ISSUES

in

THE BAYESIAN FORECASTING

of

DISPERSAL AFTER A NUCLEAR ACCIDENT

Thesis submitted in fulfilment of requirements

for the degree of

Doctor of Philosophy

Ali S. Gargoum

Department of Statistics

University of Warwick

May 1997

Contents

List of Figures

List of Tables

Acknowledgements

I am deeply indebted to my supervisor, Professor Jim Smith. His excellent guidance, sustained encouragement and friendship made this thesis possible and these years of research a very pleasant experience.

I thank Professor P. J. Harrison for many discussions, lessons and for his help in using the two software packages, PROF and BATS.

I acknowledge the help of D. Ranyard from Leeds University in coding some of the new operators and algorithms proposed in this thesis.

I must also thank the staff of the Statistics Department at the University of Warwick and my fellow research students.

I would like to express my gratitude to the University of Garyounis (Benghazi) for giving me the opportunity to carry out this research and for their financial support.

Finally, I would like to thank my family for always supporting and encouraging me.

Declaration

I hereby declare that this thesis is entirely my own work with the exception of Sections (3.5), (5.7) and (7.2), which were developed and adapted through discussion with my supervisor, Professor Jim Smith.

In memory of my parents

Summary

This thesis addresses three main topics related to the practical problems of modelling the spread of nuclear material after an accidental release.

The first topic deals with the issue of how qualitative information (expert judgement) about the development of the emission of contamination after an accident can be coded as a Dynamic Linear Model (DLM). An illustration is given of the subsequent adaptation of the expert judgement in response to the incoming data. Moreover, the height of the release at the source can be a key parameter in the subsequent dispersal. We addressed uncertainty on the release height using the Multi-Process Models framework. That is we included several models in our analysis, each with a different release height. The Bayesian methodology uses probabilities representing their relative likelihood to weight these and updates the probabilities in the light of monitoring data. A brief illustration of testing the updating algorithm on simulated contamination readings is provided.

The second topic concerns the demands of computational efficiency. We show how the Bayesian propagation algorithms on a dynamic junction tree of cliques of variables (representing a high dimensional Gaussian process), as provided by Smith et al. (1995), can be generalised to incorporate the case when data may destroy neat dependencies (i.e. when observations are taken under more than one clique). Here we introduce two classes of new operators: exact and non-exact (approximations) which act on this high dimensional Gaussian process, modifying its junction tree by another tree which allows quicker probability propagation. We also develop fast algorithms which can be defined by approximating Gaussian systems by cutting edges on junctions. The appropriateness of the approximations is based on the Kulback-Leibler/Hellinger distances.

Some of these new operators and algorithms have been implemented and coded. Preliminary tests on these algorithms were carried out using arbitrary data, and the system proved to be highly efficient in terms of P.C. user time.

The third topic concentrates on generalisations from a Gaussian process. It proposes, as a good approximation, an adaptation of the Dynamic Generalised Linear Models (DGLMs) of West, Harrison, and Migon (1985) for updating algorithms on a dynamic junction tree. The Hellinger distance is used to check the accuracy of the dynamic approximation.

The analysis of these topics involves a review and extension of some useful theory and results on Bayesian forecasting and dynamic models, graphical modelling, and information divergence.

Chapter 1

Introduction

This introduction gives a brief background to the original application which has motivated the research, presents the aims of the thesis, and then provides an outline of the whole work.

1.1 Background

In the event of an accidental release of radioactive pollutants or chemical gases, countries will be concerned about the danger of environmental contamination, and in this respect interest centres on the prediction of the distribution of the radioactive emissions reaching these countries and the identification of regions where contamination is most likely to exceed certain prescribed levels.

Nuclear accidents such as those that happened at Three Mile Island and Chernobyl have focused attention on the need for the development of emergency response systems which are able to support decisions on countermeasures. This view has led the Commission of the European Communities (CEC) to establish a number of projects to build Decision Support Systems (DSSs) and methodologies for use in the event of a future accident. One of these is the RODOS project.

RODOS is a Real-time On-line DecisiOn Support system for nuclear emergencies in Europe, being jointly developed by some European institutions with support of the CEC. It is designed as an integrated software environment, which allows the implementation of software (external programs) developed by the contractors.

RODOS comprises three subsystems (see Ehrhardt et al., 1993).

(i) The analysis subsystem (ASY). The main task of this component is to provide continually updated forecasts of the spread of any contamination. The ASY consists of an atmospheric dispersion model such as Rimpuff model–will be discussed later– which is the prediction model; it predicts concentration of contamination both at present and in the future. The predictions are based on different sources of information:

- An estimate of the source term. This comprises the mass, and the height of the release, and is most likely to be expert judgement.
- Meteorological data.
- Geographical data from a Geographical Information System (GIS).

The output of the ASY is in a form of a grid of concentrations with an associated uncertainty distribution. The grid becomes one of the inputs to the next subsystem, the CSY.

(ii) The countermeasures subsystem (CSY). The main task of CSY is to identify possible countermeasures (sheltering and evacuation for the local population; food bans; etc.) and to quantify the benefits and drawbacks of various countermeasure combinations. The CSY then outputs a list of all possible countermeasures for input to the next subsystem, the ESY.

(iii) The evaluation subsystem (ESY). The main task of ESY is to evaluate the different countermeasures strategies. Figure 1.1 below shows the structure of RODOS.

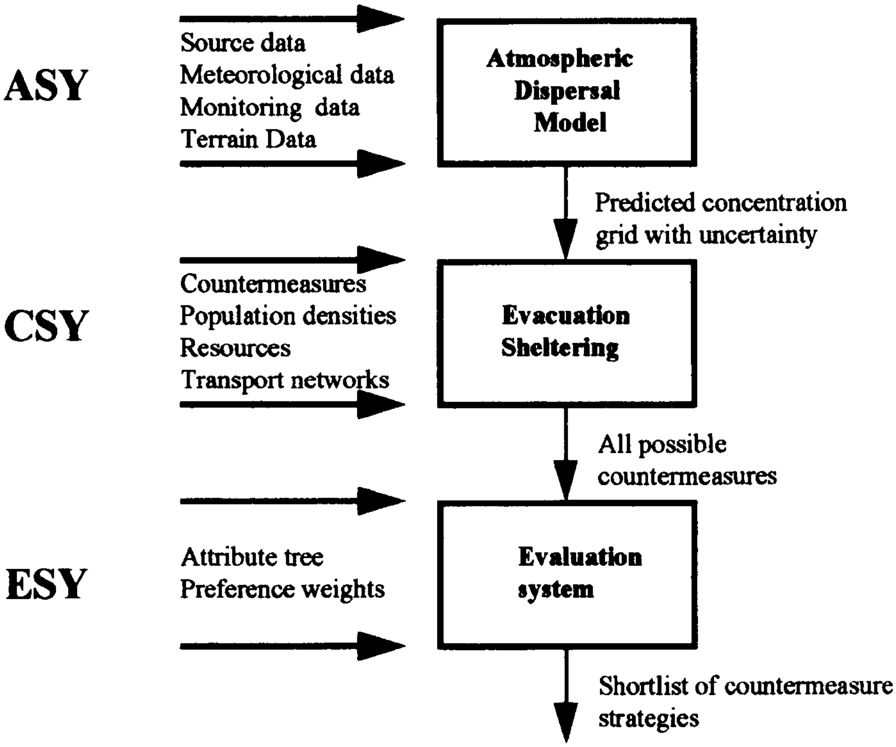


Figure 1.1: Model Structure of RODOS

In this thesis we will only be concerned with ASY, and it will be discussed and interpreted within a Bayesian framework.

All the sources of information mentioned in ASY above are not considered to be one hundred per cent accurate. There are several uncertainties associated with them. For instance, RODOS contains an algorithm which describes the dispersal of radiation in the atmosphere. Such algorithms can only ever be approximations of what is happening, and other methods must be incorporated to control them effectively. Also, uncertainty may arise from the lack of knowledge of source term characteristics (the release characteristics) and the surface characteristics that affect the pollutant behaviour; etc.

Within the framework of RODOS, Smith & French (1993) investigated the feasibility of managing the uncertainties relating to the key variables which are important for decision making in the short term, and of assimilating data as they became available using Bayesian methodology. The Bayesian methodology can tackle a number of issues of interest, for example: data assimilation, expert judgement, model uncertainty, etc. The authors designed the ASY module based upon the RIMPUFF model with Bayesian updating in the light of monitoring data in order to address the following questions:

- What is the likely spread of contamination ?
- How can the prediction be updated in the light of monitoring data?
- What are the uncertainties in the prediction ?

The authors have been able to combine the Dynamic Linear Model (DLM) methodology with the RIMPUFF model to address all three questions.

RIMPUFF is the RIso-Mesoscale-PUFF developed at the Risø Research Institute, Denmark (see Thykier-Neilsen & Mikkelsen, 1991), and is the atmospheric dispersal model used within RODOS. RIMPUFF approximates the continuous release of airborne substances (such as radiation) by a discrete series of puffs. Puffs may contain different masses reflecting the uneven pattern release from the source. Different parameters can be associated with each puff, enabling characteristics of the windfield or local information gathered for monitoring data to be incorporated. The concentration distribution in each individual puff is assumed to be Gaussian, and puffs grow in size over time.

A further feature of the RIMPUFF model is pentification. When a puff reaches a certain diameter, it is approximated by five smaller child puffs. Certain percentages of the mass of the parent are distributed amongst its children. These percentages can be adjusted in the light of observational data. The adjustment is controlled by incorporating a Bayesian probability distribution over the model. As new data arrive (ground and air contamination readings), the puff mass estimates will be revised.

It was shown by Smith et al. (1995) that the relevant uncertainties could be modelled by describing the evolution of puffs and puff fragments within the system by a high dimensional Gaussian process. This process exhibites many conditional independences that can be utilised to speed up the revision of the probability distributions of the quantities of interest, in this case puff and puff fragments masses, in the light of incoming data. Smith et al. express the variables of interest and their conditional independence structure in a dynamic junction tree over cliques which develop over time, where cliques represent collections of puffs which effect each other, and each clique becomes a single node in a junction or clique tree. The authors gave an exact algorithm for the quick absorption of information on the junction tree cliques when information sets arrive on one clique at a time.

1.2 Aims

The basic issues of this research are as follows:

Firstly, to model the qualitative information of the experts of how they believe the emission of contamination will develop over time.

Secondly, to build a probabilistic expert system which is able to:

- generalise current exact Bayesian propagation algorithms on dynamic junction trees of cliques as described in Smith et al. (1995) to even faster algorithms (computational efficiency).
- modify these exact algorithms (to approximation algorithms) when the conditional independence structures of the junction trees may be destroyed.

Thirdly, to generalise these linear updating algorithms to allow dispersal concentration readings to be a non-Gaussian process. The Dynamic Generalised Linear Model (DGLM) of West, Harrison and Migon (1985) will be modified to give closed form updating solutions which can be used on dynamic junction trees for a variety of non-Gaussian processes.

1.3 Outline of the Thesis

Chapter (2) introduces review material on the theory of the Dynamic Linear Models, presenting the basic concepts that will be necessary for the development of Chapters (3), (6) and (8).

In Chapter (3) we introduce the traditional dispersal models in general, with particular emphasis on puff models. A statistical model which embeds physical models to handle uncertainties from different sources is defined and described.

Chapters (4) and (5) provide a necessary background for the development of the new material presented in Chapters (7) and (8). Chapter (4) looks at some graph-theoretical results on graphical modelling; Chapter (5) is devoted to information divergence, where definitions and properties of some separation measures are given.

Chapter (6) deals with describing the source emission process by considering several scenarios of expert judgement about the development of the release. An illustration of adaptation with simulated data is given for different emission profile shapes. Also in this chapter we discussed uncertainty management of key parameters of the emission process namely the release height.

Chapter (7) introduces two classes of new operators (exact and non- exact) which act on a high dimensional Gaussian process transforming its junction tree to another tree which accommodates data induced dependences (conditional independences implicit in the clique structure before the data arrived are no longer necessarily valid after the data are observed). Also we proposed a new approximation scheme using edge deletion (neglecting weak dependences) in order to achieve computational efficiency.

In Chapter (8) we adapt the updating algorithms of Dynamic Generalised Linear Models (DGLMs) to updating algorithms on dynamic junction trees. The appropriateness of the dynamic approximation is based on the Hellinger metric.

Finally, Chapter (9) consists of conclusions and discussion. Some further issues are suggested as possible topics for future research.

Chapter 2

Dynamic Linear Models: A Brief Review

2.1 Introduction

The Dynamic Linear Model (DLM) offers many facilities which will be used throughout this thesis. These include the use of expert judgement to start up the system; the construction of complex models from simple components using the superposition principle; the intervention analysis; simple sequential updating recursions; and Dynamic Generalised Linear Models (DGLM).

In this chapter we briefly review the main concepts from the DLM, Harrison & Stevens (1976) and West & Harrison (1989) that will be necessary for the discussion of some of the basic ideas in the following chapters.

2.2 Bayesian Forecasting and Dynamic Models

In the control literature, linear dynamic systems are used by control engineers to monitor and control the state of a system or a process as it changes over time. To determine whether a system is operating satisfactorily, it is necessary to know the behaviour of the system at any time. For example, in the case of a nuclear accident, the state of the system may be the position of the plume of the contaminated material. The state of the system is random because physical systems are usually subject to random disturbances and the observations taken on the system are often noisy observations.

Making inference about the state of the system from noisy measurements and consequently deriving the predictive distributions of future observations is an important problem. Explicitly, the problem can be expressed in a general dynamic form as

$$Y_t = f_t(\boldsymbol{\theta}_t) + \nu_t \tag{2.1}$$

$$\boldsymbol{\theta}_t = g_t(\boldsymbol{\theta}_{t-1}) + \boldsymbol{\omega}_t \tag{2.2}$$

where the first equation relates the observations Y_t with the state $\boldsymbol{\theta}_t$ of the system, and the second relates the state variables at time t and those at time $t-1$. In the first equation, called the observational equation, the quantity ν_t reflects observational error. The second equation, called the state equation, assumes that the state at time t cannot be determined exactly by the state at time $t-1$ because of the effect of many unknown factors summarised in the random error $\boldsymbol{\omega}_t$.

This state space representation of the problem is based on the so-called Markov property, which implies the independence of the future of a process from its past, given the present state. In this case the state of the system summarises all the information from the past that is needed in order to predict the future.

The objective of the analysis is to make inference about $\boldsymbol{\theta}_{t+k}, k \geq 0$ given the set of observations $y_1, \ldots, y_{t-1}$. Originally Kalman (1960) derived a recursive algorithm to estimate the state of the system using the properties of orthogonal projection on linear spaces. Mehra (1979) summarised the key aspects of this approach to forecasting (known as the filtering approach) with particular emphasis on the original work done by Kalman. The idea of using the state-space engineering representation and the principle of recursive updating of information in statistics did not appear until the early to mid 1970's. Harrison & Stevens (1976) adopted a state–space representation in the context of Bayesian inference to describe the Dynamic Linear Model (DLM). An intrinsic difference between Harrison & Stevens's approach and that of Kalman was that Kalman expressed his recurrence relationships in terms of moments, whereas Harrison & Stevens used the same updating equations to describe the evolution of a fully probabilistic Gaussian process.

Bayesian methods are often useful in forecasting problems where there is little or no useful historical information available at the time the initial forecast is required. In this situation the early forecasts must be based on subjective assessment and experienced judgement. As the time series information becomes available, we then use Bayes theorem to combine our prior information with the observed data through the likelihood function - the joint probability of the data under the stated model assumptions- to give the posterior distribution or information. This prior to posterior process can be expressed as

$$\textbf{posterior} \propto (\textbf{prior} \times \textbf{observed likelihood})$$

An example of this process is the forecasting of contaminated material which has a short life after a nuclear accident. An experienced judgement is needed at the beginning of the emission process. This is a process based on using Bayes theorem for updating a degree of belief expressed by a probability distribution in the light of new information. Bayesian forecasting uses Bayesian inference to study systems through dynamic models.

2.3 The Univariate DLM

2.3.1 Definition

Following the notation and terminology of West & Harrison (1989), the standard normal DLM is described as follows.

Let $Y_t, t = 1, 2, \ldots$ represent a time series of scalar observations. At time t we have the following defining equations

$$Y_t = \boldsymbol{F}_t^T \boldsymbol{\theta}_t + \nu_t \tag{2.3}$$

$$\boldsymbol{\theta}_t = G_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t \tag{2.4}$$

where the quantity $\boldsymbol{\theta}_t$ is an n_t-dimensional state parameter vector evolving through time according to the evolution or system equation (2.4). In the application we have in mind, the dimension of $\boldsymbol{\theta}_t$ will have the novel feature that it will depend on t as a function of the parameter of the model. G_t is the known $n_t \times n_t$ system matrix of the model defining the systematic component of the evolution and $\boldsymbol{\omega}_t$ is a stochastic term which describes random changes in the state vector, and which provides an increase in uncertainty over the time interval as $\boldsymbol{\theta}_{t-1}$ changes to $\boldsymbol{\theta}_t$. Conditional on the past information D_{t-1} available prior to time t, it is typically assumed that $(\boldsymbol{\omega}_t | D_{t-1}) \sim N[\mathbf{0}, W_t]$ where W_t

is a known covariance matrix which provides the measure of increased uncertainty or loss of information. The structure of this matrix is defined using discount factors as discussed in Section (2.4). The state vector relates to the observation at time t via equation (2.3), systematically through the known, n_t-dimensional regression vector $\boldsymbol{F}_t$ and stochastically via the observational noise term ν_t. Typically $(\nu_t|D_{t-1}) \sim N[0, V_t]$ with V_t known apart from a constant or a scalar precision parameter which may be estimated (Harrison & West 1986, 1987). The mean response at t is $\mu_t = \boldsymbol{F}_t^T\boldsymbol{\theta}_t$, which is simply the expected value of Y_t in equation (2.3) and defines the *level* of the series at time t. Finally, the sequences $\{\nu_t\}$ and $\{\omega_t\}$ are usually assumed uncorrelated and mutually uncorrelated. The univariate DLM can then be characterised by a quadruple

$$\{F, G, V, W\}_t = \{F_t, G_t, V_t, W_t\}$$

which is known at time t. This quadruple, together with the initial information $(\boldsymbol{\theta}_0|D_0) \sim N[\boldsymbol{m}_0, C_0]$, where $\boldsymbol{m}_0$ is the prior expectation for $\boldsymbol{\theta}_0$ and C_0 is its covariance matrix, define the DLM, assuming the initial distributions for the state to be independent of ν_t and $\boldsymbol{\omega}_t$.

2.3.2 Model Updating

Suppose that at each time t an observation $Y_t = y_t$ is made such that $D_t = \{y_t, D_{t-1}\}$. Then, with $\{F, G, V, W\}_t$ known for all t the, DLM can be updated as follows.

1. At time $t-1$ we have a posterior distribution for the state vector given as

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim N[\boldsymbol{m}_{t-1}, C_{t-1}]$$

for some mean $\boldsymbol{m}_{t-1}$ and covariance matrix C_{t-1}.

2. Using equation (2.4), prior distribution on the state vector for time t, $(\boldsymbol{\theta}_t|D_{t-1})$ will be normally distributed with defining moments

$$\begin{aligned} E[\boldsymbol{\theta}_t|D_{t-1}] &= \boldsymbol{a}_t &= G_t\boldsymbol{m}_{t-1} \\ Var[\boldsymbol{\theta}_t|D_{t-1}] &= R_t &= G_tC_{t-1}G_t^T + W_t \end{aligned}$$

3. At this stage we forecast the expected new observation according to equation (2.4). The one-step ahead forecast distribution for $(Y_t|D_{t-1})$ will be normally distributed with defining moments

$$\begin{aligned} E[Y_t|D_{t-1}] &= f_t &= \boldsymbol{F}_t^T\boldsymbol{a}_t \\ Var[Y_t|D_{t-1}] &= Q_t &= \boldsymbol{F}_t^T R_t \boldsymbol{F}_t + V_t. \end{aligned}$$

4. When the new observation $Y_t = y_t$ is taken, the state vector is updated according to Bayes' rule, and from the joint normal distribution of $Y_t, \boldsymbol{\theta}_t$ we obtain the posterior distribution at time t for $\boldsymbol{\theta}_t$

$$(\boldsymbol{\theta}_t|D_t) \sim N[\boldsymbol{m}_t, C_t]$$

where

$$\boldsymbol{m}_t = \boldsymbol{a}_t + A_t(y_t - f_t)$$

$$C_t = R_t - A_tQ_tA_t^T.$$

with $A_t = R_t\boldsymbol{F}_tQ_t^{-1}$ known as the Kalman filter or adaptive coefficient.

This closed form of updating can be used simply for obtaining the k-step ahead distribution of a future observation $Y_{t+k}, k \geq 1$. For this we have to make inferences about the future parameters $\boldsymbol{\theta}_{t+k}$. These distributions are summarised in the following subsection.

2.3.3 Forecast Distributions

Definition

The *forecast function* $f_t(k)$ is defined for all integers $k \geq 0$ and for any current time t as

$$f_t(k) = E[\mu_{t+k}|D_t] = E[\boldsymbol{F}^T_{t+k}\boldsymbol{\theta}_{t+k}|D_t]$$

where

$$\mu_{t+k} = \boldsymbol{F}^T_{t+k}\boldsymbol{\theta}_{t+k}$$

is the mean response function at time $t + k$.

For k strictly greater than 0, the forecast function is defined as

$$f_t(k) = E[Y_{t+k}|D_t] \quad (k \geq 1).$$

The form of the forecast function in k plays a major role in defining DLMs. It is considered as a guide to the appropriateness of a particular model in any application. The following theorem provides the full forecast distributions for the series Y_t and the state vector $\boldsymbol{\theta}_t$. Examples of this will be given in Chapter (6).

Theorem 2.1. For each time t and $k \geq 1$, the k-step ahead distributions for $\boldsymbol{\theta}_{t+k}$ and Y_{t+k}, given D_t are given by

(a) State distribution: $(\boldsymbol{\theta}_{t+k}|D_t) \sim N[\boldsymbol{a}_t(k), R_t(k)]$,

(b) Forecast distribution: $(Y_{t+k}|D_t) \sim N[f_t(k), Q_t(k)]$,

with moments recursively defined by

$$f_t(k) = \boldsymbol{F}^T_{t+k}\boldsymbol{a}_t(k)$$

and

$$Q_t(k) = \boldsymbol{F}_{t+k}^T R_t(k) \boldsymbol{F}_{t+k} + V_{t+k},$$

where

$$\begin{aligned} \boldsymbol{a}_t(k) &= G_{t+k}\boldsymbol{a}_t(k-1), \\ R_t(k) &= G_{t+k}R_t(k-1)G_{t+k}^T + W_{t+k}, \end{aligned}$$

with starting values $\boldsymbol{a}_t(0) = \boldsymbol{m}_t$ and $R_t(0) = C_t$. In the special case that the system matrix G_t is constant $G_t = G$ for all t, then for $k \geq 0$,

$$\boldsymbol{a}_t(k) = G^k \boldsymbol{m}_t,$$

so that

$$f_t(k) = \boldsymbol{F}_{t+k}^T G^k \boldsymbol{m}_t.$$

Proof. See West & Harrsion (1989), p.115.

If both F_t and G_t are constants, then the DLM is known as time series DLM or TSDLM. The forecast function of the TSDLM has the form

$$f_t(k) = \boldsymbol{F}^T G^k \boldsymbol{m}_t.$$

2.4 The Discounting Concept

In implementing a DLM, the setting of the evolution error variance W_t is often not easy. This difficulty has been overcome by introducing the concept of discounting which was developed by Brown (1962) using only one discount factor for the global time series model. Harrison (1965) and Ameen & Harrison (1984) have used different discount factors for different components in the DLM. The idea is based on the fact that W_t

models a decay of information between observations. Recalling that

$$R_t = Var(\boldsymbol{\theta}_t|D_{t-1}) = G_tC_{t-1}G_t^T + W_t.$$

when there is no system error, then $R_t = G_tC_{t-1}G_t^T$. This means that W_t introduces uncertainty so that the quantity $G_tC_{t-1}G_t^T$ can be considered as a discounted R_t with a discount factor $\delta, 0 < \delta < 1$. Thus $R_t = \frac{G_tC_{t-1}G_t^T}{\delta}$ so that $W_t = G_tC_{t-1}G_t^T\frac{(1-\delta)}{\delta}$. In particular, in the case of a steady model with $\boldsymbol{F}_t = G_t = 1$, we have
$R_t = \frac{C_{t-1}}{\delta}$ so that $W_t = C_{t-1}(\delta^{-1} - 1)$.

2.5 Dynamic Estimation of Variance

So far we have assumed that the observation variance V_t is known. But in most applications this is not the case. Several approaches to learning on line about V_t have been suggested (see, for example, Ameen & Harrison, 1985). A tractable fully Bayesian learning mechanism when the variance is constant is available.

In the context of our application, a conjugate prior to posterior analysis with an unknown variance is possible in a limited sense if we use the following parametrisation of the distribution of a noisy observation Y_t and a vector $\boldsymbol{Q}_t$ of masses of contamination. This is an obvious modification of the algorithms of De Groot (1971) and West & Harrison (1989).

Observation

$$(Y_t|\boldsymbol{Q}_t, \nu_t) \sim N[\boldsymbol{F}^T\boldsymbol{Q}_t, \nu_tV_t] \tag{2.5}$$

where we assume that the observational variance known up to a scalar multiple, with

ν_t an unknown parameter whilst, V_t and the constant regression vector $\boldsymbol{F}$ are assumed to be known. It is obvious that the larger the value of ν_t, the larger the conditional variance of Y_t. In more complex models, V_t will be a function of various characteristics of the distribution of $\boldsymbol{Q}_t$.

State Information

Assume that

$$(\boldsymbol{Q}_t|D_{t-1},\nu_t) \sim N[\boldsymbol{a}_t,\nu_t R_t] \tag{2.6}$$

$$(\nu_t^{-1}|D_{t-1}) \sim G[\alpha_{t-1},\beta_{t-1}] \tag{2.7}$$

Updating Equations

$$\begin{aligned}(\boldsymbol{Q}_t|D_t,\nu_t) &\sim N[\boldsymbol{m}_t,\nu_t C_t]\\ (\nu_t^{-1}|D_t) &\sim G[\alpha_t,\beta_t]\end{aligned}$$

where

$$\boldsymbol{m}_t = \boldsymbol{a}_t + A_t e_t \tag{2.8}$$

$$C_t = R_t - A_t A_t^T (V_t + \boldsymbol{F}^T R_t \boldsymbol{F}) \tag{2.9}$$

with $R_t, V_t, \boldsymbol{F}, \boldsymbol{a}_t$ defined as before and

$$\begin{aligned}e_t &= y_t - \boldsymbol{F}^T \boldsymbol{a}_t\\ A_t &= \frac{R_t\boldsymbol{F}}{[V_t+\boldsymbol{F}^T R_t \boldsymbol{F}]}\end{aligned}$$

The parameters of ν_t are updated using the equations

$$\alpha_t = \alpha_{t-1} + \frac{1}{2} \tag{2.10}$$

$$\beta_t = \beta_{t-1} + \frac{1}{2}\epsilon_t^2 \tag{2.11}$$

where $\epsilon_t = \frac{y_t - \boldsymbol{F}^T \boldsymbol{m}_t}{(V_t + \boldsymbol{F}^T R_t \boldsymbol{F})^{1/2}}$.

Note that this distribution is modified in the light of the normalised residual associated with y_t. The expectation $E[\nu_t|D_t]$, in particular, is given by $\frac{\beta_t}{(\alpha_t - 1)}$ which, by repeated substitution in (2.10) and (2.11) gives

$$E[\nu_t|D_t] = \frac{\beta_0 + 1/2 \sum_{t=1}^{n} \epsilon_t^2}{\alpha_0 + 1/2n - 1}$$

Setting $\alpha_0 = 1$ and letting $\beta_0 \to 0$ (a vague prior distribution on ν_t initially) we obtain

$$E[\nu_t|D_t] = \frac{1}{n} \sum_{t=1}^{n} \epsilon_t^2$$

which is the naive estimate of $(\nu_t|D_t)$ based on normalised residuals.

2.6 Model Specification and Design

As discussed above, a TSDLM can be described by a quadruple $\{F, G, V_t, W_t\}$ with the constant pair $\{F, G\}$ known. The form of the forecast function $f_t(k)$ as a function of the step ahead index k plays an important role in determining this pair. Based on his experience, the expert may have a certain idea about the expected behaviour of the series. This idea may be expressed as a form of forecast function of the model. The pair $\{F, G\}$ can then be deduced to match this form. The following are some examples of some basic forecast functions which may represent the expert's view of the expected developments of the series. First we state the following results which follow directly from West & Harrison (1989).

Theorem 2.2. *If a TSDLM has a 2-dimensional state space, the state space can be linearly transformed so that its forecast function can be written as*

$$f_t(k) = \boldsymbol{F}^T G^k \boldsymbol{m}_t$$

where $\boldsymbol{m}^T = (m_{t1}, m_{t2})$, and $\boldsymbol{F}$ and G having one of the following forms

i) $\boldsymbol{F} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad G = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$

ii) $\boldsymbol{F} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad G = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$

iii) $\boldsymbol{F} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad G = \lambda \begin{pmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{pmatrix}$

where $\lambda_1, \lambda_2, \lambda$ and ω are real. The forecast functions associated with these are respectively

i) $f_t(k) = \lambda_1^k m_{t1} + \lambda_2^k m_{t2}$

ii) $f_t(k) = \lambda^{k-1}(\lambda m_{t1} + k m_{t2})$

iii) $f_t(k) = [m_{t1}\cos(k\omega) + m_{t2}\sin(k\omega)]\lambda^k$

An alternative expression for the forecast function (iii) above is

$$f_t(k) = \lambda^k r_t \cos(k\omega + \phi_t)$$

where $r_t^2 = m_{t1}^2 + m_{t2}^2$, r is the *amplitude* of the periodic component $m_{t1}\cos(k\omega) + m_{t2}\sin(k\omega)$, and $\phi_t = arctan(-m_{t2}/m_{t1})$ is the *phase* of the periodic component.

Example 1. This is an example of a damped growth model where the forecast function rises from zero to a maximum height at time k^* and then goes down exponentially. Here the forecast function has the form

$$f_t(k) = m_{t1}\lambda^k + m_{t2}k\lambda^{k-1}$$

with $\boldsymbol{F}^T = (1,0), \boldsymbol{m}_t = (m_{t1}, m_{t2})$ and G as in (ii) above (a 2×2 Jordan block matrix with λ diagonals) where $0 < \lambda < 1$.

Notice that the prior values of the initial state vector $\boldsymbol{\theta}_0$ can be chosen so as to put $m_{t2} = 0$.

Example 2. In this case the forecast function rises to an asymptotic value. It has the form

$$f_t(k) = \lambda_1^k m_{t1} + m_{t2}$$

with $\boldsymbol{F}^T = (1,1)$ and G as in (i) above where $0 < \lambda_1 < 1, \lambda_2 = 1$.

Usually we set $m_{01} = -m_{02}$ so that this gives a (non-negative) expected exponential rise from an initial value of zero to an asymptote m_{t2}.

2.7 The Superposition Principle

The superposition principle (West & Harrison, 1989, p.182) is a powerful statement from the model design point of view. It states that a linear combination of DLMs is itself a DLM. This principle is dependent on additivity properties associated with linear models in general. Moreover, it is important to notice that the statement of this principle would assume that the observational error terms, as well as the system error terms, are jointly normally distributed.

Consider the h time series Y_{it} for integer $h > 1$ generated by the DLM $\{\boldsymbol{F}_i, G_i, V_i, W_i\}_t$ with state vectors $\boldsymbol{\theta}_{it}$ of dimensions n_i for $i = 1, \dots, h$. Assume that for all distinct i and j $(1 \leq i, j \leq h)$ the series ν_{it} and ω_{it} are mutually independent of the series ν_{jt} and ω_{jt}. Then the series

$$Y_t = \sum_{i=1}^{h} Y_{it}$$

follows a DLM $\{\boldsymbol{F}, G, V, W\}_t$ with a state vector given by

$$\boldsymbol{\theta}_t^T = (\boldsymbol{\theta}_{1t}^T, \dots, \boldsymbol{\theta}_{ht}^T),$$

of dimension $n = n_1 + \dots + n_h$ such that

$$\begin{aligned}
\boldsymbol{F}_t^T &= (\boldsymbol{F}_{1t}^T, \dots, \boldsymbol{F}_{ht}^T),\\
G_t &= \textit{block diag}\{G_{1t}, \dots, G_{ht}\},\\
V_t &= \textstyle\sum_{i=1}^{h} V_{it},\\
W_t &= \textit{block diag}\{W_{1t}, \dots, W_{ht}\}
\end{aligned}$$

Proof. See West & Harrison (1989), p.184.

This principle is very useful since it provides a way of building up complex models for simple components. For instance a superposition of the models in examples (1) and (2) above has a regression vector

$$\boldsymbol{F}^T = (1, 1, 1, 0)$$

and a system matrix

$$G = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{pmatrix}$$

With $\boldsymbol{m}^T = (m_{t1}, m_{t2}, m_{t3}, m_{t4})$ we have for $k \geq 0$,

$$f_t(k) = \lambda_1^k m_{t1} + m_{t2} + m_{t3}\lambda^k + k m_{t4}\lambda^{k-1}.$$

In Chapter (6) we will discuss a range of these models to use in different cases.

2.8 Dynamic Generalised Linear Models (DGLM)

To generalise the concept of the DLM to non-normal distribution on the observed series and the state vector, West, Harrison & Migon (1985) proposed the Dynamic Generalised Linear Models (DGLM) which are essentially a dynamic and Bayesian version of the Generalised Linear Models (GLM) in Nelder & Wedderburn (1972). In their generalisation, West, Harrison and Migon made no distributional assumptions about the n- dimensional state vector $\boldsymbol{\theta}_t$ apart from the first two moments, whilst the observables $\{Y_t\}$ were assumed to have a full distribution specification. Explicitly, the authors assumed that observations come from exponential family distribution –although their algorithm holds in general – with a density

$$p(Y_t|\eta_t, V_t) = \exp\{V_t^{-1}[\eta_t Y_t - a(\eta_t)]\}b(Y_t, V_t) \tag{2.12}$$

where η_t and $V_t > 0$ are respectively the natural parameter and the scale parameter of the distribution, $b(y_t, V_t)$, and $a(\eta)$ are known functions. Here we assume that V_t or equivalently $V_t^{-1} = \phi_t$ is a known precision parameter for all t. The system equation is as

$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \qquad \boldsymbol{\omega}_t \sim [\mathbf{0}, W_t] \tag{2.13}$$

where G_t, is an $n \times n$ known system matrix, $\boldsymbol{\omega}_t$ is a zero mean uncorrelated sequence, with $\boldsymbol{\omega}_t$ uncorrelated with $\boldsymbol{\theta}_{t-1}$, and W_t is a known covariance matrix for all t.

Define $\lambda_t = \boldsymbol{F}_t^T \boldsymbol{\theta}_t$ as the linear score where $\boldsymbol{F}_t$ is a known $n \times 1$ regression vector as defined in the standard DLM, and the link function $g(.)$ as

$$g(\eta_t) = \lambda_t = \boldsymbol{F}_t^T \boldsymbol{\theta}_t \tag{2.14}$$

where $g(.)$ is a known, continuous and monotonic function mapping η_t to the real line. Then the observation model is defined by (2.12) and (2.14) while the system equation is defined by (2.13). Here we note that the standard DLM defined by (2.3) and (2.4) is a special case for which the distribution in (2.12) is $N[\lambda_t, V_t]$, the distributions of $\boldsymbol{\theta}_t$ and $\boldsymbol{\omega}_t$ are normal, and $g(.)$ is the identity mapping.

West, Harrison and Migon (1985) reformulates the standard sequential procedure for the DLM and extended it to the non-normal case. In their analysis, the authors drop the normality assumption of the state vector. Similarly, the distribution of the error in the system equation is now only partially specified in terms of its first two moments. An alternative interpretation of their approach assumes that the random quantities $\{\boldsymbol{\theta}_t\}$ defined in the system equation (2.13) are Gaussian and treats the whole process as approximate (see Smith, 1992 and Chapter 8). Under either interpretation the model updating is the same.

2.8.1 The DGLM updating

Suppose that

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim [\boldsymbol{m}_{t-1}, C_{t-1}],$$

and the evolution of the model is defined by assuming that

$$(\boldsymbol{\theta}_t|D_{t-1}) \sim [\boldsymbol{a}_t, R_t]$$

where $\boldsymbol{a}_t = G_t\boldsymbol{m}_{t-1}$ and $R_t = G_tC_{t-1}G^T + W_t$. A full system of recursions is summarised in the following steps:

i) Under (2.14), λ_t and $\boldsymbol{\theta}$ have a joint prior distribution which is partially specified in terms of the first two moments so that

$$(g(\eta_t)|D_{t-1}) \sim [f_t, q_t]$$

where $f_t = \boldsymbol{F}_t^T\boldsymbol{a}_t$ and $q_t = \boldsymbol{F}_t^T R_t \boldsymbol{F}_t$.

ii) In order to obtain the one-step ahead forecast distribution we need the distribution of $(\eta_t|D_{t-1})$. But this distribution is only partially specified since the full distribution form of $\lambda_t = g(\eta_t)$ is not necessarily known. Therefore further assumptions about the prior distribution of λ_t are needed. A conjugate prior for λ_t is supposed to have the form

$$p(\eta_t|D_{t-1}) = c(\alpha_t, \beta_t)\exp\{\alpha_t\eta_t - \beta_t a(\eta_t)\} \qquad (2.15)$$

for some defining parameters α_t, β_t and a normalising constant $c(\alpha_t, \beta_t)$. The defining parameters are chosen such that

$$\begin{aligned} E[g(\eta_t)|D_{t-1}] &= f_t, \\ Var[g(\eta_t)|D_{t-1}] &= q_t. \end{aligned}$$

The one-step ahead forecast distribution can now be obtained via

$$p(Y_t|D_{t-1}) = \int p(Y_t|\eta_t)p(\eta_t|D_{t-1})d\eta_t.$$

From (2.12) and (2.15) we have

$$p(Y_t|D_{t-1}) = \frac{c(\alpha_t, \beta_t)b(Y_t, V_t)}{c(\alpha_t + \phi_t Y_t, \beta_t + \phi_t)}.$$

iii) Observing $Y_t = y_t$, find the posterior for η_t in the conjugate form

$$p(\eta_t|D_t) = c(\alpha_t + \phi_t Y_t, \beta_t + \phi_t)\exp\{[(\alpha_t + \phi_t Y_t)\eta_t - (\beta_t + \phi_t)a(\eta_t)]\} \quad (2.16)$$

Denote the posterior mean and variance for $\lambda_t = g(\eta_t)$ by

$$\begin{aligned} f_t^* &= E[g(\eta_t)|D_t] \\ q_t^* &= Var[g(\eta_t)|D_t] \end{aligned}$$

iv) Using the joint posterior distribution for λ_t and $\boldsymbol{\theta}$ we can obtain the posterior moments for $(\boldsymbol{\theta}_t|D_t)$ as

$$\begin{aligned} p(\lambda_t, \boldsymbol{\theta}_t|D_t) &\propto p(\lambda_t, \boldsymbol{\theta}_t|D_{t-1})p(Y_t|\lambda_t) \\ &\propto [p(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})p(\lambda_t|D_{t-1})]p(Y_t|\lambda_t) \\ &\propto p(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})[p(\lambda_t|D_{t-1})p(Y_t|\lambda_t)] \\ &\propto p(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})p(\lambda_t|D_t). \end{aligned}$$

Here $\boldsymbol{\theta}_t$ is conditionally independent of Y_t given λ_t and D_{t-1}. The posterior distribution of $\boldsymbol{\theta}_t$ is

$$p(\boldsymbol{\theta}_t|D_t) = \int p(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})p(\lambda_t|D_t)d\lambda_t \quad (2.17)$$

The second probability density in the integrand can be obtained from step (iii) above. The first probability density is not always fully specified, but we only need its first two moments in order to obtain the first two moments for the posterior distribution of $(\boldsymbol{\theta}_t|D_t)$. The first two moments for $(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})$ can be estimated from standard Bayesian techniques. A detailed discussion of the Linear Bayesian Estimation (LBE) of moments of $(\boldsymbol{\theta}_t|\lambda_t, D_{t-1})$ is given in West & Harrison (1989), p.561. The LBE of the conditional mean $E[\boldsymbol{\theta}_t|\lambda_t, D_{t-1}]$ is given by

$$\hat{\boldsymbol{E}} = \boldsymbol{a}_t + R_t \boldsymbol{F}_t(\lambda_t - f_t)/q_t,$$

for all λ_t. The estimate of the variance is given by

$$\hat{V} = R_t - R_t \boldsymbol{F}_t \boldsymbol{F}_t^T R_t / q_t$$

for all λ_t. Now from (2.17) we have

$$\begin{aligned} E[\boldsymbol{\theta}_t | D_t] &= E[E\{\boldsymbol{\theta}_t | \lambda_t, D_{t-1}\} | D_t] \\ Var[\boldsymbol{\theta}_t | D_t] &= Var[E\{\boldsymbol{\theta}_t | \lambda_t, D_{t-1}\} | D_t] + E[Var\{\boldsymbol{\theta}_t | \lambda_t, D_{t-1}\} | D_t] \end{aligned}$$

and these can be estimated by substituting the LBE estimates as follows:

$$\begin{aligned} \boldsymbol{m}_t &= E[\hat{E}] \\ &= E[\boldsymbol{a}_t + R_t \boldsymbol{F}_t (\lambda_t - f_t)/q_t | D_t] \\ &= \boldsymbol{a}_t + R_t \boldsymbol{F}_t (f_t^* - f_t)/q_t \end{aligned}$$

and

$$\begin{aligned} C_t &= E[\hat{V}] + Var[\hat{E}] \\ &= E[R_t - R_t \boldsymbol{F}_t \boldsymbol{F}_t^T R_t / q_t] + Var[\boldsymbol{a}_t + R_t \boldsymbol{F}_t (\lambda_t - f_t)/q_t | D_t] \\ &= R_t - R_t \boldsymbol{F}_t \boldsymbol{F}_t^T R_t / q_t + R_t \boldsymbol{F}_t \boldsymbol{F}_t^T R_t q_t^* / q_t^2 \\ &= R_t - R_t \boldsymbol{F}_t \boldsymbol{F}_t^T R_t (1 - q_t^*/q_t)/q_t \end{aligned}$$

This step completes the updating procedure.

