O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-3}\right).$$

By Assumption B3, we have $s_{0,+}\frac{\sqrt{\log(n)}}{\sqrt{n}\rho_n^2} \to 0, n \to \infty$, which implies that the above is $o_P(1)$. In particular, we have now managed to get for $k = 1, \ldots, p+1$,

- $\|\hat{\Theta}_{\xi,k} - \Theta_{\xi,k}\|_1 = o_P(1)$,
- $\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k} + o_p(1)$.

A.2.4 Problem 1

We can now take a look at the problems (1) - (3) outlined above. For problem (1), we want to show:

$$\sqrt{N}\frac{\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \to \mathcal{N}(0,1).$$

Step 1: Show that

$$\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0} = \Theta_{\xi,k}P_n\nabla_\xi l_{\theta_0} + o_P\left(N^{-1/2}\right). \tag{A.21}$$

We have

$$\begin{aligned}|(\hat{\Theta}_{\xi,k} - \Theta_{\xi,k})P_n\nabla_\xi l_{\theta_0}| &\leq \|\hat{\Theta}_{\xi,k} - \Theta_{\xi,k}\|_1 \left\|\frac{1}{N}\sum_{i\neq j}\begin{pmatrix}1\\ Z_{ij}\end{pmatrix}(p_{ij}(\theta_0) - A_{ij})\right\|_\infty \\ &\leq \|\hat{\Theta}_\xi - \Theta_\xi\|_\infty \left\|\frac{1}{N}\sum_{i\neq j} D_{\xi,ij}(p_{ij}(\theta_0) - A_{ij})\right\|_\infty.\end{aligned}$$

Consider the vector $\sum_{i\neq j} D_{\xi,ij}(p_{ij}(\theta_0) - A_{ij}) \in \mathbb{R}^{p+1}$. The k-th component of it has the form $\sum_{i\neq j}(p_{ij}(\theta_0) - A_{ij})$ for $k = 1$ and $\sum_{i\neq j} Z_{ij,k-1}(p_{ij}(\theta_0) - A_{ij}), k = 2, \ldots, p+1$. Notice that for these components are all centred:

$$\mathbb{E}[D_{\xi,ij,k}(p_{ij}(\theta_0) - A_{ij})] = \mathbb{E}[D_{\xi,ij,k}\mathbb{E}[(p_{ij}(\theta_0) - A_{ij})|Z_{ij}]] = \mathbb{E}[D_{\xi,ij,k}\cdot 0] = 0,$$

as well as $|D_{\xi,ij,k}(p_{ij}(\theta_0) - A_{ij})| \leq c$, where $c > 1$ is a universal constant bounding $|Z_{ij,k}|$ for all i, j, k. Thus, by Hoeffding's inequality, for any $t > 0$,

$$P\left(\left|\frac{1}{N}\sum_{i\neq j} D_{\xi,ij,k}(p_{ij}(\theta_0) - A_{ij})\right| \geq t\right) \leq 2\exp\left(-2\frac{Nt^2}{c^2}\right)$$

and thus,

$$\frac{1}{N}\sum_{i\neq j} D_{\xi,ij}(p_{ij}(\theta_0) - A_{ij}) = O_P\left(N^{-1/2}\right).$$

Since we have $\|\hat{\Theta}_Z - \Theta_Z\|_\infty = o_P(1)$, by Section A.2.3, Step 1 is now concluded.

Step 2: Show that

$$\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k} + o_P(1).$$

Since $\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty = o_P(1)$, by Section A.2.3, for all k

$$|\hat{\Theta}_{\xi,k,k} - \Theta_{\xi,k,k}| \leq \|\hat{\Theta}_\xi - \Theta_\xi\|_\infty = o_P(1)$$

and Step 2 is concluded.

