Nuclear Inst. and Methods in Physics Research, A 1040 (2022) 167270

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, A

journal homepage: www.elsevier.com/locate/nima

Custom Orthogonal Weight functions (COWs) for event classification

Hans Dembinski [a], Matthew Kenzie [b],*, Christoph Langenbruch [c], Michael Schmelling [d]

[a] TU Dortmund, Germany
[b] University of Warwick, United Kingdom
[c] RWTH Aachen, Germany
[d] Max Planck Institute for Nuclear Physics, Heidelberg, Germany

ARTICLE INFO

Keywords:
Data analysis
Statistics
High energy physics
Event classification

ABSTRACT

A common problem in data analysis is the separation of signal and background. We revisit and generalise the so-called *sWeights* method, which allows one to calculate an empirical estimate of the signal density of a control variable using a fit of a mixed signal and background model to a discriminating variable. We show that *sWeights* are a special case of a larger class of Custom Orthogonal Weight functions (COWs), which can be applied to a more general class of problems in which the discriminating and control variables are not necessarily independent and still achieve close to optimal performance. We also investigate the properties of parameters estimated from fits of statistical models to *sWeighted* data and provide closed formulas for the asymptotic covariance matrix of the fitted parameters. To illustrate our findings, we discuss several practical applications of these techniques.

1. Introduction

This article takes a fresh look at the *sWeights* (or *sPlot*) formalism discussed by Barlow [1] and popularised more recently by Pivk and Le Diberder [2]. The *sWeights* method is used to infer properties of a signal distribution in a mixed data set containing signal and background events. The signal distribution is extracted non-parametrically by applying weights to individual events. Inference is then done on the weighted data set. The method is applicable, when individual points from the data distribution consist of a discriminating variable(s), here called m, and one or more statistically independent control variables, here called t. Both m and t can be vectors and of different dimensions without changing any of the conclusions. We will only refer to the one-dimensional case to simplify the discussion, but the general case is always implied. By fitting parametric models to the signal and background in the discriminating variable m, one can compute the weighted distribution that represents the signal density in the control variable t. The advantage of this method, compared to a fully parametric fit to the (m, t) distribution, is that one avoids the need to parameterise the background density in the control variable t, which is often challenging.

In Section 2 we re-derive the established *sWeights* method from the starting point of orthonormal functions. We show several ways of calculating the weights and compare their trade-offs, and emphasise that *sWeights* can easily be computed without some of the restrictions seen previously.

In Section 3 we then discuss a generalisation of the *sWeights* method which we dub *Custom Orthogonal Weight functions* (COWs). COWs relax most of the requirements of the *sWeights* formalism and can be applied to a larger class of problems than *sWeights*, at a small loss in precision.

In Section 4 we then discuss the properties of estimates obtained when fitting models to weighted data. We give an asymptotically correct formula for the covariance matrix of the parameters obtained from such a fit.

Finally in Section 5 we perform a variety of studies on simulated Monte Carlo which deploy *sWeights* and COWs on various applications and show comparisons of their performance. We also discuss a test of independence that can be used to determine if *sWeights* are applicable or whether the more general COWs method is needed.

2. *sWeights* as orthonormal functions

To compute the weights for the signal distribution in the control variable t, we use a discriminant variable m (often the invariant mass of some particle's decay products). The signal and background density only need to be parameterised in the discriminant variable m. In the *sWeights* formalism, the variables m and t must be statistically independent in each component, so that the respective p.d.f.s of the variables factorise. The total p.d.f. then has the following form

$$f(m, t) = z\, g_s(m)\, h_s(t) + (1 - z)\, g_b(m)\, h_b(t), \tag{1}$$

* Corresponding author.
E-mail addresses: hans.dembinski@tu-dortmund.de (H. Dembinski), matthew.kenzie@cern.ch (M. Kenzie), christoph.langenbruch@cern.ch (C. Langenbruch), michael.schmelling@mpi-hd.mpg.de (M. Schmelling).