Other approaches which accommodate general forms of observational distributions are MCMC methods, one example of which is the Gibbs sampler. These can be used to obtain the desired posterior distributions in non-algebraic form to any degree of accuracy. Such methods were introduced by Hastings (1970) and Metropolis et al. (1953) and have since been developed vigorously in a number of Bayesian applications (see Gelfand & Smith, 1990, Tanner & Wong, 1987 and West, 1992). Their greatest advantage is their relatively straightforward implementation which is attained at the cost of computational efficiency. Typically these methods, though accurate, are relatively slow and so are not currently applicable to the real time problem we have in mind here.

2.9 Multi-Process Models

In many cases, the modeller has in mind not just a single model, but several models corresponding to different possible scenarios which might explain a time series, and not just a single model. Harrison & Stevens (1976) introduced the *multi-process model* methodology for discriminating between rival DLMs $M^{(i)}(i = 1, 2, \ldots, m)$ and considering them for the series simultaneously. The authors distinguished two classes of multi-process models, Class I and Class II, that are different in structure. In this thesis we are only concerned with Class I, where the class of alternative models is constant in time. A brief discussion of this class follows (see also West & Harrison, 1989, p.439). Let the process $\{Y_t\}$, $(t = 1, 2, ...)$ follow a DLM $M_t(\boldsymbol{\alpha})$, where $\boldsymbol{\alpha}$ are uncertain defining parameters of the model (e.g. discount factors, elements of a constant system matrix G, ... etc.). The precise value $\boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ is uncertain. Let the finite discrete set $\mathcal{A}= \{\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_m\}$ denote the parameter space. Define

i) The probability that the process follows the model $M^{(i)}$ given D_{t-1} as

$$p_{t-1}^{(i)} = p(M_{t-1}^{(i)}|D_{t-1}) = p(\boldsymbol{\alpha}_i|D_{t-1}), \qquad (i = 1, ..., m)$$

ii) $\boldsymbol{Z}$ as a vector of quantities of interest, for example the state vector $\boldsymbol{\theta}_t$, a future observation $Y_{t+k}, k > 0$, ...etc.

Then

1. The posterior probability of $M^{(i)}$ is updated via

$$p_t^{(i)} \propto p(Y_t|\alpha_i, D_{t-1})p_{t-1}^{(i)}, \quad i = 1, ..., m$$

where $p(Y_t|\alpha_i, D_{t-1})$ is the predictive density for Y_t assuming the model $M^{(i)}$.

2. The marginal posterior densities are

$$p(Z|D_t) = \sum_{i=1}^{m} p(Z|\alpha_i, D_t)p_t^{(i)}$$

which are discrete probability mixtures of the standard T or normal distributions.

For example, in our application the emission process of the contaminated material after an accident can be modelled as a DLM which will lead to estimates of source term profile and predictions of the contamination spread. However, the model does not consider the uncertainty of the release height. One solution to this problem is to run a multi-process model which lets the mixing value α be the height. Each model in this set will be given a prior probability representing expert judgement on the likely height. These probabilities will then be updated in the light of monitoring data. By this means the data will give weights to most likely models. Details are given in Chapter (6).

Class I models provide a quick and often reliable model, especially if, as is often the case in their forecasting situations, the models are different in character, e.g. if they relate to different physical theories associated with different dispersal models. Details are provided in Chapter (6).

Chapter 3

A Bayesian Forecasting of Atmospheric Dispersion

3.1 Introduction

In the case of environmental disasters of different types (e.g. nuclear or chemical accidents) appropriate countermeasures must be taken to mitigate the consequences to the population and environment. A quick and accurate prediction of the dispersion of the contaminated material is crucial.

Conventional atmospheric dispersion models (physical models) are widely used for forecasting toxic contamination and obtaining results in real-time with varying degrees of accuracy. These models are deterministic, and one of the most significant problems associated with their use in prediction is the large degree of uncertainty inherent in their predictions.

Atmospheric dispersion is a stochastic phenomenon and, in general, the concentration observed at a given time and location downwind of a source cannot be predicted with precision (Chatwin, 1982). Concentration is a random quantity which should be described statistically or in a probabilistic framework rather than deterministically. Of primary importance is the need to define and reduce the uncertainty associated with predictions. The main sources of uncertainty are :

1. Uncertainties or errors in model input data such as meteorological data; source data; topographical data (surface characteristics, hills, coast lines, etc.). Source information, including the grid reference of the release point, and height of the release, also represents a source of uncertainty (see Eckman et al. 1992).

2. Errors in the field concentration measurements and incomplete knowledge of the expert judgmental data.

3. The uncertainty arising from the poorness of the physical model.

4. Uncertainty may be due to natural (stochastic) variability, (Fox, 1984).

These uncertainties may lead to a destabilisation of the decision process when environmental survey results disagree with the model predictions. Because of this, the uncertainty element must be studied as an integral part of any comprehensive model performance evaluation. Evaluation and identification of the range of model uncertainties provide a deep insight into model capabilities and increase our confidence in decision- making based on models (Rao & Hosker, 1993).

Several different but complementary approaches to defining and reducing uncertainties have been investigated. Until recently, none of the operational systems has handled uncertainty explicitly and most of the approaches are merely academic (see

Govaerts 1993); so the assessment of uncertainties and their communication to the decision- maker remains an important challenge.

Recently, and within the framework of building a decision support system to improve emergency management of any future accidental release of radioactivity, Smith & French (1993) addressed this problem and considered an atmospheric system design for the following purposes:

i) The assessment of uncertainties and their communication to the decision-maker in an operationally flexible system.

ii) The implementation of a data assimilation procedure to update predictions through the use of Bayesian methodology which:

1. Combines information from different sources.
2. Gives a probabilistic measures of uncertainty associated with the combination of information.

This chapter is concerned with describing a Bayesian statistical model (see Smith & French, 1993) which embeds the dispersal model in a description of the uncertainties mentioned above. This both allows the assimilation of data to update current forecasts and also expresses an appropriate degree of uncertainty associated with any forecasts or estimates. The statistical model is carried out within a Bayesian Paradigm (see Box & Taio, 1973; French, 1986; Smith, 1988 and West & Harrison, 1989).

3.2 Atmospheric Dispersion Models

This section gives a brief idea about the dispersion models used in modelling atmospheric transport by wind currents (advection) and turbulent diffusion. The task of the atmospheric dispersion module is to calculate space- and time- dependent air and ground concentrations of radionuclides.

Deterministic mathematical models are widely used in atmospheric studies. Differential equations are usually employed to describe the atmospheric dispersion process, and the system is summarised in terms of the solution of the differential equations. Several dispersion models have been developed which are basically classified as Lagrangian and Eulerian models.

Lagrangian models of atmospheric dispersion processes are usually numerical, and are trajectory models which simulate the release as a sequence of particles following the history of material in time and space (see ApSimon et al. 1989). It was soon recognised that these models had several limitations. For example, experience suggested that they cannot provide accurate results quickly enough because they are dependent on detailed information about source term, atmospheric parameters and terrain data. Generally this information cannot be used in real-time because complex numerical models also require considerable computational time and capabilities. All these factors make these models unreliable in real-time.

Eulerian models describe a plume as a diffusion/advection equation. They are used to calculate finite difference solutions of this equation. These solutions can take account of time-dependent wind fields and realistic vertical profiles of the wind velocity and the diffusion (see Pasler-Sauer, 1985). Again basing a statistical analysis directly on these equations looks unpromising (see Smith & French, 1993). However, under certain conditions it is sometimes possible to obtain analytic steady state solutions. The Gaussian plume model is one solution that arises on assuming desirable physical features such as stationarity, a constant wind vector and homogeneous terrain. It applies an analytical solution of the steady state diffusion advection equation. (For details of Gaussian plume models see Pasler-Sauer, 1985).

Unfortunately, because these solutions are suitable only for a “stable” environment (i.e. constant wind/terrain), they are not expected to perform well in turbulent flows over complex terrain. Furthermore, their use in the early phase of the release is obviously suspect since any sense of steady state will not yet have been reached.

3.3 Puff Models

The puff models have been proposed by many authors (e.g. Mikkelsen et al., 1984) to overcome the shortcomings of a standard plume model which are revealed in its inappropriateness in handling non-stationary, non-homogeneous flow and turbulence situations.

The basic principle for a computational puff model for prediction of atmospheric dispersion is the simulation of the continuous emission from the source by a proper distribution of discrete sequence of small puffs of different sizes. These are released at regular time intervals and then diffuse and disperse independently. Figure 3.1 shows the plume as represented by means of a puff model in which the circles represent individual puffs. Each individual puff represents an ellipsoidal spatial concentration distribution which is

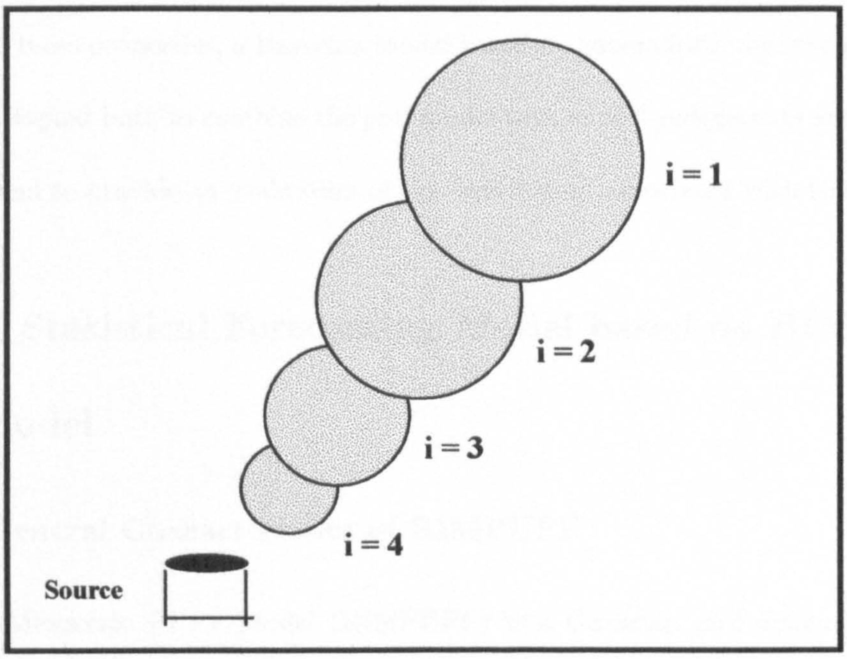


Figure 3.1: The plume as represented by puffs

often hypothesised to be Gaussian–usually truncated around an outer ellipse to describe the bounded nature of the puff better. The puff model has the following properties.

1. The model can handle the non-stationary flow associated with source emissions because different masses under puffs can reflect the often uneven pattern of an accidental release.

2. The local meteorological parameters and the resulting dispersion parameters associated with each puff dispersal can be made different, thus reflecting the characteristics of the wind field at the location of that puff.

3. Various tracer experiments (e.g. Pasler-Sauer, 1985) have suggested that puff models work well in practice in the short term.

Because of these properties, a Bayesian model based on generalisation of the puff model has been adopted both to combine the puff model with expert judgements and monitoring data, and to provide an evaluation of the uncertainty associated with the forecasts.

3.4 A Statistical Forecasting Model based on RIMPUFF Model

3.4.1 General Characteristics of RIMPUFF

The RIso-Mesoscale PUFF-Model (RIMPUFF) is a Gaussian puff dispersion model developed at Riso in Denmark (see Mikkelsen et al., 1984 and Thykier-Nielsen & Mikkelsen, 1991). It is a fast operational computer code suitable for real-time simulation of hazards from radioactivity released to the atmosphere. It has recently been adopted for inclusion into many decision support systems including RODOS.

RIMPUFF consists of an algorithm that models a continuous release by a series of consecutively released puffs. At each time step the model advects and diffuses the individual puffs in accordance with local meteorological parameter values. The relationship between the movement and expansion of a puff and the local input parameters is extremely complex and non-linear. Concurrently, the model also monitors the resulting concentrations in selected grid points. The local meteorological parameters are organised in subprograms which can be readily changed or modified according to the needs and opportunities in the actual model situations.

The puff model is structured such that it handles multiple simultaneous sources and its monitoring grid can contain several hundreds of puffs. The puffs are generated with specific release rates in the specified grid. The individual puffs are advected by the wind field. RIMPUFF calculates the locations of puffs on the specified grid by computing their movements during finite time steps, using an interpolated wind field which is based on data from the wind measurement stations.

To compute the growth of the puffs, it is necessary to have simultaneous specifications of the turbulence and/or the atmospheric stability. Once the advection and size of all puffs have been calculated, updated grid concentrations are obtained at each grid point summing up all the contributions from the puffs in the grid.

3.4.2 Stochastic Modification of the RIMPUFF

Smith & French (1993) have made use of the RIMPUFF model. The dispersion of time-dependent atmospheric plume is described by a sequence of directly released puffs whose superposition pattern approximates the concentration distribution of a continuous plume. The puffs are indexed such that puff i is released at time $t = i$. Assume that the mass under puff i is $Q(i)$. i.e. $Q(i)$ is an uncertain quantity which represents the total number of contaminated particles under the ith puff. We define $\boldsymbol{Q}_t = (Q(1), \ldots, Q(t))^T$ which approximates the release profile of the source term. Standard priors are used on the shape of the time profile (the time series) of the release. Such priors can model any uncertainty about the mass released and its duration. This gives a prior mean. Also we can encode "smoothness" in the release profile through the covariances between the $Q(i)$'s.

The spatial concentration of contamination from the ith puff at time t and location $\boldsymbol{s} = (s_1, s_2, s_3)$ where (s_1, s_2) define the horizontal direction of the grid point and

s_3 the vertical is given by the product $F_t(i, s)Q(i)$. The stochastic multiplier $F_t(i, s)$ determines how that emission is distributed over space and time. It is a proportion of the total contaminated particles under the ith puff at site s and time t. Typically $F_t(i,\ s)$ is a complicated deterministic function of parameters, themselves calculated from uncertain meteorological inputs. For example, one of the simplest of such dispersal models is a Gaussian puff (see Pasler-Sauer, 1985), which sets

$$F_t(.,s) = \frac{1}{(2\pi)^{3/2}\sigma_t(1)\sigma_t(2)\sigma_t(3)} \exp\left\{-1/2\left[\sum_{j=1}^{2}\frac{(s_j - u_t(j))^2}{\sigma_t^2(j)} + \frac{(s_3 - h)^2}{\sigma_t^2(3)}\right]\right\} \quad (3.1)$$

where $(u(1), u(2))$ is a wind velocity vector possibly depending on t, and h is the height of the emission. The radial growth of puffs during dispersion as a result of "internal turbulence" is described by the parameters $(\sigma_t(1), \sigma_t(2))$ and $\sigma_t(3)$ which denote puff sizes in horizontal and vertical directions respectively. These last parameters relating to the diffusion are in part functions of meteorological data such as low frequency fluctuations in wind direction. The parameters of $F_t(i, s)$ are calculated in rather complicated ways to take account of heterogeneity in the system. The function $F_t(i, s)$ is often truncated and set to zero for s lying outside a contour with parameters $(\sigma_t(1) = \sigma_t^*(1), \sigma_t(2) = \sigma_t^*(2), \sigma_t(3) = \sigma_t^*(3))$ (say). This explains why only a certain small number of puffs lie over a site s at any time.

Initially the stochastic multipliers $F_t(i, s)$ are assumed to be known, and we only consider uncertainty on the masses. However in Chapter (6) we will address uncertainty on certain parameters of $F_t(i, s)$ such as the release height of the emission. The initial introduction of uncertainty handling through distributions on masses under puffs brings the following advantages:

- Relating an observation to puffs.

  Instantaneous concentrations at monitoring sites are linear functions of $\boldsymbol{Q}_t$. Let $Y(t, \boldsymbol{s})$ denote an observation taken under some overlapping puffs at time t at location $\boldsymbol{s}$. Here $Y(t, \boldsymbol{s})$ represents the total number of contaminated particles at $(t, \boldsymbol{s})$. Now $Y(t, \boldsymbol{s})$, the concentration of contamination, is simply the sum of concentrations of all puffs where the ith puff contributes a proportion $F_t(i, \boldsymbol{s})$ of its total mass $Q(i)$. Thus $Y(t, \boldsymbol{s})$ will be a linear function of the combinations of masses $Q(i)$s.

  In practice, because puffs are typically bounded, it is found that for many dispersion models and scenarios that arise, only a few number of puffs will contribute to contamination at a given detector site at time t. This number of related puffs will be determined by the physical dispersion model. In terms of our formulation, it is implied that all but a few of the multipliers $F_t(i, \boldsymbol{s})$ will be non-zero at a fixed point $(t, \boldsymbol{s})$

- It is possible to use analogues of Bayesian DLM algorithms to assimilate a time series of monitoring data whose states are the uncertain masses.

In our application $Y(t, \boldsymbol{s})$ will be a noisy function of the true contamination $\theta(t, \boldsymbol{s})$ at site $\boldsymbol{s}$ at time t originating from $Q(1), \ldots, Q(t)$. In practice the dispersal of the contaminated material is patchy (see Smith & French, 1993), and this needs to be modelled stochastically as

$$\theta(t, \boldsymbol{s}) = \sum_{i=1}^{t} F_t(i, \boldsymbol{s}) Q(i) + \epsilon(t, \boldsymbol{s})$$

For simplicity we assume that $\epsilon(t, s)$ are each Gaussian with mean zero and variance $U(t, \boldsymbol{s})$, and $\epsilon(t, \boldsymbol{s}_1), \epsilon(t, \boldsymbol{s}_2)$ are independent for sites $\boldsymbol{s}_1, \boldsymbol{s}_2$. As a simple process $(Y(t, \boldsymbol{s})|\theta(t, \boldsymbol{s}))$ is defined to have a Gaussian distribution with mean $\theta(t, \boldsymbol{s})$ and a fixed variance $V(t, \boldsymbol{s})$ where $V(t, \boldsymbol{s})$ is assumed known and represents the "observation and modelling" error. Assume that $Y(t, \boldsymbol{s})$ is independent of all other variables in the system given $\theta(t, \boldsymbol{s})$ then

$$Y(t, \boldsymbol{s})|\theta(t, \boldsymbol{s})) \sim N[\theta(t, \boldsymbol{s}), V(t, \boldsymbol{s})] \tag{3.2}$$

The methods can be generalised to assimilate non-normal data (see Chapter 8).

The general observation process where $\boldsymbol{Y}(t, \boldsymbol{s})$ is a vector of observations taken at time t at a selection of sites $\boldsymbol{s}$ is discussed later in Subsection 3.6.1.1.

Now, conditioning on everything else other than masses, the model provides elegant algorithms to: update distributions of the source term in time; predict contamination over space and time; and hence to obtain predictive distributions of data and also to admit data assimilation.

However, in practice many of the variables coditioned on will be unknown. For example, we may be uncertain about a parameter like the release height and we know that this parameter has a significant effect on the multipliers.

To solve this problem we run mixed models (see Harrison and Stevens, 1976). That is we parallel process several models in a mixture, each with a different release height and update their associated probabilities according to Bayes' rule (for detailed discussion, see Chapter 6).

3.4.3 Modelling Uncertainties about Meteorological Input and the Dispersion Model

When the dispersal model is inadequate and its meteorological inputs are inaccurate then the statistical model described above cannot be expected to work well in practice. Fortunately, however, a splitting feature called pentification which is coded within the deterministic models of RIMPUFF can be adapted to manage, indirectly, much of the uncertainty indicated above.

Within the deterministic code when a puff reaches a certain diameter, the puff splits horizontally into five smaller puffs with associated multipliers also having Gaussian shapes in such a way that the centroid of the spread of the contamination of the five new puffs is the same as that of the puff that splits. To match the original Gaussian shape of the multiplier to the shape of the mixture of the five Gaussian shapes associated with the new puffs, the second moment of the concentration taken over by the new puffs is chosen to match the second moment of the original puff. The total mass of the contaminant associated with the five new puff masses is chosen to equal the original puff mass. In this way the total contamination contributed by the original puff is conserved. Figure 3.2 shows the splitting scheme : one central and four siblings approximate the original single Gaussian puff. Each of the four siblings carries 23.53 per cent of the total amount of mass. The fifth puff is assigned to the remaining 5.8 per cent while, still being located at the origin. Using this pentification concept, it is possible to build a dynamic linear model which may adapt in a simple way to monitoring data and demonstrate the working of such a model. The idea is to let the state variables be the masses of contaminant in each puff keeping all other elements in this splitting

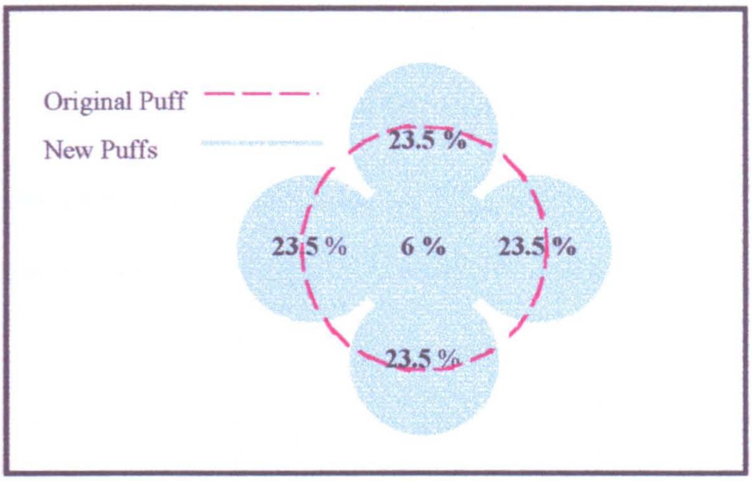


Figure 3.2: Puff pentification

algorithm deterministic. In the statistical model we allow for the possibility that reality may be better modelled by a different percentage split, i.e. that one or more puffs may receive more than their expected share of the contaminant, and that correspondingly others may receive less. As monitoring data are assimilated, the model may learn that such an asymmetric pentification would be more appropriate, and it will adjust the masses of contaminant in each puff accordingly. One effect of this is to shift the overall plume somewhat to take account of such things as misspecification of the wind field and dispersion model.

Explicitly the sibling vector of the children of the i th puff/ puff fragment $\boldsymbol{Q}(i) = (Q(i,1),\ldots,Q(i,5))$ can be expressed as

$$\boldsymbol{Q}(i) = \boldsymbol{\alpha} Q(i) + \boldsymbol{\omega}(i) \tag{3.3}$$

where

$$\begin{aligned}\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) \\ &= (0.235, 0.235, 0.235, 0.235, 0.058)\end{aligned}$$

and

$$\omega(i) = (\omega(i,1), \ldots, \omega(i,5))$$

is a system error chosen to conserve mass i.e. $\sum_{j=1}^{5} \omega(i,j) = 0$ with $\text{Var}[\omega(i,j)] = W$. One simple example sets $\omega(i) \sim N[\mathbf{0}, W^*]$, where W^* is a covariance matrix of shape

$$W^* = \begin{pmatrix} 1 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & 1 & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & 1 & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 1 & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 1 \end{pmatrix} W \qquad (3.4)$$

Now equations (3.2), (3.3) and (3.4) specify a simple linear stochastic system. This system is rich enough to exhibit sensible learning procedures for a variety of plausible scenarios. Also it faithfully mirrors in its structure a dispersal model that a physicist understands. It requires as prior inputs only:

i) The first two moments of the mass under each puff on emission.

ii) A measurement error variance.

iii) A variance parameter to define the stochastic pentification.

3.5 Model Adaptation

As a simple illustration of how the model adapts, consider the trajectory of puff emission as depicted in Figure 3.3. For simplicity, assume that there are only horizontal movements of the plume, that source emissions are independent, and that covariances

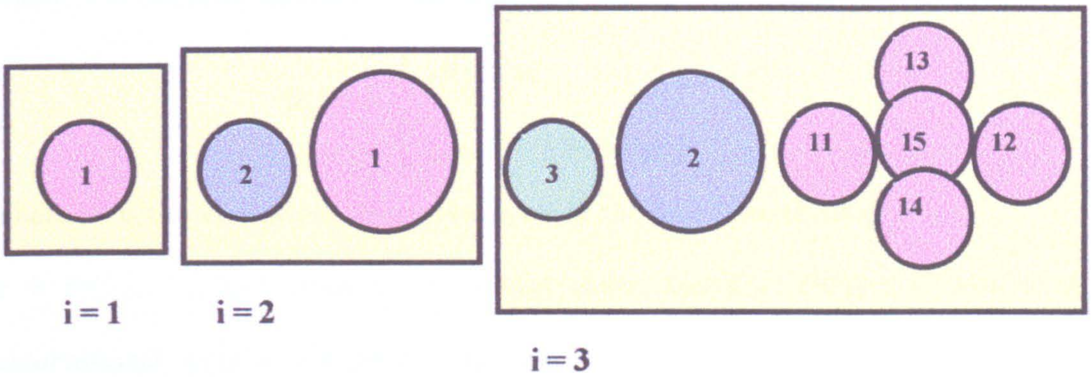


Figure 3.3: Plume trajectory and pentification

on the pentification are set as in (3.4). Consider the case where we observe contamination at a site predicted only to be contaminated by the fragment labelled (1, 5). Then it easily shown from the related equations defining the statistical model and the pentification covariance that the joint Gaussian distribution of $(Q(1), \boldsymbol{Q}(1), Y)$ has a mean vector $(\hat{q}(1), \boldsymbol{\alpha}\hat{q}(1), f\alpha_5\hat{q}(1))$ where

$\boldsymbol{Q}(1) = (Q(1,1), \ldots, Q(1,5))$, $\hat{q}(1)$ is the prior mean of the mass under the first puff,

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5) = (0.235, 0.235, 0.235, 0.235, 0.058)$$

and $f = F_t(., \boldsymbol{s})$ with $F_t(., \boldsymbol{s})$ as defined in (3.1). The covariance matrix of $(Q(1), \boldsymbol{Q}(1), Y)$ is given in a block matrix form as

$$\begin{pmatrix} S & \boldsymbol{\alpha}^T S & f\alpha_5 S \\ \boldsymbol{\alpha}^T S & C & \boldsymbol{c} \\ f\alpha_5 S & \boldsymbol{c}^T & d \end{pmatrix}$$

where S is the prior variance of the mass $Q(1)$ and $C = \{C_{ij}\}$ where

$$\{C_{ij}\} = \begin{cases} \alpha_i^2 S + W, & 1 \leq i = j \leq 5 \\ \alpha_i \alpha_j S - \frac{1}{4} W, & i \neq j \end{cases}$$

where W is the variance of the system error in the first pentification.

$\boldsymbol{c} = f(C_{15}, \ldots, C_{55})^T$ where C_{ij} is defined above, and $d = f^2 C_{55} + V$, with V the observational variance of Y conditional on $Q(1,5)$.

Now, using the usual normal theory, it is easy to derive the revised distribution of $\boldsymbol{Q}(1)$ given Y, which is a multivariate normal with mean vector $\hat{\boldsymbol{q}}_t(1)$ and covariance matrix $\hat{Q}_t(1)$ where

$$\hat{Q}_t(1) = C - \frac{\boldsymbol{c}\boldsymbol{c}^T}{d} \tag{3.5}$$

and $\hat{\boldsymbol{q}}_t(1) = (\hat{q}_t(1,1), \ldots, \hat{q}_t(1,5))^T$ where for $1 \leq j \leq 5$

$$\hat{q}_t(1,j) = \alpha_j \hat{q}(1) + a(1,j) e_t(5) \tag{3.6}$$

where

$$e_t(5) = \frac{y}{f} - \alpha_5 \hat{q}(1) \tag{3.7}$$

$$a(1,j) = \frac{f[\alpha_j \alpha_5 S - 1/4W]}{f^2[\alpha_5^2 S + W + V/f^2]}, \quad j \neq 5 \tag{3.8}$$

$$a(1,5) = \frac{1}{f}\left[1 - \frac{V/f^2}{\alpha_5^2 S + W + V/f^2}\right] \tag{3.9}$$

where

$\alpha_j \hat{q}(1)$ is the expected contamination under puff $(1,j)$ before observing y.

$e_t(5)$ is the difference between the naive estimate of $q(1,5)$ using y and its prior expected value.

$a(1,j)$ is the usual adaptive coefficient of $Q(1,j)$ to y/f.

From the previous equations we notice the following.

1. The adaptation of the fragment (1, 5) associated with the observation y pulls the mean towards the naive estimate y/f.

2. The larger the uncertainty in the source (S) and the uncertainty in the pentification (W) relative to the observational error variance V, the greater the adaptation towards the naive data based estimate.

3. The adaptation of beliefs associated with sibling fragments is interesting. Adaptation of the mean towards or away from the naive estimate y/f of $q(1,5)$ will depend on whether the ratio S/W is large or small. Thus we adjust towards the naive estimate if the source uncertainty is large (assuming this has been misestimated) and away from the naive estimate if the source reading is accurate. In the later case, if more contamination than expected has been observed under puff (1, 5), then less must exist under puffs $(1,j), 1 \leq j \leq 4$. This illustrates the critical role of the settings of prior variances on the subsequent management of uncertainty.

To adjust beliefs about the source emission quantity $Q(1)$, notice that, given y, this has a Gaussian distribution with mean $\hat{q}_t(1)$ and variance $\hat{Q}_t(1)$ where for $f\alpha_5 \neq 0$

$$\hat{q}_t(1) = \hat{q}(1) + a(1)\left[\frac{y}{f\alpha_5} - \hat{q}(1)\right]$$

and

$$a(1) = \frac{1}{f\alpha_5}\left[1 - \frac{V/f^2 + W}{\alpha_5^2 S + W + V/f^2}\right].$$

Very plausibly, remote sites where the value of f^2 is very small will give readings which adapt the estimate of $Q(1)$ very little.

3.6 A Dynamic Fragmenting Puff Model

Although we have argued that the above statistical model might be appropriate from a theoretical standpoint, there is a practical requirement that we have to meet. The covariance matrices within the model can become very large in dimension, and computational efficiency thus declines dramatically. Fortunately, this fragmenting puff model can be restructured as a dynamic junction tree, (see Chapter 7). In fact Smith et al., (1995) address this problem. They have shown how the Bayesian propagation algorithms (Lauritzen & Speigelhalter, 1988) can be modified when the trees evolve dynamically. Their methodology is defined and illustrated within a stochastic version of a fragmenting puff model. In this section we describe briefly their methodology, starting with a notation and some distributional assumptions which we will follow in later chapters.

3.6.1 Model Description

With reference to the puff splitting scheme of section (3.4), fragments arising directly from another puff or puff fragments will be called *children* of a *parent* puff/ puff fragment. In RIMPUFF each parent puff has five children and is said to be pentificate.

Let $m(t, \mathbf{l}) = m(t, l_1, \ldots, l_k)$ be the puff fragment which is the l_k^{th} child of the l_{k-1}^{th} child, ..., of the l_1^{th} child of the puff released at time t. In RIMPUFF $1 \leq l_i \leq 5, 1 \leq i \leq k$. The index k relates to the number of fragmentations that have taken place before fragment $m(t, \mathbf{l})$ appears. Let:

I_T denote the set of all puffs/puff fragments appearing on or before time T.

$Q(\mathbf{l})$ denote the true mass under $m(t, \mathbf{l})$

$\bar{\boldsymbol{Q}}(\mathrm{l})$ denote the vector of true masses under the set of the children of $m(t, \mathrm{l})$.

$\boldsymbol{Q}(\mathrm{l}) = (Q(\mathrm{l}), \bar{\boldsymbol{Q}}(\mathrm{l}))^T$

3.6.1.1 The Observation Process

Let $\boldsymbol{Q}_T$ be the vector of masses of all puffs and puff fragments emitted on or before time T. Let $\boldsymbol{Y}(t, \boldsymbol{s})$ denote a vector of observations taken at time t at a selection of sites $\boldsymbol{s}$. Assume that $\boldsymbol{Y}(t, \boldsymbol{s})|\boldsymbol{\theta}(t, \boldsymbol{s})$ is independent of all other variables in the system. Here $\boldsymbol{\theta}(t, \boldsymbol{s})$ can be interpreted as a random vector relating to the actual mass at time t on site $\boldsymbol{s}$ unconfounded with the observational errors in $\boldsymbol{Y}(t, \boldsymbol{s})$. As a simple process, $\boldsymbol{Y}(t, \boldsymbol{s})|\boldsymbol{\theta}(t, \boldsymbol{s})$ is defined as having a Gaussian distribution with mean $\boldsymbol{\theta}(t, \boldsymbol{s})$ and a fixed covariance matrix V. As explained in Section (3.4), an important feature of puff models is that at all points $(t, \boldsymbol{s})$ of the observation grid, $\boldsymbol{\theta}(t, \boldsymbol{s})$ can be written as

$$\boldsymbol{\theta}(t, \boldsymbol{s}) = F(t, \boldsymbol{s})\boldsymbol{Q}_t + \boldsymbol{\epsilon}(t, \boldsymbol{s}).$$

The matrix $F(t, \boldsymbol{s})$ is a very complicated but known function of $(t, \boldsymbol{s})$ which defines the density of contamination contributed at sites $\boldsymbol{s}$ by each puff or puff fragment at time t. Each row of this matrix corresponds to the weightings used in a dispersal model at a site which is a component of the vector of sites. Notice that $F(t, \boldsymbol{s})$ has non-zero components only on fragments that still exist and have not fragmented further. In practice it is found that only a few puff fragments will be observed at a site at any given time, which implies that for most $(t, \boldsymbol{s})$ many components of each row of $F(t, \boldsymbol{s})$ will be zeros (see Subsection 3.2.4). As before the error process $\boldsymbol{\epsilon}(t, \boldsymbol{s})$ will be Gaussian with zero mean and fixed covariance matrix U. In the particular case of observations at source $\boldsymbol{s} = \mathbf{0}$, where $\theta(t, 0)$ is a scalar we set $\boldsymbol{\theta}(t, \mathbf{0}) = Q(t)$ and hence $\boldsymbol{\epsilon}(t, \mathbf{0}) = 0$.

To specify the joint distribution of $\boldsymbol{Q}_T$, at any time T we need to specify the following processes:

3.6.1.2 The Fragmentation Process

This process assumes that a vector of mass fragments (children) $\bar{\boldsymbol{Q}}(\mathbf{l})$ of a parent $m(t,\mathbf{l})$ is independent of all masses $\boldsymbol{Q}_t$ given the mass $Q(\mathbf{l})$. Using Dawid's (1979), notation this can be written as

$$\bar{\boldsymbol{Q}}(\mathbf{l}) \perp\!\!\!\perp \{\boldsymbol{Q}_t \setminus Q(\mathbf{l})\} | Q(\mathbf{l}).$$

Thus, the masses inherited by fragments depend only on the mass of the parent unfragmented puff and no other puff. Thus, to specify the joint distribution of puff fragments it is only necessary to specify the conditional distribution of $\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})$ for each puff/puff fragment $m(t,\mathbf{l})$.

To model the dispersal of a gas, these conditional distributions are usually chosen to conserve mass. For example in the RIMPUFF model we set

$E[\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})] = \boldsymbol{\alpha} Q(\mathbf{l})$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_5)^T, \sum_{i=1}^5 \alpha_i = 1, \alpha_i > 0$,

and

$Var[\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})] = W^*$, where $\mathbf{1}^T W^* \mathbf{1} = 0$ and $\mathbf{1}$ denotes a vector of ones.

Obviously if $\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})$ is chosen to be conditionally Gaussian, then these equations uniquely define its distribution.

3.6.1.3 The Emission Process

The emission process is modelled as a Dynamic Linear Model (DLM) with state space $(Q(t), \psi_t)^T$ where ψ_t is a vector of dummy variables. Explicitly

$$\begin{pmatrix} Q(t) - \mu(t) \\ \psi_t \end{pmatrix} \mid \begin{pmatrix} Q(t-1) - \mu(t-1) \\ \psi_{t-1} \end{pmatrix} \sim N \left[G \begin{pmatrix} Q(t-1) - \mu(t-1) \\ \psi_{t-1} \end{pmatrix}, W \right]$$

where G and W are fixed square matrices and $\mu(t)$ is a trend term which is a function of time t. This is just a standard state space model on the univariate process $\{Q(t), t = 1, 2, \ldots, \}$. Special cases of these models (e.g. set $\psi(t)$ as null when the process becomes 1-dimensional, $Q(t)|Q(t-1) \sim N[Q(t-1) + \mu(t) - \mu(t-1), W]$) are discussed in Chapter (6).

Chapter 4

An Introduction to Graphical Models

4.1 Introduction

Graphical models as statistical models embodying a collection of marginal and conditional independencies which may be summarised by means of a graph are quickly becoming an integral part of modern statistics. The graphical representation of a statistical model can help in many ways: the graph provides an effective means for the elicitation and simplification of a problem; it depicts the dependency structure posited in the model; and it may be transformed into a structure that can be used for efficient calculations of various quantities of interest, as we shall see in Chapter (7).

Graphical methods were used in the early 1980's for the analysis of statistical problems where no decision variables or utilities are explicitly represented. In a series of papers by Darroch et al. (1980), Wermuth and Lauritzen (1983), Lauritzen et al.

(1984), Kliveri et al. (1984) and Lauritzen and Wermuth (1987), the authors addressed the problem of how graphs such as influence diagrams can help in understanding the conditional independence properties that a given factorisation of a probability density implies.

Another issue of great importance is how graphs can be used to perform efficient probability calculations in high dimensional problems (computational efficiency). This issue is discussed in a number of papers by Kim and Pearl (1983), Pearl (1986), Lauritzen and Spiegelhalter (1988) and Spiegelhalter et al. (1993).

This chapter deals with some graph-theoretical results and offer a background amount of graphs such as influence diagrams and junction trees, which is necessary for the further development of the thesis.

4.2 Notation and Terminology

This section introduces some graph-theoretical terms which will be used frequently in the thesis. Here I use the terminology given by Whittaker (1990).

A network or *graph* is a pair $G = (V, E)$ that consists of a finite set of vertices $V = 1, 2, \ldots, v$ and a set of edges (arcs) $E \subseteq V \times V$ of ordered pairs of distinct vertices. An edge from vertex i (parent) to vertex j (child) is a *directed* edge (arrow) denoted by $i \rightarrow j$ if $(i, j) \in E$ and $(j, i) \notin E$. If both (i, j) and (j, i) are $\in E$, then the edge between i and j is *undirected* (line). If the graph has only undirected edges, it is an *undirected* graph, and if all edges are directed, the graph is said to be a *directed* graph.

A *path* of length $m \geq 0$ from i to j is an ordered sequence $(i = i_1, i_2, \ldots, i_m = j)$ of distinct vertices $i_1, i_2, \ldots, i_m$ such that (i_l, i_{l+1}) is in E for each $l = 1, 2, \ldots, m$. If there is a path from i to j, we say that i leads to j and write $i \longmapsto j$.

The *descendants* $de(i)$ of i are the vertices j such that $i \longmapsto j$. The vertices i that lead to j are the *ancestors* of j denoted by $an(j)$.