Step 3: Show that

$$\left|\frac{1}{\Theta_{\xi,k,k}}\right| \leq C < \infty,$$

for some universal constant $C > 0$. Then, we may conclude from Step 1 and Step 2 that

$$\sqrt{N}\frac{\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} = \sqrt{N}\frac{\Theta_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\Theta_{\xi,k,k}}} + o_P(1).$$

To prove Step 3, notice that Θ_ξ is symmetric and hence has only real eigenvalues. Therefore it is unitarily diagonalizable and for any $x \in \mathbb{R}^{p+1}$, we have $x^T\Theta_\xi x \geq \lambda_{\min}(\Theta_\xi)\|x\|_2^2$. We also know that

$$\lambda_{\min}(\Theta_\xi) = \frac{1}{\lambda_{\max}(\Sigma_\xi)}.$$

Under Assumption 4.1 we can now deduce an upper bound on the maximum eigenvalue of Σ_ξ: For any $x \in \mathbb{R}^p$,

$$x^T\Sigma_\xi x = x^T\frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi]x \leq x^T\frac{1}{N}\mathbb{E}[D_\xi^T D_\xi]x \leq (1 \vee \lambda_{\max})\|x\|_2^2,$$

where we have used that any entry in W_0^2 is bounded above by one. Since $x^T\Sigma_\xi x \leq \lambda_{\max}(\Sigma_\xi)\|x\|_2^2$ and since this bound is tight (we have equality if x is an eigenvector corresponding to $\lambda_{\max}$), we can conclude by Assumption 4.1 that $\lambda_{\max}(\Sigma_\xi) \leq (1 \vee \lambda_{\max}) \leq C < \infty$ for some universal constant $C > 0$.

In particular, since $\Theta_{\xi,k,k} = e_k^T \Theta_\xi e_k$, we get

$$\Theta_{\xi,k,k} \geq \lambda_{\min}(\Theta_\xi)\|e_k\|_2^2 = \frac{1}{\lambda_{\max}(\Sigma_\xi)} \geq C > 0,$$

uniformly for all n. Consequently,

$$0 < \frac{1}{\Theta_{\xi,k,k}} \leq C < \infty.$$

Step 3 is thus concluded.

Step 4: Finally, show that

$$\sqrt{N}\frac{\Theta_{\xi,k} P_n \nabla_\xi l_{\theta_0}}{\sqrt{\Theta_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Such that by all the above

$$\sqrt{N}\frac{\hat{\Theta}_{\xi,k} P_n \nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

For brevity, we write p_{ij} for the true link probabilities $p_{ij}(\theta_0)$. Also keep in mind that $\Theta_{\xi,k}$ denotes the k-th *row* of Θ_ξ, while $D_{\xi,ij}$ denote $((p+1)\times 1)$-*column* vectors. We want to apply the Lindeberg-Feller Central Limit Theorem (CLT). The random variables we study are the summands in

$$\sqrt{N}\Theta_{\xi,k} P_n \nabla_\xi l_{\theta_0} = \sum_{i \neq j} \left\{ \frac{1}{\sqrt{N}} \Theta_{\xi,k} D_{\xi,ij}(p_{ij} - A_{ij}) \right\}.$$

These random variables are centred:

$$\mathbb{E}\left[\frac{1}{\sqrt{N}}\Theta_{\xi,k} D_{\xi,ij}(p_{ij} - A_{ij})\right] = \mathbb{E}\left[\frac{1}{\sqrt{N}}\Theta_{\xi,k} D_{\xi,ij}\mathbb{E}[p_{ij} - A_{ij}|Z_{ij}]\right] = 0.$$

For the Lindeberg-Feller CLT we need to sum up the variances of these random variables. We claim that

$$\sum_{i \neq j} \mathrm{Var}\left(\frac{1}{\sqrt{N}}\Theta_{\xi,k} D_{\xi,ij}(p_{ij} - A_{ij})\right) = \Theta_{\xi,k,k}.$$

Indeed, consider the vector-valued random variable $\sum_{i\neq j}\left\{\frac{1}{\sqrt{N}} D_{\xi,ij}(p_{ij} - A_{ij})\right\} \in$