https://doi.org/10.1016/j.nima.2022.167270
Received 27 January 2022; Received in revised form 24 June 2022; Accepted 13 July 2022
Available online 23 July 2022

where z is the signal fraction, $g_s(m)$ and $h_s(t)$ are the signal p.d.f.s in the discriminating and control variables, respectively, and $g_b(m)$ and $h_b(t)$, the corresponding background p.d.f.s. The *sWeights* method allows one to obtain an asymptotically efficient non-parametric estimate of $z\,h_s(t)$ while only requiring parametric models for $g_s(m)$ and $g_b(m)$.

We stress that the *sWeights* method is only applicable when the p.d.f.s in m and t factorise for both the signal and the background, which is conditional on their independence. Independence is a stronger condition than lack of correlation; tests which demonstrate a lack of correlation between m and t provide necessary, but not sufficient, evidence for the applicability of the *sWeights* method. We come back to proper tests of independence in Section 5.

2.1. Construction of an optimal weight function

We postulate that a weight function, $w_s(m)$, exists which extracts the signal component, $z\,h_s(t)$, when $f(m,t)$ is multiplied by it and integrated over m[1]:

$$\begin{aligned} z\,h_s(t) &\overset{!}{=} \int \mathrm{d}m\; w_s(m)\, f(m,t) \\ &= \int \mathrm{d}m\; w_s(m)\left[z\,g_s(m)\,h_s(t) + (1-z)\,g_b(m)\,h_b(t)\right] \\ &= z\,h_s(t) \int \mathrm{d}m\; w_s(m)\,g_s(m) + (1-z)\,h_b(t) \int \mathrm{d}m\; w_s(m)\,g_b(m). \end{aligned} \tag{2}$$

The left and the right-hand sides of Eq. (2) are equal in general only if the following conditions hold:

$$\int \mathrm{d}m\; w_s(m)\,g_s(m) = 1 \text{ and } \int \mathrm{d}m\; w_s(m)\,g_b(m) = 0. \tag{3}$$

If we regard $\int \mathrm{d}m\,\phi(m)\,\psi(m)$ as the inner product of a vector space over functions, then these conditions define $w_s(m)$ as the vector orthogonal to $g_b(m)$ and normal to $g_s(m)$. In other words, $w_s(m)$ is an orthonormal function in this space.

Since the vector space over m is infinite-dimensional, there are infinitely many orthonormal functions $w_s(m)$ that satisfy these conditions. For example, the classic sideband subtraction method can be regarded as a special case where $w_s(m)$ is a piece-wise constant function which is positive in the signal region and negative in the background region.

In order to obtain a unique solution for $w_s(m)$ we can chose to minimise its variance. Since $f(m,t)$ factorises and $w_s(m)$ is only a function of m, we can obtain all information about w_s from the density $g(m)$, computed by integrating Eq. (1) over t,

$$g(m) = \int \mathrm{d}t\; f(m,t) = z\,g_s(m) + (1-z)\,g_b(m). \tag{4}$$

The expectation of w_s over $g(m)$ is

$$\mathrm{E}[w_s] = \int w_s(m)\,g(m)\,\mathrm{d}m\; = z, \tag{5}$$

and the variance of w_s over $g(m)$ is given by

$$\mathrm{Var}(w_s) = \mathrm{E}[w_s^2] - \mathrm{E}[w_s]^2 = \int w_s(m)^2\,g(m)\,\mathrm{d}m\; - z^2. \tag{6}$$

Minimising the variance, $\mathrm{Var}(w_s)$, guarantees that the sample estimate $\hat{z} = 1/N \sum_i^N \hat{w}_s(m_i)$ asymptotically has minimum variance. As a byproduct, this choice also produces minimum variance for the estimated background fraction $(1-\hat{z})$, and generally smooth functions, $w_s(m)$, since oscillating solutions have larger variance.