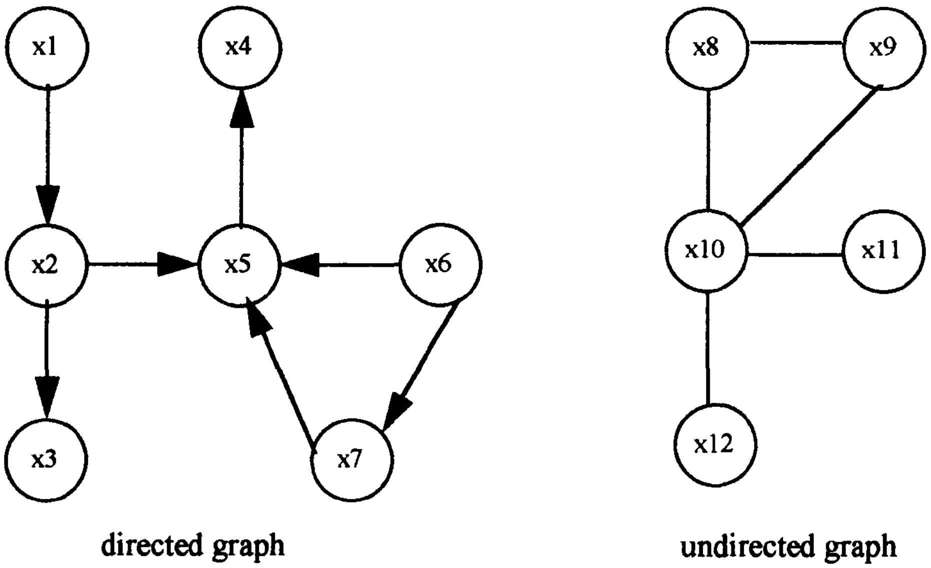
A subset $C \subseteq V$ is said to be a (i, j) *separator* if all paths from i to j intersect C. The subset C is said to separate A from B if it is an (i, j) separator for every $i \in A, j \in B$.

For $A \subseteq V$, the set of parents of A denoted by $P_a(A)$ is the set of all these vertices in V, but not in A, that have a child in A.

An *m-cycle* is a path of length m with the exception that the end points are equal; that is $i = j$. A graph is *acyclic* if it has no cycles.

A *directed acyclic graph* (DAG) is a directed graph without cycles.

Figure 4.1 illustrates some graph-theoretical terms.



directed graph

undirected graph

Figure 4.1: An illustration of some graph-theoretical terms

Note that $x_5 \rightarrow x_4, x_3 \not\rightarrow x_5$, the set of parents of x_5 is $\{x_2, x_6, x_7\}$ and x_2 has children

$\{x_3, x_5\}$.

4.3 Influence Diagrams

An influence diagram (ID) is a schematic representation of conditional independence relationships which is used for deducing new independencies from those employed in the construction of the diagram. Influence diagrams were first developed in the mid 1970's by Miller et al. (1976). Howard & Matheson (1981) extended the theory to decision analysis. Olmsted (1983) and Shachter (1986, 1988) gave a procedure for evaluating a decision problem using an influence diagram. In this section we present a brief introduction on how to use influence diagrams as a modelling framework that underpins a probability distribution in order to learn about and efficiently calculate various quantities of interest. We begin by defining a chance influence diagram.

4.3.1 Chance Influence Diagrams

In graph-theoretical terms a chance influence diagram or influence diagram (ID) is a directed graph $G = (V, E)$ where V is a set of nodes represented by circles and called chance nodes, and E is the set of directed edges or arrows joining these nodes. Chance nodes label random variables/uncertain quantities relevant to the problem being modelled, and directed edges represent probabilistic dependencies. A chance node which labels a random variable X_1 must be a *parent* of a chance node which labels a random variable X_2 if and only if the distribution of the random variable X_2 is calculated conditional on the value of the random variable X_1, and on the assumption that X_1 and X_2 are not independent. The generalisation to higher dimensions is given below.

Let $\boldsymbol{X} = (X_1, \ldots, X_m)$ be an ordered set of m random variables with a joint probability function

$$p(\boldsymbol{x}) = p(x_1) \prod_{r=2}^{m} p(x_r | x_1, \ldots, x_{r-1}) \tag{4.1}$$

Suppose $p(x_r|x_1, \ldots, x_{r-1})$ is a function of x_r and the parent set $P(r) \subseteq \{x_1, \ldots, x_{r-1}\}$ only. This will imply that given $P(r)$, X_r is independent of $R(r)$ where

$$R(r) = \{X_1, \ldots, X_{r-1}\} \setminus P(r)$$

is the set of random variables listed before X_r which do not appear explicitly in the conditional probability function $p(x_r|x_1, \ldots x_{r-1})$. Using Dawid's (1979) notation this can be written as

$$X_r \perp\!\!\!\perp R(r)|P(r) \qquad r = 2, \ldots, m \tag{4.2}$$

Then the graph of an influence diagram over $X_1, \ldots, X_m$ is any directed graph with nodes representing random variables $X_1, \ldots, X_m$ satisfying the property (4.2).

Influence diagrams are clearly acyclic because only nodes of lower index can be connected to nodes of higher index. The graph of an influence diagram together with the c.i. statements in (4.2) is called an influence diagram.

As a simple illustration, suppose $\boldsymbol{X} = \{X_1, \ldots, X_8\}$. Then from (4.1)

$$p(\boldsymbol{x}) = p(x_1) \prod_{r=2}^{8} p(x_r | x_1, \ldots, x_{r-1}).$$

Suppose the parents are: $P(2) = \{X_1\}, P(3) = \{X_1, X_2\}, P(4) = \{X_3\}, P(5) = \{X_3, X_4\}, P(6) = \{\phi\}$ (the empty set), $P(7) = \{X_5, X_6\}, P(8) = \{X_7\}$.

The influence diagram I of this example is given in Figure 4.2

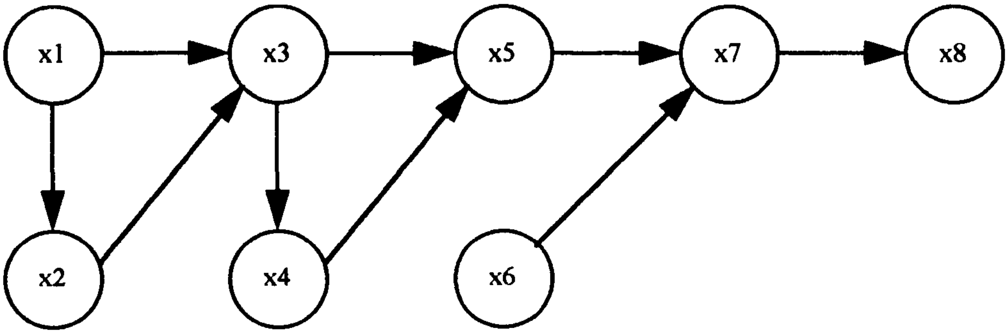


Figure 4.2: An ID I

4.3.2 The Clique Marginal Representation

The clique marginal representation is one of many ways of specifying a joint probability distribution(see, for example, Lauritzen & Spiegelhalter, 1988 and Smith, 1988). We start by identifying the cliques of an influence diagram G and $p(\boldsymbol{x})$ by looking at the small sets of variables called ***precliques*** (see Smith, 1995a) of the form

$$\tilde{C}(r) = \{X_r, P(r)\} \quad (P(1) = \phi), \quad 1 \leq r \leq m. \tag{4.3}$$

Then we delete from this collection any preclique $\tilde{C}(r)$ for which there exists a $\tilde{C}(k) \quad (k > r)$ such that

$$\tilde{C}(r) \subseteq \tilde{C}(k).$$

The remaining sets of variables after such deletions are called the *cliques* of $p(\boldsymbol{x})$ and G. This set of cliques will be denoted by $\boldsymbol{C} = \{C(1), \ldots, C(n)\}, \quad 1 \leq n \leq m-1.$

After identifying the cliques, we can determine $p(\boldsymbol{x})$ in terms of the joint probability functions $p_1(\boldsymbol{x}), \ldots, p_n(\boldsymbol{x})$ over the cliques $\{C(1), \ldots, C(n)\}$. A sufficient condition for this is that $p(\boldsymbol{x} \in P(r)) > 0$ for each $\boldsymbol{x} \in P(r), \quad 2 \leq r \leq m$ whenever $P(r) \neq \phi$. Then

(4.1) can be expressed as:

$$p(\boldsymbol{x}) = \frac{\prod_{r=1}^{m} p(\boldsymbol{x} : \boldsymbol{x} \in \tilde{C}(r))}{\prod_{r=2}^{m} p(\boldsymbol{x} : \boldsymbol{x} \in P(r))} \tag{4.4}$$

where $p(\boldsymbol{x} \in P(r)) = 1$ if $P(r) = \phi$, the empty set.

Since by definition $p(\boldsymbol{x} : \boldsymbol{x} \in \tilde{C}(r))$ (and hence also $p(\boldsymbol{x} : \boldsymbol{x} \in P(r))$ can be obtained from a $p(\boldsymbol{x} : \boldsymbol{x} \in C(k))$ where $C(k)$ is a clique of $p(\boldsymbol{x})$ such that $\tilde{C}(r) \subseteq C(k), \quad 2 \leq r \leq m$, then (4.4) can be simplified to

$$p(\boldsymbol{x}) = \frac{\prod_{k=1}^{n} p_k(\boldsymbol{x})}{\prod_{k=2}^{n} q_k(\boldsymbol{x})} \tag{4.5}$$

where $p_k(\boldsymbol{x})$ is as defined above and $q_k(\boldsymbol{x}) = p(\boldsymbol{x} : \boldsymbol{x} \in P(r))$ for a $\tilde{C}(r)$ remaining in the clique set, such that $\tilde{C}(r) = C(k), 1 \leq k \leq n$. A set of parents $P(r)$ associated with a clique $C(k)$ is called a ***preseparator*** and is denoted by $\tilde{S}(k), 2 \leq k \leq n$. The clique representation (4.5) of $p(\boldsymbol{x})$ has many computational advantages as we shall see below.

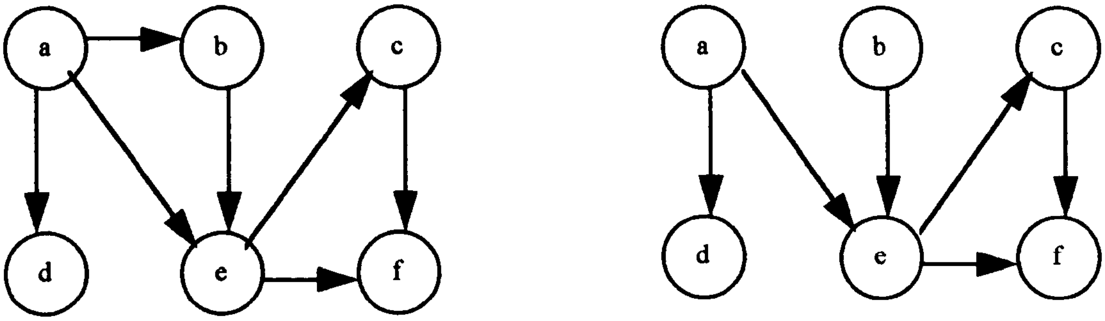
As a simple illustration of how we express $p(\boldsymbol{x})$ as in (4.5), consider the influence diagram of Figure 4.2 where we identify the cliques and the preseparators as

Precliques	**Cliques**	**Preseparators**
$\tilde{C}(1) = (X_1)$		
$\tilde{C}(2) = (X_1, X_2)$		
$\tilde{C}(3) = (X_1, X_2, X_3)$	$C(1) = (X_1, X_2, X_3)$	
$\tilde{C}(4) = (X_3, X_4)$		
$\tilde{C}(5) = (X_3, X_4, X_5)$	$C(2) = (X_3, X_4, X_5)$	$\tilde{S}(2) = (X_3)$
$\tilde{C}(6) = (X_6)$		
$\tilde{C}(7) = (X_5, X_6, X_7)$	$C(3) = (X_5, X_6, X_7)$	$\tilde{S}(3) = (X_5)$
$\tilde{C}(8) = (X_7, X_8)$	$C(4) = (X_7, X_8)$	$\tilde{S}(4) = (X_7)$

and $p(\boldsymbol{x}) = \frac{p(x_1,x_2,x_3)p(x_3,x_4,x_5)p(x_5,x_6,x_7)p(x_7,x_8)}{p(x_3)p(x_5)p(x_7)}$. The separators of $p(\boldsymbol{x})$ can be obtained from its list of preseparators by deleting any duplicated sets.

4.3.3 Decomposable Influence Diagrams

An ID G is called *decomposable* if the set $P(X)$ of direct predecessors of X is completely connected (i.e. each node in $P(X)$ is connected by an edge to another node), this being true for all X in G. Figure 4.3 illustrates two graphs: graph G is decomposable and graph H is not, since the parent nodes a and b are not joined. Decomposable influence diagrams have several properties which make them useful to study. One property is that their structure helps in propagating probabilities as the joint distribution of the system can be stored as margins of cliques (see Section 5 below).



decomposable ID G

non-decomposable ID H

Figure 4.3: Graphs of decomposable and non-decomposable IDs

The cliques of a decomposable influence diagram can be ordered (see Tarjan & Yannakaskis, 1984 for a simple technique for ordering nodes called the maximum cardinality search MCS) so that the cliques satisfy the so called *running intersection property* (RIP) (Beeri et al., 1981, 1983; Lauritzen et al, 1984 and Tarjan & Yannakakis, 1984) which states that: *there exists an ordering* $C[1],\ldots,C[n]$ *of the cliques* $C(1),\ldots,C(n)$ *such that for all* $2 \le i \le n$

$$C[i] \cap [\cup_{j=1}^{i-1} C[j]] = S(i) \subseteq C(b_i)$$

for some $b_i, 1 \le b_i \le i-1$.

This means that the intersection of the i th clique with all the preceding ones is a subset of one of the preceding cliques. For example, the cliques of the undirected graph in Figure 4.4 $C(1) = \{X_1, X_2, X_3, X_4\}, C(2) = \{X_3, X_4, X_5, X_6\}, C(3) = \{X_6, X_7\}, C(4) = \{X_3, X_6, X_8\}$ are satisfying the (RIP), since

$$\begin{aligned} S(2) &= C(2) \cap C(1) &&= \{X_3, X_4\} \subseteq C(1) \\ S(3) &= C(3) \cap (C(1) \cup C(2)) &&= \{X_6\} \subseteq C(2) \\ S(4) &= C(4) \cap (C(1) \cup C(2) \cup C(3)) &&= \{X_3, X_6\} \subseteq C(2) \end{aligned}$$

where $b_2 = 1, \quad b_3 = 2, \quad b_4 = 2$.

4.4 Junction Tree, Junction Forest and Probability Propagation

The clique representation (4.5) of the $p(\boldsymbol{x})$ can be used efficiently to propagate information through the system, working indirectly with the margins $p_k(\boldsymbol{x})$ and $q_k(\boldsymbol{x})$, successively updating them rather than updating the whole joint probability function $p(\boldsymbol{x})$ directly. This can be done by passing "simple messages" along the edges of a new graph called a *junction tree* constructed from the influence diagram of $p(\boldsymbol{x})$.

However, in the applications cited above, distributions will not always remain decomposable. Because of this we need to define a new graph called *junction graph*

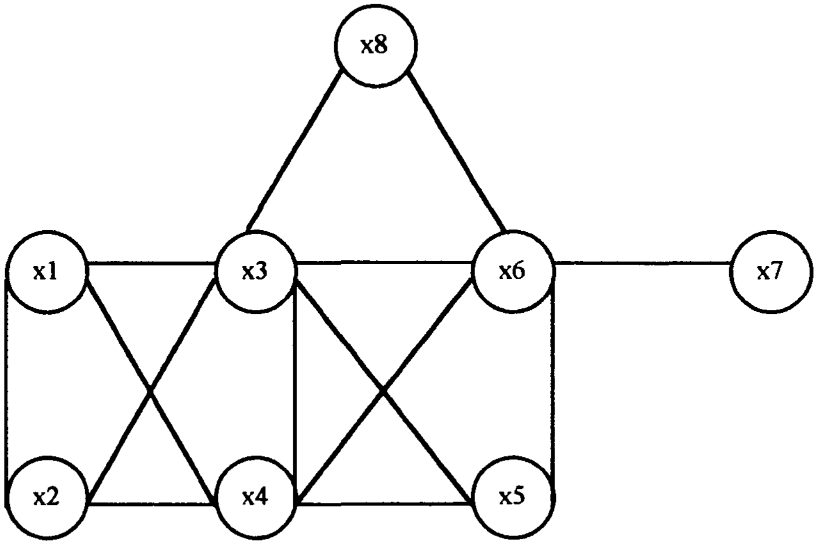


Figure 4.4: An undirected graph with cliques satisfying the RIP

which is an influence diagram on vectors of variables in the original influence diagram of the process. We then show that the definition of a junction tree is just a special case of the (undirected version of) a junction graph. The use of junction graphs will become apparent later in this thesis. A formal definition of a junction graph follows.

Definition

A junction graph $\mathcal{G}$ of any density satisfying (4.5) is a directed graph with n nodes labelling the n cliques $C(1), \ldots, C(n)$. There is an edge to node $C(i)$ from node $C(j), i > j$ if and only if

i) $S(i) \cap C(j) \neq \phi$

ii) there exists no $j' < j$ such that

$$S(i) \cap C(j') \supseteq S(i) \cap C(j).$$

A minimal junction graph $\mathcal{G}$ is a junction graph which has no other junction graph $\mathcal{G}'$ as a proper subgraph.

In general a joint probability function will have several junction graphs and minimal junction graphs over a chosen ordering of its cliques. An influence diagram J and its junction graph are shown in Figure 4.5. The undirected versions of junction graphs are called *junction trees* when the separator of any clique is contained in exactly one previously listed clique or separator. The undirected version of the junction graph of Figure 4.5 is in fact a junction tree.

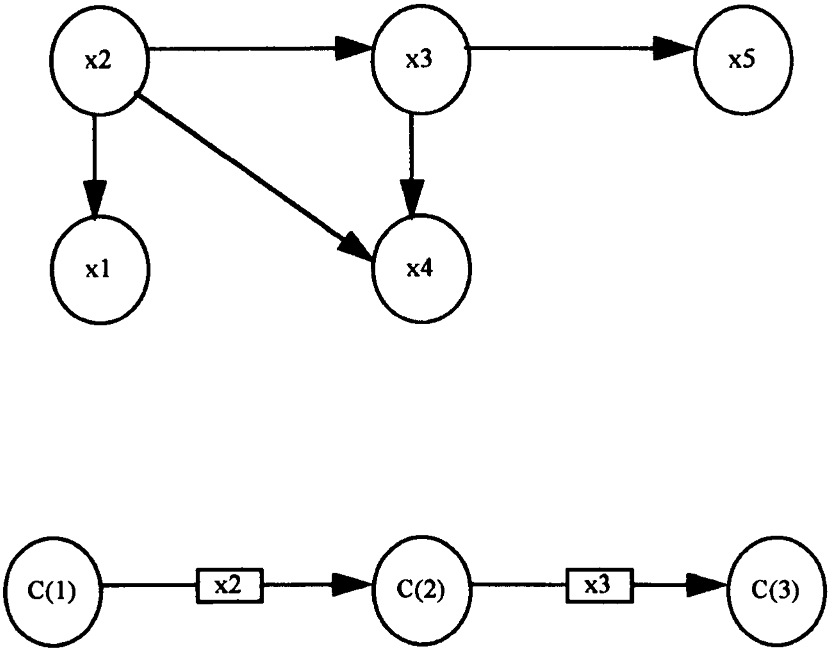


Figure 4.5: An ID J and its junction graph

In the case where $p(\boldsymbol{x})$ is decomposable, a collection of disconnected junction trees will be called *a junction forest.*

4.4.1 The Propagation of Information on Junction Trees

Let $C = \{C(1), \ldots, C(n)\}$ denote the set of cliques of the joint probability function $p(\boldsymbol{x})$. Suppose we learn the values of some or all of the variables lying in some arbitrary clique $C(1) \in C$, and we want to compute the conditional distribution of all variables in the system given a subset of variables in $C(1)$. Smith (1995a) described a propagation algorithm paralleling that given in Lauritzen & Spiegelhalter (1988).

Now, it is clear that we can obtain the new probability function $p^*(\boldsymbol{x}(1))$ of the variables $\boldsymbol{x}(1)$ in $C(1)$ from $p(\boldsymbol{x}(1))$ its original probability function using Bayes' rule.

Smith (1995a) shows how to update probabilities over the variables in the other cliques given the values of some of the variables in $C(1)$. The updating is possible using the junction tree of the system. For detailed discussions, see the above references.

4.5 Graphical Representation of the Fragmenting Puff Models

In Chapter (3) we described fragmenting puff models and some distributional assumptions concerning models for the instantaneous, emission readings and for the fragmenting process were discussed.

In this section we show a graphical representation of the conditional probability breakdown of puffs and puff fragments.

4.5.1 Clique Representation of Puff Distributions

Following Smith et al.'s (1995) notation, let $\boldsymbol{X}_T$ denote a vector of state random variables of interest (vector of mass emissions and their fragments in our context) existing on or before time T. In the model defined in Chapter (3) it is easy to check that because of the conditional independencies in the system, the joint density $p_T(\boldsymbol{x})$ of $\boldsymbol{X}_T$ can be written as

$$p_T(\boldsymbol{x}) = p(Q(1),\psi(1))\prod_{t=2}^{T} p(Q(t),\psi(t)|Q(t-1),\psi(t-1))\prod_{I_T} p(\bar{\boldsymbol{Q}}(\mathbf{l})|Q(\mathbf{l})) \qquad (4.6)$$

where $Q(t),\psi(t),\bar{\boldsymbol{Q}}(\mathbf{l}),Q(\mathbf{l})$ and I_T are as defined in Chapter (3). The density can be expressed in a suitable form, namely the clique marginal representation form of equation (4.5), for an efficient propagation of probabilities (see Smith et al., 1995).

Let

$$C^*(t) = \{Q(t),\psi(t),Q(t+1),\psi(t+1)\}, \qquad 1 \leq t \leq T-1 \qquad (4.7)$$

$$C(\mathbf{l}) = \{\boldsymbol{Q}(\mathbf{l})\} = \{Q(\mathbf{l}),Q(\mathbf{l},l_1),\ldots,Q(\mathbf{l},l_5)\}, \quad \mathbf{l} \in I_T \qquad (4.8)$$

where $C^*(t), C(\mathbf{l})$ are cliques.

Applying equation (4.5), $p_T(\boldsymbol{x})$ can be written as

$$p_T(\boldsymbol{x}) = \frac{\prod_{1\leq t\leq T-1} p(C^*(t))\prod_{\mathbf{l}\in I_T} p(C(\mathbf{l}))}{\prod_{2\leq t\leq T-1} p(S(t))\prod_{\mathbf{l}\in I_T}[p(Q(\mathbf{l}))]^{r_T(\mathbf{l})}} \qquad (4.9)$$

where $p(C^*(t))$ and $p(C(\mathbf{l}))$ denote respectively the joint densities of the variables in the cliques $C^*(t)$ and $C(\mathbf{l}), S(t) = \{Q(t),\psi(t)\}$ and $r_T(\mathbf{l})$ is the number of offsprings of $Q(\mathbf{l})$ produced before or at time T. Using this simplified representation, the joint density $p_T(\boldsymbol{x})$ can be stored as a moderate number of joint densities of low dimension instead of a single density of a high dimension.

4.5.2 A Junction Tree of the Fragmentation Process

The structure of the joint density $p_T(\boldsymbol{x})$ can be represented by a dynamic influence diagram (see, for example, Queen, 1991; 1993 and Smith et al., 1995). The nodes of the ID are the random variables (or vectors) defined on the cliques. For example the ID given in Figure 4.6 represents the conditional probability breakdown of puff and puff fragments in the early stages of an accidental release.

A source has emitted 4 puffs at time T. The first puff has pentificated, the 2nd and 5th fragments have then pentificated, and further fragmentation has occurred on the 2nd offspring of the 2nd fragment. The second puff has also pentificated and its 2nd puff also splits into 5. The 3rd and 4th puffs have yet to fragment.

Here we note that it is easy to check that the ID of Figure 4.6 is *decomposable* (all parents of a given child are connected) with its cliques having the running intersection property (RIP) of Section (6), that is at any time T, the cliques can be ordered as $C[1], \ldots, C[9]$ such that

$$C[i] \cap [\cup_{j=1}^{i-1} C[j]] = S[i] \subseteq C[b_i] \quad 2 \leq i \leq 9.$$

for some $b_i, 1 \leq b_i \leq i-1$. Also we note the following:

(i) If $C[i] = C^*(t)$ then $C[b_i] = C^*(t-1)$ and $S[i] = Q(t)$.

(ii) If $C[i] = C(\mathbf{l})$, if $\mathbf{l} = t$ then $C[b_i] = C^*(t)$ and $S[i] = \{Q(t)\}$, if $\mathbf{l} = (t, l_1, \ldots, l_k)$ then $C[b_i] = C(t, l_1, \ldots, l_{k-1})$ and $S[i] = \{Q(\mathbf{l})\}$.

Since the ID is decomposable, we can form a junction tree whose nodes are the cliques of $p_T(\boldsymbol{x})$. Figure 4.6 shows an ID of the example of the early emission and its junction tree.

A typical clique $\bar{C}[i]$ of this junction tree will have a probability defined conditionally in terms of a particular separator $\bar{S}[i]$ of the junction tree. That separator will take one of the following forms:

(a) If $\bar{C}[i] = C^*(t)$ it will take the form $S(t)$ of equation (4.7).

(b) If $\bar{C}[i] = C(\mathbf{l})$ it will take the form $Q(\mathbf{l})$.

4.5.3 Relating Observations to Cliques

As we indicated in Sections (3.4) and (3.5), measurements will be taken as a single observation $Y(t, \mathbf{0})$ at the source mass (reading on chimney stack) at time t with error or as a vector of linear combination of mass fragments with error at different sites.

In practice, when true contamination $\theta(t, \boldsymbol{s})$ is regressed on $\boldsymbol{Q}_t$, only a small number of the stochastic multipliers $F_t(i, \boldsymbol{s})$'s are non-zeros. These regression coefficients are usually on fragments (children) of the same parent (i.e. the observation lies under a single clique). This happens when the wind field is not turbulent, and the terrain is not too heterogeneous. In this case the observation can be expressed as a linear function of the combination of masses $Q(i)$s exist at time t in the same clique C.

$$Y(t, \boldsymbol{s}) = \sum_{Q(i) \in C} F_t(i, \boldsymbol{s}) Q(i) + \epsilon(t, \boldsymbol{s})$$

where $\epsilon(t, \boldsymbol{s})$ is an independent error term. However, in some cases, the rate of pentification is too great and the windfield is turbulent, so that the non-zero regression coefficients are on children of several cliques (i.e. the observation will be "divided" between several cliques). The number of these cliques is determined by the dispersal (physical) model and the observation under these cliques is related to the masses under fragments such that

$$Y(t,s) = \sum_{j \in J} \left[\sum_{Q(i) \in C_j} F_j(i,s) Q_j(i) \right] + \epsilon(t,s)$$

where $1 \le i \le t$ and J is the index set of cliques under which the observation is taken. In Chapter (7) we consider the propagation of information as in the types above. An exact algorithm for quick absorption of information on such junction trees which evolve dynamically and some approximation schemes for efficient propagation, will be discussed. (See further Smith et al., 1995 and Gargoum and Smith, 1994a.)

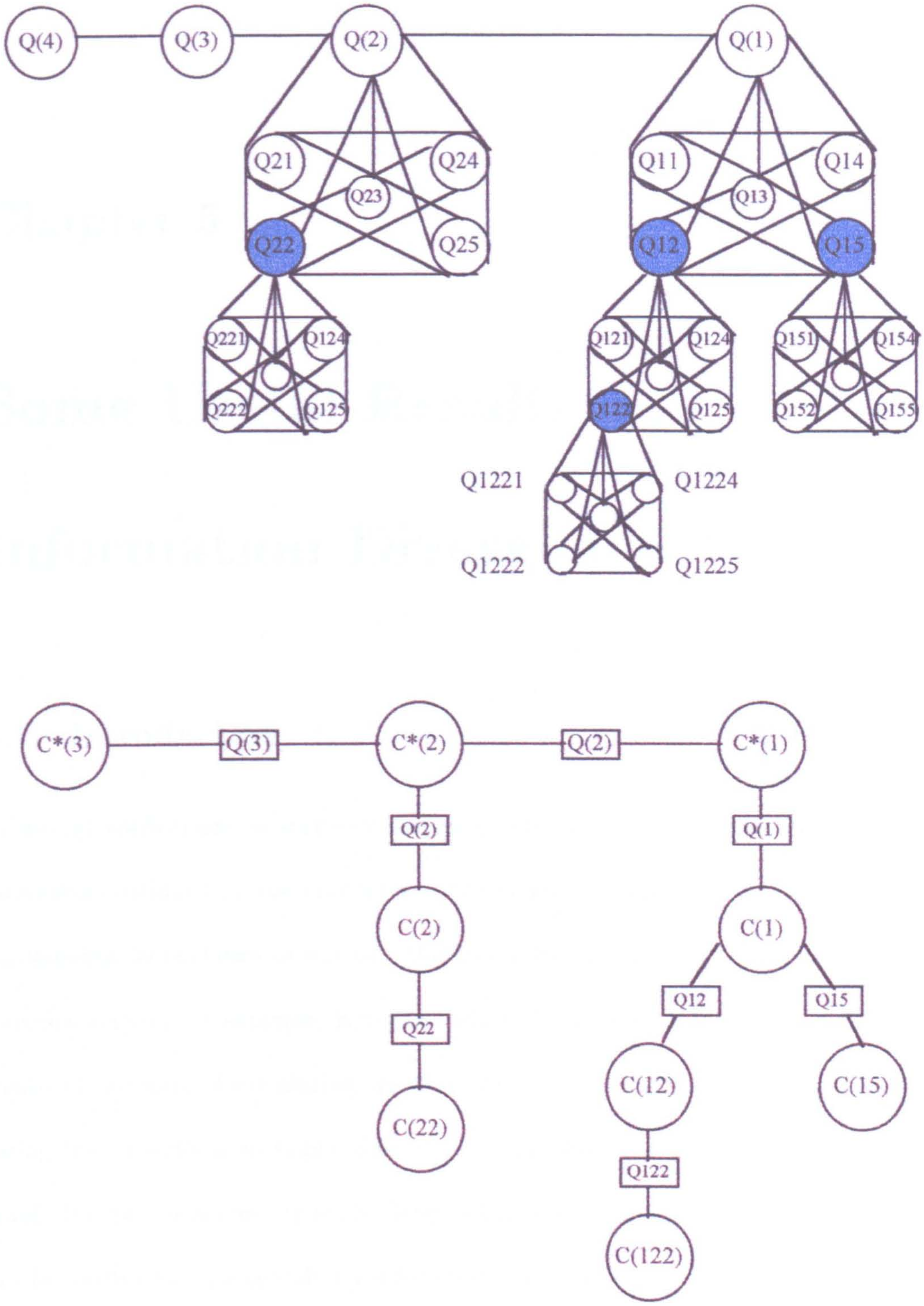


Figure 4.6: An ID of early emissions and its junction tree

Chapter 5

Some Useful Results on Information Divergence

5.1 Introduction

In several applications in statistics and graphical models we need to measure the information contained in one random variable about the value of another. It is also of fundamental importance to examine the proximity of one density function to its approximation(see, for example, Kjaerulf 1992 and Gargoum et al., 1995). There is a choice of methods of calculating the "distance" between two densities and for measuring the strength of an edge connecting two variables in a conditional independence graph, but two measures are particularly important in parametric set up because they can be written in an algebraic form for most common families of distributions. These are the Kullback-Leibler (K-L) measure of separation and the Hellinger distance. In this chapter we discuss several interesting properties and results, both old and new, associated with these measures and which are needed in later chapters of this thesis. (see, for example, Devroye, 1987).

5.2 Distances between Densities

Let $(\mathcal{X}, \mathcal{A})$ be a measurable space, with P and $\hat{P}$ as two probability measures on $(\mathcal{X}, \mathcal{A})$ (P may represent a Bayesian probability model and $\hat{P}$ is an approximation for P on a σ-field $\mathcal{A}$); and let $p, \hat{p}$ be the densities of P and $\hat{P}$ with respect to μ (a positive σ - finite measure dominating P and $\hat{P}$). Several separation measures can be used to measure the distance between P and $\hat{P}$. For example we may define the **variation distance** between the two probability measures as

$$d_V(P; \hat{P}) = \sup_{A \in \mathcal{A}} |P(A) - \hat{P}(A)|$$

where A denotes an arbitrary measurable set for which the probability under P or $\hat{P}$ is defined. The relation between the variation and the L_1-distances (see, for example, Reiss, 1989) is given by

$$d_V(p, \hat{p}) = 1/2 \int |p - \hat{p}| \tag{5.1}$$

It should be noted that $0 \leq d_V \leq 1$.

The interpretation of the L_1 criterion in terms of the difference between probabilities makes it unique. It is easily interpreted: when we report, for example, that L_1 error is 0.010, then we know that all probabilities of all sets are off by at most 0.005. From the relation (5.1), when d_V is 0.005, then we know that for any set A, the probability assigned to it by p differs at most by 0.005 from the probability assigned to it by $\hat{p}$.

In this chapter we shall also introduce further distances and show their relation to the variation distance. For example the distance between p and $\hat{p}$ can be measured in terms of the entropy related quantity, the **Kullback-Leibler**(K-L) divergence which is defined as

$$d_K(p;\hat{p}) = \int p \log \left[\frac{p}{\hat{p}}\right] d\mu,$$

or in terms of the **Hellinger distance** which is defined by

$$d_H(p;\hat{p}) = \left[\frac{1}{2}\int (p^{1/2} - \hat{p}^{1/2})^2 d\mu\right]^{1/2}.$$

In this chapter, we will explore some of the most important properties of the Kullback-Leibler divergence and the Hellinger distance in more detail.

5.3 The Kullback-Leibler Divergence

Definition

The Kullback-Leibler divergence between two densities p and $\hat{p}$ for the random variable (or vector) X is

$$d_K(p;\hat{p}) = E[\log \frac{p}{\hat{p}}] \tag{5.2}$$

where the expectation is taken with respect to the density p. The Kullback-Leibler is, perhaps, the most well-known separation measure; however it is not symmetric, i.e. in general $d_K(p;\hat{p}) \neq d_K(\hat{p};p)$. Sometimes it is appropriate to use the notation $d_K(p;\hat{p}|X)$ or $d_K(p_X, \hat{p}_X)$ to indicate the random variable (vector) for which the divergence is taken.

5.4 The Kullback-Leibler between Gaussian Distributions

(i) *The univariate case.* Suppose that the random variable X is normally distributed under both f_1 and f_2 where under $f_1 : X \sim N[\mu_1, V_1]$ and under $f_2 : X \sim N[\mu_2, V_2]$. Then

$$\begin{aligned}\log f_1 - \log f_2 &= -\tfrac{1}{2}[\log V_1 + V_1^{-1}(x-\mu_1)^2 - \log V_2 + V_2^{-1}(x-\mu_2)^2] \\ &= -\tfrac{1}{2}\log\tfrac{V_1}{V_2} - \tfrac{1}{2}R(x)\end{aligned}$$

where $R(x) = V_1^{-1}(x-\mu_1)^2 - V_2^{-1}(x^2 - 2\mu_2 x + \mu_2^2)$. So

$$E[R(x)] = 1 - V_2^{-1}[V_1 + (\mu_1 - \mu_2)^2],$$

and

$$2d_K(f_1; f_2) = V_2^{-1}(\mu_1-\mu_2)^2 + \frac{V_1}{V_2} - 1 - \log\frac{V_1}{V_2}$$

Letting $V_1 = V_2 + \Delta$

$$2d_K(f_1; f_2) = V_2^{-1}(\mu_1-\mu_2)^2 + [\frac{\Delta}{V_2} - \log(1+\frac{\Delta}{V_2})] \tag{5.3}$$

Since the second term in (5.3) is always positive

$$\log(1+\frac{\Delta}{V_2})\begin{cases} < \frac{\Delta}{V_2} & \text{if } \Delta > 0 \\ > \frac{\Delta}{V_2} & \text{if } \Delta < 0\end{cases}$$

Notice in particular that if $(\frac{\Delta}{V_2}) \to 0$ then

$$d_K(f_1; f_2) \to \frac{1}{2}V_2^{-1}(\mu_1-\mu_2)^2 \tag{5.4}$$

(ii) *The multivariate case.* More generally, if the k- variate vector $\boldsymbol{X}$ is normally distributed under f_1 and f_2 where under $f_1 : \boldsymbol{X} \sim N[\boldsymbol{\mu}_1, \Sigma_1]$ and under $f_2 : \boldsymbol{X} \sim N[\boldsymbol{\mu}_2, \Sigma_2]$, then,

$$d_K(f_1; f_2) = \frac{1}{2}\log\frac{|\Sigma_2|}{|\Sigma_1|} - \frac{k}{2} + \frac{1}{2}tr(\Sigma_1\Sigma_2^{-1}) + \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{5.5}$$

where $\Sigma = \frac{1}{2}(\Sigma_1 + \Sigma_2)$ (see, for example, Kullback, 1968).

5.4.1 Properties of the Kullback-Leibler Divergence

In this section we discuss briefly some of the most important properties of the Kullback-Leibler divergence. A comprehensive account of the properties and proofs are given in Kullback (1968); Whittaker (1990) and Kjaerulf (1992).

1. $d_K(f_1; f_2) \geq 0$, with equality if and only if $f_1 = f_2$.

   This property follows directly from Jensen's inequality,

$$-d_K(f_1; f_2) = \int f_1 \log \frac{f_2}{f_1} \leq \log(\int f_1 \frac{f_2}{f_1}) = 0.$$

2. If X and Y are independent random variables under both f_1 and f_2, then

$$d_K(f_1; f_2|X, Y) = d_K(f_1; f_2|X) + d_K(f_1; f_2|Y).$$

   This property still exists even if X and Y are not independent but in terms of a conditional information divergence as follows. Suppose that (X_1, Y_1) has a joint density/ mass function f_1, and that (X_2, Y_2) is absolutely continuous with respect to f_1 and has a joint density/ mass function f_2. Denote by

$$d_K(f_1(x_1, y_1); f_2(x_2, y_2)) = E_{(X_1,Y_1)}[\log f_1(x_1, y_1) - \log f_2(x_2, y_2)].$$

   Since

$$\log f_j(x_j, y_j) = \log f_j(x_j) + \log f_j(y_j|x_j) \quad j = 1, 2.$$

   Then

$$d_K(f_1(x_1, y_1); f_2(x_2, y_2)) = d_K(f_1(x_1); f_2(x_2)) + E_{X_1}[d_K(f_1(y_1|x_1); f_2(y_2|x_2))] \tag{5.6}$$

where

$$\begin{aligned} d_K(f_1(x_1); f_2(x_2)) &= E_{(X_1,Y_1)}[\log f_1(x_1) - \log f_2(x_2)] \\ &= E_{X_1}[\log f_1(x_1) - \log f_2(x_2)] \end{aligned}$$

and

$$E_{X_1}[d_K(f_1(y_1|x_1); f_2(y_2|x_2)] = E_{X_1}[E_{(Y_1|X_1)}(\log f_1(y_1|x_1) - \log f_2(y_2|x_2)].$$

In particular, if (X_j, Y_j) $j = 1, 2$ are independent, equation (5.6) simplifies to

$$d_K(f_1(x_1, y_1); f_2(x_2, y_2)) = d_K(f_1(x_1); f_2(x_2)) + d_K(f_1(y_1); f_2(y_2)) \qquad (5.7)$$

which can be generalised easily.

Note from (5.6) that, under the Kullback-Leibler divergence, the marginal distributions of two random variables are always closer together than their joint distributions with other variables. Thus, because the the second term in (5.6) is always positive, then

$$d_K(f_1(x_1); f_2(x_2)) \leq d_K(f_1(x_1, y_1); f_2(x_2, y_2)) \qquad (5.8)$$

3. The Kullback-Leibler divergence can be calculated locally.

Let p_V be a probability function for a junction tree $\mathcal{T} = (\mathcal{C}, \mathcal{S})$ where $\mathcal{C}$ is a set of cliques and $\mathcal{S}$ is a set of separators with a corresponding graph $G = (V, E)$. Let $C_0 \in \mathcal{C}$ be the clique containing the variables $\{X, Y\}$ and let $\hat{p}_V$ be the probability function for $\mathcal{T}$ after cutting the edge joining X and Y. Then

$$d_K(p_V; \hat{p}_V) = d_K(p_{C_0}; \hat{p}_{C_0}).$$

For proof of this property see Kjaerulf (1992).