$\mathbb{R}^{p+1}$. It has covariance matrix

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i\neq j}\left\{\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\right\}\sum_{i\neq j}\left\{\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\right\}^T\right] \\
&= \mathbb{E}\left[\sum_{i\neq j}\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\frac{1}{\sqrt{N}}D_{\xi,ij}^T(p_{ij}-A_{ij})\right], \text{ by independence across } i,j \\
&= \frac{1}{N}\sum_{i\neq j}\left[\mathbb{E}[D_{\xi,ij,k}D_{\xi,ij,l}(p_{ij}-A_{ij})^2]\right]_{k,l=1,\ldots,p+1}, \text{ as a } ((p+1)\times(p+1))\text{-matrix} \\
&= \frac{1}{N}\mathbb{E}[D_\xi^T W_0^2 D_\xi] \\
&= \Sigma_\xi.
\end{aligned}$$

Thus, by independence across i,j,

$$\begin{aligned}
\sum_{i\neq j}\mathrm{Var}\left(\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})\right) &= \mathrm{Var}\left(\Theta_{\xi,k}\sum_{i\neq j}\frac{1}{\sqrt{N}}D_{\xi,ij}(p_{ij}-A_{ij})\right) \\
&= \Theta_{\xi,k}\Sigma_\xi\Theta_{\xi,k}^T = \Theta_{\xi,k,k},
\end{aligned}$$

where for the last equality we have used that $\Theta_\xi = \Sigma_\xi^{-1}$ and thus, $\Sigma_\xi\Theta_{\xi,k}^T = e_k$. Now, we need to show that the Lindeberg condition holds. That is, we want for any $\epsilon > 0$,

$$\begin{aligned}
\lim_{n\to\infty}\frac{1}{\Theta_{\xi,k,k}}\sum_{i\neq j}\mathbb{E}\Bigg[&\left\{\frac{1}{\sqrt{N}}\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})\right\}^2 \\
&\mathbb{1}\left(|\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})| > \epsilon\sqrt{N\Theta_{\xi,k,k}}\right)\Bigg] = 0.
\end{aligned} \tag{A.22}$$

We have

$$|\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})| \leq p\cdot c\cdot\|\Theta_{\xi,k}\|_1 \leq C\|\Theta_\xi\|_\infty \leq C\rho_n^{-1}.$$

At the same time, we know from Step 3 that $\Theta_{Z,k,k} \geq C > 0$ for some universal C. Then, as long as $\rho_n^{-1} \to \infty$ at a rate slower than n, which is enforced by Assumption B3, we must have for n large enough

$$|\Theta_{\xi,k}D_{\xi,ij}(p_{ij}-A_{ij})| < \epsilon\sqrt{N\Theta_{\xi,k,k}}$$

uniformly in i,j. Thus, the indicator function and therefore each summand in (A.22) is equal to zero for n large enough. Hence, (A.22) holds. Then, by the Lindeberg-Feller CLT,

$$\sqrt{N}\frac{\Theta_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\Theta_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Now, by the Steps 1-4,

$$\sqrt{N}\frac{\hat{\Theta}_{\xi,k}P_n\nabla_\xi l_{\theta_0}}{\sqrt{\hat{\Theta}_{\xi,k,k}}} \xrightarrow{d} \mathcal{N}(0,1).$$

This concludes solving problem 1.

A.2.5 Problem 2

For Problem 2 we must show

$$\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\hat{\Theta}_{\xi,k}\frac{1}{N}D_\xi^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0) = o_P\left(N^{-1/2}\right).$$

Since we have $\|\hat{\Theta}_\xi - \Theta_\xi\|_\infty = o_P(1)$, we do not need to worry about $\hat{\Theta}_{\xi,k,k}^{-1/2}$, because $\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k} + o_P(1)$ and $\Theta_{\xi,k,k}^{-1/2} \leq C < \infty$, i.e. $\hat{\Theta}_{\xi,k,k}^{-1/2} = O_P(1)$. By Theorem 4.4 we also have a high-probability error bound on $\|\hat{\vartheta}-\vartheta_0\|_1$. The problem will be bounding the corresponding matrix norms.