To find the function $w_s(m)$ which minimises $\mathrm{Var}(w_s)$, we have to solve a constrained minimisation problem. The solution, computed in Appendix A, is

$$w_s(m) = \frac{\alpha_s\,g_s(m) + \alpha_b\,g_b(m)}{g(m)}. \tag{7}$$

The constants $\alpha_{s,b}$ are obtained by inserting Eq. (7) into Eq. (3) and solving the resulting system of linear equations. The signal component plays no special role in the derivation so far. We could have equally postulated a weight function $w_b(m)$ to extract the background, which leads to the conditions

$$\int \mathrm{d}m\; w_b(m)\,g_s(m) = 0 \tag{8}$$

$$\int \mathrm{d}m\; w_b(m)\,g_b(m) = 1, \tag{9}$$

and

$$w_b(m) = \frac{\beta_s\,g_s(m) + \beta_b\,g_b(m)}{g(m)}. \tag{10}$$

The coefficients α_c and β_c with $c \in \{s,b\}$ can be computed by solving

$$\underbrace{\begin{pmatrix} W_{ss} & W_{sb} \\ W_{sb} & W_{bb} \end{pmatrix}}_{\boldsymbol{W}} \cdot \underbrace{\begin{pmatrix} \alpha_s & \beta_s \\ \alpha_b & \beta_b \end{pmatrix}}_{\boldsymbol{A}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \tag{11}$$

with

$$W_{cd} = \int \mathrm{d}m\; \frac{g_c(m)\,g_d(m)}{z g_s(m) + (1-z) g_b(m)}, \tag{12}$$

where $c,d \in \{s,b\}$. In other words, the matrix $\boldsymbol{A}$, formed by the coefficients to compute $w_s(m)$ and $w_b(m)$, is the inverse of the symmetric positive-definite $\boldsymbol{W}$ matrix. The discussion throughout this section assumes just two components (one signal and one background) but is equally applicable to any number of components, N, in which case the $\boldsymbol{W}$ and $\boldsymbol{A}$ matrices are not 2×2 but $N \times N$.

Applying Cramer's rule to Eq. (11), we get

$$\alpha_s = \frac{W_{bb}}{W_{ss}W_{bb} - W_{sb}^2}, \qquad \alpha_b = \frac{-W_{sb}}{W_{ss}W_{bb} - W_{sb}^2}, \tag{13}$$

$$\beta_s = \frac{-W_{sb}}{W_{ss}W_{bb} - W_{sb}^2,} \qquad \beta_b = \frac{W_{ss}}{W_{ss}W_{bb} - W_{sb}^2}. \tag{14}$$

One can further replace $g(m)$ in the denominator of Eq. (7) (or Eq. (10)) by inserting Eq. (7) into Eq. (5) to find that $z = \alpha_s + \alpha_b$, and similarly one finds $1 - z = \beta_s + \beta_b$. With these ingredients, we obtain the final equations

$$w_s(m) = \frac{W_{bb}\,g_s(m) - W_{sb}\,g_b(m)}{(W_{bb} - W_{sb})\,g_s(m) + (W_{ss} - W_{sb}) g_b(m)}, \tag{15}$$

$$w_b(m) = \frac{W_{ss}\,g_b(m) - W_{sb}\,g_s(m)}{(W_{bb} - W_{sb})\,g_s(m) + (W_{ss} - W_{sb}) g_b(m)}. \tag{16}$$

In summary, to obtain $w_s(m)$ or $w_b(m)$ one has to compute the matrix elements W_{ss}, W_{sb}, W_{bb}, which depend only on $g_{s,b}(m)$ and z.