4. The Kullback-Leibler divergence is additive over a series of approximations.

   Suppose that a sequential cutting of edges is made on a junction tree $\mathcal{T}_0$ with a density function/mass function p_0. Let $\mathcal{T}_i$ and $p_i (i = 1, ..., m)$ be respectively the junction tree and the density function after cutting the ith edge. Then the Kullback-Leibler divergence between p_0 and p_m is equivalent to the sum of the divergences between p_{i-1} and p_i for all $i = 1, ...m$. For example, to achieve computational efficiency (see Chapter 7) we need to cut a separator from a clique C containing S in the junction tree $\mathcal{T} = (\mathcal{C}, \mathcal{S})$– with a probability density/mass function p– which will be transformed to $\mathcal{T}^* = (\mathcal{C}^*, \mathcal{S}^*)$ with a density/mass function $\hat{p}$. This will be done frequently during the evolution of the emission and fragmentation processes. According to the additivity property, the overall Kullback-Leibler divergence between p and $\hat{p}$ is equal to the sum of the individual Kullback-Leibler divergences.

5.5 The Hellinger Distance

Definition

The Hellinger distance between the two densities p and $\hat{p}$ is defined as

$$\begin{aligned} d_H(p, \hat{p}) &= [\tfrac{1}{2} \textstyle\int (p^{1/2} - \hat{p}^{1/2})^2 d\mu]^{1/2} \\ &= [1 - I(p, \hat{p})]^{1/2} \end{aligned}$$

where $I(p, \hat{p}) = \int p^{1/2} \hat{p}^{1/2} d\mu$ is the affinity between p and $\hat{p}$. Also known as the Bhahacharyya coefficient, it measures the closeness of the distributions. As easily seen, $I(p, \hat{p})$ has the following properties (see Matusita, 1976):

(i) $0 \leq I(p,\hat{p}) \leq 1$ with equality of 1 if and only if $p = \hat{p}$

(ii) $I(p,\hat{p})$ is symmetric on distributions.

5.6 The Hellinger Distance between Gaussian Distributions

(i) *The univariate case.* The Hellinger distance between two univariate normal densities f_1 and f_2 with respective means and variances $(\mu_1, V_1); (\mu_2, V_2)$ can be calculated from

$$I^2(f_1; f_2) = \frac{2V_1^{1/2}V_2^{1/2}}{V_1 + V_2} \exp\{-\frac{1}{2}(V_1 + V_2)^{-1}(\mu_1 - \mu_2)^2\}$$

(ii) *The multivariate case.* Let f_1 and f_2 be two multivariate densities with respective means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariances matrices Σ_1 and Σ_2, respectively, then

$$I = \int f_1^{1/2} f_2^{1/2} = \frac{|\Sigma_1|^{1/4}|\Sigma_2|^{1/4}}{|\Sigma|^{1/2}} . \exp\{-\frac{1}{8}S\}$$

where

$$\begin{aligned} \Sigma &= \tfrac{1}{2}(\Sigma_1 + \Sigma_2) \\ S &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \end{aligned}$$

Note that S depends on the distributional dimension unless $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$. The effect of this and the invariance property of the Hellinger distance are discussed in Section (5.7).

Now

$$\begin{aligned} d_H^2(f_1; f_2) &= 1 - I \\ &= 1 - \exp\{-\tfrac{1}{4}(\tfrac{1}{2}S + 2\log|\Sigma| - \log|\Sigma_1| - \log|\Sigma_2|)\} \\ &= 1 - \exp\{-\tfrac{1}{4}R(f_1; f_2)\} \end{aligned}$$

where

$$R(f_1; f_2) = \frac{1}{2}S + T$$

and

$$T = 2\log|\Sigma| - \log|\Sigma_1| - \log|\Sigma_2|$$

Note that

$$R(f_1; f_2) = -4\log[1 - d_H^2(f_1; f_2)] \Longleftrightarrow d_H^2(f_1, f_2) = (1 - \exp\{-\frac{1}{4}R(f_1; f_2)\}$$

In Chapter (7) we will approximate a normal density function f of a random vector $\boldsymbol{X} = (\boldsymbol{X}(1), ..., \boldsymbol{X}(p))$ where $\boldsymbol{X}(j)$ has a mean vector $\boldsymbol{\mu}(j), 1 \leq j \leq p$ and covariance matrix $\Sigma(j,k) = Cov(\boldsymbol{X}(j), \boldsymbol{X}(k)), 1 \leq j, k \leq p$ by the normal density $\hat{f}$ of a random vector with the same $\boldsymbol{\mu}(j)$ and $\Sigma(j,j), 1 \leq j \leq p$ but with all covariances block $\Sigma(j,k), j \neq k$ set to 0.

The following result shows the connection between the variation distance and Hellinger distance (see, for example, Devroye, 1987 and Smith, 1995b).

For any two densities p and $\hat{p}$

$$\sqrt{2}d_H(p;\hat{p}) \geq d_V(p;\hat{p}) \geq d_H^2(p;\hat{p})$$

Proof.

Here we will omit the arguments $(p, \hat{p})$

$$\begin{aligned} 2d_V &= \int |p - \hat{p}| \\ &= \int |p^{1/2} - \hat{p}^{1/2}||p^{1/2} + \hat{p}^{1/2}| \\ &\geq \int (p^{1/2} - \hat{p}^{1/2})^2 \\ &= 2d_H^2 \end{aligned}$$

Also, by Cauchy-Schwartz inequality

$$\begin{aligned} 4d_V^2 = [\int |p-\hat{p}|]^2 &\le \int (p^{1/2}-\hat{p}^{1/2})^2 \int (p^{1/2}+\hat{p}^{1/2})^2 \\ &= 2d_H^2(2+2\int p^{1/2}\hat{p}^{1/2}) \\ &= 2d_H^2(4-2d_H^2) \\ &= 4d_H^2(2-d_H^2) \end{aligned}$$

which implies that $d_V^2 \le d_H^2(2-d_H^2)$. i.e $d_V \le d_H\sqrt{2-d_H^2} \le \sqrt{2}d_H$.

For any two densities, the Hellinger distance is topologically equivalent to the variation distance. Small values in d_V are equivalent to small values in d_H, so if an approximation is good with respect to d_H, it will also be good with respect to d_V (see, for example, Smith 1995b).

A bound for the Hellinger distance (and thus for the variation distance) can be constructed by using the Kullback-Leibler divergence. We have

$$d_H(p,\hat{p}) \le (d_K(p,\hat{p}))^{1/2} \tag{5.9}$$

For the proof of this result see Reiss (1989).

As an example, the three distances $d_K, d_H, d_V = \sqrt{2}d_H$ between the standard normal density and the densities $N[0,\sigma^2]$ where $0 \le \sigma^2 \le 2.5$ are plotted in Figure 5.1.

5.7 Prerequisite Results for Hellinger Distances

In this section we outline some useful results needed for the discussion of the algebraic forms of some Hellinger distances between certain distributions.

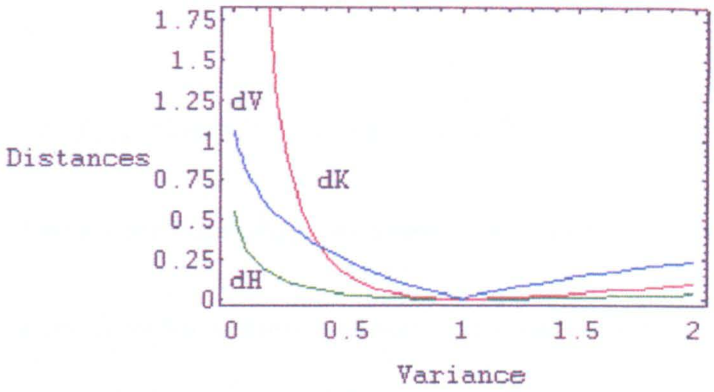


Figure 5.1: A plot of distances between $N[0,1]$ and $N[0,\sigma^2]$

1. First we notice from the definition of $d_H(f;g)$ that if we denote

$$R(f;g) = -4\log[1 - d_H^2(f;g)] \tag{5.10}$$

and

$$f_i = \prod_{j=1}^{k} f_i^{(j)} \quad i = 1, 2.$$

Then

$$R(f_1;f_2) = \sum_{j=1}^{k} R(f_1^{(j)}; f_2^{(j)}) \tag{5.11}$$

which implies that the distance between two densities of random vectors each comprising independent components can be calculated easily from the distance between their component densities.

2. Hellinger distances can sometimes be calculated explicitly between two densities which come from different families. For instance, let f be a normal density with mean μ and variance V and g be a Gamma density $G(\alpha,\beta)$ with the same mean and variance i.e. $\alpha = \mu^2/V$ and $\beta = \mu/V$, then (see, for example, Smith, 1995b)

$$I(f;g) = (2\pi)^{-\frac{1}{2}} 2^{\frac{1}{2}(\alpha-1)} \alpha^{\frac{\alpha}{4}} \frac{\Gamma(\frac{1}{4}(\alpha+1))}{\sqrt{\Gamma(\alpha)}} e^{-\alpha/4}$$

or

$$R(f;g) = \log(\pi/2) - \alpha(\log\alpha + 2\log 2 - 1) + 4\log\Gamma(1/4(\alpha+1)) - 2\log\Gamma(\alpha)$$

Here we note that $d_H(f;g)$ depends only on the parameter α

3. Again from the definition of $d_H(f;g)$ it can be shown that d_H is invariant to the representation of a probability density. This means that for $i = 1, 2$, if the random vector $\boldsymbol{X}_i$ has density f_i and the vector Y_i has density g_i where $Y_i = \xi(X_i)$ and ξ can be inverted, then

$$d_H(f_1; f_2) = d_H(g_1; g_2) \tag{5.12}$$

Of course, if there is a "natural scale" to a problem - for example as introduced by a utility function- then this invariance which is shared by the variation distance and the K-L separation is something of a liability, since it constrains how approximations between variables which take small values are judged compared with approximations between variables which take large values. However, in our application this is not a severe problem because the metric is only used to determine whether it is safe to approximate, so inappropriately large values of separation only have the consequence of delaying an approximation unnecessarily.

We note that the family of appropriate utilities is bounded and often only takes the value 0 or 1 (see Chapter 7), i.e. expected utilities are probabilities on sets. So the variation distance and (by topological equivalence) the Hellinger distance give a natural upper bound discrepancy. Also we note that in our application, because of the physics of the process, the approximation using the Hellinger distance seems to be best when the values of the contamination are large and hence provides good approximations to the expected utility function because of the form of the utility functions used here (see Chapter 7). This therefore tends to make the decision to approximate a conservative one.

4. In dealing with graphical models, we may need to break down a large joint density into smaller components using conditional independence relationships in order to make probability manipulation more manageable (see, for example, Lauritzen & Speigelhalter, 1988; Lauritzen, 1992; Speigelhalter et al., 1993; Smith et al., 1995 and Gargoum et al., 1995). To achieve computational efficiency in these models, it is often necessary to approximate the joint density by substituting either a marginal density by an approximation and keeping the conditional density on the rest of the variables fixed or approximating a conditional density whilst holding the margins on the rest fixed. The following results are very useful for obtaining Hellinger distances on these approximations

   Let

$$\begin{aligned} f(x,y) &= f_X(x) f_{Y|X}(y|x) \\ g(x,y) &= g_X(x) g_{Y|X}(y|x) \end{aligned}$$

   and

$$h(x) = I(f_{Y|X}(.|x), g_{Y|X}(.|x))$$

   Then, by definition

$$I(f,g) = \int f_X^{1/2}(x) g_X^{1/2}(x) h(x) dx$$

   When $f_{Y|X}$ and $g_{Y|X}$ are equal, $h(x) = 1$, then

$$I(f,g) = I(f_X, g_X)$$

which implies that the Hellinger distance between f and g is the same as that between the margins.

When $f_X = g_X$ then

$$I(f,g) = E[h(X)]$$

which gives $d_H^2(f,g)$ in terms of the average conditional square distance.

5.8 Posterior Hellinger Distance and Approximate Bayesian analysis

Let Y be an observation like Poisson counts of a Bayesian model and X represent the states or parameters of this model so that $f_{Y|X}$ and $g_{Y|X}$ can be thought of as likelihoods of X given Y. In many cases it is convenient to approximate a Bayesian analysis by substituting a likelihood which is close (see, for example, West & Harrison, 1989 and Gargoum & Smith, 1995) where interest lies in the distance between the posterior densities of X, $f_{X|Y}$ and $g_{X|Y}$. Here we would not expect to obtain a general result which said that the distance between $f_{X|Y}$ and $g_{X|Y}$ would always be less than the distance between $f_{Y|X}$ and $g_{Y|X}$. However results exist which show that the expectation of the Hellinger distance between posterior densities is smaller than the distances between likelihoods (see Smith, 1995b). We must then arise: How close are posterior densities with common prior and different likelihoods ? This topic will be discussed in Chapter (8).

Chapter 6

Bayesian Dynamic Models for Emission Profiles

6.1 Introduction

In this chapter we outline how qualitative information about the shape of the development of the emission of contamination after an accident can be coded as a Dynamic Linear Model (DLM). Expert judgement about the profile of future emissions is extremely informative and can be accommodated into Bayesian uncertainty management in puff models (see Gargoum & Smith, 1994b). Some examples, using simulated data (model generated data), are given to illustrate how the model's prediction of the emission profile changes as stack monitoring readings are taken.

6.1.1 Background

In Chapter (3) we stated that the true source emissions $\{Q(1), Q(2), ...\}$, where $Q(i)$ denotes the mass of contamination under the ith emitted puff, can be modelled as a DLM with state parameters $\boldsymbol{\theta}_t = (Q(t) - \mu(t), \boldsymbol{\psi}_t)^T$ where $\boldsymbol{\psi}_t$ is a vector of dummy variables whose interpretation will be given later and $\mu(t)$ is the mean of $Q(t)$ given the past. Explicitly,

$$\begin{pmatrix} Q(t) - \mu(t) \\ \boldsymbol{\psi}_t \end{pmatrix} \mid \begin{pmatrix} Q(t-1) - \mu(t-1) \\ \boldsymbol{\psi}_{t-1} \end{pmatrix} \sim N \left[G \begin{pmatrix} Q(t-1) - \mu(t-1) \\ \boldsymbol{\psi}_{t-1} \end{pmatrix}, W \right]$$

where G and W are fixed square matrices.

The discussion in this chapter will focus on the prior specification and subsequent estimation of the source emissions $\boldsymbol{F}^T \ \boldsymbol{\theta}_t + \mu(t) = Q(t)$. The structure of the forecast function and hence the prescription of G, $\boldsymbol{m}_0 = E[\boldsymbol{\theta}_0]$, the initial prior mean, and $\mu(t), \quad t = 1, 2, \ldots$ gives critical prior inputs for designing a DLM consistent with an expert's view concerning the future development of the shape of the emission of contamination after an accident.

In a few scenarios an expert may have good knowledge of the trend term $\mu(t)$, but typically this will not be the case and therefore this term is often set to zero. On the other hand it will often be the case that quite good prior information about the shape of the emission -for example that a release will rise to a peak and then decline to an asymptote- will be available. The problem in examples like the one given above is that the height of the peak, the time it will be reached and the asymptotic value to which the release will converge are all a priori very uncertain. However these quantities can usually be elicited with appropriate confidence bounds.

In Section (2) we outline how qualitative information about the shape of the development of emissions can be coded as a DLM. In Section (3) we describe how prior information about features of this process can be written in terms of prior means and variances of its states. In fact most shapes that arise in this context can be represented by the superposition of one or two canonical 2-dimensional state evolutions. In Section (4) we give an example of how such a profile adapts its estimates to incoming data and in Section (5) we discuss how we manage uncertainty on some key parameters (such as release height) in the dispersal models.

6.2 Examples of the Forecast Functions of the Emission Process

The forecast function of a DLM as a function of the step ahead index k, is determined by the powers of the system matrix. Here we focus on some basic forecast functions which provide the expert's view of the expected development of the series of emissions. Based on theorem (2.2) of Chapter (2), we consider some scenarios of the forecast functions of the emission process.

Case 1. Here we set $\boldsymbol{\psi}_t$ as null where the process becomes 1-dimensional with $\boldsymbol{\theta} = (Q(t) - \mu(t))$, $\boldsymbol{F} = 1$, $G = \lambda$, $|\lambda| < 1$ so that

$$\begin{aligned} Y_t &= Q(t) - \mu(t) + \nu_t, & \nu_t \sim N[0, V] \\ Q(t) &= \mu(t) + \lambda(Q(t-1) - \mu(t-1)) + \omega_t, & \omega_t \sim N[0, W] \end{aligned}$$

or alternatively

$$Q(t)|Q(t-1) \sim N[\mu(t) + \lambda(Q(t-1) - \mu(t-1)), W].$$

If V is small compared with W, this assumes that source readings $Y(t,0), t = 1, 2, \ldots$ are extremely accurate, and observing $Y(1,0), Y(2,0), \ldots$, then $\{Q(1), Q(2), \ldots\}$ are approximately independent. The forecast function is

$$f_t(k) = E[Q(t+k) - \mu(t+k)|D_t]$$

which implies that the future emissions $Q(t+k), k = 1, 2, \ldots$ have expectation

$$E[Q(t+k)|D_t] = \mu(t+k) + \lambda^k(E[Q(t)] - \mu(t)).$$

The values of λ clearly determine the values of the forecast function.

Case 2. The same as case 1 with $0 \leq \lambda \leq 1$, and $\mu(t) = 0, t = 1, 2, \ldots$

When the shape of the emission profile is very vague, it is useful to model this case by setting $\lambda = 1$ which is a steady model (Harrison & Stevens, 1976). For this model the forecast future emission at any time t is $f_t(k) = E[Q(t+k)|D_t] = E[Q(t)|D_t] = m_t, \ k = 1, 2, \ldots$ i.e. constant. If past source emissions have been measured very accurately under this model, then, $f_t(k) = y_t$ where y_t is the last reading of the source emission. In general with an appropriate prior distribution on $Q(1), m_t$ is an exponentially weighted moving average of the past observations, with an adaptive coefficient which decreases as the observation variance increases.

Case 3. Here we consider the 2-dimensional process so that the canonical model has the form

$$\left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix}, \cdot, \cdot \right\},$$

where $0 < \lambda < 1$ and $\mu(t) = 0, \ t = 1, 2, \ldots.$

With state vector $\boldsymbol{\theta}_t = (Q(t) \quad \psi_t(1))^T$ and $E[\boldsymbol{\theta}_t|D_t] = \boldsymbol{m}_t = (m_{t,1}, m_{t,2})^T$

This case is particularly useful for modelling a release which is expected monotonically to rise to an asymptotic value and then drift like a random walk. Here the forecast function has the form, $f_t(k) = E[Q(t+k)|D_t] = \lambda^k m_{t,1} + m_{t,2}$, where usually we set $m_{0,1} = -m_{0,2}$, thus giving an expected exponential rise in emission to an asymptote. Figure 6.1 illustrates the expected profile for the case when $\lambda = 1/2$ and $m_{t,2} = 1$.

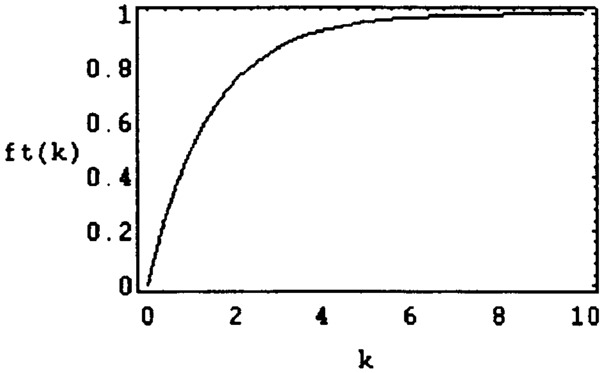


Figure 6.1: Forecast function $f_t(k) = 1 - (0.5)^k$

Case 4. Here the canonical model has the form

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}, ., ., \right\},$$

where $0 < \lambda < 1$, and $\mu(t) = 0$.

With state vector $\boldsymbol{\theta} = (Q(t), \psi_t(1))^T$ and evolution error $\boldsymbol{\omega}_t = (\omega_{t1}, \omega_{t2})^T$ so that

$$\begin{aligned} Y_t &= Q(t) + \nu_t \\ Q(t) &= \lambda Q(t-1) + \psi_{t-1}(1) + \omega_{t1} \\ \psi_t(1) &= \lambda \psi_{t-1}(1) + \omega_{t2} \end{aligned}$$

With $\boldsymbol{m}_t = (m_{t,1}, m_{t,2})^T$, the forecast function is

$$f_t(k) = m_{t,1}\lambda^k + km_{t,2}\lambda^{k-1}$$

This gives another very useful emission profile. The expert expects that the emission will rise from zero to a maximum height at time k^* (say) and then it decays exponentially. Typically we would set $m_{0,1} = 0$, modelling the initial release as zero. The profile when $\lambda = 0.90$ and $m_{t,2} = 1$ is given in Figure 6.2.

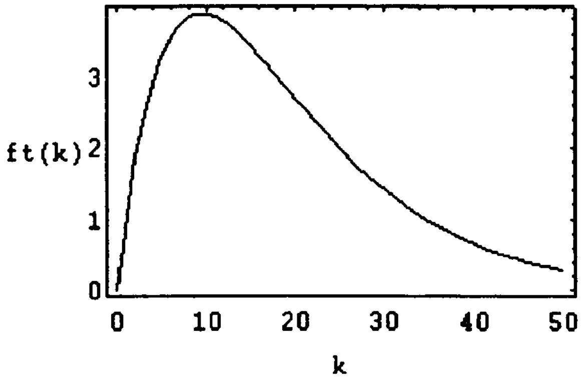


Figure 6.2: Forecast function $f_t(k) = (k/0.90)(0.90)^k$

Now, based on the superposition principle (see Chapter 2), we give some examples of building up complex models for simple components. In fact we find that for most emission shapes that might occur, it is sufficient to consider at most 4-state component models.

Example 1. This expected emission profile is particularly common. Starting from zero it is expected that emissions will rise to a maximum height and then reduce to a leakage. This leakage is assumed to drift like a random walk. Figure 6.3 illustrates this case with

$$\begin{aligned} F &= (1 \quad 1 \quad 0)^T, \\ G &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}, \\ f_t(k) &= m_{t,1} + \lambda^k m_{t,2} + k\lambda^{k-1} m_{t,3}. \end{aligned}$$

where $m_{t,2} = -m_{t,1}$, $m_{t,1} = 1$, $m_{t,3} = 1$ and $\lambda = 0.70$.

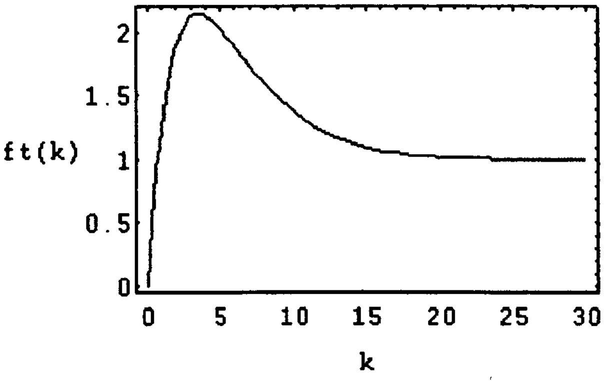


Figure 6.3: Forecast function $f_t(k) = 1 - (0.7)^k + k(0.7)^{k-1}$

Example 2. In this example it may be expected that emission will start from zero, taking a form of a sine/cosine wave, and then dampen to a steady level. Hence the form of the forecast function is determined by the frequency of the periodic component ω that defines the number of time intervals over which the harmonic completes a full cycle. Note that the frequency is modified by the multiplicative term λ^k determined by λ. where $0 < \lambda < 1$. Figure 6.4 illustrates this case with

$$\begin{aligned} F &= (1 \quad 1 \quad 0)^T, \\ G &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \lambda\cos\omega & \lambda\sin\omega \\ 0 & -\lambda\sin\omega & \lambda\cos\omega \end{pmatrix}, \\ f_t(k) &= m_{t1} + \lambda^k r_t \cos(k\omega + \phi_t). \end{aligned}$$

where $r_t^2 = m_{t,2}^2 + m_{t,3}^2$, r is the amplitude of the oscillation and $\phi_t = \arctan(-m_{t,2}/m_{t,3})$ is the phase. In the diagram, $m_{t,1} = 1$, $m_{t,3} = -m_{t,1}$, $m_{t,2} = 0$, $\lambda = 0.70$ and $\omega = \pi/2$.

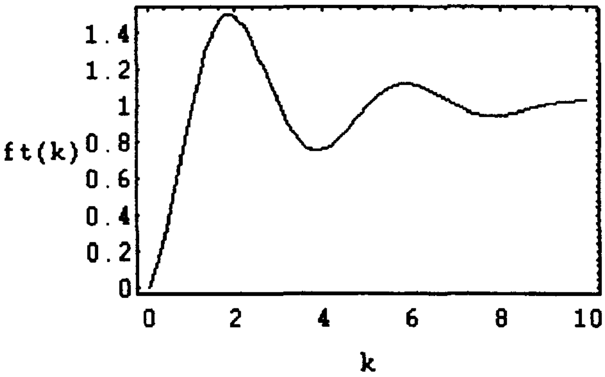


Figure 6.4: Forecast function $f_t(k) = 1 - (0.7)^k \cos(k\pi/2)$

6.3 Prior Information Settings

Most of the time expert judgements do not relate directly to the states but to various features of the distribution of the forecast profile of the emission. However the expert's implicit beliefs about states can be deduced from his or her estimates of those features and associated uncertainty related to them. These deductions are often easy to make and will be illustrated using Example 1.

Specifying the posterior for $\boldsymbol{\theta}_{t-1}$ as

$$(\boldsymbol{\theta}_{t-1}|D_{t-1}) \sim N[\boldsymbol{m}_{t-1}, C_{t-1}]$$

where

$$\boldsymbol{m}_{t-1} = \begin{pmatrix} m_{t-1,1} \\ m_{t-1,2} \\ m_{t-1,3} \end{pmatrix}$$

and

$$C_{t-1} = \begin{pmatrix} c_{t-1}(11) & c_{t-1}(12) & c_{t-1}(13) \\ c_{t-1}(12) & c_{t-1}(22) & c_{t-1}(23) \\ c_{t-1}(13) & c_{t-1}(23) & c_{t-1}(33) \end{pmatrix}.$$

Write the evolution matrix as

$$W_t = \begin{pmatrix} W_t(11) & W_t(12) & W_t(13) \\ W_t(12) & W_t(22) & W_t(23) \\ W_t(13) & W_t(23) & W_t(33) \end{pmatrix}.$$

The sequential analysis has the following steps.

1. $(\boldsymbol{\theta}_t|D_{t-1}) \sim N[\boldsymbol{a}_t, R_t]$

   where

$$\boldsymbol{a}_t = \begin{pmatrix} m_{t-1,1} \\ \lambda m_{t-1,2} + m_{t-1,3} \\ \lambda m_{t-1,3} \end{pmatrix}, \qquad R_t = \begin{pmatrix} R_t(11) & R_t(12) & R_t(13) \\ R_t(12) & R_t(22) & R_t(23) \\ R_t(13) & R_t(23) & R_t(33) \end{pmatrix}$$

with

$$\begin{aligned} R_t(11) &= c_{t-1}(11) + W_t(11) \\ R_t(22) &= \lambda(\lambda c_{t-1}(22) + c_{t-1}(23)) + \lambda c_{t-1}(23) + c_{t-1}(33) + W_t(22) \\ R_t(33) &= \lambda^2 c_{t-1}(33) + W_t(33) \\ R_t(12) &= \lambda c_{t-1}(12) + c_{t-1}(13) + W_t(12) \\ R_t(13) &= \lambda c_{t-1}(13) + W_t(13) \\ R_t(23) &= \lambda^2 c_{t-1}(23) + \lambda c_{t-1}(33) + W_t(23) \end{aligned}$$

2. The one-step forecast distribution is $(Y_t|D_{t-1}) \sim N[f_t, Q_t]$,

where

$$\begin{aligned} f_t &= m_{t-1,1} + \lambda m_{t-1,2} + m_{t-1,3} \\ Q_t &= R_t(11) + 2R_t(12) + R_t(22) + R_t(13) + R_t(23) + V \end{aligned}$$

3. The adaptive vector is given by

$$A_t = \begin{pmatrix} A_{t1} \\ A_{t2} \\ A_{t3} \end{pmatrix} = \begin{pmatrix} R_t(11) + R_t(12)/Q_t \\ R_t(12) + R_t(22)/Q_t \\ R_t(13) + R_t(23)/Q_t \end{pmatrix}.$$

4. The posterior moments are

$$\boldsymbol{m}_t = \begin{pmatrix} m_{t,1} \\ m_{t,2} \\ m_{t,3} \end{pmatrix}, \qquad C_t = \begin{pmatrix} c_t(11) & c_t(12) & c_t(13) \\ c_t(12) & c_t(22) & c_t(23) \\ c_t(13) & c_t(23) & c_t(33) \end{pmatrix}$$

with

$$\begin{aligned} \boldsymbol{m}_t &= \boldsymbol{a}_t + A_t e_t \\ C_t &= R_t - A_t A_t^T Q_t \end{aligned}$$

where $e_t = Y_t - f_t$.

We now need to specify the prior mean vector and the covariance matrix of $\boldsymbol{\theta}_0$ as

$$\boldsymbol{m}_0 = \begin{pmatrix} m_{0,1} \\ m_{0,2} \\ m_{0,3}) \end{pmatrix} \qquad C_0 = \begin{pmatrix} c_0(11) & c_0(12) & c_0(13) \\ c_0(12) & c_0(22) & c_0(23) \\ c_0(13) & c_0(23) & c_0(33) \end{pmatrix},$$

respectively, the covariance matrix W and λ. Experience dictates that we set $W = diag(W_{11}, 0, 0)$.

It is clear that $m_{0,1}, c_0(11) + W_{11}$ are respectively the mean and variance of the asymptote of emission for large t. It is known that at time zero there was no emission that is

$$\theta_{0,1} + \theta_{0,2} = 0$$

which implies $m_{0,2} = -m_{0,1}, c_0(22) = c_0(11), c_0(12) = -c_0(11)$ and $c_0(13) = -c_0(23)$.

6.4 Example of Adaptation of the Estimates to Incoming Data

Using simulated data (model generated data on Example 1 above)- the data are given in Section A.1 of Appendix A- Figures 6.5 and 6.6 of the Appendix illustrate how the model predictions of the emission profile change as stack monitoring readings are taken. Although the observational variance here is relatively high, reflecting the fact that the stack measurements are not that precise, it can be seen that the model quickly and realistically adjusts its forecasts to take account of the lower than predicted observed emission masses.

The upper frame in the graph of Figure 6.5 shows the forecast emission after 4 observations were taken forecasting the next 12 and the lower frame illustrates the forecast after 8 observations forecasting the next 8. Figure 6.6 shows the retrospective values of the concentration after 16 observations were taken.

6.5 Uncertainty of Release Height

As we discussed above, the DLM can be combined with a puff model to estimate the source term profile and predict the contamination spread as long as we believe the model. To allow for modelling error, we can inflate the diagonal of the evolution matrix W_t. However, there are omissions from the model which must be dealt with directly, such as uncertainty of release height or of wind direction.

The height of release at source is a key parameter in the subsequent dispersal of contamination (e.g. the higher the release goes, the faster it spreads). When setting the initial parameters of the model, it is difficult to estimate the height of the release and this will obviously effect the consequences. One solution to this problem is to reduce the risk of setting an erroneous height value by running mixed models. That is, we include several models in our analysis, each with a different release height (see Chapter 2, Section 6). The Bayesian methodology assigns probabilities to each model representing its relative likelihood and updates these probabilities in the light of monitoring data. This has the effect that the data gives most weight to the most likely model, and thus models which consistently perform badly can be discarded.

6.5.1 The Bayesian Updating Algorithm

Suppose that we have m dispersal models $M^{(h_i)}$, $(i = 1, \ldots, m)$ where the dispersal algorithms were the same but whose parameters were different (e.g. the initial height parameter h of source emissions). Suppose that one of the models (as yet uncertain to us) is assumed to be true. Let

$$p(M^{(h_i)} \text{ is true}) = p_{t-1}^{(h_i)} = p(M_{t-1}^{(h_i)}|D_{t-1}) = \pi_i$$

where $\sum_{i=1}^{m} \pi_i = 1$ and $\pi_i > 0, 1 \leq i \leq m$.

The h_i and the π_i are chosen to give approximates in the prior of the release height. Then the probability of an event A (e.g. A might be an observation of contamination at site s lies in the interval [a, b]) is given by

$$p(A) = \sum_{i=1}^{m} \pi_i p_i(A)$$

where $p_i(A)$ is the probability attributed to the event A by the model $M^{(h_i)}$.

Note that if $\theta(t, s)$ is the density of contamination at site s and time t then

$$E[\theta(t, s)] = \sum_{i=1}^{m} \pi_i E_i[\theta(t, s)]$$

where $E_i[\theta(t, s)]$ is the expected contamination under model $M^{(h_i)}$ at site s and time t.

6.5.2 Updating Model Probabilities in the Light of Observations

The Bayesian algorithm allows the updating of $\pi = (\pi_1, \ldots, \pi_m)$ in a simple manner, following the principles of parallel processing of multi-process models, Class I, as introduced by Harrison & Stevens (1976) and described in Section (2.9) of Chapter (2).

Suppose an event $B = \{Y = y\}$ has a density value under model $M^{(h_i)}$ of $p_i(y)$. Then

$$p(A \cap B) = \sum_{i=1}^{m} \pi_i p_i(A \cap B)$$

where p is the probability of the combined model and p_i the probability coming from $M^{(h_i)}$. So

$$p(A|Y = y)p(y) = \sum_{i=1}^{m} \pi_i p_i(A|Y = y)p_i(y)$$

where $p(y) = p(B) = \sum_{i=1}^{m} \pi_i p_i(y)$.

So we have that, given $Y = y$, our updated probability $p^*(A) = p(A|Y = y)$ for an arbitrary event A is given by

$$p^*(A) = \sum_{i=1}^{m} \pi_i^* p_i^*(A)$$

where $\pi_i^* = \frac{p_i(y)\pi_i}{\sum_{j=1}^{m} p_j(y)\pi_j}$.

Implementation

Figure 6.7 of Section A.2, Appendix A shows the Bayesian updating of the dispersal of contamination based on running the sequential learning with the RIMPUFF atmospheric dispersion model used on a real site (Lundtofte Nord1) under real atmospheric conditions but with simulated observational data. The expected dispersal is measured in the unit of radioactivity, Becquerel (Bq), where 1 Bq = 1 atomic disintegration. A detailed documentation of the simulated observations used is given in French & Smith (1992).

Assuming that $\mathcal{A} = \{h_1 = 200m, h_2 = 400m, h_3 = 600m\}$ are taken as representing the a priori plausible range of height values and their initial probabilities are assigned as

$$\pi_i = \frac{1}{3} \quad (i = 1, 2, 3)$$

Figure 6.7 shows the posterior probabilities for the three heights, each with its corresponding expected dispersal together with the marginal expected dispersal. The models are clearly rather different: Model 3 with height = 600m has a higher posterior probability at the time of interest compared with models (1) and (2). Note that if one of the models has a very high posterior probability at the time of interest, then it can be adopted alone for inference. Otherwise, the full unconditional mixture will be used for this purpose.

This direct approach is well known in the statistics literature, since for sufficient many grid points it can produce results as accurate as we like. In the control literature, similar approaches have been used under the heading of parallel processing and Gaussian sums (e.g. Anderson & Moore, 1979, Chapter 10; Alspach & Sorenson, 1972). More refined techniques of numerical approximation are suggested by many authors. Efficient, quadrature– based techniques of numerical integration, which are specifically designed to provide good numerical approximations to posterior functions of interest based on grids of reasonably small numbers of points, are described in Pole & West (1988).

In a Bayesian framework, Draper (1995) has offered an important contribution towards model uncertainty by averaging over competing models rather than just one model. He points out several weaknesses in the conventional approach to model uncertainty, which is dominated by the use of a single model. His approach has much in common with the multi-process models of Harrison & Stevens (1976) (see also Chatfield 1995).

Mixture modelling has been the focus of much research in recent years as the use of mixture distributions has been explored. The Bayesian approach to mixture estimation, which relies on MCMC methods, is even more recent (see Tanner & Wong 1987, and Gelfand & Smith, 1990). Diebolt & Robert (1994) and Escobar & West (1994) proposed Gibbs implementation of mixture estimation.

In the last decade, progress has been made in developing suitable approximation techniques in dealing with the computational problems which arise with Bayesian methods. (See Smith & Roberts, 1992; Smith, 1991; Mengersen & Robert, 1993 and Shaw, 1987, 1988) for excellent overviews and detailed references.

Since our methods need to be almost instantaneous, we have not used the numerical methodologies. Instead we have adopted the basic direct approach of mixture modelling using multi-process Class I as introduced by Harrison & Stevens (1976). But again, these are all useful techniques for verifying the approximation techniques that we are developing.

6.6 Conclusion

Expert judgement about the profile of future emissions is extremely informative and can be simply accommodated into Bayesian uncertainty management in puff models. Even when the parameters of the profile are extremely uncertain a priori, the systems quickly and automatically fit to their empirically justified values. Bayesian software is currently available and can be justified by providing this extra facility with no significant increase in associated running times.

To improve the model effectiveness and manage uncertainty about key parameters such as source height and wind direction, it is necessary to include several models in our analysis to reflect potential errors in these parameters. Hence, the model as described above estimates and provides distributions for source term and release height at the source.

These techniques are now being coded into RODOS during the next three years. Expert judgement will be incorporated to set forecast functions appropriate to different accident scenarios. Each scenario will be given a prior probability and the marginal forecast distributions will be updated as outlined in Section (6).

The key point of interest here is how expert judgement of a qualitative nature can be incorporated into the software using slightly non-standard DLM methodology. Before the feasibility of these techniques was demonstrated, practitioners believed that such coding would be impossible.

6.7 Appendix A

A.1

Tables of simulated data (model generated data on Example 1) together with the prior setting table assuming readings are taken every quarter of an hour

Model generated data on Example 1:

i) Taking 4 observations: forecasting the next 12.

Time	1	2	3	4
Observation	0.75	1.2	1.6	1.8

ii) Taking 8 observations: forecasting the next 8.

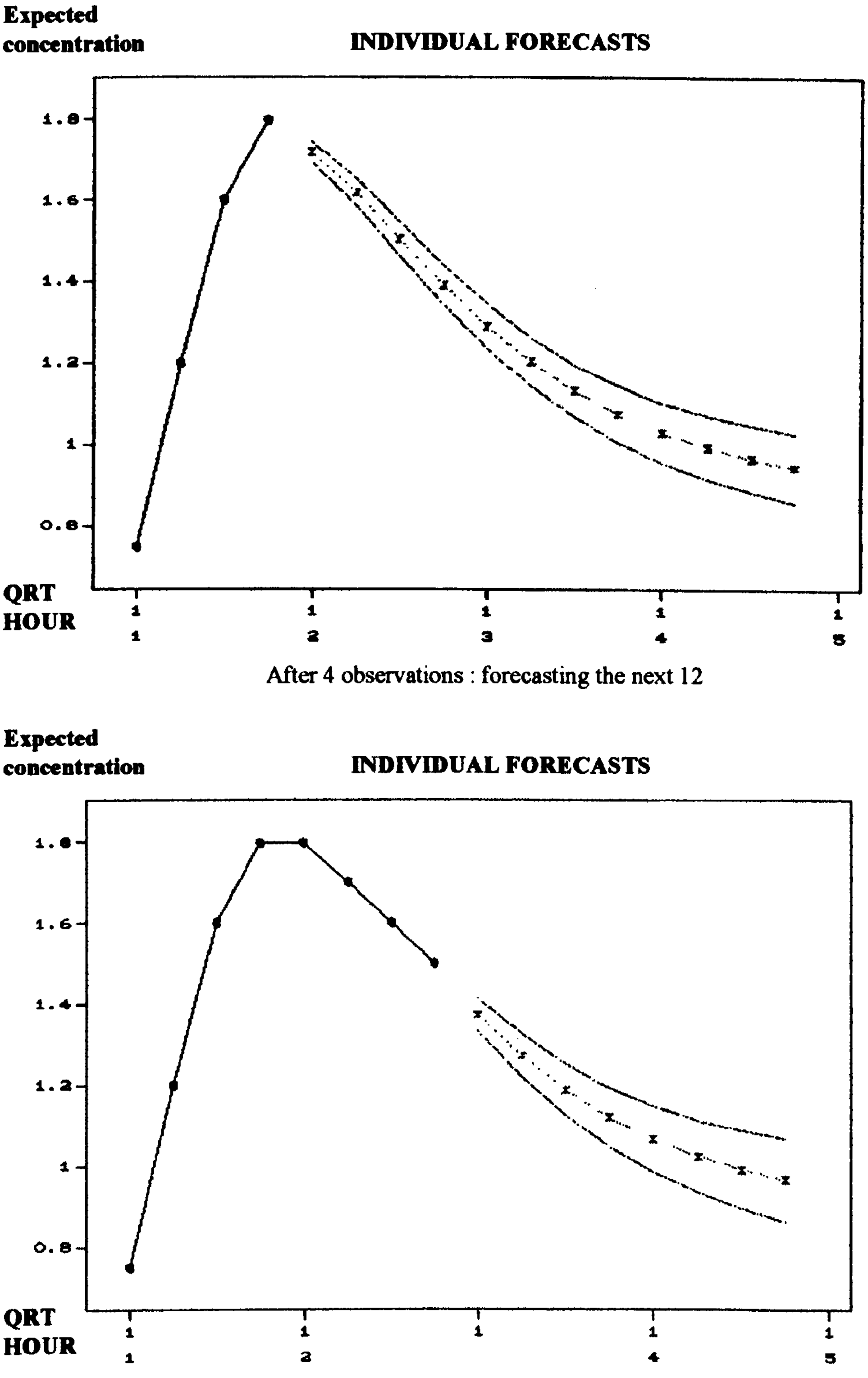
Time	1	2	3	4	5	6	7	8
Observation	0.75	1.2	1.6	1.8	1.8	1.7	1.6	1.5

iii) Taking 16 observations.