$$\left|\hat{\Theta}_{\xi,k}\frac{1}{N}D_\xi^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0)\right| \leq \left\|\frac{1}{N}X^T\hat{W}^2D_\xi\hat{\Theta}_{\xi,k}^T\right\|_\infty \|\hat{\vartheta}-\vartheta_0\|_1.$$

Notice that in the display above we have the vector ℓ_∞-norm. Also,

$$\left\|\frac{1}{N}X^T\hat{W}^2D_\xi\hat{\Theta}_{\xi,k}^T\right\|_\infty \leq \left\|\frac{1}{N}X^T\hat{W}^2D_\xi\right\|_\infty \|\hat{\Theta}_{\xi,k}^T\|_\infty.$$

Here we used the compatibility of the matrix ℓ_∞-norm with the vector ℓ_∞-norm. The first term is the matrix norm, the second the vector norm. We know,

$$\|\hat{\Theta}_{\xi,k}^T\|_\infty \leq \|\hat{\Theta}_\xi\|_\infty \leq C\rho_n^{-1},$$

where on the left hand side we have the vector norm and in the middle display the matrix norm. Finally, $1/N \cdot X^T\hat{W}^2D_\xi$ is a $(2n \times (p+1))$-matrix. The (k,l)-th element looks like $1/N \cdot S_{k,l}$, where $S_{k,l}$ is the sum of $n-1$ terms of the form $D_{\xi,il,k}\hat{w}_{il}^2$, summed over the appropriate indices i, j, all of which are uniformly bounded. Thus,

$$\left|\left(\frac{1}{N}X^T\hat{W}^2D_\xi\right)_{k,l}\right| \leq \frac{1}{N}\cdot(n-1)\cdot C = \frac{C}{n}.$$

Thus, the ℓ_1-norm of any row of $\frac{1}{N}X^T\hat{W}^2D_\xi$ is bounded by pC/n and thus

$$\left\|\frac{1}{N}X^T\hat{W}^2D_\xi\right\|_\infty \leq \frac{C}{n}.$$

Recall that $\|\hat{\vartheta} - \vartheta_0\|_1 = O_P\left(s_{0,+}\frac{\sqrt{\log(n)}}{\sqrt{n}}\rho_n^{-1}\right)$ by Theorem 4.4. Then,

$$\begin{aligned}\left|\hat{\Theta}_{\xi,k}\frac{1}{N}X^T\hat{W}^2D_\xi(\hat{\vartheta}-\vartheta_0)\right| &\leq \|\hat{\Theta}_{\xi,k}^T\|_\infty \left\|\frac{1}{N}D_\xi^T\hat{W}^2X\right\|_\infty \|\hat{\vartheta}-\vartheta_0\|_1 \\ &= O_P\left(\frac{s_{0,+}}{\rho_n^2\cdot n}\cdot\frac{\sqrt{\log(n)}}{\sqrt{n}}\right).\end{aligned}$$

Multiplying by $\sqrt{N} = O(n)$, gives

$$\sqrt{N}\left|\hat{\Theta}_{\vartheta,k}\frac{1}{N}D_\vartheta^T\hat{W}^2X(\hat{\vartheta}-\vartheta_0)\right| = O_P\left(\frac{s_{0,+}}{\rho_n^2}\cdot\frac{\sqrt{\log(n)}}{\sqrt{n}}\right),$$

which is $o_P(1)$ under Assumption B3.

A.2.6 Problem 3

Finally, we must show

$$O\left(\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}\begin{pmatrix}1\\Z_{ij}\end{pmatrix}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right) = o_P\left(N^{-1/2}\right).$$

Again, since $\hat{\Theta}_{\xi,k,k} = \Theta_{\xi,k,k} + o_P(1)$ and $\Theta_{\xi,k,k} \geq C > 0$ uniformly in n, we do not need to worry about the factor $\frac{1}{\sqrt{\hat{\Theta}_{\xi,k,k}}}$ and it remains to show

$$O\left(\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}D_{\xi,ij}|D_{ij}^\top(\hat{\theta}-\theta_0)|^2\right) = o_P\left(N^{-1/2}\right).$$