2.2. Application to finite samples

The calculations so far were carried out for the true p.d.f.s, $g_{s,b}(m)$, and true signal fraction, z, on which the matrix elements W_{cd} depend. In practice, these need to be replaced by sample estimates $\hat{g}_{s,b}(m)$ and $\hat{z}$, typically obtained from a maximum-likelihood fit, although any kind of estimation can be used. The plug-in estimate [3] of Eq. (15) is

$$\hat{w}_s(m) = \frac{\widehat{W}_{bb}\,\hat{g}_s(m) - \widehat{W}_{sb}\,\hat{g}_b(m)}{(\widehat{W}_{bb} - \widehat{W}_{sb})\,\hat{g}_s(m) + (\widehat{W}_{ss} - \widehat{W}_{sb})\,\hat{g}_b(m)}. \tag{17}$$

For the computation of the estimates $\widehat{W}_{cd}$ we face a choice between several options.

- *Variant A:* We replace the true quantities in Eq. (12) with their plug-in estimates and compute the integral analytically or numerically,

$$\widehat{W}^A_{cd} = \int \mathrm{d}m\; \frac{\hat{g}_c(m)\,\hat{g}_d(m)}{\hat{z}\,\hat{g}_s(m) + (1-\hat{z})\,\hat{g}_b(m)}. \tag{18}$$

Solving the one-dimensional integral numerically is not an issue; this is a standard problem for which efficient and robust algorithms exist. The computation is independent of the number of

[1] Throughout this paper the symbol, $\overset{!}{=}$, is used to indicate that the equation should be solved.

data points. A numerical integration typically requires around 100 to 1000 function evaluations.

If the p.d.f.s $g_c(m)$ have no shape parameters that need to be estimated from the data, we have $\hat{g}_c(m) = g_c(m)$ and only $\hat{z}$ has to be estimated. In this special case, it is shown in Appendix C that the sum over *sWeights* in a bin k of t are uncorrelated and their variance is estimated by the sum of weights squared. In practice, however, the component p.d.f.s are often not known, and thus in general, the sums of weights in different bins are not uncorrelated and the sum of weights squared is not a proper estimate of the bin-wise variance. Calculation of the covariance then requires the sandwich estimator described in Appendix C.

- *Variant B*: The integral in Eq. (18) can be replaced by a sum over the observations in the data sample. In general, an integral over a function $\phi(m)$ can be written as an expectation value over the p.d.f. $g(m)$,

$$\int \mathrm{d}m\ \phi(m) = \int \mathrm{d}m\ g(m) \frac{\phi(m)}{g(m)} = \mathrm{E}\left[\frac{\phi(m)}{g(m)}\right]. \tag{19}$$

In a finite sample, the arithmetic mean is an unbiased estimate of the expectation due to the law of large numbers, thus we can construct an unbiased estimator by replacing the expectation with the arithmetic mean,

$$\mathrm{E}\left[\frac{\phi(m)}{g(m)}\right] \longrightarrow \frac{1}{N}\sum_i \frac{\phi(m_i)}{g(m_i)}, \tag{20}$$

where m_i is the ith observed value of m and N is the sample size. We obtain

$$\widehat{W}^B_{cd} = \frac{1}{N}\sum_i \frac{\hat{g}_c(m_i)\,\hat{g}_d(m_i)}{\left(\hat{z}\hat{g}_s(m_i) + (1-\hat{z})\hat{g}_b(m_i)\right)^2}, \tag{21}$$

which is the formula to compute *sWeights* given in Ref. [2]. We will refer to *sWeights* computed with variant B as *classic sWeights* throughout the paper.

The computation with variant B is straight-forward, there is no need for a numerical integration algorithm. For large samples that exceed 1000 items, variant A will in general be faster than variant B, but the difference is hardly noticeable in practice. Sums over *sWeights* in bins of t computed with Eq. (21) are always correlated, even if the component p.d.f.s are parameter-free (in contrast to variant A), as shown in Ref. [4].

In general, both variant A and B produce correlations between sums of *sWeights* in bins of t (a weighted histogram), which makes further analysis more complicated. Ref. [2] states that such bins are uncorrelated and have simple variance estimates, but this is correct only under the special circumstances discussed in the following.