Time	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Observation	0.75	1.2	1.6	1.8	1.8	1.7	1.6	1.5	1.4	1.3	1.2	1.1	1.0	0.98	0.95	0.93

Prior setting:

NAMES	MEAN	S.E	DISCOUNT
S.E: dof	0.01	2.00	0.99
LEVEL	1.00	0.01	0.99
REAL 11	-1.0	0.01	0.49
REAL 12	1.00	0.01	0.49



After 4 observations : forecasting the next 12

After 8 observations : forecasting the next 8

Figure 6.5: Forecast emission from simulated noisy stack readings using example (1)

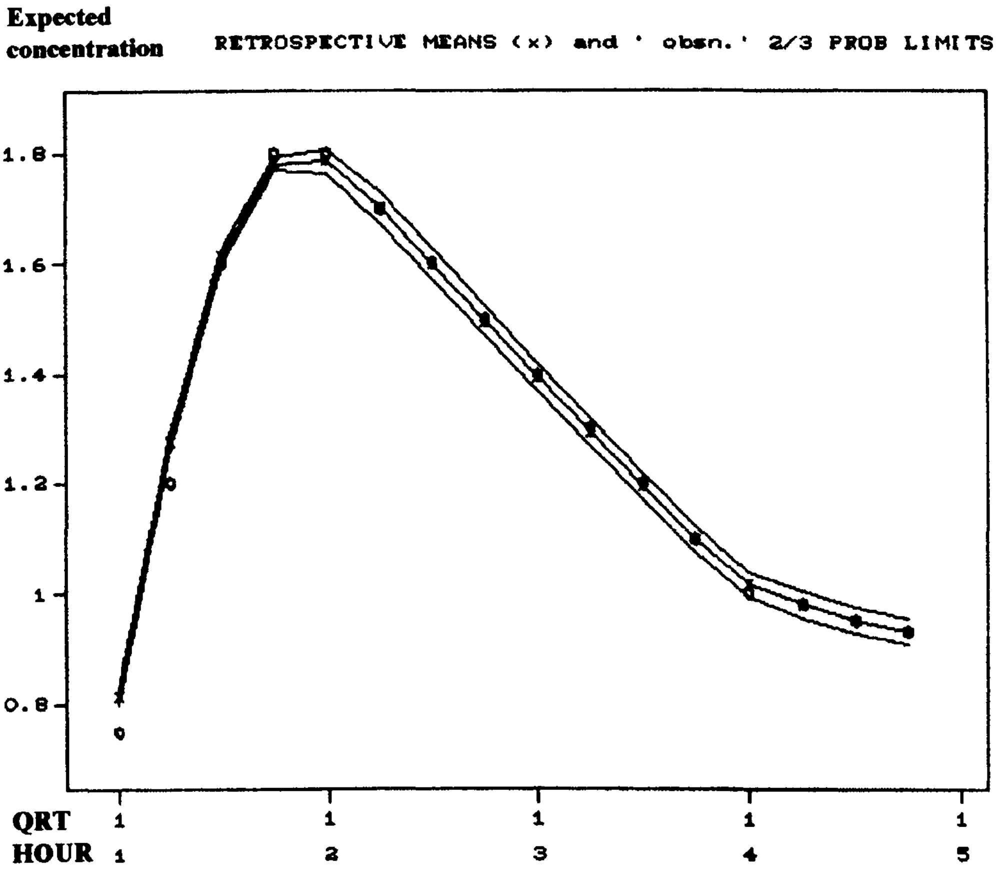
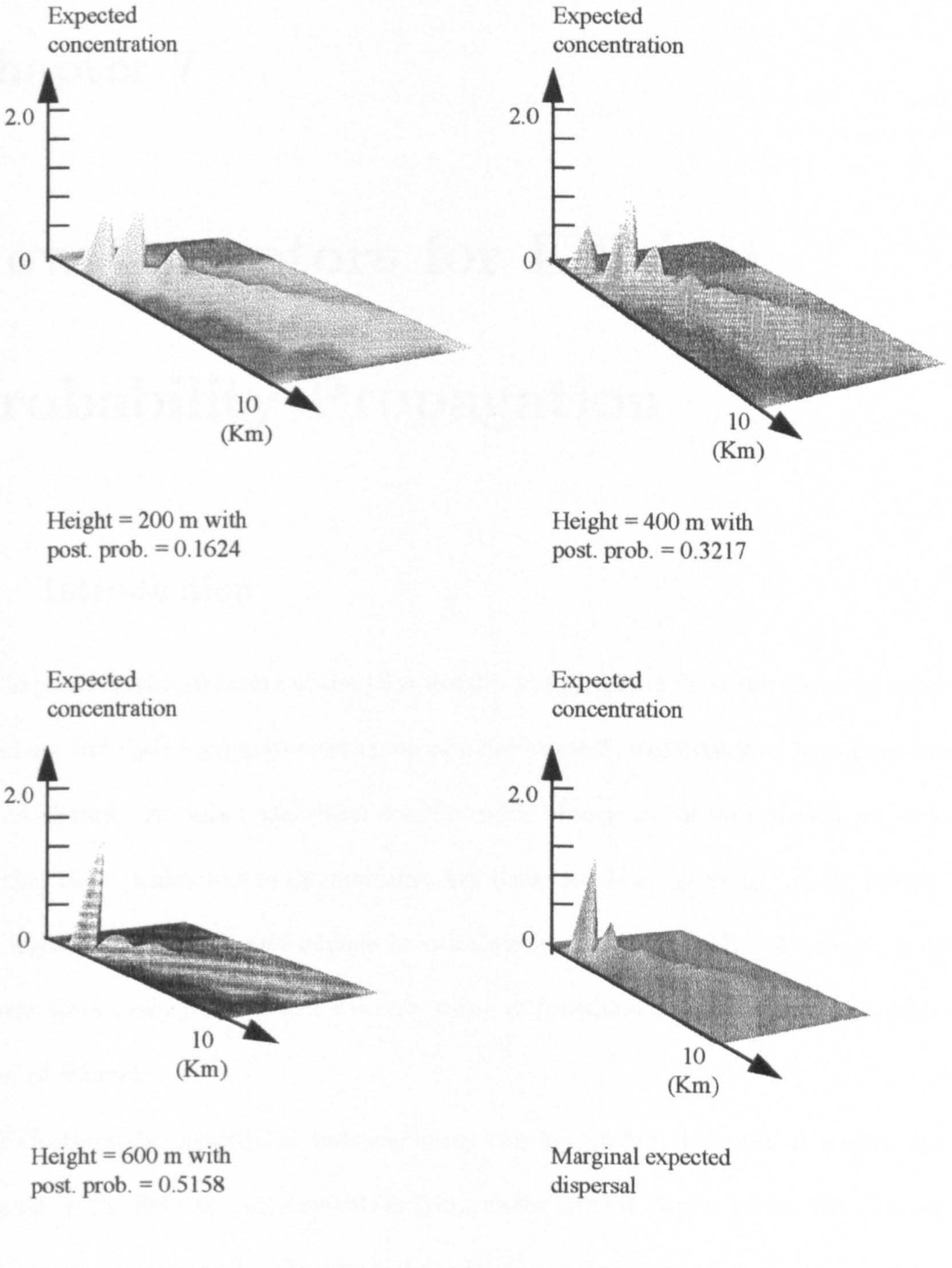


Figure 6.6: Retrospective fitted concentration

A.2

Posteriors for expected dispersals (in Becquerel Bq) in cross 2-D spatial grid (a horizontal scale of 0 -10 Km) associated with 3 different release heights at the source



Height = 200 m with post. prob. = 0.1624

Height = 400 m with post. prob. = 0.3217

Height = 600 m with post. prob. = 0.5158

Marginal expected dispersal

Figure 6.7: Mangement of uncertainty of release height

Chapter 7

New Operators for Efficient Probability Propagation

7.1 Introduction

In Chapter (4), the structure of the joint density $p_T(\boldsymbol{x})$, where $\boldsymbol{X}$ is the vector of mass emissions and their fragments existing on or before time T, was described by a junction tree of cliques. An exact algorithm for the quick absorption of information on such junction trees, which evolve dynamically, has been provided by Smith et al. (1995). The algorithm is simple and flexible in calculating the joint density of quantities of interest when noisy observations are only taken of functions of masses under the same clique of masses.

Unfortunately, conditional independences can be quickly destroyed if masses are updated in the light of the observation lying under several cliques where the number of cliques is determined by the physical model (see Subsection 4.5.3 of Chapter 4 for relating an observation to several cliques). In practice, there are many scenarios when such situations arise. When this happens, the junction tree formally changes, the associated cliques of variables are enlarged, and the local computations become less efficient each time such an observation arrives. A possible solution to this problem is to combine the dynamic evolution with an approximation whose junction tree has cliques which have a limited number of component puff fragments. Kjaerulf (1992) considers an analogous approximation for use with discrete time series. In this chapter we generalise the work of Smith et al. (1995) on Gaussian networks by developing new approximation schemes which are based upon the Kullback-Leibler divergence measure / the Hellinger distance between the true and approximating distribution in Gaussian processes. Initial simulations have shown that the algorithm is fast enough to provide forecasts within the requirements of the RODOS decision support system.

This chapter proceeds as follows . In Section (2) we present some notation and background results. In Section (3) we discuss two new operators which act on a high dimensional Gaussian process, approximating its junction tree by another which allows quicker probability propagation. As these operators are not exact, expressions for the Kullback-Leibler divergence and the Hellinger distance between the true and approximating processes are derived. The operators would not be used unless this divergence measure is small. In Sections (4) and (5) we define exact operators which modify a junction tree to accommodate a new observation vector. In Section (6) we discuss a coarse approximation method which effectively ignores a posteriori, the covariances being induced by observations under different cliques.

7.2 Some Notation and Background Results

Following the notation and terminology of Chapter (4), let $p(\boldsymbol{x})$ denote a Gaussian density over a random vector $\boldsymbol{X}$. Let $\boldsymbol{X}(i)$ denote a sub-random vector of $\boldsymbol{X}$, $1 \leq i \leq n$, and let $\boldsymbol{X}^s(i)$ and $\boldsymbol{X}^r(i)$ denote sub-random vectors of $\boldsymbol{X}(i)$, $2 \leq i \leq n$. Denote the collection of component random variables in $\boldsymbol{X}(i)$, $\boldsymbol{X}^s(i)$ and $\boldsymbol{X}^r(i)$, by $C(i)$, $S(i)$, and $R(i)$ respectively where $C(i) = R(i) \cup S(i), R(i) \cap S(i) = \phi$. As explained in Chapter (4), $C(i)$ is called a *clique*, $S(i)$ a *separator* of $C(i)$ and $R(i)$ a *remainder* of $C(i)$, $1 \leq i \leq n$, where $S(i)$ may be the empty set ϕ.

Let $p_i(\boldsymbol{x}(i))$ denote the density of $\boldsymbol{X}(i)$, $1 \leq i \leq n$, and $q_i(\boldsymbol{x}^s(i))$ denote the density of $\boldsymbol{X}^s(i)$, $2 \leq i \leq n$ where we set $q_i(\boldsymbol{x}^s(i))$ to 1 if $S(i)$ is empty. The density $p(\boldsymbol{x})$ is said to be ***decomposable*** on $(C(1), \ldots, C(n))$ if it is possible to find cliques $C(i)$, $1 \leq i \leq n$, and separators $S(i) \subseteq C(i)$, $2 \leq i \leq n$, such that

(i)

$$p(\boldsymbol{x}) = \frac{\prod_{i=1}^{n} p_i(\boldsymbol{x}(i))}{\prod_{i=2}^{n} q_i(\boldsymbol{x}^s(i))} \tag{7.1}$$

(ii) the cliques $C(1), \ldots, C(n)$ have ***the running intersection property*** i.e.

$$C(i) \cap [\cup_{j=1}^{i-1} C(j)] = S(i) \subseteq C(b_i) \quad 2 \leq i \leq n.$$

where $C(b_i)$ is any clique listed before $C(i)$ i.e. $1 \leq b_i < i$. Henceforth let b_i denote the least index with the property above.

In many applications it is possible to write $p(\boldsymbol{x})$ in a decomposable form where the dimension of $C(i)$ is small for each $i, 1 \leq i \leq n$. When this is the case, it is much more efficient to propagate information through the system working indirectly with the margins $p_i(\boldsymbol{x}(i))$ and $q_i(\boldsymbol{x}^s(i)), 1 \leq i \leq n$, successively updating these, than to update the whole joint density $p(\boldsymbol{x})$ directly (see, for example, Lauritzen and Spiegelhalter, 1988 and Jensen, 1988).

In this chapter we assume that just before new information is accommodated into the system, $p(\boldsymbol{x})$ is decomposable and is stored in terms of a *junction forest* $\mathcal{T}$, over a set of cliques $\mathcal{C} = \{C(1), \ldots, C(n)\}$ and a set of separators $\mathcal{S} = \{S(2), \ldots, S(n)\}$. It is convenient to label the edges of the sub-trees of the forest by the separators $S(i)$, which are not null, $2 \leq i \leq n$. The structure of the joint probability density $p(\boldsymbol{x})$ can be conveniently represented by the triple $(\mathcal{T}, \mathcal{C}, \mathcal{S})$, together with the marginal densities on the cliques $\mathcal{C}$. Because $S(i) \subseteq C(b_i), 1 \leq i \leq n$, the clique margins over $\mathcal{S}$ can be calculated from those over $\mathcal{C}$ and so, in turn, can $p(\boldsymbol{x})$.

7.2.1 Specifying the Evolution of a Decomposable Gaussian Structure

In the past, most interest in the study of probabilistic networks has centred around problems where the junction tree (or at least the variables in that tree) are fixed. However there are a whole class of spatial-temporal processes on which the efficient probability propagation algorithms developed for static networks can be used. One example, developed in our application, has as its variables puffs of contaminated gas which are continually emitted from a chimney stack. In such examples, new variables are being continually added so that, at any time in the process, the joint density of all variables up to that time satisfy equation (7.1)

To specify such a process in terms of its junction graph it is necessary to explain how cliques of variables and their distributions now relate to cliques of variables and their distributions in the immediate past. This motivates the following formal definitions.

Call a Gaussian process $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots)$ *dynamic decomposable* hereafter (d.d.) if the following three conditions hold:

(i) For all stages $t = 1, 2, \ldots$ the density $p_t(\boldsymbol{x}_t)$ of $\boldsymbol{X}_t$ is decomposable with cliques $\mathcal{C}_t = (C_t(1), \ldots, C_t(n_t))$ and separators $\mathcal{S}_t = (S_t(2), \ldots, S_t(n_t))$. Here we denote by $\mathcal{T}_t$ and $\mathcal{S}_t$ a junction forest and a set of separators of $p_t(\boldsymbol{x}_t)$, respectively.

(ii) **(a)** $n_{t+1} = n_t + 1 \quad t = 1, 2, \ldots$

(b) $C_{t+1}(i) = C_t(i) \quad t = 1, 2, \ldots\ 1 \leq i \leq n_t.$

(c) $\mathcal{T}_t$ is the sub-forest of $\mathcal{T}_{t+1}$ on the nodes $\mathcal{C}_t$

(iii) If $\boldsymbol{\mu}_t(i)$, $\Sigma_t(i)$ denote respectively the mean vector and covariance matrix of the clique $C_t(i), 1 \leq i \leq n_t$, $t = 1, 2 \ldots.$

Then

(a) $\boldsymbol{\mu}_{t+1}(i) = \boldsymbol{\mu}_t(i) \qquad 1 \leq i \leq n_t,$

(b) $\Sigma_{t+1}(i) = \Sigma_t(i) \qquad 1 \leq i \leq n_t.$

A d.d. process is usually specified inductively. The inputs required are :

(i) A prior specification of mean, covariance matrix pairs $(\boldsymbol{\mu}_1(i)\ , \Sigma_1(i))$ of each clique of variables $C_1(i)$, $1 \leq i \leq n_1$, in $\mathcal{C}_1$. This is sufficient to specify the joint density $p_1(\boldsymbol{x}_1)$, using equation (7.1).

(ii) The specification of $p_{t+1}(\boldsymbol{x}_{t+1} \setminus \boldsymbol{x}_t | \boldsymbol{x}_t)$, which is obviously Gaussian, has a conditional mean and a covariance matrix denoted by $(\boldsymbol{\mu}^0_{t+1}(n_{t+1})\ , \Sigma^0_{t+1}(n_{t+1}))$ (say).

This pair is sufficient completely to specify the joint distribution of variables in $C_{t+1}(n_{t+1})$ given the variables in C_t. Again using equation (7.1), we can then obtain $p_{t+1}(\boldsymbol{x}_{t+1})$ in terms of its clique margins by calculating $(\boldsymbol{\mu}_{t+1}(n_{t+1})\ ,\ \Sigma_{t+1}(n_{t+1}))$, using the usual formulae, from $\boldsymbol{\mu}^0_{t+1}(n_{t+1})$, $\Sigma^0_{t+1}(n_{t+1})$ and the mean and the covariance matrix of $\boldsymbol{X}_t$ (see, for example, Mardia et al., 1979).

In fact, from the decomposability of $(\mathcal{T}, \mathcal{C}, \mathcal{S})$,

$$p_{t+1}(\boldsymbol{x}_{t+1} \setminus \boldsymbol{x}_t | \boldsymbol{x}_t) = p_{t+1}(\boldsymbol{x}_{t+1} \setminus \boldsymbol{x}_t | S_{t+1}(n_{t+1}))$$

and since by definition, $p_{t+1}(\boldsymbol{x}_{t+1})$ is decomposable, then,

$$S_{t+1}(n_{t+1}) \subseteq C_{t+1}(b(n_{t+1}))$$

with

$$b(n_{t+1}) \leq n_{t+1}$$

By (ii) of the definition of d.d. we have that $C_{t+1}(b(n_{t+1})) \in \mathcal{C}_t$. It follows that to calculate $(\boldsymbol{\mu}_{t+1}(n_{t+1}), \Sigma_{t+1}(n_{t+1}))$ we need only $(\boldsymbol{\mu}^0_{t+1}(n_{t+1}), \Sigma^0_{t+1}(n_{t+1}))$ and $(\boldsymbol{\mu}_t(b(n_{t+1})), \Sigma_t(b(n_{t+1})))$, the latter pair being provided from a clique mean and covariance pair that has already been calculated. The explicit formulae for constructing $p_{t+1}(\boldsymbol{x}_{t+1})$ from $p_t(\boldsymbol{x}_t)$ in this way are given in for example Smith et al. (1995).

In the class of problems we consider here, either because an approximation scheme is used or because the process evolves, one Gaussian probability model $p(\boldsymbol{x})$ maps to another Gaussian probability model $p^*(\boldsymbol{x})$. Following the representation formulated in Section (2), we therefore need to specify :

(a) the map $(\mathcal{T}, \mathcal{C}, \mathcal{S}) \longrightarrow (\mathcal{T}^*, \mathcal{C}^*, \mathcal{S}^*)$

relating the original forest, clique set and separator set triple to a new forest, clique set and separator set triple.

(b) the pair $(\boldsymbol{\mu}_i^*, \Sigma_i^*)$ of $C^*(i^*) \in \mathcal{C}^* = \{C^*(1), \ldots, C^*(n^*)\}$ as a function of $(\boldsymbol{\mu}_i, \Sigma_i)$, $1 \leq i \leq n$ for each $i^*, 1 \leq i^* \leq n^*$.

For a d.d. process on $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots)$ a relevant density over a random vector $\boldsymbol{X}_{t+1}$ needs to be calculated from the random vector $\boldsymbol{X}_t$ of variables alone at time t, $\boldsymbol{X}_t$ being a sub-vector of $\boldsymbol{X}_{t+1}$. Then $(\mathcal{T}, \mathcal{C}, \mathcal{S})$ associated with $p_t(\boldsymbol{x}_t)$ evolves to $(\mathcal{T}^*, \mathcal{C}^*, \mathcal{S}^*)$, the clique structure of $p_{t+1}(\boldsymbol{x}_{t+1})$.

From the notation above we have that $\mathcal{C}^* = \mathcal{C} \cup \{C_{t+1}(n_{t+1})\}$ and $\mathcal{S}^* = \mathcal{S} \cup \{S_{t+1}(n_{t+1})\}$. The forest $\mathcal{T}^*$ is obtained from $\mathcal{T}$ by adding a new node labelled $C_{t+1}(n_{t+1})$ connected by an edge labelled by $S_{t+1}(n_{t+1})$ to $C_{t+1}(b(n_{t+1}))$, provided $S_{t+1}(n_{t+1})$ is not null. We obtain the clique means and covariances of $\mathcal{C}^*$ in the way described above. This is our first example of the specification of the evolution of a process using (a) and (b).

7.2.2 The Exact Updating (The ADJUST Operator)

Smith et al. (1995) provide an algorithm based on one by Dawid (1992) for the quick absorption of information on cliques of a junction forest when information vector $\boldsymbol{Y}$ arrives with direct relevance to only one arbitrary clique $C(1) \in \mathcal{C}$ of that forest. Thus, in the notation defined above, $\boldsymbol{Y}$ is assumed to have the property that $\boldsymbol{Y}$ is independent of $\boldsymbol{X}_{t+1} \setminus \boldsymbol{x}(1)$ given $\boldsymbol{X}(1) = \boldsymbol{x}(1)$. This algorithm keeps $(\mathcal{T}^*, \mathcal{C}^*, \mathcal{S}^*) = (\mathcal{T}, \mathcal{C}, \mathcal{S})$. It then chooses an ordering of those cliques $\mathcal{C}(1) = \mathcal{C}^*(1)$ in the sub-tree $\mathcal{T}(1)$ of $\mathcal{T}$ which contains $C(1)$ satisfying the *running intersection property* and beginning with $C(1)$, the clique about which $\boldsymbol{Y}$ is informative. This is always possible (see Lauritzen and Spiegelehalter, 1988). It then updates the mean and the covariance matrix, $\boldsymbol{\mu}(1)$ and $\Sigma(1)$ to $\boldsymbol{\mu}^*(1)$ and $\Sigma^*(1)$ respectively in the light of $\boldsymbol{Y}$, using Bayes' rule. All distributions outside $\mathcal{C}(1)$ are left unchanged. The distribution of variables in $\mathcal{C}(1)$ is updated as follows.

Let the vector $\boldsymbol{X}(i)$ of variables in $C(i)$, $2 \leq i \leq n$ have a joint normal distribution specified by the mean vector

$$\boldsymbol{\mu}(i) = (\boldsymbol{\mu}_1(i), \boldsymbol{\mu}_2(i))^T$$

and covariance matrix

$$\Sigma(i) = \begin{bmatrix} \Sigma_{11}(i) & \Sigma_{12}(i) \\ & \Sigma_{22}(i) \end{bmatrix}$$

where the vector $\boldsymbol{X}^s(i)$ of the variables in its separator $S(i)$ has mean and covariance matrix $\boldsymbol{\mu}_1(i)$, $\Sigma_{11}(i)$ respectively. We now proceed inductively.

Suppose all the means and covariance matrices in the cliques $C(1), \ldots, C(i-1)$ $\in \mathcal{C}(1)$ have already been updated, $2 \leq i \leq n$. In particular this will mean that, in this pass, the mean and covariance matrix of the vector $\boldsymbol{X}^s(i)$ of the variables in $S(i) \subset C(b_i), 1 \leq b_i \leq i-1$ have been updated from $(\boldsymbol{\mu}_1(i)$, $\Sigma_{11}(i))$ to $(\boldsymbol{\mu}_1^*(i)$, $\Sigma_{11}^*(i))$. We can now define the ADJUST operator $\mathrm{Adj}(C(i), S(i))$ as a function of $(\boldsymbol{\mu}(i)$, $\Sigma(i))$ and $(\boldsymbol{\mu}_1^*(i)$, $\Sigma_{11}^*(i))$ which maps

$$\boldsymbol{\mu}(i) = (\boldsymbol{\mu}_1(i), \boldsymbol{\mu}_2(i))^T \longrightarrow (\boldsymbol{\mu}_1^*(i), \boldsymbol{\mu}_2^*(i))^T$$

$$\Sigma(i) = \begin{bmatrix} \Sigma_{11}(i) & \Sigma_{12}(i) \\ & \Sigma_{22}(i) \end{bmatrix} \longrightarrow \begin{bmatrix} \Sigma_{11}^*(i) & \Sigma_{12}^*(i) \\ & \Sigma_{22}^*(i) \end{bmatrix}$$

where

$$\boldsymbol{\mu}_2^*(i) \quad = \quad \boldsymbol{\mu}_2(i) + A(i)(\boldsymbol{\mu}_1^*(i) - \boldsymbol{\mu}_1(i))$$

$$\Sigma_{21}^{*}(i) = A(i)\Sigma_{11}^{*}(i)$$

$$\Sigma_{22}^{*}(i) = \Sigma_{22}(i) - A(i)[\Sigma_{11}(i) - \Sigma_{11}^{*}(i)]A^{T}(i)$$

with $A(i) = \Sigma_{21}(i)\Sigma_{11}^{-1}(i)$. Proceeding in this way, we eventually update all the means and variances of the cliques in $\mathcal{C}(1)$. With the original comment we can assert that if an observation $\boldsymbol{Y}$ is taken and $\boldsymbol{Y}$ is independent of $\boldsymbol{X}_{t+1} \setminus \boldsymbol{X}(1)$, then all the mean vectors and covariance matrices of $p(\boldsymbol{x})$ can be updated successively to the clique means and covariance matrices of $p(\boldsymbol{x}|\boldsymbol{y})$.

Typically an observation vector $\boldsymbol{Y}$ is partitioned into k sub-vectors $(\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(k))$ such that $\boldsymbol{Y}(j)$ is independent of $\boldsymbol{X}_{t+1}$ given $\boldsymbol{X}^{(j)}(1)$, where $\boldsymbol{X}^{(j)}(1)$ is the vector of components of a clique $C^{(j)}(1) \in \mathcal{C}_{t+1}$, $1 \leq j \leq k$. The ADJUST operator is then used successively on $\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(k)$ to find $p(\boldsymbol{x}|\boldsymbol{y})$.

A special case of the use of this operator in conjunction with a d.d. Gaussian process $(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots\ )$ is given in Smith et al. (1995).

7.3 Approximating to a More Efficient Clique Structure

If an observation vector $\boldsymbol{Y}$ has a distribution which depends on variables contained in several cliques, the cliques and the corresponding tree structure of $p(\boldsymbol{x}|\boldsymbol{y})$ are usually different to those of $p(\boldsymbol{x})$. Typically cliques become larger and hence the propagation algorithms like the one outlined above become slower. Many observations of this type can make some cliques extremely large and exact algorithms then become very unwieldy. Fortunately, it is common for many of the partial correlations between components of $\boldsymbol{X}$ after conditioning on $\boldsymbol{Y}$ to be very close to zero. Efficiency can then be preserved, by approximating a probability distribution associated with the true complicated tree structure by a close probability distribution with an associated simpler tree structure. In this section we present some approximation schemes defined through two new operators, namely the CUT operator and the TEAR operator. These approximation schemes would be activated only when it is known that the Kullback-Leibler divergence / the Hellinger distance between the true and the approximating distribution is sufficiently small. Before discussing these schemes, some consideration of how and why certain thresholds for distributional approximations in our particular application is provided.

In the context of our application, decisions are made on whether or not to

- evacuate a population to avoid relatively high short-term exposures by protecting the population against the inhalation of radioactive material and external exposure from radioactive material in the air and on the grounds;
- administer stable (non-radioactive) iodine tablets to prevent iodine from concentration in the thyroid gland;
- shelter a population to provide short-term protection from external irradiation from radioactive material in the air and on the ground, and from inhalation of radioactive material.

Any of these actions will have benefits (radiation dose averted, reassurance, etc.) but also harm (risks, disruption, anxiety, cost, etc.). Action is taken when the benefits outweigh the disadvantages. Henceforth we shall discuss the component of the utility U associated with harms.

Decisions are based on whether contamination in a region is "dangerously large", where a dangerously large value is determined to be greater than C (say). This upper value of radiation dose (called Emergency Reference Level) for the introduction of particular countermeasures is technically defined in handbooks and varies from country to country.

To define the application utility function, let A be the set of actions or acts (e.g. evacuation or non-evacuation) and denote a typical member of this set (that is an action) by a. The uncertainty of the problem concerns the uncertain quantity (state of the world) θ (contamination). The set of possible states of the world or the possible values of θ is labelled S, the probability distribution of θ is denoted by $P(\theta)$, and $U(a,\theta)$ is the utility function that associates a utility with each pair (a,θ).

Given these inputs, one must make a decision which amounts to selecting an act a from the set A that has the highest expected utility of all acts or of all members of A. The expected utility of any act a is denoted by $\bar{U}(a)$ and can be calculated as

$$\bar{U}(a) = E[U(a)] = \int U(a,\theta)p(\theta)d\theta$$

where the integration is over the set S. An act a^* is optimal if

$$E[U(a^*)] \geq E[U(a)] \quad \text{for all} \quad a \in A$$

In terms of the notation presented above, the zero-one utility function is defined as

$$U(a,\theta) = \begin{cases} 1 & \text{if contamination} > C \\ 0 & \text{otherwise} \end{cases}$$

Taking expectation with respect to this distribution, the expected utility is

$$\bar{U}(a) = p[\text{contamination} > C|a]$$

Thus, the optimal act a^* is the act such that

$$\bar{U}(a^*) \geq \bar{U}(a) \qquad \text{for all} \qquad a \in A$$

is

$$p[\text{contamination} > C|a^*] \geq p[\text{contamination} > C|a] \qquad \text{for all} \qquad a \in A$$

Now, if we replace the density p by $\hat{p}$, a question that arises is how we set a threshold value α for the Hellinger distance/K-L which ensures that probabilities like those above are sufficiently accurate. In a sense, the setting of such a threshold is highly dependent on the model being used and its environment. Upper bounds can be obtained from the relationship between Hellinger distances and variation distances given in Section (5.6). However these upper bounds are very coarse and tend to make the model unnecessarily reluctant to offer approximations.

In this application technicians in Leeds have run many analyses based on different models and data sets and using:

i) the (slow) exact propagation algorithm.

ii) the quick approximation algorithm based on different thresholds of Hellinger distances.

In terms of the types of probabilities we need, it was apparent that a value of $\alpha = 0.01$ ensured that differences between predictions (errors) using the two algorithms are off by at most 10 per cent of the predictions obtained by using the exact algorithm.

7.3.1 Approximation by Using Edge Deletion (The CUT Operator)

In this section we discuss the CUT operator which separates the original distribution with a connected junction tree into a collection of independent distributions having an associated set of disconnected junction trees. More specifically, this removes an edge from a junction tree by making the variables in a separator S and those in its complement R in a clique C independent.

The removal of a single edge may speed up the propagation of information and imply an enormous reduction of complexity of inference. This is because data $\boldsymbol{Y}$, accommodated using the ADJUST operator, need only to be propagated through parts of a junction graph which are connected to the cliques immediately influenced by that information (see Kjaerulf, 1992 and the Adjust updating equations given above).

The CUT operator CUT$(S;C)$ will "cut" the separator S from a clique C containing S in the triple $(\mathcal{T}, \mathcal{C}, \mathcal{S})$. It acts as follows.

Let Cut$(S;C)$: $(\mathcal{T}, \mathcal{C}, \mathcal{S}) \longrightarrow (\mathcal{T}^*, \mathcal{C}^*, \mathcal{S}^*)$, where

1. $\mathcal{T}^*$ is obtained by deleting an edge labelled by $S \subseteq C$. When $\mathcal{T}$ is a tree, then, $\mathcal{T}^* = (\mathcal{T}_1, \mathcal{T}_2)$ where $\mathcal{T}_1$ (containing C) and $\mathcal{T}_2$ are the two disconnected subtrees of $\mathcal{T}$.

2. $\mathcal{C}^* = \{\mathcal{C} \setminus \{C\}\} \cup \{R\}$ where $R = C \setminus S$.

3. The means and variances of cliques in $\mathcal{C}^* \setminus \{R\}$ are copied from their counterparts in $\mathcal{C}$. The mean and covariance matrix of R can be read directly from the mean and covariance matrix of C in $\mathcal{C}$. These are given respectively by the mean vector of components in R as given in C and the covariance matrix of R found in the

covariance matrix of C as a block in that matrix associated with components R in C.

As an approximation we need the Kullback-Leibler divergence associated with the cut. This is often very easy to calculate.

Suppose that the underlying joint probability density function $p(\boldsymbol{x})$ over the random vector $\boldsymbol{X} = (\boldsymbol{X}(1), \ldots, \boldsymbol{X}(n))^T$ can be expressed as in (7.1).

Let $C(i^*) = R(i^*) \cup S(i^*)$ be the clique for which the separator $S(i^*)$ is cut and let $R(i^*)$ be the remainder after cutting $S(i^*)$ so that $R(i^*) \cap S(i^*) = \phi$.

Consider the situation when for all cliques $C(i), 1 \leq i \neq i^* \leq n$, either $C(i) \cap C(i^*) \subseteq R(i^*)$ or $C(i) \cap C(i^*) \subseteq S(i^*)$. This condition will hold for all models developed in Smith et al. (1995), for example, and will be automatic if $C(i^*)$ contains exactly two variables.

Suppose we approximate $p(\boldsymbol{x})$ by $\hat{p}(\boldsymbol{x})$, where $\hat{p}(\boldsymbol{x})$ is the density obtained after the cutting in which the variables in $R(i^*)$ and $S(i^*)$ are independent. Then in $\hat{p}(\boldsymbol{x})$, $\frac{p(\boldsymbol{x}(i^*))}{q(\boldsymbol{x}^s(i^*))}$ is replaced by $q(\boldsymbol{x}^r(i^*))$. Because of the condition above, all other clique margins will be the same. So, in particular from equation (7.1) we have that

$$\frac{p(\boldsymbol{x})}{\hat{p}(\boldsymbol{x})} = \frac{p(\boldsymbol{x}(i^*))}{q(\boldsymbol{x}^s(i^*))q(\boldsymbol{x}^r(i^*))}$$

Because this ratio only involves variables in the vector $\mathbf{X}(i^*)$, the Kullback-Leibler divergence in the Gaussian case is (see Section B.1 of Appendix B)

$$\begin{aligned} d_K(p(\boldsymbol{x}); \hat{p}(\boldsymbol{x})) &= \int_{\boldsymbol{x}(i^*) \in C(i^*)} [\log p(\boldsymbol{x}(i^*)) - \log q(\boldsymbol{x}^s(i^*))q(\boldsymbol{x}^r(i^*))] p(\boldsymbol{x}(i^*)) d\boldsymbol{x}(i^*) \\ &= -\frac{1}{2} \log \left[\frac{detV}{detV^s detV^r} \right] \end{aligned}$$

where V is the covariance matrix of $(\boldsymbol{X}^s(i^*), \boldsymbol{X}^r(i^*))$ and V^s, V^r are respectively the covariance matrices of $X^s(i^*)$ and $X^r(i^*)$. The Hellinger distance in this case can be obtained using the formulae of Section (5.6).

7.3.2 The Steady Model and CUT Operator

An example of using the CUT operator at the source

Here we illustrate how to set the K-L divergence for the CUT operator in the case when we have steady emission and only single readings at a given source. Consider the univariate process $\{Q(t), t = 1, 2, \ldots\}$ where as stated in Chapter (3) the puffs are indexed such that puff i is released at time $t = i$ where $Q(i)$ denotes the mass of contamination under the ith emitted puff. As explained in Chapters (3) and (6), the steady emission process is modelled as

$$\begin{aligned} Y_t &= Q(t) + \nu_t, \quad \nu_t \sim N[0, V] \\ Q(t) &= Q(t-1) + \omega_t, \quad \omega_t \sim N[0, W] \end{aligned}$$

From Section B.1 of Appendix B we have

$$I = Inf(Q(t), Q(t-1)|Y_1, \ldots, Y_{t-1}) = -\frac{1}{2}\log[1 - corr^2(Q(t), Q(t-1)|Y_1, \ldots, Y_{t-1}]$$

where

$$Corr^2(Q(t), Q(t-1)|Y_1, \ldots, Y_{t-1}) = \frac{Cov^2(Q(t-1), Q(t-1) + \omega_t|Y_1, \ldots, Y_{t-1})}{Var(Q(t)|Y_1, \ldots, Y_{t-1})Var(Q(t-1) + \omega_t|Y_1, \ldots, Y_{t-1})}$$

Let $C = Var(Q(t-1)|Y_1, \ldots, Y_{t-1})$

Then

$$Corr^2(Q(t), Q(t-1)|Y_1, \ldots, Y_{t-1}) = \frac{C^2}{C(C+W)} = \left(1 + \frac{W}{C}\right)^{-1}$$

So we "cut" $Q(t)$ from $Q(t-1)$ after observing Y_{t-1} if the information divergence

$$\begin{aligned} I &= -\tfrac{1}{2}\log[1 - corr^2(Q(t), Q(t-1)|Y_1, \ldots, Y_{t-1}) \\ &= -\tfrac{1}{2}\log[1-(1+\tfrac{W}{C})^{-1}] \\ &= \tfrac{1}{2}\log[\tfrac{C+W}{W}] \\ &= \tfrac{1}{2}[\log(C+W) - \log(W)] \end{aligned}$$

is small.

Notice that this is increasing in C and takes the value 0 when $C = 0$.

In the long run, it is easy to check (see West & Harrison, 1989, p51) that

$$C \longrightarrow \frac{1}{2}W[(1+4S^{-1})^{1/2} - 1]$$

where $S = (W/V)$. So in the limit

$$\begin{aligned} \log(C+W) &= \log\tfrac{1}{2}W[(1+4S^{-1})^{1/2}+1] \\ &= \log W + \log\tfrac{1}{2}[1+(1+4S^{-1})^{1/2}] \end{aligned}$$

Thus

$$I = \frac{1}{2}\log\left[\frac{1}{2}((1+4S^{-1})^{1/2}+1)\right]$$

Notice that when I is small which corresponds to the case when S^{-1} is small and when this is the case then, $(1+4S^{-1})^{1/2} \simeq (1+2S^{-1})$ which implies that

$$I \simeq \frac{1}{2}\log(1+S^{-1}) \simeq \frac{1}{2}S^{-1} = \frac{1}{2}(W/V)^{-1}.$$

Typical values of $S^{-1} = (V/W)$ and the corresponding I are illustrated in the following table:

S^{-1}	50	10	1	0.5	0.2	0.1	0.01
I	1.01	0.65	0.24	0.16	0.08	0.04	0.0049

Notice that small values of I are associated with accurate source term readings (i.e. when the observation variance V is very small).

Implementation

The CUT operator was implemented in the RODOS computer code using the Lundtofte Nord1 data set (see Section 6.5) and the results reported here were based on running the sequential learning with this data. Observational data were collected from several sites in the plume at time steps t = 7200 secs, t = 8400 secs and t = 9600 secs after the accident via detector points (see Table 7.1 of Section B.2, Appendix B). These observations were used in Bayesian updating of the expected instantaneous contamination by two models: model 1 where no cuts between the cliques were allowed (the Hellinger distance = 0), and model 2 where cuts are allowed (the Hellinger distance = 0.01). The input parameters were as follows:

- prior variance of mass = 1.0e+6
- the "system" error in the pentification variance (the symmetry of pentification variance) = 0.15
- stack observation variance = 1.0e-10

Here we are assuming a steady emission of mass from the source whose associated variance was large relative to the variance of the instantaneous readings remote from the site. The accuracy by which the Normal-Cut model calculates the expected concentration has been tested by comparison with the exact Normal model.

Table 7.2 of Section B.2, Appendix B shows comparisons of the expected concentrations in Becquerels (1 Becquerel (1 Bq) = 1 atomic disintegration per second) at different time steps using the two models. For instance at detector point number 34 at time step $t = 7200$ secs, the expected concentration using model 1 (without cuts) was 8.319278e-02, and using model 2 (with cuts) it was 8.319239e-02 with difference = 3.9e-07. In fact, as indicated in the table, the maximum absolute deviation of predictions is less than 0.005 at all the detector points.