We have for each $i \neq j$, $|\hat{\Theta}_{\xi,k}D_{\xi,ij}| \leq C\|\hat{\Theta}_{\xi,k}\|_1$. Thus,

$$\begin{aligned}\left|\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j}D_{\xi,ij}|D_{ij}^\top(\hat{\theta}-\theta_0)|^2\right| &\leq \frac{1}{N}\sum_{i\neq j}|\hat{\Theta}_{\xi,k}D_{\xi,ij}||D_{ij}^T(\hat{\theta}-\theta_0)|^2 \\ &\leq C\|\hat{\Theta}_{\xi,k}\|_1\frac{1}{N}\sum_{i\neq j}|D_{ij}^T(\hat{\theta}-\theta_0)|^2 \\ &\leq C\frac{1}{\rho_n}\frac{1}{N}\sum_{i\neq j}|D_{ij}^T(\hat{\theta}-\theta_0)|^2,\end{aligned}$$

where for the last inequality we have used that $\|\hat{\Theta}_{\xi,k}\|_1 \leq \|\hat{\Theta}_\xi\|_\infty \leq C\frac{1}{\rho_n}$. Now remember from (A.11) that

$$\frac{1}{N}\sum_{i\neq j}|D_{ij}^T(\hat{\theta}-\theta_0)|^2 \leq C\|\hat{\bar{\theta}}-\bar{\theta}_0\|_1^2,$$

where we make use of the fact that $\bar{D}\bar{\theta} = D\theta$. From Theorem 4.4 we know that under the assumptions of Theorem 4.5, $\|\hat{\bar{\theta}} - \bar{\theta}_0\|_1 = O_P\left(s_{0,+}\sqrt{\frac{\log(n)}{N}}\rho_n^{-1}\right)$. Thus,

$$\sqrt{N}\left|\hat{\Theta}_{\xi,k}\frac{1}{N}\sum_{i\neq j} D_{\xi,ij}|D_{ij}^T(\hat{\theta}-\theta_0)|^2\right| = O_P\left((s_{0,+})^2\frac{\log(n)}{\sqrt{N}}\rho_n^{-3}\right).$$

We see that this is $o_P(1)$ by applying Assumption B3 twice. Problem 3 is solved.

Proof of Theorem 4.5. Theorem 4.5 now follows from the solved problems (1) - (3). □

Bibliography

Abbe, E. (2018), 'Community detection and stochastic block models: recent developments', *Journal of Machine Learning Research* **18**, 1–86.

Adamic, L. A. & Glance, N. (2005), The political blogosphere and the 2004 us election: divided they blog, *in* 'Proceedings of the 3rd international workshop on Link discovery', pp. 36–43.

Agarwal, A., Tewary, U., Pettersson, F., Das, S., Saxén, H. & Chakraborti, N. (2010), 'Analysing blast furnace data using evolutionary neural network and multiobjective genetic algorithms', *Ironmaking & Steelmaking* **37**(5), 353–359.

Amini, A. A., Chen, A., Bickel, P. J. & Levina, E. (2013), 'Pseudo-likelihood methods for community detection in large sparse networks', *Annals of Statistics* **41**(4), 2097–2122.

Anderson-Cook, C. M., Lu, L. & Parker, P. A. (2019), 'Effective interdisciplinary collaboration between statisticians and other subject matter experts', *Quality Engineering* **31**(1), 164–176.

Bertsekas, D. (1995), *Nonlinear Programming*, Athena Scientific.

Bickel, P. J. & Chen, J. (2009), 'A nonparametric view of network models and Newman-Girvan and other modularities', *Proceedings of the National Academy of Science* **106**, 21068–21073.

Bickel, P. J. & Sarkar, P. (2016), 'Hypothesis testing for automated community detection in networks', *Journal of the Royal Statistical Society: Series B: Statistical Methodology* pp. 253–273.

Bousquet, O. (2002), 'A bennett concentration inequality and its application to suprema of empirical processes', *Comptes Rendus Mathematique* **334**(6), 495–500.

Brimacombe, J. K. (1999), 'The challenge of quality in continuous casting processes', *Metallurgical and Materials Transactions A* **30**(8), 1899–1912.