In the case of variant A, the correlations vanish only if the true shapes in the component p.d.f.s, $g_c(m)$, are known, which is almost never the case in practice. In the case of variant B, correlations are present even if the component p.d.f.s are known. The correlations are usually small, but do not vanish as the sample size grows.

In general, the covariance matrix for a weighted histogram has to be computed with a sandwich estimator, irrespective of whether variant A or B are used. Variant A has a slight advantage over B, since the computation of the sandwich estimator is a bit simpler. Sandwich estimators for binned and unbinned fits to *sWeighted* data are given in Ref. [4]. The sandwich estimator for an unbinned fit is described in Section 4 and an outline for the computation of a sandwich estimator for a weighted histogram is given in Appendix C. The computation of these sandwich estimators can be automated by software. Altogether, we recommend variant B for practical applications, since it produces estimates with smaller variance than variant A, as described in Section 4, but we also note that the difference between variant A and B is negligible in most toy examples that we tried.

Finally, we note that the sum over *sWeights* computed with either variant is exactly equal to the previously estimated signal yield $\hat{N}_s = N\hat{z}$ that is used in their calculation,

$$T = \sum_i \hat{w}_s(m_i) = N\hat{z} = \hat{N}_s. \tag{22}$$

The proofs are provided in Appendix D. For variant B this is generally true, and for variant A, it is true, if $\hat{N}_s$, $\hat{N}_b$, and the component p.d.f.s $\hat{g}_c(m)$ are estimated with the extended maximum-likelihood method, described in the next section.

2.3. Connection to extended maximum-likelihood fit

There is a curious connection between Eq. (21) and the results of an extended maximum-likelihood fit in which $\hat{g}_s$ and $\hat{g}_b$ are maximum-likelihood estimates and the respective signal and background yields, N_s and N_b, are regarded as independent variables [2]. In such a fit, one maximises the extended log-likelihood function [5] which, without constant terms, is

$$\ln\mathcal{L}(N_s, N_b) = -(N_s + N_b) + \sum_i \ln[N_s\,\hat{g}_s(m_i) + N_b\,\hat{g}_b(m_i)]. \tag{23}$$

The extremum is determined by solving the score functions

$$\frac{\partial \ln\mathcal{L}}{\partial N_c} = -1 + \sum_i \frac{\hat{g}_c(m_i)}{N_s\,\hat{g}_s(m_i) + N_b\,\hat{g}_b(m_i)} \overset{!}{=} 0, \tag{24}$$

with $c \in \{s, b\}$. The maximum-likelihood estimates obtained from these score functions are $\hat{N}_s = N\hat{z}$ and $\hat{N}_b = N(1-\hat{z})$, where $\hat{z}$ is the estimated signal fraction as before. The elements of the Hessian matrix, of second derivatives of the log-likelihood function, are given by

$$\frac{\partial^2 \ln\mathcal{L}}{\partial N_c\,\partial N_d} = -\sum_i \frac{\hat{g}_c(m_i)\,\hat{g}_d(m_i)}{\left(N_s\,\hat{g}_s(m_i) + N_b\,\hat{g}_b(m_i)\right)^2}. \tag{25}$$

We note the similarity between Eq. (25) and Eq. (21) and evaluate the second derivative at the maximum of $\ln\mathcal{L}$ to find

$$\begin{aligned} -\left.\frac{\partial^2 \ln\mathcal{L}}{\partial N_c\,\partial N_d}\right|_{N_s=N\hat{z},\,N_b=N(1-\hat{z})} &= \sum_i \frac{\hat{g}_c(m_i)\,\hat{g}_d(m_i)}{\left(N\,\hat{z}\,\hat{g}_s(m_i) + N\,(1-\hat{z})\,\hat{g}_b(m_i)\right)^2} \\ &= \frac{1}{N}\widehat{W}^B_{cd}. \end{aligned} \tag{26}$$