Section B.3 of Appendix B includes 3- dimensional plots which show the difference between the forecasts of mean levels of contamination at three different time steps for the following cases:

1. by using the deterministic model, i.e. no updating in the light of instantaneous concentration observations (using the RIMPUFF model with only a priori data);
2. by using the Normal model;
3. by using the Normal-Cut model (the approximation)

Here it can be seen that the forecasts of the concentration using the Normal model are very similar to those of the Normal-Cut model with significantly less computational times of the complete run (i.e. all three time steps) in the later case. It should also be noted that because of the setting of the ratio of the relative uncertainty between the mass of contamination and the symmetry of pentification, the data suggest that the overall release was rather greater than initially input to RIMPUFF. The effect of this can be seen in the difference between the forecasts using the contamination readings (in the Normal and Normal–Cut models) and not (in the deterministic model). This can be noted in the dramatic increase in the predicted levels of contamination.

The main point to emphasise here is that the approximating model is extremely useful in obtaining systems with disconnected junction trees to achieve efficiency. In this example the clique update used with RIMPUFF without approximation took longer in user time compared with the clique update with approximation . In case of large data sets the approximation is even more favourable.

7.3.3 Approximation by Splitting the Cliques (The TEAR Operator)

When data arrives about more than one clique at a time, these cliques may be joined together (see Section 7.4 below) and an efficient exact absorption of information can be achieved as long as the cliques are small. However, when the clique sizes become large, exact algorithms for propagation become inefficient and an approximation which returns the clique structure and junction trees to their original forms together with its associate K-L/Hellinger distances is needed. The TEAR operator, TEAR$(\bar{C}; \bar{C}^*(1), \ldots, \bar{C}^*(q))$, is designed to address this problem by breaking up a large clique into smaller and more manageable fragments. It approximates a clique $\bar{C}$ by replacing it by a set of sub-cliques $\bar{C}^*(1), \ldots, \bar{C}^*(q)$ where $\bar{C} = \{\bar{C}^*(1) \cup \ldots \cup \bar{C}^*(q)\}$ and $\bar{C}^*(i) \cap \bar{C}^*(j) \subseteq \bar{S}, \quad 1 \leq i \neq j < q$, where $\bar{S}$ is the separator of $\bar{C}$ in the original junction forest $\mathcal{T}$. TEAR$(\bar{C}; \bar{C}^*(1), \ldots, \bar{C}^*(q)$ acts as follows:

1. Define $(\mathcal{T}, \mathcal{C}, \mathcal{S}) \rightarrow (\mathcal{T}^*, \mathcal{C}^*, \mathcal{S}^*)$ where

   (a) $\mathcal{C}^* = \{\mathcal{C} \setminus \bar{C}\} \cup \{\bar{C}^*(1), \ldots, \bar{C}^*(q)\}$.

   (b) The separator $\bar{S}$ associated with $\bar{C}$ is replaced by $\bar{S}^*(1), \ldots, \bar{S}^*(q)$, where $\bar{S}^*(i) = \bar{S} \cap \bar{C}^*(i)$ become the separators of $\bar{C}^*(i), \quad 1 \leq i \leq q$ in $\mathcal{T}^*$. All other separators $S \in \mathcal{S}$ are unchanged.

(c) The graph $\mathcal{T}^*$ is obtained from $\mathcal{T}$ by removing from $\mathcal{T}$ the edge labelled $\bar{S}$ and adding the edges labelled $\bar{S}^*(1), \ldots, \bar{S}^*(q)$ from the grandparent node $C(0)$ to the nodes $\bar{C}^*(1), \ldots, \bar{C}^*(q)$.

2. Specify $(\boldsymbol{\mu}_i^*, \Sigma_i^*)$ for cliques other than $\bar{C}^*(1), \ldots, \bar{C}^*(q)$ directly from their corresponding cliques in $\mathcal{C}$. Each $(\boldsymbol{\mu}_i^*, \Sigma_i^*)$ associated with the vector of variables $\boldsymbol{X}^*(i)$ of $\bar{C}^*(i)$ can be obtained directly from the mean vector and covariance matrix of the random vector $\bar{\boldsymbol{X}}$ of variables associated with $\bar{C}$, since $\bar{C}^*(i) \subset \bar{C}, \quad 1 \leq i \leq q$.

Then, using the same argument as for the CUT operator, it is easily checked that the divergence $d_K(p(\boldsymbol{x}); \hat{p}(\boldsymbol{x}))$ can be written in terms of the variance-covariance matrix of the original clique $\bar{C}$ in $\mathcal{C}$. Explicitly the Kullback-Leibler divergence is

$$\begin{aligned} d_K(p(\boldsymbol{x}); \hat{p}(\boldsymbol{x})) &= \int_{\bar{\boldsymbol{x}} \in \bar{C}} [\log p(\bar{\boldsymbol{x}}) - \textstyle\sum_i^q \log p(\boldsymbol{x}^*(i))] p(\bar{\boldsymbol{x}}) d\bar{\boldsymbol{x}} \\ &= -\tfrac{1}{2} \log[\tfrac{det V}{\prod_{i=1}^q det V_i}] \end{aligned}$$

where $p(\bar{\boldsymbol{x}})$ denotes a Gaussian density over a random vector $\bar{\boldsymbol{X}}$ such that $\bar{C}$ represents the collection of component random variables of $\bar{\boldsymbol{X}}$ and $p(\boldsymbol{x}^*(i))$ denotes a Gaussian density over a random vector $\boldsymbol{X}^*(i)$ associated with $\bar{C}^*(i), \quad 1 \leq i \leq q$, represents the collection of component random variables of $\boldsymbol{X}^*(i)$.

Here, V is the covariance matrix of $\bar{\boldsymbol{X}}$ and V_i is the covariance matrix of $\boldsymbol{X}^*(i), \quad 1 \leq i \leq q$.

EXAMPLE

To show how to implement the TEAR operator using the Hellinger distance, consider the simple case where we approximate a clique $\bar{C}$, which consists of only two components, by replacing it by two sub-cliques $\bar{C}^*(1)$ and $\bar{C}^*(2)$, each consists of one component.

Let $\bar{X} = (X(1), X(2))$ be a vector of measurements whose components make up the clique $\bar{C}$ with a bivariate normal density $p(\bar{x})$ with mean μ and covariance matrix $V_1 = P_1^{-1}$. Let $X^*(i) = X(i), (i = 1, 2)$ are independent sub-vectors(variables) of measurements whose components make up the sub-cliques $\bar{C}^*(i)$. Under the approximation, the density of $\bar{X}$ will be a bivariate normal $\hat{p}(\bar{x})$ with the same mean μ and variance-covariance matrix of the form

$$V2 = P_2^{-1} = \begin{pmatrix} VarX(1) & 0 \\ 0 & VarX(2) \end{pmatrix}$$

For simplicity, and because of the scale invariance of $d_H(p; \hat{p})$ we can, without loss of generality, assume that if $X(1), X(2)$ have the same variances then these can be set to unity.

Suppose that

$$V_1 = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \Rightarrow P_1 = (1-\rho^2)^{-1} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix},$$

$$V_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \Rightarrow P_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Then

$$P_1 + P_2 = \begin{pmatrix} 2-\rho^2 & -\rho \\ -\rho & 2-\rho^2 \end{pmatrix} (1-\rho^2)^{-1}$$

This gives

$$\begin{aligned} |P_1| &= (1-\rho^2)^{-1} \\ |P_2| &= 1 \end{aligned}$$

$$
\begin{aligned}
|P_1 + P_2| &= \frac{(2-\rho^2)^2 - \rho^2}{(1-\rho^2)^2} \\
&= \left(\frac{4-\rho^2}{1-\rho^2}\right)
\end{aligned}
$$

From Chapter (5), we have

$$I(p;\hat{p}) = \int p^{1/2}\hat{p}^{1/2} = \left[\frac{|P_1||P_2|}{(1/4)^2|P_1+P_2|^2}\right]^{1/4}$$

Thus

$$\frac{|P_1||P_2|}{(1/4)^2|P_1+P_2|^2} = \frac{(1-\rho^2)}{(1/4)^2(4-\rho^2)^2}$$

So

$$I(p;\hat{p}) = \left[\frac{(1-\rho^2)}{(1/4)^2(4-\rho^2)^2}\right]^{1/4}$$

and

$$d_H^2(p;\hat{p}) = 1 - I(p;\hat{p})$$

Clearly, whenever the correlation ρ between two new cliques is close to zero, so is $d_H^2(p;\hat{p})$ and the original clique of two variables can then be torn apart.

In general we approximate a normal density f of a vector $\boldsymbol{X} = (\boldsymbol{X}(1), \ldots, \boldsymbol{X}(p))$ where $\boldsymbol{X}(j)$ has a mean vector $\boldsymbol{\mu}(j), 1 \leq j \leq p$ and covariance matrix $\Sigma = \{\Sigma(j,k)\} = Cov(\boldsymbol{X}(j), \boldsymbol{X}(k)), 1 \leq j, k, \leq p$ by the normal density $\hat{f}$ of a random vector with the same $\boldsymbol{\mu}(j)$ and $\hat{\Sigma} = \{\Sigma(j,j)\}, 1 \leq j \leq p$ but with all covariances block $\Sigma(j,k), j \neq k$ set to zero.

Let $D(j,k) = |\Sigma(j,k)|$, $1 \leq j, k \leq p$, D denote the $p \times p$ matrix $\{D(j,k)\}$ and $D^* = \{D^*(j,k)\}$ where

$$
\begin{aligned}
D^*(j,j) &= D(j,j) \\
D^*(j,k) &= \frac{1}{2}D(j,k)
\end{aligned}
$$

Then from Subsection (5.6) we have

$$R(f;\hat{f}) = 2\log|D^{*}| - \log|D| - \sum_{j=1}^{p}\log D(j,j)$$

and

$$d_H^2(f;\hat{f}) = 1 - \exp\{-\frac{1}{4}R(f;\hat{f})\}$$

whenever this quantity is small, we tear the original clique of the p random vectors

7.4 The Exact Absorption of Information on Cliques of Cousins (The JOIN Operator)

When an observation is taken which is informative about variables in more than one clique, the conditional independences implicit in the clique structure before the data arrived are no longer necessarily valid after the data has been observed. This will lead to the change of the structure of the junction trees, and the system needs to be adapted so that the ADJUST operator defined above is still valid.

In this section we show how to adapt the system using the JOIN operator for an exact absorption of information on cliques of "cousins" and illustrate the corresponding change of the clique structure and junction trees with a simple example.

Let $C(1),\ldots,C(q)$ represent a subset of the cliques containing "cousins" which are descendants of the same grandparent $C(0)$ in a certain junction forest $\mathcal{C}$.

The JOIN operator JOIN$(C(1),\ldots,C(q))$ can be used to obtain a new junction forest by combining q cliques $C(1),\ldots,C(q)$ into a merged clique $\bar{C}$ in order to retain valid dependences (e.g. after data observation) and revising the forest accordingly. The JOIN operator acts as follows:

1. Set $(\mathcal{T}, \mathcal{C}, \mathcal{S}) \longrightarrow (\mathcal{T}^*, \mathcal{C}^*, \mathcal{S}^*)$ where

   a) The set of cliques $C(1), \ldots, C(q)$ in $\mathcal{C}$ are replaced by the single clique $\bar{C} = \{C(1) \cup, \ldots, \cup C(q)\}$.

   Explicitly $\mathcal{C}^* = \{\mathcal{C} \setminus \{C(1), \ldots, C(q)\}\} \cup \bar{C}$

   b) All the separators in $\mathcal{S}^*$ are the same as in $\mathcal{S}$ except that the separator of $\bar{C}$ from $C(0)$ namely $\bar{S} = \{S(1) \cup, \ldots, \cup S(q)\}$, which has a q-dimensional vector of means and $q \times q$ variance-covariance matrix, will replace the individual separators $S(i), (1 \leq i \leq q)$.

   Explicitly $\mathcal{S}^* = \{\mathcal{S} \setminus \{S(1), \ldots, S(q)\}\} \cup \bar{S}$.

   c) The graph $\mathcal{T}^*$ is obtained from $\mathcal{T}$ by removing from $\mathcal{T}$ the edges labelled $S(1), \ldots, S(q)$ and then adding an edge labelled $\bar{C}$ from the node $C(0)$ to a new node labelled $\bar{C}$.

   d) The grandparent clique $C(0)$ will remain as it was.

   Notice that the separator is now a q-dimensional one, but the analogous updates of the dynamic tree algorithm "ADJUST operator" are valid for updating $C(0)$.

2. Given an observation Y, it is now necessary only to specify the pair $(\boldsymbol{\mu}_{\bar{C}}, \Sigma_{\bar{C}})$, the posterior mean vector and the covariance matrix of the variables in $\bar{C} \in \mathcal{C}^*$, since for those cliques in $\mathcal{C} \cap \mathcal{C}^*$ we demand that they have the same means and covariance matrices in $\mathcal{C}^*$ as in $\mathcal{C}$

In our application each clique $C(i)$, $(i = 1, \ldots, q)$ consists of 6 components (a parent

$Q(i)$ and five children) so that

$$C(i) = \boldsymbol{Q}(i) = (Q(i), Q_{i1}, \ldots, Q_{i5}), \quad (i = 1, 2, \ldots, q).$$

Now using the expression for the siblings vector of the ith puff (see Chapter 3)

$$(Q_{i1}, \ldots, Q_{i5}) = \boldsymbol{\alpha} Q(i) + \boldsymbol{\omega}(i)$$

where $\boldsymbol{\alpha}^T = (\alpha_1, \ldots, \alpha_5)$ represents the proportions of the distribution of the mass of a puff $Q(i)$ to its children.

i.e.

$$Q_{ij} = \alpha_j Q(i) + \omega_{ij} \quad (i = 1, \ldots, q; \;\; j = 1, \ldots, 5).$$

So that

$$\begin{aligned} VarQ_{ij} &= \alpha_j^2 VarQ(i) + Var\omega_{ij} \\ Cov(Q(i), Q_{ij}) &= \alpha_j VarQ(i) + \text{zero term.} \end{aligned}$$

Then the posterior covariance matrix of $\bar{C}$ can be written in terms of the separators. Explicitly the posterior distribution of $\bar{C}$ is normal with mean $\boldsymbol{\mu}_{\bar{C}}$ and covariance matrix

$$\Sigma_{\bar{C}} = \{\Sigma_{mn}\}, \quad (1 \leq m, n \leq q),$$

where Σ_{mm} is the variance-covariance matrix of variables in $C(m)$ and

$$\Sigma_{mn} = \begin{pmatrix} Cov(Q(m), Q(n)) & \boldsymbol{\phi}^T \\ \boldsymbol{\phi} & B \end{pmatrix}, \qquad m \neq n$$

where

$$\begin{aligned} \boldsymbol{\phi}^T &= \boldsymbol{\alpha}^T Cov(Q(m), Q(n)) &&= \boldsymbol{\alpha}^T Cov(S(m), S(n)) \\ B &= \boldsymbol{\alpha}\boldsymbol{\alpha}^T Cov(Q(m), Q(n)) &&= \boldsymbol{\alpha}\boldsymbol{\alpha}^T Cov(S(m), S(n)) \end{aligned}$$

EXAMPLE

Joining two cliques:

As a simple illustration, consider using the JOIN operator with $q = 2$ where a puff $C(0)$ (grandparent) pentificates to 5 children, 2 of them (aunts) pentificate to 5 children each to give cliques $C(1) = \boldsymbol{Q}(1) = (Q(1), Q_{11}, \ldots, Q_{15})$ and $C(2) = \boldsymbol{Q}(2) = (Q(2), Q_{21}, \ldots, Q_{25})$ (see Figure 7.1)

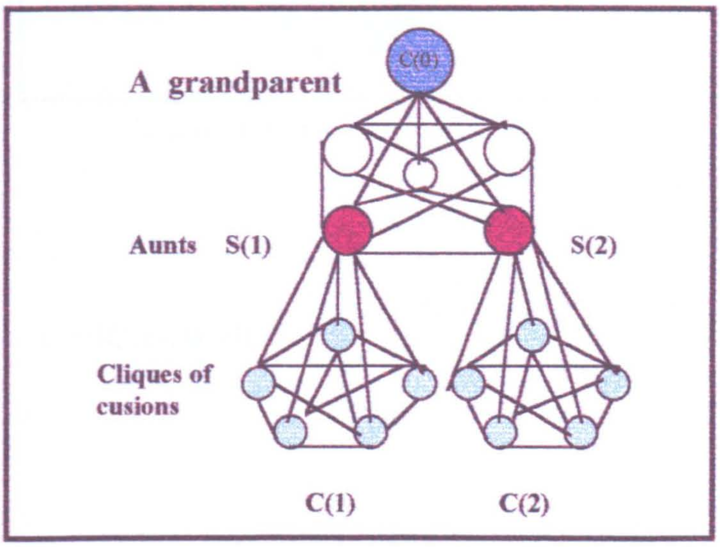


Figure 7.1: Tip of a decomposable graph

When an observation is taken under the two cliques $C(1)$ and $C(2)$, we can obtain a new clique by joining them thus we obtain a clique tree tip as in Figure 7.2

The covariance between the 6-dimensional vector $\boldsymbol{Q}(1)$ and the 6-dimensional vector

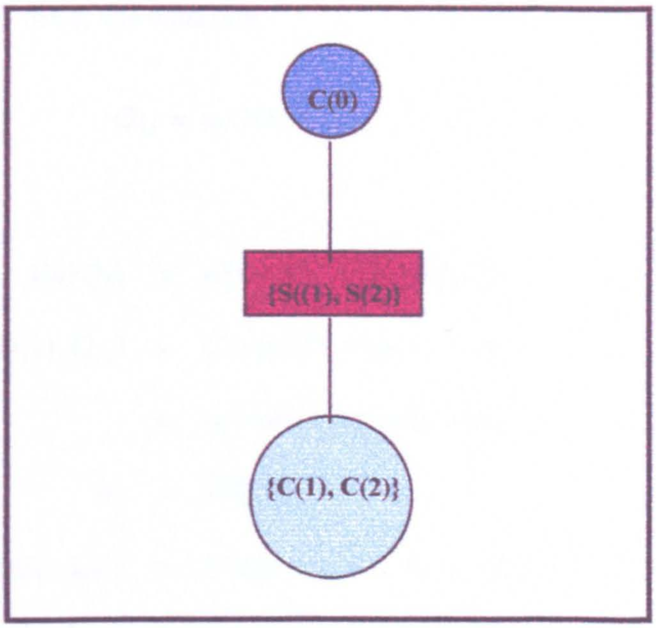


Figure 7.2: Tip of a clique tree

$\boldsymbol{Q}(2)$ is

$$\Sigma_{\bar{C}} = Cov(\boldsymbol{Q}(1), \boldsymbol{Q}(2)) = \begin{pmatrix} Var\boldsymbol{Q}(1) & Cov(\boldsymbol{Q}(1), \boldsymbol{Q}(2)) \\ & Var\boldsymbol{Q}(2) \end{pmatrix}$$

$$= \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ & \Sigma_{22} \end{pmatrix}$$

where

$$\Sigma_{11} = \begin{pmatrix} VarQ(1) & Cov(Q(1), Q_{11}) & \cdots & Cov(Q(1), Q_{15}) \\ & VarQ_{11} & & \\ \vdots & \vdots & \ddots & \vdots \\ & & & VarQ_{15} \end{pmatrix}$$

Note that all the elements in $\Sigma_{11}(\Sigma_{22})$ can be written in terms of the variance of the parent $Q(1)(Q(2))$ using the equation

$$Q_{1j} = \alpha_j Q(1) + \omega_{1j} \quad j = 1, \ldots, 5$$

where

$$\begin{aligned}
VarQ_{1j} &= \alpha_j^2 VarQ(1) + Var(\omega_{1j}) \\
Cov(Q(1), Q_{1j}) &= Cov(Q(1), \alpha_j Q(1) + \omega_{1j}) \\
&= \alpha_j VarQ(1) + \text{zero term} \\
\alpha_j &= \frac{Cov(Q(1), Q_{1j})}{VarQ(1)} \\
Cov(Q_{11}, Q_{1j}) &= Cov(\alpha_1 Q(1) + \omega_{11}, \alpha_j Q(1) + \omega_{1j}) \\
&= \alpha_1 \alpha_j VarQ(1) + \text{zero terms} \quad j = 2, \ldots, 5 \\
\vdots \\
Cov(Q_{14}, Q_{1j}) &= \alpha_4 \alpha_j VarQ(1) + \text{ zero terms} \quad j = 5
\end{aligned}$$

Similarly we can write the elements of Σ_{22} in terms of the variance of the parent $Q(2)$ using the equation

$$Q_{2j} = \alpha_j Q(2) + \omega_{2j}, \quad j = 1, \ldots, 5$$

Now the elements of the block matrix

$$\Sigma_{12} = \left(\begin{array}{c|ccc}
Cov(Q(1), Q(2)) & Cov(Q(1), Q_{21}) & \cdots & Cov(Q(1), Q_{25}) \\
\hline
Cov(Q_{11}, Q(2)) & Cov(Q_{11}, Q_{21}) & \cdots & \cdots \\
\vdots & \cdots & \ddots & \cdots \\
Cov(Q_{15}, Q(2)) & \cdots & \cdots & Cov(Q_{15}, Q_{25})
\end{array} \right)$$

$$= \begin{pmatrix} Cov(S(1), S(2)) & \phi^T \\ \phi & B \end{pmatrix}$$

can be written in terms of the covariance of the separators $S(1), S(2)$ where $Cov(Q(1), Q(2)) = Cov(S(1), S(2))$.

Thus

$$\begin{aligned}\boldsymbol{\phi}^T &= (Cov(Q(1), Q_{21}), \ldots, Cov(Q(1), Q_{25})) \\ &= (Cov(Q(1), \alpha_1 Q(2) + \omega_{21}), \ldots, Cov(Q(1), \alpha_5 Q(2) + \omega_{25})) \\ &= (\alpha_1 Cov(Q(1), Q(2)), \ldots, \alpha_5 Cov(Q(1), Q(2))) \\ &= \boldsymbol{\alpha}^T Cov(Q(1), Q(2))\end{aligned}$$

and

$$B = \boldsymbol{\alpha} Cov(S(1), S(2)) \boldsymbol{\alpha}^T$$

If $\boldsymbol{\alpha} = (0.235, 0.235, 0.235, 0.235, 0.058)$ and the system errors ω_{ij} are very small, then the posterior covariance matrix $\Sigma_{\tilde{C}}$ for the variables of the merged clique $\tilde{C}$ can be written in terms of the covariances of the separators $Cov(S(i), S(i)) = Cov(Q(i), Q(i))$, $(i = 1, 2)$ as

$$\Sigma_{\tilde{C}} = \begin{pmatrix} GVarQ(1) & GCov(Q(1), Q(2)) \\ & GVarQ(2) \end{pmatrix}$$

where

$$G = \begin{pmatrix} 1 & 0.235 & 0.235 & 0.235 & 0.235 & 0.058 \\ & (0.235)^2 & (0.235)^2 & (0.235)^2 & (0.235)^2 & 0.014 \\ & & (0.235)^2 & (0.235)^2 & (0.235)^2 & 0.014 \\ & & & (0.235)^2 & (0.235)^2 & 0.014 \\ & & & & (0.235)^2 & 0.014 \\ & & & & & (0.058)^2 \end{pmatrix}$$

7.5 The OBLINK Operator

In this section we discuss the exact absorption of information into disconnected cliques (i.e when observations are taken under cliques which belong to disconnected sub-trees) by suggesting the OBLINK operator.

The operator OBLINK $(B_Y; C(1),\ldots,C(q))$ introduces a dummy clique B_Y consisting of a set of variables, not all contained in one clique, on which a data vector depends. It has as its argument the set B_Y and a minimal set of cliques $C(1),\ldots,C(q)$ containing variables in B_Y.

OBLINK $(B_Y; C(1),\ldots,C(q))$ acts as follows :

1. It acts on the junction graph $\mathcal{T}$ by replacing $\mathcal{T}$ by a graph $\mathcal{T}^*$ and on the cliques set $\mathcal{C}$ by replacing it by $\mathcal{C}^*$ where $\mathcal{C}^* = \mathcal{C} \cup B_Y$. The graph $\mathcal{T}^*$ joins B_Y to a clique $C \in \mathcal{C}$ by an edge iff C is one of $(C(1),\ldots,C(q))$. These new edges are labelled by separators $S(i) = C(i) \cap B_Y$, $1 \leq i \leq q$.

2. It acts on means and covariances on cliques by :

   (i) transferring means and covariances of $C \in \{\mathcal{C} \cap \mathcal{C}^*\} \rightarrow \mathcal{C}^*$;

   (ii) calculating means in B_Y from $S(i) \subseteq C(i)$, $C(i) \in \mathcal{C}$ and obtaining covariances in B_Y either :

   (a) directly from $C(i)$ (if two variables X_1, X_2 are in $C(i) \cap B_Y$) .

   (b) by setting the covariance terms to zero if the cliques $C(1),\ldots,C(q)$ are disconnected in $\mathcal{T}$ or equivalently if $\mathcal{T}^*$ is a forest.

Notice that, as defined above, $\mathcal{T}^*$ is not a forest in general; so to employ the ADJUST operator of Section (7.2), $\mathcal{T}^*$ must be approximated by a forest by successive use of the "Cut" and "Tear" operations.

As a simple illustration, suppose that a tip of a clique tree has been cut in several places as shown in Figure 7.3

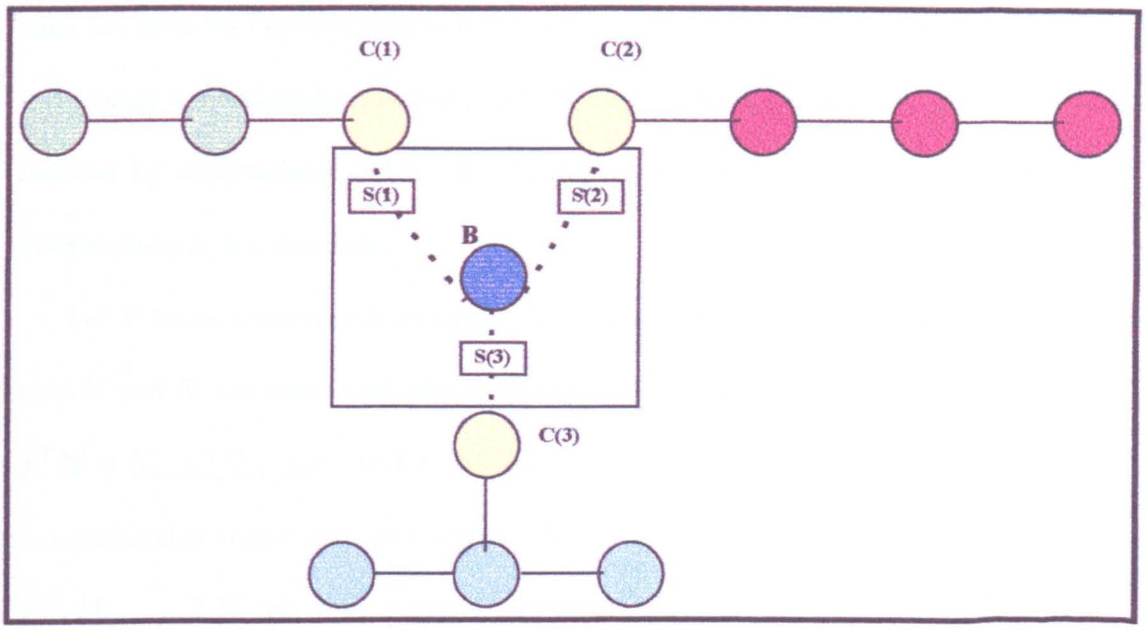


Figure 7.3: An illustration of the OBLINK operator

Let Y be an observation taken under the cliques $C(1), C(2), C(3)$. Denote a new clique B to be the collection of all masses or variables on which the observation Y depends. Assume these variables are $S(1), S(2), S(3)$ where $S(i)$ are variables in clique $C(i), 1 \leq i \leq 3$. Clearly the separator of $C(i)$ from B is just $S(i)$.

The main point here is to note that if $C(1), \ldots, C(q)$ in the clique tree(s) of the problem are originally disconnected, then there is an *exact update* on the new tree created by connecting B to $C(1), \ldots, C(q)$ using separators $S(1), \ldots, S(q)$ to label newly created edges (These are depicted by dotted lines in Figure 7.3). In this update we use B as the root clique ordering all edges away from it.

7.6 The DivObs Operator

As explained above, if observations are taken under descendants of different sources, then the induced dependencies will be complex. Here we discuss an alternative quick but coarser approximation method which effectively ignores a posteriori, the covariances induced by observations under different cliques, if the K-L divergence between joint distributions is not too large.

Let Y be an observation which is taken under the cliques $C(1),\ldots,C(q)$. Suppose that Y and $\boldsymbol{X}$ are related via the conditional distribution $Y|\boldsymbol{X} \sim N[\mu, V]$ where $\mu = \boldsymbol{f}^T\boldsymbol{X} = \sum_{i=1}^{q}\sum_{j=1}^{n_i} f_{ij}x_{ij}$ and V is a scalar variance, where $\boldsymbol{X} = (\boldsymbol{X}(1),\ldots,\boldsymbol{X}(q))^T$ is a parameter vector with components $\boldsymbol{X}(i) = (x_{i1},\ldots,x_{in_i})^T \in C(i), 1 \leq i \leq q$ and $\boldsymbol{f} = (\boldsymbol{f}_1,\ldots,\boldsymbol{f}_q)^T$ where $\boldsymbol{f}_i = (f_{i1}\ldots,f_{in_i})^T, 1 \leq i \leq q$.

The operator DivObs $(Y;Y_1,\ldots,Y_q)$ replaces (divides) the observation Y into an approximately equivalent vector of dummy observations $Y_1,\ldots,Y_q$, which would give approximately the same posterior distribution on the parameter vector $\boldsymbol{X}$.

Let $\boldsymbol{\mu}(i)$ and $\Sigma(i)$ represent the mean and the covariance matrix of $\boldsymbol{X}(i), 1 \leq i \leq q$.

Then set

$$Y_i = Y - \sum_{i^* \neq i}^{q} \boldsymbol{f}_{i^*}^T \boldsymbol{x}(i^*).$$

We make the approximation assumption that $Y_i|\boldsymbol{X}, 1 \leq i \leq q$ are normally distributed with means

$$\mu_i = \mu - \sum_{i^* \neq i}^{q} \boldsymbol{f}_{i^*}^T \boldsymbol{\mu}(i^*)$$

and variances

$$V_i = V + \sum_{i^* \neq i}^{q} \boldsymbol{f}_{i^*}^T \Sigma(i^*) \boldsymbol{f}_{i^*}$$

Now, assuming that $\boldsymbol{X}(i), 1 \leq i \leq q$ are mutually independent a priori, the approximation then substitutes the vector $\boldsymbol{Y} = (Y_1, \ldots, Y_q)$ of dummy observations for the original observed random variable, choosing $\boldsymbol{Y}$ so that the posterior mean vectors and covariance matrices of $\boldsymbol{X}(i)|Y$ and $\boldsymbol{X}(i)|\boldsymbol{Y}$ agree, $1 \leq i \leq q$ (but not necessarily the posterior covariances between $\boldsymbol{X}(i), 1 \leq i \leq q$).

The Kullback-Leibler divergence associated with the approximation is

$$Inf(\perp\!\!\!\perp_{i=1}^{q} \mathbf{X}(i)|Y) = \frac{1}{2}\left[\sum_{i=1}^{q} \log(det\Sigma^*(ii)) - \log det(\Sigma^*)\right]$$

(see Section B.1 of Appendix B), where

$$\begin{array}{rcccl} \Sigma^* & = & Cov(\boldsymbol{X}|Y) & = & \{\Sigma^*(ij)\} \quad 1 \leq i, j \leq q \\ \Sigma^*(ii) & = & Var(\boldsymbol{X}(i)|Y) & = & \Sigma(i) - \frac{\Sigma(i)\boldsymbol{f}_i \boldsymbol{f}_i^T \Sigma(i)}{\sigma_Y^2} \\ \Sigma^*(ij) & = & Cov(\boldsymbol{X}(i), \boldsymbol{X}(j)|Y) & = & \frac{-\Sigma(i)\boldsymbol{f}_i \boldsymbol{f}_j^T \Sigma(j)}{\sigma_Y^2} \quad i \neq j \end{array}$$

and $\sigma_Y^2 = \sum_{i=1}^{q} \boldsymbol{f}_i^T \Sigma(i) \boldsymbol{f}_i + V$.

Implementation

The DivObs operator was implemented in the RODOS computer code using the Lundtofte Nord1 data supplied within the RIMPUFF code. The K-L divergence value associated with this approximation was 0.01. Observations were collected from different sites (different detector points) in the plume at each time step between 1200 and 9600 seconds.

RIMPUFF was tested by varying two of the input parameters, namely the height and the strength of the release. The observations were taken by running RIMPUFF with a height of 0.6 km and a source strength of 1.0e+7. Many tests were carried out with variations on these input parameters. Section B.4 of Appendix B shows one of these tests.

The essential characteristic of the system that has been tested was that the DivObs operator which gives an approximation was being tested against the full matrix update (i.e we approximate a density p of a random vector (say) $\boldsymbol{X} = (\boldsymbol{X}(1), \ldots, \boldsymbol{X}(p))$ where $\boldsymbol{X}(i)$ has mean $\boldsymbol{\mu}(i), 1 \leq i \leq p$ and covariances $\Sigma(i,j) = Cov(\boldsymbol{X}(i), \boldsymbol{X}(j)), 1 \leq i,j \leq p$ by the density $\hat{p}$ of a random vector with the same mean and $\Sigma(i,i), 1 \leq i \leq p$ but with all covarince block $\Sigma(i,j), i \neq j$ are zeros).

Results

At each detector point where observations were taken, a graph of the predicted concentration over time at that particular point was produced. Each graph shows plots of the DivObs update and the full matrix update where we can see that in all the cases the DivObs operator matches the full matrix update very closely. However the approximate predictions are lower than they should be because the observations are much higher than expected a priori, and because the approximate model ignores some dependencies, it does not adjust as quickly as it should in the light of the mismatch.

7.7 Appendix B

B.1 Basic results on information divergence

Definition. Suppose that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are two random vectors with a joint density function $f_{\boldsymbol{X}_1\boldsymbol{X}_2}$ and marginal densities $f_{\boldsymbol{X}_1}$ and $f_{\boldsymbol{X}_2}$, respectively, then the information in one random vector about the other (the information proper) is defined by

$$Inf(\boldsymbol{X}_1, \boldsymbol{X}_2) = d_K(f_{\boldsymbol{X}_1\boldsymbol{X}_2}; f_{\boldsymbol{X}_1} f_{\boldsymbol{X}_2})$$

where d_K is the Kullback-Leibler information divergence.

This measures the divergence between the joint density of $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ and the "independent" density given by the product of the marginal densities of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

The divergence against independence

suppose that

$$f : (\boldsymbol{X}_1, \boldsymbol{X}_2) \sim N[\boldsymbol{\mu}, V]; \quad g : g_{12} = f_1 f_2$$

where f_i is the marginal density of $\boldsymbol{X}_i$, $\quad i = 1, 2$. Then the information against independence (information divergence) is given by

$$Inf(\boldsymbol{X}_1 \perp\!\!\!\perp \boldsymbol{X}_2) = -\frac{1}{2} \log \frac{detV}{detV_{11} detV_{22}}$$

where V is the variance covariance matrix of $\boldsymbol{X}_1 \cup \boldsymbol{X}_2$ and V_i is the variance covariance matrix of $\boldsymbol{X}_i$, $\quad i = 1, 2$.

The divergence against conditional independence

Suppose that

$$f : (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3) \sim N[\boldsymbol{\mu}, V]; \quad g : g_{23|1} = f_{2|1} f_{3|1}, g_1 = f_1$$

Then the divergence against conditional independence (conditional information measure) is given by

$$Inf(\boldsymbol{X}_2 \perp\!\!\!\perp \boldsymbol{X}_3|\boldsymbol{X}_1) = -\frac{1}{2}\log\frac{detV_{2\cup 3|1}}{detV_{22|1}detV_{33|1}}$$

where $V_{2\cup 3|1}$ is the variance-covariance matrix of $\boldsymbol{X}_2 \cup \boldsymbol{X}_3|\boldsymbol{X}_1$.

When $\boldsymbol{X}_2 = X_i$ and $\boldsymbol{X}_3 = X_j$ are one dimensional, the conditional information simplifies to

$$Inf(X_i \perp\!\!\!\perp X_j|\boldsymbol{X}_1) = -\frac{1}{2}\log(1 - corr^2(X_i, X_j|\boldsymbol{X}_1)).$$

The divergence between standard bivariate Normal distributions with differing correlation coefficients

Suppose that under $f_{X_1X_2}$ the correlation between X_1 and X_2 is ρ while under $g_{X_1X_2}$, the correlation is zero. So under $g_{X_1X_2}, X_1 \perp\!\!\!\perp X_2$ and consequently $g_{X_1X_2}$ is the product of its marginals: $g_{X_1X_2} = f_{X_1}f_{X_2}$. Then

$$Inf(f_{X_1X_2}; g_{X_1X_2}) = I(\rho; 0) = -\frac{1}{2}\log(1 - \rho^2)$$

For proofs of these results see Whittaker (1990).

B.2 Tables of detector points and comparison of concentraton forecasts

Table 7.1: **Positions of detector points on a grid**

.	1	8	16	.	.	.	.	.	.	.	.
.	2	9	17	.	.	.	.	.	.	.	.
.	3	10	18	27	41	.	.	.	.	.	.
.	4	11	19	28	42	.	.	.	.	.	.
.	5	12	20	29	43	55	69	.	.	.	.
.	6	13	21	30	44	56	70	83	96	109	123
.	7	14	22	31	45	57	71	84	97	110	124
.	.	15	23	32	46	58	72	85	98	111	125
.	.	.	24	33	47	59	73	86	99	112	126
.	.	.	25	34	48	60	74	87	100	113	127
.	.	.	26	35	49	61	75	88	101	114	128
.	.	.	.	36	50	62	76	89	102	115	129
.	.	.	.	37	51	63	77	90	103	116	130
.	.	.	.	38	52	64	78	91	104	117	131
.	.	.	.	39	53	65	79	92	105	118	132
.	.	.	.	40	54	66	80	93	106	119	133
.	.	.	.	.	.	67	81	94	107	120	134
.	.	.	.	.	.	68	82	95	108	121	135

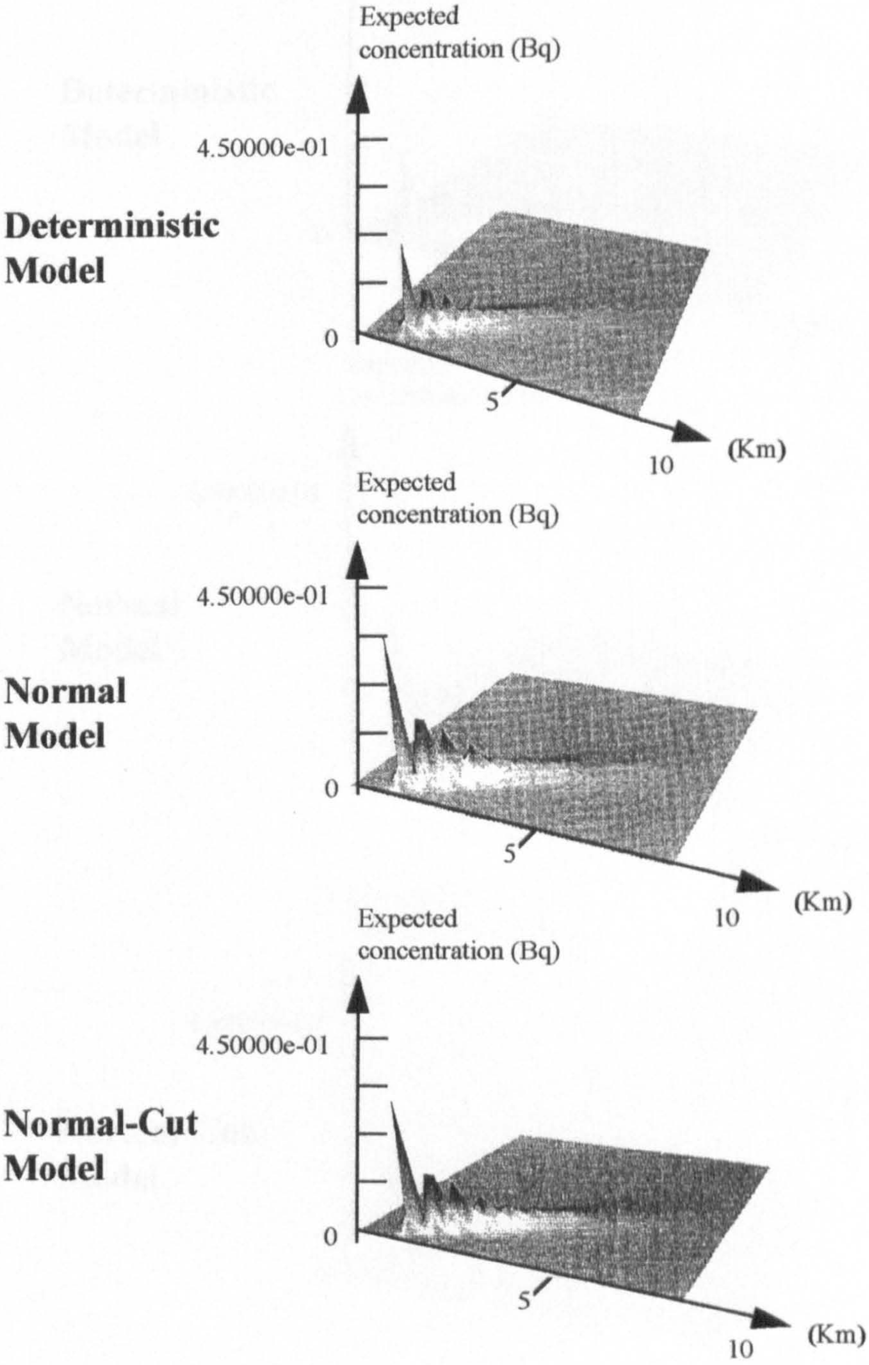
A dot indicates no detection point.