Buena, F. (2008), 'Honest variable selection in linear and logistic regression models via l1 and l1 + l2 penalization', *Electronic Journal of Statistics* **2**, 1153–1194.

Cai, T. T. & Li, X. (2015), 'Robust and computationally feasible community detection in the presence of arbitrary outlier nodes', *Annals of Statistics* **43**(3), 1027–1059.

Caron, F. & Fox, E. (2017), 'Sparse graphs using exchangeable random measures (with discussion)', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 1295–1366.

Chatterjee, S. & Diaconis, P. (2013), 'Estimating and understanding exponential random graph models', *Ann. Statist.* **41**(5), 2428–2461.

Chatterjee, S., Diaconis, P. & Sly, A. (2011), 'Random graphs with a given degree sequence', *Annals of Applied Probability* **21**(4), 1400–1435.

Chen, K. & Lei, J. (2018), 'Network cross-validation for determining the number of communities in network data', *Journal of the American Statistical Association* **113**(521), 241–251.

Chen, M., Kato, K. & Leng, C. (2020), 'Analysis of networks via the sparse beta

model'. arXiv:1908.03152.

Chen, Y., Li, X. & Xu, J. (2018), 'Convexified modularity maximization for degree-corrected stochastic block models', *Annals of Statistics* **46**(4), 1573–1602.

Creative Commons Attribution 4.0 International License (2013).
URL: *https://creativecommons.org/licenses/by/4.0/*

Erdős, P. & Rényi, A. (1959), 'On random graphs I', *Publicationes Mathematicae (Debrecen)* **6**, 290–297.

Erdős, P. & Rényi, A. (1960), 'On the evolution of random graphs', *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17–60.

Fienberg, S. E. (2012), 'A brief history of statistical models for network analysis and open challenges.', *Journal of Computational and Graphical Statistics* **21**, 825–839.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.

Geerdes, M., Chaigneau, R. & Kurunov, I. (2015), *Modern Blast Furnace Ironmaking: An Introduction*, 3 edn, IOS Press.

Gilbert, E. G. (1959), 'Random graphs', *Annals of Mathematical Statistics* **30**, 1141–1144.

Girvan, M. & Newman, M. E. (2002), 'Community structure in social and biological networks', *Proceedings of the national academy of sciences* **99**(12), 7821–7826.

Goldenberg, A., Zheng, A. X., Feinberg, S. E. & Airoldi, E. M. (2009), 'A survey of statistical network models', *Foundations and Trends in Machine Learning* **2**, 129–233.

Graham, B. S. (2017), 'An econometric model of network formation with degree heterogeneity', *Econometrica* **85**, 1033–1063.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparametrization', *Bernoulli* **10**, 971–988.

Holland, P. W., Laskey, K. & Leinhardt, S. (1983), 'Stochastic blockmodels: First steps', *Social Networks* **5**, 109–137.

Holland, P. W. & Leinhardt, S. (1981), 'An exponential family of probability distributions for directed graphs', *Journal of the American Statistical Association* **76**, 33–50.

Holm, S. (1979), 'A simple sequentially rejective multiple test procedure', *Scandinavian Journal of Statistics* **6**(2), 65–70.

Hu, J., Qin, H., Yan, T. & Zhao, Y. (2020), 'Corrected bayesian information criterion for stochastic block models', *Journal of the American Statistical Association* **115**(532), 1771–1783.

Huang, S. & Feng, Y. (2018), 'Pairwise covariates-adjusted block model for community detection'. arXiv:1807.03469.

Javanmard, A. & Montanari, A. (2014*a*), 'Confidence intervals and hypothesis testing for high-dimensional regression', *Journal of Machine Learning Research* **15**, 2869–2909.

Javanmard, A. & Montanari, A. (2014*b*), 'Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory', *IEEE Transactions on Information Theory* **60**(10), 6522–6554.

Jensen, W. A. (2020), 'Statistics = analytics?', *Quality Engineering* **32**(2), 133–144.