This offers another opportunity to compute estimates of W_{cd}, which was already pointed out in Ref. [2]. The second derivatives of the log-likelihood for $\hat{N}_s$, $\hat{N}_b$, and the shape parameters $\boldsymbol{\theta}_{s,b}$ of $\hat{g}_{s,b}(m;\boldsymbol{\theta}_{s,b})$ are routinely computed by MINUIT [6] by calling the routine HESSE to estimate the covariance matrix $\mathbf{C}$ of the parameters once a minimum is found by the routine MIGRAD. The matrix $\mathbf{C}$ obtained in this way is the negative inverse of the Hessian,

$$\mathbf{C}^{-1} = -\begin{pmatrix} \frac{\partial^2 \ln\mathcal{L}}{\partial N_s^2} & \frac{\partial^2 \ln\mathcal{L}}{\partial N_s \partial N_b} & \cdots \\ \frac{\partial^2 \ln\mathcal{L}}{\partial N_s \partial N_b} & \frac{\partial^2 \ln\mathcal{L}}{\partial N_b^2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}. \tag{27}$$

The dotted parts of the matrix correspond to derivatives that contain one or two shape parameters of $\boldsymbol{\theta}_{s,b}$.

Variant C to compute the elements of $\widehat{W}^C_{cd}$ consists of the following steps:

- Invert the covariance matrix $\mathbf{C}$ of the fit of yields $N_{s,b}$ and shape parameters $\boldsymbol{\theta}_{s,b}$.
- Isolate the 2 × 2 sub-matrix of the Hessian which contains the derivatives with respect to the yields $N_{s,b}$.
- Use Eq. (26) on these matrix elements to obtain $\widehat{W}^C_{cd}$.

It would be incorrect to switch steps 1 and 2, *i.e.* isolate the 2 × 2 sub-matrix of $\mathbf{C}$ that contains the yields and invert it, because this does not restore the derivatives.

An equivalent alternative is to do a second fit which leaves only the yields free while keeping shape parameters fixed. In this case, the covariance matrix computed by MINUIT can be scaled to yield an estimate of the coefficient matrix from Eq. (11):

$$\begin{pmatrix} \hat{\alpha}_s & \hat{\beta}_s \\ \hat{\alpha}_b & \hat{\beta}_b \end{pmatrix} = \frac{1}{N} \begin{pmatrix} C_{ss} & C_{sb} \\ C_{sb} & C_{bb} \end{pmatrix} . \tag{28}$$

Mathematically, variant B and C should produce the exact same result. In practice, this is not exactly true, because the Hessian matrix is not calculated analytically with Eq. (25). The second derivatives in Eq. (27) are instead computed approximately by numerical differentiation of Eq. (23). The accuracy of a numerically computed second derivative is of the order of $\epsilon^{1/3}$, where ϵ is the round-off error of floating point arithmetic on a computer ($\epsilon \approx 10^{-12}$ for double precision). The accuracy of the W matrix computed with variant C is much lower than one computed with variant B or A, so that Eq. (22) only holds approximately.

3. Custom orthogonal weight functions

So far we considered the special case where the p.d.f. is a mixture of two components that each factorise in both the discriminant and control variables. We now generalise to an arbitrary number of factorising components, and also allow for a non-factorising function of frequency weights, $\epsilon(m,t)$, which are usually identified with an efficiency. In High Energy Physics applications, such a function may arise from a non-uniform acceptance introduced by the detector or the selection requirements applied to the data in order to improve the signal-to-background ratio. The p.d.f. for the observed data then is

$$\rho(m,t) = \frac{1}{D}\epsilon(m,t) f(m,t) \quad \text{with} \quad D = \int \mathrm{d}m\, \mathrm{d}t\, \epsilon(m,t)\, f(m,t) \,. \tag{29}$$