Table 7.2: **Comparisons of concentration forecasts using two models**

Point	t = 7200 secs		t = 8400 secs		t = 9600 secs	
	Normal	Normal-cut	Normal	Normal-cut	Normal	Normal-cut
34	8.319278e-02	8.319239e-02	8.308697e-02	8.308696e-02	8.321272e-02	8.321271e-02
35	1.889749e-02	1.889613e-02	1.887943e-02	1.887943e-02	1.890700e-02	1.890699e-02
36	1.756954e-03	1.736629e-03	1.849823e-03	1.849528e-03	1.833936e-03	1.833876e-03
37	2.545246e-04	2.496480e-04	2.684541e-04	2.666556e-04	2.650214e-04	2.631713e-04
38	2.068095e-05	1.200508e-05	1.867422e-05	1.061398e-05	4.596202e-05	4.493275e-05
39	1.762685e-05	1.723211e-05	7.161732e-06	4.070558e-06	7.931329e-06	4.604057e-06
48	3.592030e-01	3.590053e-01	3.596814e-01	3.596786e-01	3.600480e-01	3.600474e-01
49	9.929284e-02	9.890854e-02	1.009579e-01	1.009415e-01	1.007603e-01	1.007498e-01
50	8.166903e-02	8.033992e-02	8.849554e-02	8.822625e-02	8.726874e-02	8.711474e-02
51	2.008350e-02	1.968763e-02	2.026915e-02	1.983933e-02	2.008777e-02	1.963160e-02
52	7.260875e-03	7.098353e-03	2.974480e-03	1.707038e-03	3.289420e-03	1.925197e-03
54	8.651568e-05	8.457825e-05	3.704380e-05	2.195070e-05	3.892836e-05	2.259752e-05
60	2.920820e-01	2.911242e-01	2.962689e-01	2.962525e-01	2.958360e-01	2.958359e-01
61	1.498958e-01	1.480570e-01	1.569295e-01	1.565033e-01	1.556118e-01	1.551866e-01
62	1.029113e-01	1.008580e-01	9.176869e-02	8.640490e-02	9.193601e-02	8.619222e-02
63	7.301547e-02	7.142982e-02	4.364576e-02	4.364576e-02	4.563762e-02	4.563013e-02
65	3.558554e-02	3.478864e-02	1.646579e-02	1.030907e-02	1.689781e-02	9.459005e-03
66	1.143459e-03	1.117854e-03	4.834370e-03	4.816080e-03	2.436222e-03	6.569522e-04
74	3.674485e-02	3.656198e-02	3.748341e-02	3.748341e-02	3.738343e-02	3.737574e-02
75	1.054851e-01	1.035432e-01	1.098397e-01	1.085846e-01	1.086536e-01	1.073369e-01
77	2.302454e-01	2.253201e-01	1.627306e-01	1.626334e-01	1.377701e-01	1.375596e-01
86	3.834858e-04	3.834228e-04	3.832872e-04	3.832863e-04	3.838144e-04	3.838142e-04
88	5.834661e-02	5.721184e-02	6.195809e-02	6.150037e-02	6.116239e-02	6.069001e-02
89	6.383308e-02	6.242843e-02	3.394453e-02	3.334260e-02	3.549034e-02	3.537487e-02
90	1.496550e-01	1.463213e-01	7.817259e-02	7.374822e-02	7.764670e-02	7.653873e-02
91	1.425462e-01	1.393567e-01	9.533554e-02	9.193875e-02	8.121030e-02	8.089660e-02
93	9.135805e-03	8.931479e-03	6.675320e-02	6.576584e-02	5.265130e-02	4.968873e-02
94	6.774712e-05	6.623195e-05	1.057525e-01	1.015328e-01	9.526104e-02	9.128656e-02
100	3.012261e-05	2.956600e-05	3.074256e-05	3.073506e-05	3.031339e-05	3.031186e-05
105	3.764815e-02	3.680614e-02	1.188004e-01	1.157924e-01	7.986379e-02	7.670377e-02
106	3.367801e-03	3.292480e-03	9.553026e-02	9.232928e-02	8.107294e-02	7.967312e-02
107	2.528139e-05	2.471597e-05	4.178783e-02	4.012068e-02	9.005766e-02	8.900801e-02
114	7.392442e-05	7.227122e-05	3.000512e-05	2.705996e-05	3.323003e-05	2.929516e-05
116	8.855480e-03	8.657440e-03	3.067673e-03	2.536665e-03	4.189586e-03	3.350490e-03
117	3.296382e-03	3.222661e-03	2.826642e-02	2.774505e-02	3.338833e-03	3.319885e-03
118	1.184547e-03	1.158054e-03	4.105739e-02	3.963541e-02	3.381222e-02	3.030637e-02

B.3 A comparison of expected concentrations in Becquerels (Bq) at time intervals after the first emission in cross 2-D spatial grid (a horizontal scale of 0 - 10 Kilometers)

Time: 7200 seconds

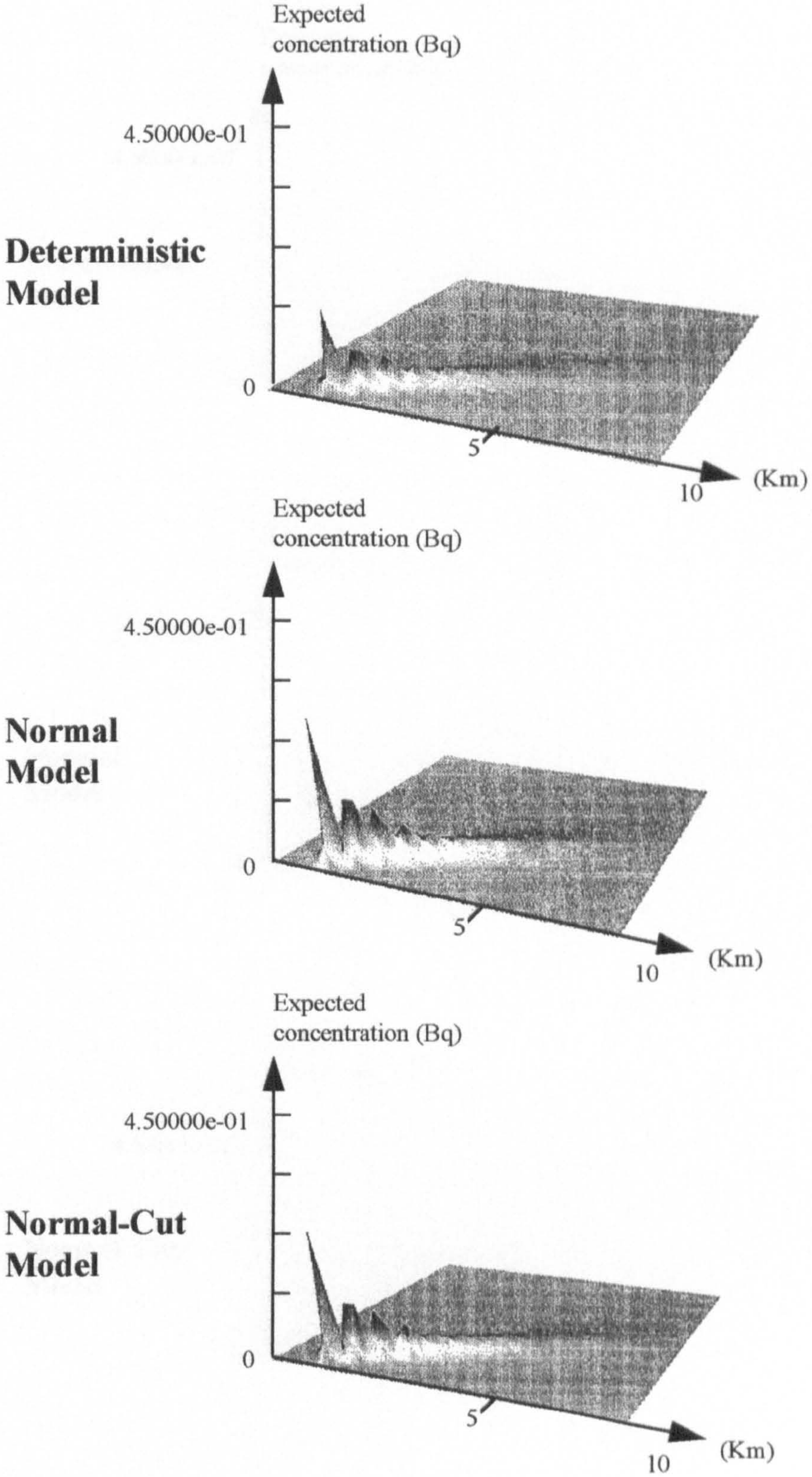


Deterministic Model

Normal Model

Normal-Cut Model

Time: 8400 seconds

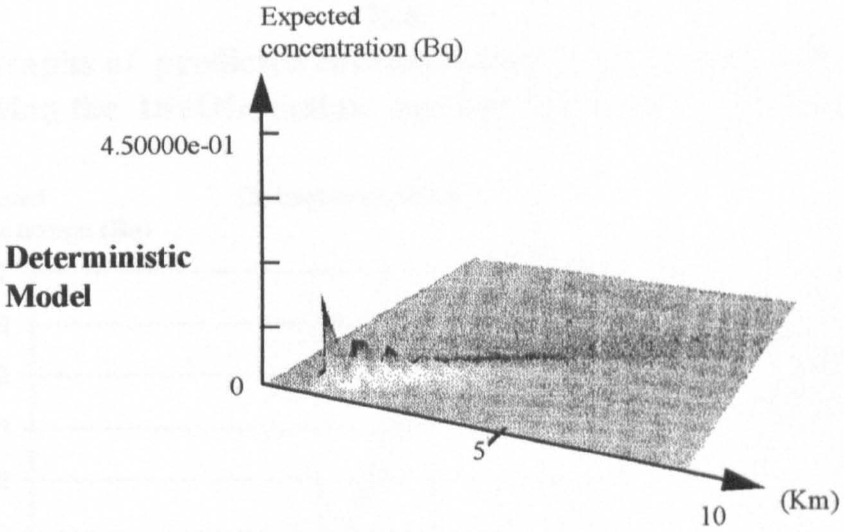


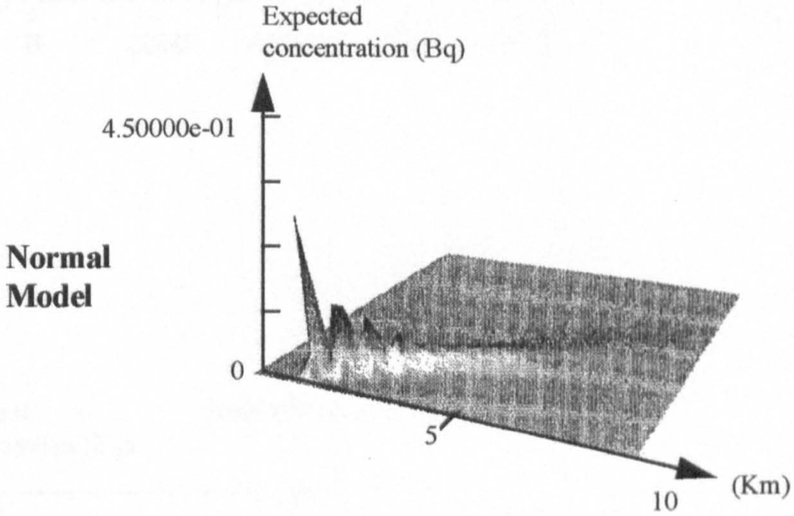
Deterministic Model

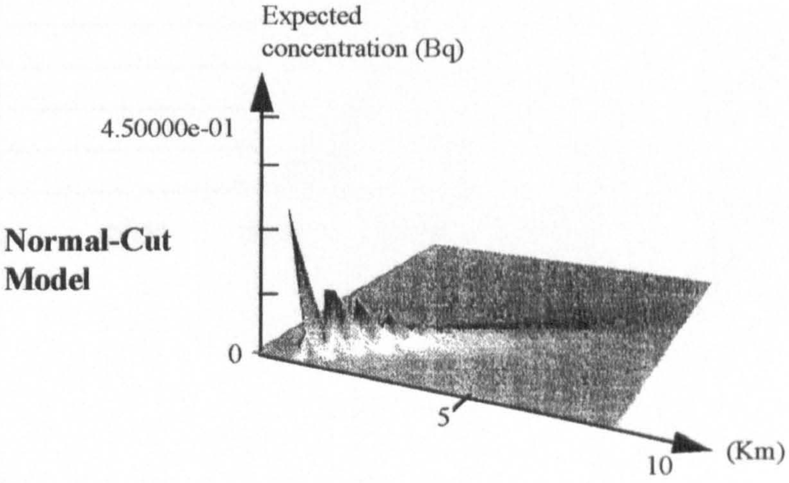
Normal Model

Normal-Cut Model

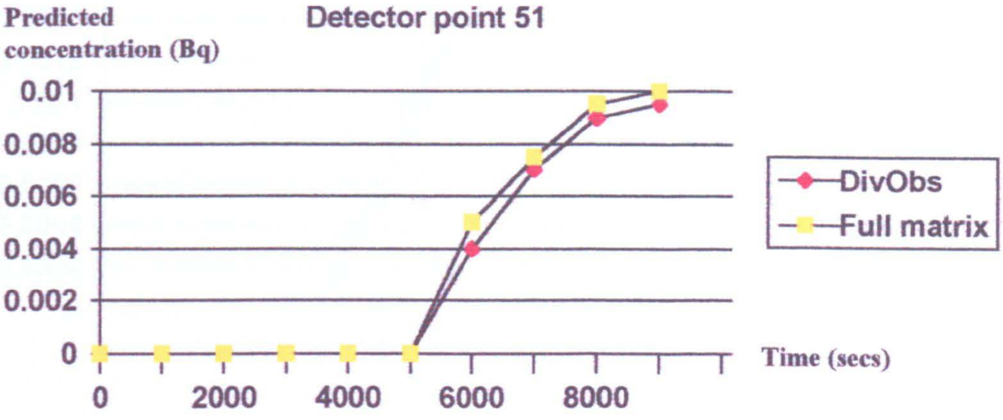
Time: 9600 seconds

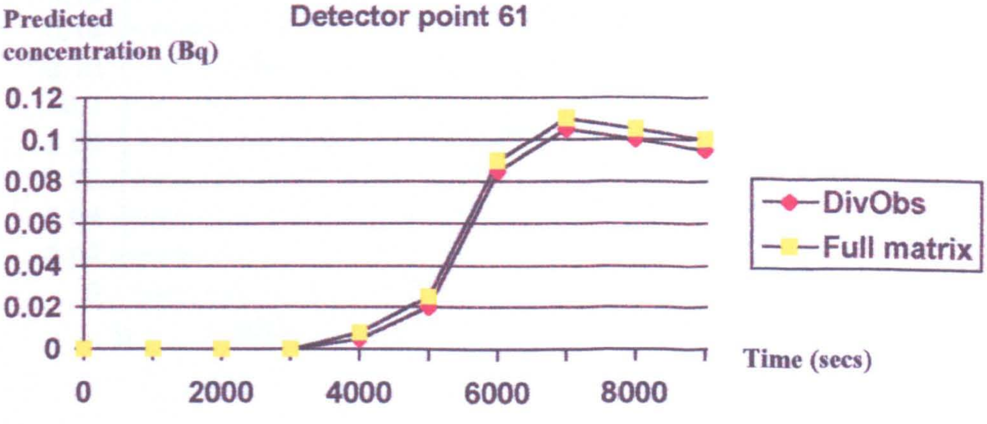


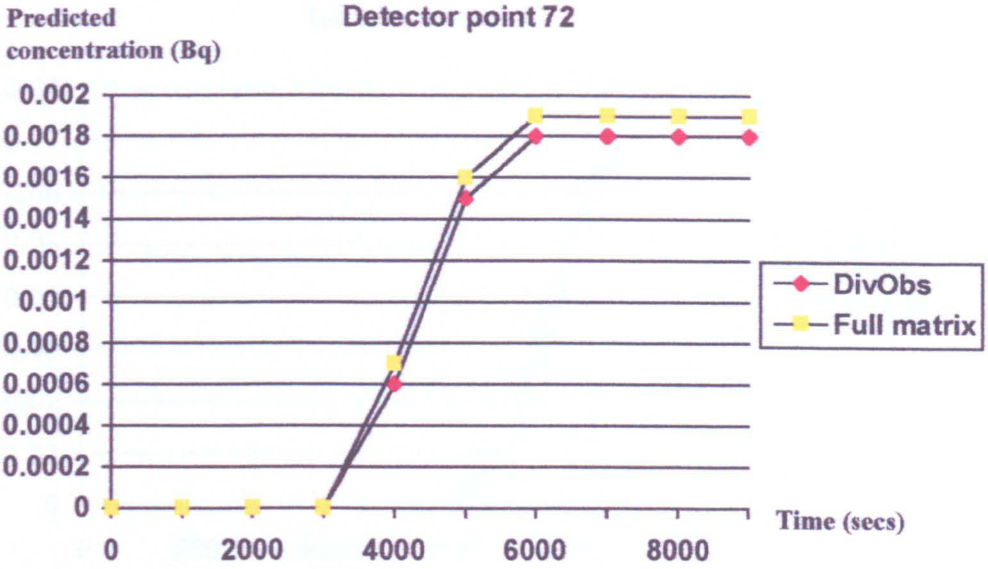


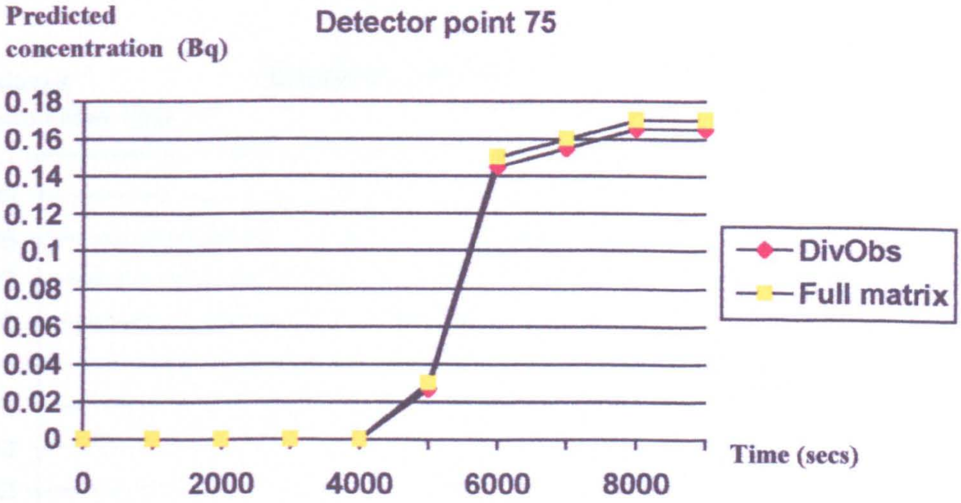


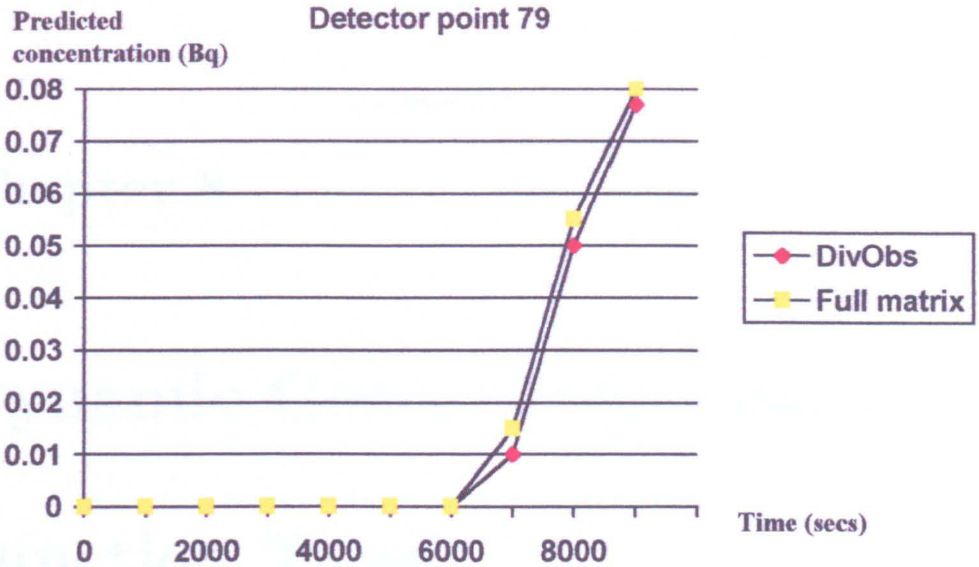
B.4
Graphs of predicted concentration in Becquerels (Bq) using the DivObs update and the Full matrix update



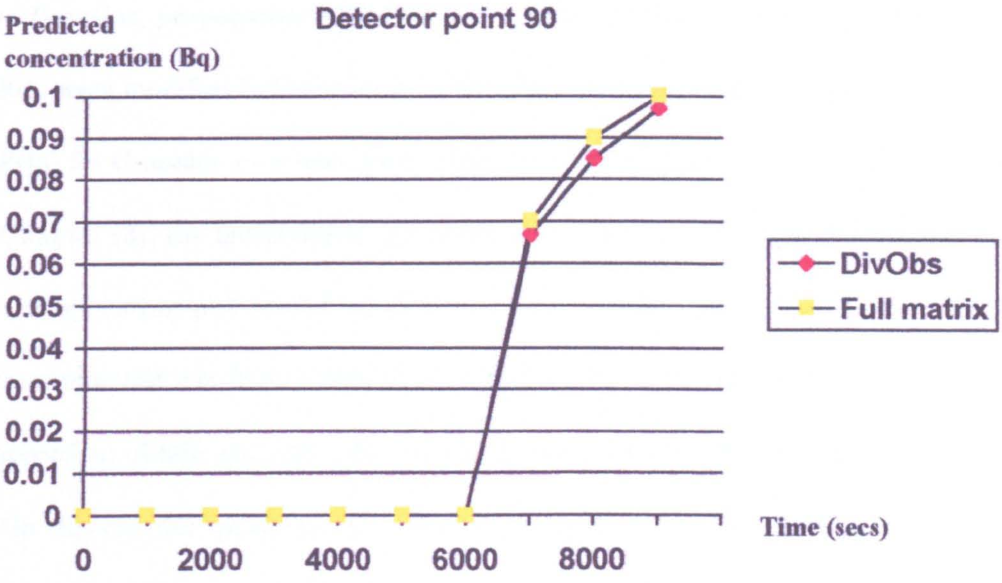



Predicted concentration (Bq)
Detector point 72
0.002
0.0018
0.0016
0.0014
0.0012
0.001
0.0008
0.0006
0.0004
0.0002
0
0
2000
4000
6000
8000
Time (secs)
DivObs
Full matrix


Predicted concentration (Bq)
Detector point 75
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
2000
4000
6000
8000
Time (secs)
DivObs
Full matrix



Predicted concentration (Bq)
Detector point 79
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
2000
4000
6000
8000
Time (secs)
DivObs
Full matrix



Predicted concentration (Bq)
Detector point 90
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
2000
4000
6000
8000
Time (secs)
DivObs
Full matrix

Chapter 8

Dynamic Generalized Linear Junction Trees

8.1 Introduction

The Bayesian propagation algorithms on fixed junction trees as described in Dawid (1992) were modified in Smith et al. (1995) and more generally in Gargoum and Smith (1994a) for Gaussian processes where the junction trees evolved with time. As described in Chapter (3), the methodology was defined and illustrated within a stochastic version of a fragmenting puff model which was used to predict the diffusion and dispersal of a contaminated gas from a source. In Chapter (7), new approximation schemes were proposed to obtain systems with smaller clique sizes and disconnected junction trees.

In this chapter we address the issue of the non-Gaussianity of the process. Now the propagation algorithm of Dawid (1992) is quite general and does not require processes to be Gaussian. The problem is then however to find quick methods for calculating the marginal distributions needed in the propagation. This is of course always possible using intensive numerical techniques, for example, the Gibbs sampler (see Geman and Geman, 1984). However this may take a long time. An alternative approximation methodology is to accommodate non-Gaussian distributions on observations using a slight generalisation of the class of Dynamic Generalised Linear Models (DGLMs) of West et al. (1985) and Smith (1992). The advantages of this method are:

1. It is very quick and updating is achieved in a closed form.

2. Approximate dynamic models can easily be specified and then to interact just as in their original time series setting (see West et al., 1985).

In Section (2) we provide general background information and basic results. Section (3) describes dynamic generalised linear models on junction trees. In Section (4) we discuss the closeness of dynamic approximation using the Hellinger metric. An example is provided in Section (5) and a conclusion follows in Section (6).

8.2 Background

As is now discussed in a several papers (see Lauritzen & Speigelhalter 1988; Speigelhalter et al., 1993 and Dawid; 1992), if a joint density can be written in the form (8.1) given below, then quick and efficient methods of probability propagation in probabilistic expert systems or high dimensional Bayesian statistical models are possible. As described in Chapter (4), a density $p(\boldsymbol{x}) > 0$ is said to be decomposable if it can be written in the algebraic form

$$p(\boldsymbol{x}) = \frac{\prod_{i=1}^{n} p_i(\boldsymbol{x}(i))}{\prod_{i=2}^{n} q_i(\boldsymbol{x}^s(i))} \tag{8.1}$$

Here $\boldsymbol{x}(i)$ is a vector of measurements whose components make up a clique $C(i)$ (say), $\boldsymbol{x}^s(i)$ is a subvector of both $\boldsymbol{x}(i)$ and some other vector of measurements $\boldsymbol{x}(j)$, and its components make up a separator $S(i)$ between cliques $C(i)$ and $C(j)$. Here p_i just denotes the density of the random vector $\boldsymbol{X}(i)$ and q_i the density of the random vector $\boldsymbol{X}^s(i)$.

Fast propagation algorithms are available for problems with decomposable densities where the cardinality of any clique is small. These algorithms are usually based on junction trees.

All Markov processes have junction trees whose nodes all lie on a single line. But a wide range of other more complicated dependence structures can be represented in terms of a junction tree (see, for example, Lauritzen & Speigelhalter, 1988; Goldstein, 1993; and Smith et al., 1995)

The type of learning we consider in this chapter is as follows.

Suppose we observe $Y_1, Y_2, \ldots, Y_T$ and assume:

(i) Given $\boldsymbol{X}, Y_1, \ldots, Y_T$ are all independent of each other i.e. $\perp\!\!\!\perp_{t=1}^{T} Y_t | \boldsymbol{X}$.

(ii) The density or mass function of $Y_t | \boldsymbol{X}$ for $1 \leq t \leq T$ is an explicit function of $\boldsymbol{X}$ only through the function $\eta_t(\boldsymbol{X})$, and $\eta_t(\boldsymbol{X})$ is itself a function only of variables lying in a single clique $C(t) = C(i_t)$ (say), i.e., $Y_t \perp\!\!\!\perp \boldsymbol{X} | \eta_t(\boldsymbol{X}_{i_t})$.

This situation arises very often. For example in a Dynamic Linear Model (Harrison & Stevens, 1976) conditional on the values of states, $Y_1, \ldots, Y_T$ are independent, and furthermore Y_t only depends on current states which lie in the same clique. In the spatio-temporal process described in this thesis, an observation is taken at time t whose expectation is linear in states, the linear combination depending upon the site at which the observation is taken.

We now ask two questions: How in general should we update $p(\boldsymbol{x})$ in the light of an observation Y using the probability density breakdown given in equation (8.1), and how do we obtain the predictive distribution of the next observation ? Note that the answer to the first question is sufficient to answer the second question for a sequence of observations $Y_1, \ldots, Y_T$ because the procedure is just iterated. For this reason we will henceforth suppress the time index t.

When Y is assumed to come from an exponential family with parameter $\eta(\boldsymbol{x})$ which is linear in $\boldsymbol{x}$ and $Y|\eta(\boldsymbol{x})$ is normal, then the propagation procedure can be written down explicitly as follows (see Lauritzen, 1992 and Smith et al., 1995).

(a) Label by $C(1)$ the clique that receives the new information and containing all components explicitly in $\eta(\boldsymbol{x})$.

(b) Pick an ordering of the cliques $C(1), \ldots, C(n)$ starting from $C(1)$ which satisfies the running intersection property such that $C(i)$ is connected by an edge of the junction tree to one of $C(1), \ldots, C(i-1)$, $\quad 2 \leq i \leq n$.

An algorithm which constructs such an ordering is given in Tarjan & Yannakakis (1984).

Suppose the vector of variables in $C(i)$ is $\boldsymbol{X}(i) = (X_1(i), \ldots, X_{r_i}(i))^T$ and that

$$\eta(\boldsymbol{X}(i)) = \sum_{j=1}^{r_i} f_j X_j(i) = \boldsymbol{F}^T \boldsymbol{X}(i) \tag{8.2}$$

where $\boldsymbol{F} = (f_1, \ldots, f_{r_i})^T$ is a known regression vector.

When we generate this procedure later it will be useful to introduce some redundant notation at this point, so

1. Set $\eta = \lambda = \boldsymbol{F}^T \boldsymbol{X}(1)$ where $\boldsymbol{X}(1) \sim N[\boldsymbol{\mu}(1), \Sigma(1)]$.

2. Calculate from the joint normal prior distribution of λ and and $\boldsymbol{X}(1)$ the normal distribution of $\boldsymbol{X}(1)|\lambda$ with mean and covariance matrix given by

$$\begin{aligned}\boldsymbol{\mu}^0(1) &= \mu(1) + \Sigma(1)\frac{\boldsymbol{F}(\lambda - \boldsymbol{F}^T\mu(1))}{\boldsymbol{F}^T\Sigma(1)\boldsymbol{F}},\\ \Sigma^0(1) &= \Sigma(1) - \Sigma(1)\frac{\boldsymbol{F}\boldsymbol{F}^T\Sigma(1)}{\boldsymbol{F}^T\Sigma(1)\boldsymbol{F}}.\end{aligned}$$

3. Observing $Y = y$, use Bayes' rule to calculate the distribution of $\eta|y$.

4. From this obtain the Gaussian distribution of $\lambda|y$ with mean m and variance V.

5. From steps 2 and 4 above calculate the distribution of $p(\boldsymbol{X}(1)|Y = y)$ as Gaussian with mean $\boldsymbol{\mu}^*(1)$ and variance $\Sigma^*(1)$ given by

$$\begin{aligned}\boldsymbol{\mu}^*(1) &= \mu(1) + \Sigma(1)\frac{\boldsymbol{F}(m - \boldsymbol{F}^T\mu(1))}{\boldsymbol{F}^T\Sigma(1)\boldsymbol{F}}\\ \Sigma^*(1) &= \Sigma(1) - \Sigma(1)\boldsymbol{F}\boldsymbol{F}^T\Sigma(1)\frac{(1 - \frac{V}{\boldsymbol{F}^T\Sigma(1)\boldsymbol{F}})}{\boldsymbol{F}^T\Sigma(1)\boldsymbol{F}}\end{aligned}$$

6. Having obtained the Gaussian marginal for the vector of random variables in $C(1)$, we now simply order the cliques $(C(1), \ldots, C(n))$ to update the distribution of their components sequentially through this list. It is obvious (see Smith et al., 1995 and Chapter 7) that given $Y = y$, each clique has a joint Gaussian margin, if the mean vector and covariance matrix of $\boldsymbol{X}(i)$ which are given by

$$\boldsymbol{\mu}(i) = (\boldsymbol{\mu}_1(i), \boldsymbol{\mu}_2(i))^T$$

$$\Sigma(i) = \begin{pmatrix} \Sigma_{11}(i) & \Sigma_{12}(i) \\ & \Sigma_{22}(i) \end{pmatrix}$$

respectively are updated to $(\boldsymbol{\mu}^*(i), \Sigma^*(i))$ given y using the formulae of Section (7.2).

This is possible because the mean vector $\boldsymbol{\mu}_1^*(i)$ and covariance matrix $\Sigma_{11}^*(i)$ are associated with the subvector of $\boldsymbol{X}(i)$ whose distribution has already been updated because it lies in a previously listed clique. For more details of this algorithm see Lauritzen (1992) and Smith et al. (1995).

In this way we can update all marginal densities on $C(1), \ldots, C(n)$ and hence the whole distribution $p(\boldsymbol{x})$. We are then ready to receive the next observation where the procedure is repeated.

Now unfortunately, outside the purely discrete or Gaussian systems like the one described above, there are only a very few distributions for which an algorithm like the one above works exactly and for which the vector $\boldsymbol{X}$ continues to lie in a recognised family of distributions (see Laurtizen 1992 for some exceptions). To side-step these problems Thomas et al. (1992) uses a numerical integration method to update the clique margins. This method, which can calculate numerical distributions to arbitrary degrees of accuracy, has much to recommend it. However it tends to be slow, and because the solutions cannot be given in algebraic form it is often difficult to understand why the distributions turn out the way they do.

In this chapter we suggest a different route. This uses the updating techniques of Dynamic Generalized Linear Models (DGLMs) of West et al. (1985) and West & Harrison (1989). We treat their algorithm as if it were a dynamic approximation technique of a full Bayesian analysis as developed in Smith (1992).

8.3 Dynamic Generalised Linear Models on Junction Trees

Suppose a random variable η belongs to some parametrised family of densities Ψ which is closed under sampling of an observation Y whose distribution conditional on η lies in a family Υ. For simplicity of exposition West & Harrison (1989) choose to restrict Υ to be the exponential family although this condition is not strictly necessary for their algorithm to work. Now assume

$$\lambda = g(\eta) \tag{8.3}$$

where g is a known, continuous and monotonic function and λ is a linear function of the normally distributed uncertain state vector $\boldsymbol{X}$ as in equation (8.2). Because the distribution of λ is Gaussian, it is unlikely that the density of $\eta = g^{-1}(\lambda)$ will lie in Ψ. However, provided that the function g is chosen appropriately and Ψ is two dimensional, it will often be possible to find a density $\hat{p}(\eta) \in \Psi$ which is very close to the transformed Gaussian density $p(\eta)$ of η. And conversely it should be possible to find a transformed Gaussian density $\hat{q}(\eta)$ which is close to any density $q(\eta) \in \Psi$ that might arise in our analysis. Appropriate definitions of closeness will be discussed in the next section.

To ensure not only that recurrence equations can be written in a closed form, but also that the states of the process retain relationships between each other which are straightforward to explain, we usually choose simple functions g and elementary ways of approximating $p(\eta)$ by $\hat{p}(\eta)$, for example by equating moments.

Assume that a good simple approximation of this type is available. We can now make a straightforward modification to the procedure defined in Section (2). First replace step 1 by 1^*

1^*. Find the density $\hat{p}(\eta)$ which approximates to the density $p(\eta)$ under the link function relationship $\eta = g^{-1}(\lambda)$ defined above.

We keep step 2 and also 3 identical, noting that, because Ψ is closed under sampling to Y, the posterior density associated with $\eta, p(\eta|y) \in \Psi$. We now again use our approximation, replacing step 4 by 4^*.

4^*. Find a density $\hat{p}(\eta|y)$ which approximates $p(\eta|y)$ and for which $g(\eta)$ is Gaussian with mean m and variance V.

We now retain the propagation steps 5 and 6 to obtain the approximate updating equations of the full density $p(\boldsymbol{x})$ expressed as a function of its clique marginals.

A few examples of the updating equations derived from following the sequences of steps 1^*, 2, 3, and 4^*, when g is a linear link function, are given below and summarised in Table 8.1.

8.3.1 Some Illustrative Examples

The Poisson Model

Let $(Y_t|\eta_t(\boldsymbol{x}_t(1)) \sim \text{Poisson}(\eta_t(\boldsymbol{x}_t(1))$ where $\boldsymbol{X}_t(1)$ is the state vector of the random variables in a clique $C(1)$ which receives the new information Y_t and let the prior for $\boldsymbol{X}_t(1)$ be

$$(\boldsymbol{X}_t(1)|D_{t-1}) \sim N[\boldsymbol{\mu}_t(1), \Sigma_t(1)].$$

The linear function of the state vector $\lambda_t = \boldsymbol{F}_t^T \boldsymbol{X}_t(1) \sim N[\mu_0, \sigma_0^2)]$.

Now, omitting the arguments, the adaptive procedure briefly comprises the following steps.

1. **Approximating λ_t by η_t.**

   Approximate the distribution of λ_t which is a univariate Gaussian $N[\mu_0, \sigma_0^2]$ by a variable η_t with a Gamma$[\alpha_0, \beta_0]$ distribution where we identify means and variances so that $\alpha_0 = \frac{\mu_0^2}{\sigma_0^2}$, $\beta_0 = \frac{\mu_0}{\sigma_0^2}$.

2. From the joint normal prior distribution of λ_t and $\boldsymbol{X}_t(1)$

   $$\begin{pmatrix} \lambda_t & \\ & |D_{t-1} \\ \boldsymbol{X}_t(1) & \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_0 \\ \boldsymbol{\mu}_t(1) \end{pmatrix}, \begin{pmatrix} \sigma_0^2 & \boldsymbol{F}_t^T \Sigma_t(1) \\ \Sigma_t(1)\boldsymbol{F}_t & \Sigma_t(1) \end{pmatrix}\right]$$

   Calculate the normal distribution of $\boldsymbol{X}_t(1)|\lambda_t$ with mean and covariance matrix given by

   $$\begin{aligned} \boldsymbol{\mu}_t^0(1) &= \boldsymbol{\mu}_t(1) + \frac{\Sigma_t(1)\boldsymbol{F}_t(\lambda_t - \boldsymbol{F}_t^T \mu_t(1))}{\boldsymbol{F}_t^T \Sigma_t(1)\boldsymbol{F}_t} \\ \Sigma_t^0(1) &= \Sigma_t(1) - \frac{\Sigma_t(1)\boldsymbol{F}_t\boldsymbol{F}_t^T\Sigma_t(1)}{\boldsymbol{F}_t^T\Sigma_t(1)\boldsymbol{F}_t} \end{aligned}$$

3. **Updating η_t.**

   Observing Y_t use Bayes' rule to calculate the distribution of $\eta_t|y_t$ for $Y_t|\eta_t \sim$ Poisson$[\eta_t]$

   $$(\eta_t|y_t) \sim \text{Gamma}[\alpha_1, \beta_1]$$

   where $\alpha_1 = \alpha_0 + y_t$ and $\beta_1 = \beta_0 + 1$.

4. **Updating λ_t**

   Approximate the Gamma distribution in the previous step by a normal distribution with the same mean and variance to obtain $\lambda_t|y_t \sim N[\mu_1, \sigma_1^2]$, where

$$\begin{aligned}\mu_1 &= \tfrac{\alpha_1}{\beta_1} = (1-A)\mu_0 + Ay_t\\ \sigma_1^2 &= \tfrac{\alpha_1}{\beta_1^2} = A\mu_1\end{aligned}$$

where $A = \frac{\sigma_0^2}{\mu_0+\sigma_0^2}$.

5. **Updating for $X_t(1)$**

   From steps 2 and 4 above calculate the distribution of $(X_t(1)|y_t)$ as Gaussian with mean and variance

$$\begin{aligned}\mu_t^*(1) &= \mu_t(1) + \Sigma_t(1)\frac{F_t(\mu_1 - F_t^T\mu_t(1))}{F_t^T\Sigma_t(1)F_t}\\ \Sigma_t^*(1) &= \Sigma_t(1) - \Sigma_t(1)F_tF_t^T\Sigma_t(1)\frac{1-\frac{\sigma_1^2}{F_t^T\Sigma_t(1)F_t}}{F_t^T\Sigma_t(1)F_t}\end{aligned}$$

6. Having obtained the Gaussian marginal for the vector of random variables in the clique $C(1)$, we order the cliques $C(1), \ldots, C(n)$ to update the distribution of their components sequentially.