Ji, P. & Jin, J. (2016), 'Coauthorship and citation networks for statisticians', *The Annals of Applied Statistics* **10**(4), 1779–1812.

Jin, J. (2015), 'Fast community detection by score', *Annals of Statistics* **43**(1), 57–89.

Jin, J., Ke, Z. T. & Luo, S. (2021), 'Estimating network memberships by simplex vertex hunting', *Annals of Statistics* .

Jochmans, K. (2018), 'Semiparametric analysis of network formation', *Journal of Business & Economic Statistics* **36**(4), 705–713.

Karrer, B. & Newman, M. E. (2011), 'Stochastic blockmodels and community structure in networks', *Physical review E* **83**(1), 016107.

Karwa, V. & Slavković, A. (2016), 'Inference using noisy degrees: Differentially private β-model and synthetic graphs', *Annals of Statistics* **44**(1), 87–112.

Kaufman, S., Rosset, S. & Perlich, C. (2011), 'Leakage in data mining: Formulation, detection, and avoidance', *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **6**, 556–563.

Kock, A. B. & Tang, H. (2019), 'Uniform inference in high-dimensional dynamic panel data models with approximately sparse fixed effects', *Econometric Theory* **35**(2), 295–359.

Kolaczyk, E. D. (2009), *Statistical Analysis of Network Data: Methods and Models*, Springer.

Kolaczyk, E. D. (2017), *Topics at the Frontier of Statistics and Network Analysis: (Re)Visiting the Foundations*, Cambridge University Press.

Koltchinskii, V. (2011), *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. École d'été de probabilités de Saint-Flour XXXVIII-2008*, Springer.

Kra, I. & Simanca, S. R. (2012), 'On circulant matrices', *Notices of the American Mathematical Society* **59**(3), 368–377.

Larsen, K. & Becker, D. (2018), *Automated Machine Learning for Business*, Oxford University Press. (in press).

Lazega, E. (2001), *The Collegial Phenomenon: The Social Mechanisms of Cooperation Among Peers in a Corporate Law Partnership*, Oxford University Press.

Ledoux, M. & Talagrand, M. (1991), *Probability in Banach Spaces*, Springer-Verlag.

Lei, J. (2016), 'A goodness-of-fit test for stochastic block models', *The Annals of Statistics* **44**(1), 401–424.

Li, T., Levina, E. & Zhu, J. (2020), 'Network cross-validation by edge sampling', *Biometrika* **107**(2), 257–276.

Liu, Y., Guo, B., Zou, X., Li, Y. & Shi, S. (2020), 'Machine learning assisted materials design and discovery for rechargeable batteries', *Energy Storage Materials* **31**, 434 – 450.

Ma, S., Su, L. & Zhang, Y. (2020), 'Detecting latent communities in network formation models', *arXiv preprint arXiv:2005.03226* .

Ma, Z., Ma, Z. & Yuan, H. (2020), 'Universal latent space model fitting for large networks with edge covariates', *Journal of Machine Learning Research* **21**(4), 1–67.

Mann, H. B. & Whitney, D. R. (1947), 'On a test of whether one of two random variables is stochastically larger than the other', *Ann. Math. Statist.* **18**(1), 50–60.

Meinshausen, N. & Bühlmann, P. (2006), 'High-dimensional graphs and variable selection with the lasso', *The Annals of Statistics* **34**(3), 1436–1462.

Newman, M. (2018), *Networks (2nd Edition)*, Oxford University Press.

Newman, M. E. (2006), 'Modularity and community structure in networks', *Proceedings of the national academy of sciences* **103**(23), 8577–8582.

Olhede, S. C. & Wolfe, P. J. (2014), 'Network histograms and universality of blockmodel approximation', *Proceedings of the National Academy of Sciences* **111**(41), 14722–14727.

Omori, Y. (1987), *Blast Furnace Phenomena and Modelling*, Elsevier.

Ravikumar, P., Wainwright, M. J. & Lafferty, J. D. (2010), 'High-dimensional ising model selection using l1 -regularized logistic regression', *Ann. Statist.* **38**(3), 1287–1319.