The normalisation constant, D, ensures that $\rho(m,t)$ is a probability density. We expand the true density of the events into n factorising components,

$$f(m,t) = \sum_{k=0}^{n} z_k\, g_k(m)\, h_k(t) \quad \text{with} \quad \sum_{k=0}^{n} z_k = 1 \,. \tag{30}$$

The Kolmogorov–Arnold representation theorem [7,8] ensures that a finite sum of terms on the right-hand side can represent any two-dimensional function $f(m,t)$. The $g_k(m)$ and $h_k(t)$ are normalised probability densities. Normalised Bernstein [9] or B-spline [10] basis polynomials can fill this role in general. For practical applications, it is beneficial if the expansion requires only a few terms, which can be achieved with $g_k(m)$ and $h_k(t)$ suitably chosen for the specific case. For a given expansion, we will assume that the first s terms pertain to the signal density while the others describe the background, i.e.

$$f(m,t) = \underbrace{\sum_{k=0}^{s-1} z_k\, g_k(m)\, h_k(t)}_{\text{signal}} + \underbrace{\sum_{k=s}^{n} z_k\, g_k(m)\, h_k(t)}_{\text{background}} \,. \tag{31}$$

Each partial sum in general produces a non-factorising function. This makes it possible to generalise the previous results from Section 2.1 to non-factorising signal and background components.

Any single function $h_k(t)$ in $f(m,t)$ can be isolated by a weight function

$$w_k(m) = \sum_{\ell=0}^{n} \frac{A_{k\ell}\, g_\ell(m)}{I(m)} \text{ with } A = W^{-1} \text{ and } W_{k\ell} = \int \mathrm{d}m\, \frac{g_k(m)\, g_\ell(m)}{I(m)} \,. \tag{32}$$

The function $I(m)$ is an arbitrary non-zero function in the considered range of m, this is another generalisation with respect to classic *sWeights*. We dub it the *variance function*, because the optimal function $I(m)$ corresponds to the point-wise variance of a density estimate (as we will see later). It is straight-forward to show with $\sum_i A_{ki}\, W_{ij} = \delta_{kj}$ that weight functions defined in this way are orthonormal to the component p.d.f.s $g_k(m)$,

$$\int \mathrm{d}m\, w_k(m)\, g_\ell(m) = \delta_{k\ell} \,. \tag{33}$$

The *Custom Orthogonal Weight function* (COW) to extract the density $z_k\, h_k(t)$ from data, is then

$$w'_k(m,t) = D \frac{w_k(m)}{\epsilon(m,t)} \,, \tag{34}$$

which is now a function of both m and t. The expectation value of the COW in an infinitesimal bin, $\mathrm{d}t$, in the control variable is

$$\frac{\mathrm{d}}{\mathrm{d}t}\, \mathrm{E}\left[D \frac{w_k(m)}{\epsilon(m,t)} \right] = z_k\, h_k(t), \tag{35}$$

so it is an asymptotically unbiased estimate of the efficiency corrected density, $h_k(t)$, for any choice $I(m)$. The COWs that project out the entire signal or background component are given by

$$w'_s(m,t) = \sum_{k=0}^{s-1} w'_k(m,t) \quad \text{and} \quad w'_b(m,t) = \sum_{k=s}^{n} w'_k(m,t) \,. \tag{36}$$

By integrating Eq. (35) over t, one sees that $\mathrm{E}[w'_k] = z_k$. An estimate of z_k is

$$\hat{z}_k = \frac{\hat{D}}{N} \sum_{i=1}^{N} \frac{w_k(m_i)}{\epsilon(m_i,t_i)} = \sum_{i=1}^{N} \frac{w_k(m_i)}{\epsilon(m_i,t_i)} \bigg/ \sum_{i=1}^{N} \frac{1}{\epsilon(m_i,t_i)} \,. \tag{37}$$