The Lognormal Model

Let $\log Y_t|\eta_t(x_t(1)) \sim N[\log \eta_t(x_t(1)), V]$, so that $Y_t|\eta_t(x_t(1))$ is lognormally distributed with median $\eta_t(x_t(1)$, where $X_t(1)$ is the state vector of random variables in a clique $C(1)$ and let the prior for $X_t(1)$ be

$$X_t(1)|D_{t-1} \sim N[\mu_t(1), \Sigma_t(1)]$$

and

$$\log \eta_t(x_t(1) \sim N[\xi_0, \tau_0^2].$$

Now, omitting the arguments, the principal updating steps are summarised as follows.

1. **Approximating λ_t by η_t.**

   Approximate the distribution of $\log \eta_t = \lambda_t$ which is a univariate Gaussian $N[\xi_0, \tau_0^2]$ by a variable η_t with a distribution $\text{Logn}[\mu_0, \sigma_0^2]$ with mean and variance as

$$\begin{aligned} \mu_0 &= e^{\xi_0 + \frac{1}{2}\tau_0^2} \\ \sigma_0^2 &= (e^{\tau_0^2} - 1)\mu_0^2 \Longrightarrow \tau_0^2 = \log \frac{\sigma_0^2 + \mu_0^2}{\mu_0^2} > 0 \end{aligned}$$

2. As in the Poisson model.

3. **Updating η_t.**

   Observing $Y_t = y_t$, find the posterior distribution of $\log \eta_t$ with a normal prior with mean ξ_0 and variance τ_0^2 for $\log Y_t|\eta_t \sim N(\log \eta_t, V)$, so that $\log \eta_t | Y_t = y_t$ is Gaussian with moments

$$\begin{aligned} E[\log \eta_t | y_t] &= (1 - A)\xi_0 + A \log y_t &= \xi_1 \\ Var[\log \eta_t | y_t] &= AV \\ &= (1 - A)\tau_0^2 &= \tau_1^2 \end{aligned}$$

   which implies that $\eta_t | y_t \sim \text{Logn}[\mu_1, \sigma_1^2]$ with mean and variance as

$$\begin{aligned} \mu_1 &= e^{\xi_1 + \frac{1}{2}\tau_1^2} \\ \sigma_1^2 &= (e^{\tau_1^2} - 1)\mu_1^2 \end{aligned}$$

   where $A = \frac{\tau_0^2}{\tau_0^2 + V} = \frac{\log(\mu_0^2 + \sigma_0^2) - \log(\mu_0^2)}{\log(\mu_0^2 + \sigma_0^2) - \log(\mu_0^2) + V}$

4. **Updating λ_t.**

   Approximate the posterior distribution in the previous step which is $\text{Logn}[\mu_1, \sigma_1^2]$ by a normal distribution with the same mean and variance to obtain the posterior

Table 8.1: **Some illustrative examples**

Distributions **of η, λ**	**Distribution** **of $Y\|\eta$**	**Posterior mean μ_1** **of λ**	**Posterior variance σ_1^2** **of λ**
Normal $N(\mu_0, \sigma_0^2)$ $\lambda = \eta$	Normal (η, V)	$(1-A)\mu_0 + Ay$	$AV = (1-A)\sigma_0^2$ $A = \frac{\sigma_0^2}{\sigma_0^2+V}$
$\lambda \sim N(\mu_0, \sigma_0^2)$ $\eta \sim G(\alpha_0, \beta_0)$ $\alpha_0 = \frac{\mu_0^2}{\sigma_0^2}, \beta_0 = \frac{\mu_0}{\sigma_0^2}$	Poisson $P_0(\eta)$	$(1-A)\mu_0 + Ay$	$A\mu_1$ $A = \frac{\sigma_0^2}{(\sigma_0^2+\mu_0)}$
$\lambda = \log\eta \sim N(\xi_0, \tau_0^2)$ $\eta \sim \text{Logn}(\mu_0, \sigma_0^2)$ $\mu_0 = e^{(\xi_0 + \frac{1}{2}\tau_0^2)}$ $\sigma_0^2 = (e^{\tau_0^2} - 1)\mu_0^2$	Lognormal (η, V) $\log Y\|\eta \sim N(\log\eta, V)$	$\mu_0^{(1-A)}y^A$	$(e^{AV} - 1)\mu_1^2$ $A = \frac{\log(\mu_0^2+\sigma_0^2)-\log\mu_0^2}{\log(\mu_0^2+\sigma_0^2)-\log\mu_0^2+V}$

distribution of λ_t as Gaussian with moments

$$\begin{aligned} \mu_1 &= e^{(1-A)\xi_0 + A\log y_t + 1/2(1-A)\tau_0^2} &= y_t^A \mu_0^{(1-A)} \\ \sigma_1^2 &= (e^{AV} - 1)\mu_1^2 \end{aligned}$$

5. As 5 in the Poisson model

6. As 6 in the Poisson model

A summary of these examples is given in Table 8.1 . For more detailed discussion, see Gargoum & Smith (1995). Notice that the obvious difference from the Gaussian case is that now the posterior variance of $\lambda|y$ is a function of y. The point we make here is that the updating on the junction tree is just as quick and simple as the Gaussian because it is algebraic and approximate, so that Gaussianity over states is preserved.

Thus a very slight change to code allows the quick processing of data which is not Gaussian.

Of course the validity of the updating algorithm depends critically on how well the true posterior density of $\boldsymbol{x}|\boldsymbol{y}$ is approximated by the Gaussian one calculated by our algorithm. This topic is the subject of the next section.

8.4 The Closeness of Dynamic Approximations

Here, we choose the Hellinger metric to check the appropriateness of the dynamic approximation. Recall from Chapter (5) that the Hellinger distance between two densities is defined by

$$d_H(f;h) = \left(1 - \int f^{1/2}(\boldsymbol{x})h^{1/2}(\boldsymbol{x})d\boldsymbol{x}\right)^{1/2} \tag{8.4}$$

Also we defined

$$I(f;g) = 1 - d_H^2(f;g) \tag{8.5}$$

In fact $d_H^2(f;g)$ can be calculated in closed form for most densities in a standard family. It is also sometimes possible explicitly to write down the Hellinger distance between two densities from different families. For example, when f is a normal density with mean μ and variance V and g is a Gamma density $G(\alpha,\beta)$ with the same mean and variance, then $I(f;g)$ is obtained as in Section (5.7) of Chapter (5).

We note the property listed below also hold true both for the variation metric and the popular Kullback-Leibler separation measure.

Suppose that p and $\hat{p}$ are joint densities on $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ which have different margins p_1 and $\hat{p}_1$ on $\boldsymbol{X}_1$ but whose conditional densities of $\boldsymbol{X}_2|\boldsymbol{X}_1$ agree. Then, directly from (8.4) we have that

$$d_H(p, \hat{p}) = d_H(p_1, \hat{p}_1) \tag{8.6}$$

Now, within our context we approximate only the distribution of λ, as information about the states is channelled through λ. It follows from (8.6) that the closeness of the joint density over states depends only on the closeness of our approximation of the one dimensional normal posterior density of λ to the true posterior density of λ.

As an example consider the case when $p_0(\boldsymbol{x})$ is a Gaussian prior density on $\boldsymbol{X}$. Let f_1 and f_2 denote the posterior densities on $\boldsymbol{x}$ given the true normalised Gamma likelihood ℓ_1 associated with a Poisson observation Y or a normalised Gaussian approximation ℓ_2 of the DGLM, respectively. Then, by definition, omitting the arguments,

$$f_i = \frac{p\ell_i}{\int p\ell_i} \qquad i = 1, 2$$

So

$$\begin{aligned} I^2(f_1; f_2) &= \frac{(\int p\ell_1^{1/2}\ell_2^{1/2})^2}{(\int p\ell_1)(\int p\ell_2)} \\ d_H^2(f_1; f_2) &= 1 - \sqrt{I^2(f_1, f_2)} \\ &= 1 - \frac{B}{\sqrt{A}} \end{aligned}$$

where $B = \frac{\int \ell_1^{1/2}\ell_2^{1/2}p}{\int \ell_1 p}$ and $A = \frac{\int p\ell_2}{\int p\ell_1}$.

Notice that if ℓ_1 and ℓ_2 are very close, then both A and B will be close to 1 and consequently d_H will be close to zero. Appendix C gives the details of how to derive an upper bound for $d^2(f_1, f_2)$, i.e. an upper bound on B and a lower bound on A; which can be used in the particular case when l_1 and l_2 are as above.

8.5 Conclusion

This analogue of dynamic generalized linear models, when used on junction trees, gives a quick computational approach for dealing with non-normal data which is easy to understand, gives a closed form updating algorithm and provides an approximation whose validity can be checked numerically, for example by using the Hellinger distance metric. In an iterative system where quick calculation is essential and easy interpretation is paramount, it is our opinion that the methods described in this chapter provide a practical methodology for quick Bayesian inference in complex dynamical systems.

8.6 Appendix C

An upper bound for $d_H^2(f_1, f_2)$.

Let ℓ_1, ℓ_2, p, A, and B defined as before.

Write

$$\ell_1^{1/2}\ell_2^{1/2}p = (\ell_2^{1/2} - \ell_1^{1/2})\ell_1^{1/2}p + \ell_1 p.$$

Then

$$B - 1 = \frac{\int(\sqrt{2}\ell_1^{1/2}p)((\sqrt{2})^{-1}(\ell_2^{1/2} - \ell_1^{1/2}))}{\int \ell_1 p}$$

So

$$|B - 1| \leq \frac{\int(\sqrt{2}\ell_1^{1/2}p)(\sqrt{2})^{-1}|\ell_2^{1/2} - \ell_1^{1/2}|)}{\int \ell_1 p}$$

which, by Cauchy-Schwartz inequality applied to the expressions in the upper integral,

$$\leq \frac{(\int 2\ell_1 p^2)^{1/2}}{(\int \ell_1 p)}\left(\int 1/2(\ell_2^{1/2} - \ell_1^{1/2})^2\right)^{1/2} = \tau d_H(\ell_1, \ell_2)$$

where $\tau = \sqrt{2}\frac{(\int \ell_1 p^2)^{1/2}}{(\int \ell_1 p)} \leq \left(\frac{(2M)}{I_1}\right)^{1/2} = \bar{\tau}$

where $M = \text{Sup } p$ and $I_1 = \int \ell_1 p$.

Thus

$$1 - \bar{\tau} d_H(\ell_1, \ell_2) \leq B \leq 1 + \bar{\tau} d_H(\ell_1, \ell_2).$$

To derive a bound on A, first note that

$$A = \frac{\int \ell_2 p}{\int \ell_1 p} = \frac{\int(\ell_2 - \ell_1)p + \int \ell_1 p}{\int \ell_1 p} = 1 + \frac{\int(\ell_2 - \ell_1)p}{\int \ell_1 p}.$$

So

$$|A| = 1 + \frac{|\int(\ell_2 - \ell_1)p|}{\int \ell_1 p}.$$

Then

$$|\int (\ell_2 - \ell_1)p| \leq \int |\ell_2 - \ell_1|p = \int \{(\sqrt{2})^{-1}|\ell_2^{1/2} - \ell_1^{1/2}|\}\{(\sqrt{2})|\ell_1^{1/2} + \ell_2^{1/2}|p\}$$

which, by Cauchy-Schwartz inequality,

$$\begin{aligned} &\leq \left(1/2\int(\ell_2^{1/2} - \ell_1^{1/2})^2\right)^{1/2}\left[2\int(\ell_1^{1/2} + \ell_2^{1/2}p^2\right]^{1/2} \\ &= d_H(\ell_1, \ell_2)\left[2\int(\ell_1^{1/2} + \ell_2^{1/2})^2p^2\right]^{1/2} \end{aligned}$$

where, in particular

$$\begin{aligned} \int(\ell_1^{1/2} + \ell_2^{1/2})^2p^2 &= \int \ell_1 p^2 + \int \ell_2 p^2 + 2\int \ell_1^{1/2}\ell_2^{1/2}p^2 \\ &\leq 2M^2[1 + \int \ell_1^{1/2}\ell_2^{1/2}] \\ &= 4M^2[1 - 1/2d_H^2(\ell_1, \ell_2)] \end{aligned}$$

So

$$\begin{aligned} A &\leq 1 + I_1^{-1}\left[2\textstyle\int(\ell_1^{1/2} + \ell_2^{1/2})^2p^2\right]^{1/2} d_H(\ell_1, \ell_2) \\ &\leq 1 + \tfrac{2M}{I_1}[1 - 1/2d_H^2(\ell_1, \ell_2)]^{1/2}d_H(\ell_1, \ell_2) \\ &\leq 1 + \bar{\tau}^2 d_H(\ell_1, \ell_2) \end{aligned}$$

Thus, provided that M/I_1 is bounded above, the likelihoods with close Hellinger distances give rise to posterior densities also with close Helinger distances.

Now notice that, for $y > 0$

$$(1 + y)^{-1/2} \geq 1 - \frac{1}{2}y.$$

Thus

$$(1 + \bar{\tau}^2 d_H(\ell_1, \ell_2))^{-1/2} \geq 1 - 1/2\bar{\tau}^2 d_H(\ell_1, \ell_2).$$

but $A \leq 1 + \bar{\tau}^2 d_H(\ell_1, \ell_2)$ or $A^{-1/2} \geq (1 + \bar{\tau}^2 d_H(\ell_1, \ell_2))^{-1/2}$.

So

$$A^{-1/2} \geq 1 - 1/2\bar{\tau}^2 d_H(\ell_1, \ell_2).$$

It follows from the above that

$$\begin{aligned} 1 - A^{-1/2}B &\leq 1 - [1 - 1/2\bar{\tau}^2 d_H(\ell_1, \ell_2)][1 - \bar{\tau} d_H(\ell_1, \ell_2)] \\ &\leq d_H(\ell_1, \ell_2)\bar{\tau}(1 + 1/2\bar{\tau}) \end{aligned}$$

i.e.

$$d_H^2(f_1, f_2) \leq \bar{\tau}(1 + 1/2\bar{\tau}) d_H(\ell_1, \ell_2)$$

where $\bar{\tau} = \sqrt{\frac{2M}{I_1}}$ where $I_1 = \int \ell_1 p$ and $M = \text{Sup } p$.

Thus posterior densities will be close in Hellinger distance if ℓ_1 and ℓ_2 are close. This gives an upper bound for $d_H(f_1, f_2)$ which is suitable for our purposes.

Chapter 9

Discussion and Further Research

9.1 Introduction

This research was originally motivated by the practical problems involved in modelling and updating the dispersal and deposition of contaminated material emitted after a nuclear accident. The main objectives were to: model and code the expert judgement about the emission profile as a DLM; combine an existing propagation methodology for dynamical junction trees with approximation algorithms when data may destroy neat dependencies; and accommodate non - Gaussian distributions on contamination readings using DGLMs.

In this application, the learning and updating algorithm associated with the modelling process is a component within an integrated decision support system for guiding countermeasures in the first 24 hours after the accidental release. So output must be informative about the expedience of *short-term* decisions that might be taken.

9.2 Review of Chapters (3) and (4)

In Chapters (3) we revealed the inadequacy of existing deterministic models of dispersal. We then showed how the continuous release of gas can be described as a series of puffs of contaminated mass emitted sequentially at discrete times and then dispersed and diffused (puff models). We also described a stochastic version of these puff dispersal models. This version made it possible to incorporate and adjust to uncertain information about contamination readings at different sites. We then proceeded to show that all relevant uncertainties could be modelled by describing the evolution of puffs and puffs fragments within the system by a high dimensional Gaussian process exhibiting many conditional independences. These conditional independences can be utilised to speed up the revision of the probability distribution of the puff and puff fragment masses. Finally, we discussed how the stochastic puff model could incorporate complex processes defining source emission and fragmentation.

In Chapter (4), we discussed some graph–theoretical concepts and results on influence diagrams and junction trees which are necessary to describe a clique representation of these fragmentation processes. This representation is suitable for an efficient propagation of evidence as it arrives.

9.3 Review of Chapter (6)

One of the most important pieces of information that will be needed to inform very early decision–making immediately after an accident is how experts (plant designers and nuclear safety engineers) believe source emissions of contamination will develop over time. To address this issue it is important, first to code as much expert judgement as possible about the possible types and profiles of release; and, secondly, to modify these expert judgements – which are often very uncertain – in the light of any observations which do become available.

In Chapter (6) we have considered different scenarios of how experts might believe the shape of the emission will develop. One scenario is when there is very little expert judgement about the development of the release of mass at time t, $\boldsymbol{Q}_t$. Let us assume that an expert judges that the leakage gives an initial expected mass $\mu(1)$, but is very uncertain about this with large variance $\sigma^2(1)$. Suppose he or she is also uncertain about how the shape of the release profile will develop apart from the belief that the leakage will be relatively continuous. This case is modelled by a random walk such that $\boldsymbol{Q}_t|\boldsymbol{Q}_{t-1} \sim N[\boldsymbol{Q}_{t-1}, Z]$.

In another scenario the expert may believe that the emissions will rise from zero to an uncertain height h and then decline to an asymptotic leakage a which then drifts as a random walk. By giving means and variances of h and a, we have been able to build a probabilistic model which adapts its estimate of the emissions $\boldsymbol{Q}_t$ and also its estimate of the height h, in the light of any incoming data.

Figures 6.5 and 6.6 of Appendix A illustrate how the model predictions of the emission profile change as monitoring data (model–generated data) are taken. Also in Chapter (6) we dealt specifically with the uncertainty of the release height h, where we run mixed models at different release heights (200m, 400m, 600m) with certain prior probabilities and then update these probabilities (i.e. update the distributions on the release height).

Figure 6.8 of Appendix A shows posteriors for expected dispersals associated with the three source release heights together with the marginal expected dispersal.

The facility for accommodating any time series model for estimating the source emissions is now available under Leeds–Warwick software.

9.4 Review of Chapters (7) and (8)

In recent years, many practical and efficient expert systems (see, for example, Lauritzen & Speigelhalter, 1988; Dawid, 1992 and Speigelhalter et al., 1993) have been built and used in a complex environment on the basis of probability algebra. These systems are based on graphical representation provided that the following conditions hold:

(i) The expert system has a fixed number of variables in it.

(ii) The joint probability distribution is of a particular form (decomposable form)– that is, the joint density can be expressed as a product of marginal densities over sets of variables $C(1),\ldots,C(n)$ (cliques of moderate sizes) divided by probabilities on subsets of cliques.

(iii) The distributional form of all the random variables in the system is assumed to be either discrete or Gaussian.

A description of the algorithms for calculating these probabilities is given in the references above.

Unfortunately, in many practical cases one or more of the conditions (i)–(iii) are violated. Firstly, learning is dynamic and happens sequentially over time, and variables which are of direct interest at one point of time subsequently lose their relevance. For example, in the dispersal models changing atmospheric conditions can make one set of variables irrelevant and other previously unmodelled features critical to forecasts. Secondly, condition (ii) can often be violated, especially when data are collected which depend on variables in several cliques.

Thirdly, condition (iii) is a very restrictive one for use in a general statistical model. Accordingly, the system will be extended to a system with variables which develop dynamically and with the following additional features:

(a) The system will need to process very large numbers of variables at any one time.

(b) Data can destroy neat dependences.

(c) The distributional form of all random variables in the system can be generalised to a non-Gaussian.

In the context of our application, we adopted dynamic influence diagrams (and their corresponding junction trees) (see Smith et al., 1995) to describe the dispersal process. Because these diagrams are defined on a sequence of random variables, there are high demands on computational efficiency. Typically the system will need to process a very large number of variables at any one time.

In Chapter (7) we addressed this problem by approximating the process to obtain a system with disconnected junction trees through a faster algorithm which deletes "irrelevant" cliques using edge deletion (CUT operator). This approximation algorithm is based on a divergence measure between the true and the approximating distributions (i.e. before and after the cut).

Also in Chapter (7) we addressed the issue of condition (ii) above. We dealt with the situation in which an observation is taken under more than one clique. (When this happens, the conditional independences of the clique structure will be destroyed; see b above). We thus developed approximation algorithms for the Guassian process based on the Kulback-Liebler / Hellinger distances. Explicitly we proposed two classes of operators to deal with this case:

- exact operators (e.g. JOIN operator); here we use this operator to join cliques under which the observation is taken to retain valid dependences after data observation);
- non-exact operators, (e.g. DivObs operator); using this operator, an observation has to be "divided" between several cliques, so approximations are necessary which are based on measures of divergence such as the Kullback-Leibler divergence measure.

Initial simulations have shown that these algorithms are fast enough to provide forecasts within the requirements of the RODOS decision support system. Section B.3 of Appendix B of Chapter 7 shows plots of the expected concentrations using the normal and the normal-cut models. The forecasts of the concentrations using the two models are very close with significantly less computational time using the normal–cut model. Section B.4 shows the results of the implementation of the DivObs operator. At all detector points the DivObs update (approximation) matches the true update very closely.

The problem of the particular distributional forms (Gaussian or discrete) of the exact algorithms was addressed in Chapter (8). Approximate decomposable structures were found, and suitable metrics over distributions were used to ensure that the probability distributions required for making decisions are at least approximately correct both a priori and a posteriori. We proposed to accommodate non-Gaussian distributions on contamination readings using a slight generalisation of the dynamic generalised linear models of West et al. (1985), thus giving closed form solutions to updating. Some illustrative examples of accommodating non-Gaussian models (Poisson, Lognormal,...) have been provided. For example, in the Poisson case we write

$$\lambda = \sum_{Q(i) \in C} F_t(i,s) Q(i)$$

where $Q(i)$, $F_t(i,s)$ and C are defined as before.

Let $\lambda \sim N[\mu_0, \sigma_0^2]$, $Y|\lambda \sim \text{Poisson}(\eta)$ where $\eta \sim \text{Gamma}[\alpha_0, \beta_0]$

Then

$$\lambda | y \approx N[\mu_1, \sigma_1^2]$$

where

$$\begin{aligned} \mu_1 &= (1-A)\mu_0 + Ay \\ \sigma_1^2 &= A\mu_1 \\ A &= \frac{\sigma_0^2}{(\sigma_0^2 + \mu_0)} \end{aligned}$$

In this case it is easy to calculate $p(C|y) = p(C)\frac{p(\lambda|y)}{p(\lambda)}$ as an update on mean and variance. Now we propagate as in the normal case.

Note that the appropriateness of the updating algorithm depends on how well the true posterior density is approximated by the Gaussian one. Here we used the Hellinger metric to check this dynamic approximation. An upper bound on the Hellinger distance of the true posterior density and its approximation is provided in Appendix C of Chapter 8.

An alternative formal method to deal with accommodating non-Gaussian readings is to calculate exact margins numerically using Metropolis algorithms (see, for example, Gilks et al., 1993). This approach is satisfactory as long as the cliques and their separators are small, ensuring that the process does not take a long time. Moreover, such methods could be useful for validating our approximation of accommodating non–Gaussian distributions described in Chapter (8).

Although the basic motivation of this research was to overcome the practical problems of modelling the diffusion of a gas from a source in a complex stochastic windfield, the contributions in the thesis are of more general relevance and can be extended to include other applications of the diffusion and dispersal of a process from a source where we may have different junction tree and clique components, but where the algorithms for coding the qualitative information, the speeding up of the calculations of the relevant joint distributions, and the generalisation from Gaussian processes are not significantly modified.

Finally, there are other topics and unsolved problems that are not described in this thesis. Of great importance is the issue of extending the system beyond the analysis of *dispersal* to incorporate the modelling of uncertainty concerning the *deposition* of waste where the clouds lose a proportion of their mass in response, for example, to *rain fall*. This issue gives rise to two particular important topics for future research:

the prediction of how the contaminated material will spread out over the ground ; and the relationship between the spatial process of depositions and the stochastic process and dispersal models.

Bibliography

[1] Alspach, D. L. and Sorenson, H. W. (1972). Nonlinear Bayesian estimation using Gaussian sum approximation, IEEE Trans. Aut. Con., **17**, 439 - 448.

[2] Ameen, J. R. M., and Harrison, P. J. (1984). Discount weighted estimation. J. of Forecasting **3**, 285 - 296.

[3] Ameen, J. R. M. and Harison, P.J. (1985). Normal discount Bayesian models. In Bayesian Statistics *2*, J. M. Bernardo, M. H. De Groot, D. V. Lindley and A. F. M. Smith (eds.) North Holland, Amsterdam, and Valencia University Press.

[4] Anderson, B. D. O. and Moore, J. B. (1979). Optimal Filtering. Prentice- Hall, New Jersey.

[5] ApSimon, H. M., Wilson, J. J. N. and Simms, K. L. (1989). Analysis of the dispersal and deposition of radionuclides from Chernobyl across Europe. Proc. R. Soc. Lon. **A425**, 365 - 405.

[6] Beeri, C., Fagin, R., Majer, D., Mendelzon, A., Ullman, J., and Yannakakis, M. (1981). Properties of acyclic database schemes. Proc. 13th Annual ACM Symposium on the Theory of Computing, Milwaukee, J. Assoc. Comput. Mach. New York.

[7] Beeri, C. Fagin, R., Majer, D., and Yannaakakis, M. (1983). On the desirability of acyclic database schemes, J. Assoc. Comput. Mach., **30**, 479 - 513.

[8] Box, G. E. P., and Taio G. C. (1973). Bayesian Inference in Statistical Analysis. Addison - Wesley, Massachusetts.

[9] Brown, R. G. (1962). Smoothing, Forecasting and Prediction of Discrete Time Series. Prentice - Hall, Englewood Cliffs, NJ.

[10] Chatfield, C. (1995). Discussion of paper by Draper, D. (1995) given below, J. Roy. Statist. Soc. (Ser B) **57**, No. 1, 45 - 97.

[11] Chatwin, P. C. (1982). The use of statistics in describing and predicting the effect of dispersing gas clouds. J. Hazard Mater. **6**, 213 - 230.

[12] Covaerts, P. (1993). Local scale decision support systems, actual situations and trends for future. Radiation Protection Dosimetry Vol. **50** Nos 2 - 4, 141 - 144. Nuclear Technology Publishing.

[13] Darroch, J. N., Lauritzen, S. L., and Speed, T. P. (1980). Markov fields and loglinear interaction models for contingency tables. Annals of Statistics, **8**, 522 - 539.

[14] Dawid, A. P. (1979). Conditional independence in statistical theory. J. Roy. Statist. Soc. (Ser B), **41**, 1 - 31.

[15] Dawid, A. P. (1992). Applications of general propagation algorithm for probabilistic expert systems. Statistics and Computing, **2**, 25 - 36.

[16] De Groot, M. H., (1971). Optimal Statistical Decisions. McGraw- Hill, New york.

[17] Devroye, L. (1987). A Course in Density Estimation, Progress in Probability and Statistics, Vol. **14**, Birkhäuser, Boston.

[18] Diebolt, J., and Robert, C. P. (1994). Estimation of finite mixture distributions by Bayesian sampling. J. Roy. Statist. Soc. (Ser b) **2**, 362 - 375.

[19] Draper, D. (1995). Assessment and Propagation of Model Uncertainty (with discussion). J. Roy. Statist. Soc. (Ser B) **57**, No. 1, 45 - 97.

[20] Eckman, R. M., Dobosy, R. J., and Rao, K. S. (1992). Spatial variability of the wind over moderately complex terrain. Preprint, tenth symposium on turbulence and diffusion, Portland, OR, 84 - 87 (American Meteorological Society, Boston).

[21] Ehrhardt, J., Fischer, F., Pasler-Sauser, J., Schule, O., Benz,G.,Rafat, M. (1993). RODOS and RESY : Two integrated Real Time On-Line Decision Support Systems for nuclear emergencies in Europe., Proceedings of Mathematical Methods and Supercomputing in Nuclear Applications, Karlsruhe (Germany) 319 - 330.

[22] Escobar, M. D., and West, M. (1994). Bayesian prediction and density estimation. J. Amer. Statist. Association. (to appear).

[23] Feinbury, S. E., and Tsay, R. S. (1985). Comment on West et al. (1985), given below, J. Amer. Statist. Ass., **80**, 89 - 90.

[24] Fox, D. G. (1984). Uncertainty in air quality modelling. Bull. Am. Meteoral. Soc. **65**, 27 - 36.

[25] French, S. (1986). Decision theory; an introduction to the mathematics of rationality, Ellis Horwood, Chichester.

[26] French, S., and Smith, J. Q. (1992). The use of model predictions and monitoring data in decision making about off-site emergency actions in the event of an accident, Report for CEC, Grant no. B170075GB.

[27] Gargoum, A. S., and Smith, J. Q. (1994a). Approximated schemes for efficient probability propagation in evolving high dimensional Gaussian processes, Warwick Research Report, 266. Department of Statistics, University of Warwick.

[28] Gargoum, A. S., and Smith, J. Q. (1994b). Bayesian models of emission profiles of the release of a toxic gas. In S. French, J. Smith and D. Ranyard (eds.), Uncertainty in RODOS. Version 1.0 73 - 81. School of Computer Scinces, Leeds University.

[29] Gargoum, A. S., and Smith, J. Q. (1995). Dynamic Generalised Linear Junction Trees (in revision for resubmission to Biometrika).

[30] Gargoum, A. S., Smith J. Q., Ranyard, D., and Paminchail, K. N. (1995). Approximating Dynamic Junction Trees. Warwick Research Report (in preparation). Department of Statistics, University of Warwick.

[31] Gelfand, A. E., and Smith A. F. M. (1990). Sampling-based approaches to calculating marginal densities. J. Am. Statist. Ass, **85**, 398 - 409.

[32] Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Mach. Intell. **6**, 721 - 740.

[33] Gilks, W., Clayton, D. G., Spiegelhalter, D. J., Best, N. G., Sharples, L. D., and Kirby, A. J. (1993). Modelling complexity applications of Gibbs sampling in medicine. (with discussion). J. Roy. Statist. Soc. (Ser.B) **55**, 39 - 52.

[34] Goldstein, M. (1993). Prediction under the inference: Bayes linear influence diagrams for prediction in a large brewey. The Statistician **42**, 445 - 459.

[35] Harrison, P. J. (1965). Short-tem sales forecasting. Applied Statistics. **15**, 102 - 139.

[36] Harrison, P. J., and Stevens, C. F. (1976). Bayesian Forecasting (with discussion). J. Roy. Statist. Soc. (Ser B). **38**, 205 - 247.

[37] Harrison, P. J., and West, M. (1986). Bayesian forecasting in practice. Bayesian Statistics Study Year Report 13, University of Warwick.

[38] Harrison, P. J., and West, M. (1987). Practical Bayesian Forecasting. The Statistician **36**, 115 - 125.

[39] Hastings, W. K. (1970). Mont Carlo simulation methods using Markov chains and their applications. Biometrika. **57**, 97 - 109.

[40] Howard, R. A., and Matheson, J. E. (1981). Influence diagrams. In R. A. Howard and J. E. Matheson (eds.). Reading on the Principles and Applications of Decision Analysis, Vol II. Strategic Decisions Group, Menlo Park, Calif., 719 - 762.

[41] Jensen, F. V. (1988). Junction trees and decomposable hypergraphs, Research Report, Juder. Datasystemer ALS. Aalborg, Denmark.

[42] Kailath, T. (1967). The divergence and Bahttacharyya distance measures in signal selection. IEEE Trans. Commun. Tech. **COM-15**, 52-60.

[43] Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. J. of Basic Engineering, **82**, 35 - 45.

[44] Kim, J., and Pearl, J. (1983). A computational model for combined causal and diagnostic reasoning in inference systems. Proc. 8th International Conference on Artificial Intelligence, 190 - 193.

[45] Kjaerulf, U. (1992). Approximation of Bayesian networks through edge removals. Research Report, Department of Mathematics and Computer Science, Aalborg University, Denmark.

[46] Kliveri, H., Speed, T. P., and Carlin, J. B. (1984). Recursive causal models. J. Austral Math. Soc. (Ser A). **36**, 30 - 51.

[47] Kullback, S. (1968). Information Theory and Statistics. Wiley, New York.

[48] Lauritzen, S. L. (1992). Propagation of Probabilities, means and variances in mixed graphical association models. J. Amer. Statist. Assoc. **87**, 1098 - 1108.

[49] Lauritzen, S. L., Speed, T. P., and Vijayan, K. (1984). Decomposable graphs and hypergraphs. J. Austral. Math. Soc. (Ser. A), **36**, 12 - 29.

[50] Lauritzen, S. L., and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion), J. R. Statist. Soc. (Ser.B) **50**, 157 - 224.

[51] Lauritzen, S. L., and Wermuth, N. (1987). Mixed interaction models. Research Report R 84 - 8, Institute for Elektroniske Systems, Aalborg Universitets center, Denmark.

[52] Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimate for the linear model. J. Roy. Statist. Soc. B., **34**, 1 - 18.

[53] Mardia, K. V., Kent, J. T., and Bibby, J. M.(1979). Multivariate Analysis. London: Academic Press.

[54] Matusita, K. (1976). On the notion of affinity of several distributions and some of its applications. Ann. Inst. Statist. Math., **19**, 181 - 192.

[55] Mehra R. K. (1979). Kalman filters and their applications to forecasting, TIMS Studies in the Management Sciences **12**, 75 - 94, North-Holland Publishing Company.

[56] Mengresen, K. L., and Robert, C. P. (1993). Testing for mixtures: a Bayesian entropic approach. Doc. Travail No. 9340, Crest, Insee.

[57] Metropolis, N., Rosenbluth, A. W., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machine. J. Chem. Phys., **21**, 1087 - 1091.

[58] Mikkelsen, T., Larsen, S. E., and Thykier-Nielsen, S. (1984). Description of RISO Puff Diffusion Model. Nucl. Safety, **67**, 56 - 65.

[59] Mikkelsen, T., and Thykier-Nielsen, S. (1987). Atmospheric Dispersion over Complex Terrain. In: Proc. from the U.S. Army Atmospheric Sciences Workshop on Mesoscale Meteorology, RISO National Laboratory Roskild, Denmark, 103 - 110.

[60] Mikkelsen, T., Thykier-Nielsen, S., Larsen, S. E., Troen, I., De Bass, A. F., Kamada, R., Skupniewicz, C., and Schacher, G. A. (1989). Models for Accidental Releases In Complex Terrain. In: Air Pollution Modelling and its Application VII, Proc. of the 17th NATO/CCMS international meeting on Air Pollution Modelling and its Application, Cambridge, UK, 65 - 76. (Plenum Press New York).

[61] Miller, A. C., Merkhofer, M. W., Howard, R. A., Matheson, J. E. and Rice, T. R. (1976). Development of automated aids for decision analysis, Stanford Research Institute, Menlo Park, Calif.

[62] Morris, C. N. (1987). Comment on the calculation of posterior distributions by data augmentation, by M. A. Tanner and W. H. Wong. J. Amer. Statist. Assoc. **82**, 542 - 543.

[63] Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalised linear models. J. Roy. Statist. Soc. (Ser.A). **135**, 370 - 384.

[64] Olmsted, S. M. (1983). On representing and solving decision problems. Ph.D thesis, Engineering Economic systems Department, Stanford, Calif.

[65] Pasler-Sauer, J. (1985). Atmospheric Dispersion in Accident Consequence Assessments. Present modelling, future needs and comparative calculations. Proceedings of the workshop on methods for assessing the off-site consequences of nuclear accidents, Luxembourg, April 15 -19,1985, CEC, EUR-Report.

[66] Pearl, J. (1986). Fusion, Propagation and Structuring in Belief Networks. AI Journal, **29** (3), 241 - 288.

[67] Pearl, J., and Paz, A. (1985). GRAPHOIDS: a graph-based logic for reasoning about relevance relations, UCLA Computer Science Department, Technical Report 8500038 (R-53), October, also, Proceedings, EC AI-86, Brighton UK, June 86.

[68] Pearl, J., and Verma, T. (1987). The logic of representing dependencies by directed acyclic graphs, Proc. AAA-I, Seattle, Wash., July, 374 - 79.

[69] Pole, A. (1988). Transfer response models: a numerical approach. In Bayesian Statistics *3*, J. M. Bernardo, M. H. De Groot, D. V. Lindley. and A. F. M. Smith (eds). Oxford University Press.

[70] Pole, A., and West, M. (1988). Efficient numerical integration in dynamic models. Warwick Research Report 136, Department of Statistics University of Warwick.

[71] Queen, C. M. (1991). Bayesian Graphical Forecasting Models for Business Time Series. Ph.D thesis, Department of Statistics, University of Warwick.

[72] Queen, C. M., and Smith, J. Q. (1993). Multiregression Dynamic Models. J. R. Statist Soc. B. 55, No 4, 849-870.

[73] Rao, K. S., and Hosker, R. P. (1993). Uncertainty in the assessment of atmospheric concentrations of toxic contaminations from an accidental release. Radiation Protection Dosimetry. Vol. **50** Nos 2 -4, 281 - 288. Nuclear Technology Publishing.

[74] Reiss, R. J. (1989). Approximate Distributions of Order Statistics. Spring-Verlag New York.

[75] Shachter, R. D. (1986). Evaluating influence diagrams. Oper. Res., **34** (6), 871 - 882.

[76] Shachter, R. D. (1988). Probabilistic inference and influence diagrams. Oper. Res., **36**, (4), 589 - 604.

[77] Shaw, J. E. H. (1987). A strategy for reconstructing multivariate probability distributions. Research Report 123, Department of Statistics, University of Warwick.

[78] Shaw, J. E. H. (1988). Aspects of numerical integration and summation. In Bayesian Statistics *3*, J. M. Bernardo, M. H. De Groot , D. V. Lindely and A. F. M. Smith. (eds). Oxford University Press.

[79] Smith, A. F. M. (1991). Bayesian computational methods. Phil. Trans. R. Soc. Lond. A(1991). **337**, 369 - 386.

[80] Smith, A. F. M., and Roberts, G. O. (1992). Bayesian computation via Gibbs and related Markov chain Mont Carlo methods (with discussion). J. Roy. Statist. Soc. (Ser B). **55**, 3 - 24.

[81] Smith, J. Q. (1988). Decision Analysis: A Bayesian Approach, Chapman and Hall, London.

[82] Smith, J. Q. (1992). A comparison of the characteristics of some Bayesian forecasting models. International Statistical Review, **1**, 75 - 87.

[83] Smith J. Q. (1995a). Handling Multiple sources of variation using influence diagrams. European Journal of Operational Research, **86**, 189 - 200.

[84] Smith, J. Q. (1995b). Bayesian models and the Hellinger metric. University of Warwick, Department of Statistics, Research Report 279.

[85] Smith, J. Q., French, S. (1993). Bayesian updating of atmospheric dispersion models for use after an accidental release of radioactivity, The Statistician, **42**, 501-511.

[86] Smith, J. Q. French, S., and Ranyard, D. (1995). An efficient graphical algorithm for updating the estimates of the dispersal of gaseous waste after an accidental release. Proceedings of Adaptive Computing and Information Processing. Unicom Seminar Ltd., 583 - 610.

[87] Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L., and Cowell, R. G. (1993). Bayesian Analysis in Expert Systems. Statistical Science, Vol. **8**, No. 3, 219 - 283.

[88] Tanner M., and Wong, W. (1987). The calculation of posterior distributions by data augmentation (with discussion). J. Amer. Statist. Assoc., **82**, 528 - 550.

[89] Tarjan, R. E., and Yannakakis, M. (1984). Simple linear-time algorithms to test chordality of graphs, text acyclicity of hypergraphs, and selectively reduce acyclic hypergraphs. SIAM J. Comput., **13**, 566 - 579.

[90] Thomas, A., Spiegelhalter, D. J. and Gillks, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In Bayesian Statistics *4*, J. M. Bernardo, J. O, Berger, A. P. Dawid, and A. F. M. Smith (eds), 837–42, Clarendon Press, Oxford.

[91] Thykier-Neilsen, S. and Mikkelsen, T. (1991). RIMPUFF User Guide: Version 30. (National Laboratory, Roskilde, Germany).

[92] Wermuth, N., and Lauritzen, S. L. (1983). Graphical and recursive models for contingency tables. Biometrika, **70**, 537 - 557.

[93] West, M. (1992). Modelling with mixtures. In Bayesian Statistics *4*, J. M. Bernardo, J. O. Berger, A. P. Dawid. and A. F. M. Smith (eds). Oxford University Press.

[94] West, M., and Harrison, P. J. (1989). Bayesian Forecasting and Dynamic Linear Models. Spring-Verlag.

[95] West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalised linear models and Bayesian forecasting (with discussion). J. Amer. Statist. Assoc., **80**, 73 - 97.

[96] Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. John Wiley & Sons, Inc., Chichester.