Rinaldo, A., Petrović, S. & Fienberg, S. E. (2013), 'Maximum likelihood estimation in the β-model', *The Annals of Statistics* **41**(3), 1085–1110.

Sengupta, S. & Chen, Y. (2018), 'A block model for node popularity in networks with community structure', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**(2), 365–386.

Silva, J. M. C. S. & Tenreyro, S. (2006), 'The log of gravity', *The Review of Economics and Statistics* **88**(4), 641–658.

Snijders, T. A. B., Pattison, P. E., Robins, G. L. & Handcock, M. S. (2006), 'New specifications for exponential random graph models', *Sociological Methodology* **36**(1), 99–153.

Stein, S. (2019), *igate: Guided Analytics for Testing Manufacturing Parameters*, University of Warwick. R package version 0.3.3.
URL: *https://CRAN.R-project.org/package=igate*

Stein, S. & Leng, C. (2020), 'A sparse β-model with covariates for networks', arXiv 2010.13604.

Stein, S. & Leng, C. (2021), 'A sparse random graph model for sparse directed networks', arXiv 2108.09504.

Stein, S., Leng, C., Thornton, S. & Randrianandrasana, M. (2021), 'A guided analytics tool for feature selection in steel manufacturing with an application to blast furnace top gas efficiency', *Computational Materials Science* **186**, 110053.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* **58**(1), 267–288.

Tukey, J. W. (1959), 'A quick, compact, two-sample test to duckworth's specifications', *Technometrics* **1**(1), 31–48.

van de Geer, S. A. (2008), 'High-dimensional generalized linear models and the lasso', *The Annals of Statistics* **36**(2), 614–645.

van de Geer, S. & Bühlmann, P. (2011), *Statistics for High-Dimensional Data*, Springer Series in Statistics, Springer-Verlag.

van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* **42**(3), 1166–1202.

van der Hofstad, R. (2016), *Random Graphs and Complex Networks*, Cambridge University Press.

van der Hofstad, R. (2021), *Random Graphs and Complex Networks*, Vol. 2, Preprint.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

van der Vaart, A. & Wellner, J. (1996), *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer-Verlag.

Varah, J. (1975), 'A lower bound for the smallest singular value of a matrix', *Linear Algebra and its Applications* **11**(1), 3 – 5.

Wainwright, M. J. (2019), *High-dimensional statistics: A non-asymptotic viewpoint*, Cambridge University Press.

Wang, Y. R. & Bickel, P. J. (2017), 'Likelihood-based model selection for stochastic block models', *The Annals of Statistics* **45**(2), 500–528.

Wilcoxon, F. (1945), 'Individual comparisons by ranking methods', *Biometrics Bulletin* **1**(6), 80–83.

Yan, T., Jiang, B., Fienberg, S. E. & Leng, C. (2019), 'Statistical inference in a directed network model with covariates', *Journal of the American Statistical Association* **114**(526), 857–868.

Yan, T., Leng, C. & Zhu, J. (2016), 'Asymptotics in directed exponential random graph models with an increasing bi-degree sequence', *The Annals of Statistics* **44**, 31–57.

Yan, T., Qin, H. & Wang, H. (2016), 'Asymptotics in undirected random graph models parameterized by the strengths of vertices', *Statistica Sinica* **26**, 273–293.

Yan, T. & Xu, J. (2013), 'A central limit theorem in the β-model for undirected random graphs with a diverging number of vertices', *Biometrika* **100**, 519–524.

Yu, Y., Bradic, J. & Samworth, R. J. (2019), 'Confidence intervals for high-dimensional cox models', *Statistics Sinica (to appear)* . arXiv:1803.01150.

Zhang, A. Y. & Zhou, H. H. (2016), 'Minimax rates of community detection in stochastic block models', *The Annals of Statistics* **44**(5), 2252–2280.

Zhang, C.-H. & Zhang, S. S. (2014), 'Confidence intervals for low dimensional parameters in high dimensional linear models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 217–242.

Zhang, J., He, X. & Wang, J. (2021), 'Directed community detection with network embedding', *Journal of the American Statistical Association* pp. 1–11.