The estimate for D used here is derived from Eq. (29),

$$\begin{aligned} \frac{1}{D} &= \int \frac{f(m,t)}{D} \mathrm{d}m\, \mathrm{d}t = \int \frac{1}{\epsilon(m,t)} \rho(m,t)\, \mathrm{d}m\, \mathrm{d}t \\ &= \mathrm{E}\left[\frac{1}{\epsilon}\right] \longrightarrow \frac{1}{N} \sum_i \frac{1}{\epsilon(m_i,t_i)} \,. \end{aligned} \tag{38}$$

In analogy with Section 2, one also has to replace the true W matrix with an estimate. The estimates for COWs corresponding to variant A and B for *sWeights* are

$$\widehat{W}^A_{k\ell} = \int \mathrm{d}m\, \frac{\hat{g}_k(m)\, \hat{g}_\ell(m)}{\hat{I}(m)} \tag{39}$$

$$\widehat{W}^B_{k\ell} = \frac{1}{N} \sum_i \frac{\hat{g}_k(m_i)\, \hat{g}_\ell(m_i)}{\hat{\rho}_m(m_i)\, \hat{I}(m_i)}, \tag{40}$$

where all true functions are replaced with estimates and $\hat{\rho}_m(m)$ is an estimate of the observed density, $\rho_m(m) = \int \mathrm{d}t\, \rho(m,t)$, in the discriminant variable. Since $\widehat{W}^B_{k\ell}$ for COWs is more cumbersome to estimate, we consider only $\widehat{W}^A_{k\ell}$ in the following and omit the distinction.

In contrast to *sWeights*, the sum of COWs is not unity in general, even if $\epsilon(m,t) = 1$. This is the case only if $I(m)$ is a linear combination of the basis functions, $g_k(m)$. One finds for arbitrary constants a_k

$$\sum_{k=0}^{n} w_k(m) = 1 \quad \text{if} \quad I(m) = \sum_{k=0}^{n} a_k\, g_k(m) \,, \tag{41}$$

as shown in Appendix E. When $I(m)$ takes this form, every event contributes with a total weight of unity to the possible states k. This property is not of practical relevance, however. A corollary of this result is that with an increasing number of terms the sum $\sum_k w_k(m)$ will converge towards unity for any function $I(m)$, since a linear combination of sufficiently many basis functions $g_k(m)$ always allows for a good approximation of $I(m)$.

We now discuss an optimal choice for the variance function $I(m)$. We can find functions, such that

(I) the variances of the $\hat{z}_k$ are minimal,
(II) the $\hat{z}_k$ are maximum-likelihood estimates.

As shown in Appendix F, requirement I leads to

$$I(m) = q(m) = \int \mathrm{d}t\, \frac{\rho(m,t)}{\epsilon^2(m,t)} \,. \tag{42}$$

This choice minimises the variances of the $\hat{z}_k$ and is therefore optimal. An estimate of $q(m)$ can be obtained by a histogram of the $1/\epsilon^2(m,t)$ weighted m-distribution or by fitting a suitable parameterisation to it. The extreme case of a single-bin histogram is equivalent to $I(m) = 1$ (the scale of $I(m)$ is irrelevant). The exact details of the binning are not important, since a coarse binning will merely increase the variance of $\hat{z}_k$ above the minimum possible.

Appendix G shows that the alternative requirement II leads to

$$I(m) = \sum_{k=0}^{n} \hat{z}_k \, g_k(m) \,, \tag{43}$$

where the $\hat{z}_k$ are estimates for the true fractions z_k obtained from an $1/\epsilon(m,t)$ weighted unbinned maximum-likelihood fit. In general, this variance function is different from $q(m)$ and not optimal.

In the special case of constant efficiency, equivalent to setting $\epsilon(m,t) = 1$, the two variance functions converge asymptotically,

$$I_{\mathrm{I}}(m) = q(m) = \sum_{k=0}^{n} z_k \, g_k(m) \quad \text{and} \quad I_{\mathrm{II}}(m) = \sum_{k=0}^{n} \hat{z}_k \, g_k(m). \tag{44}$$

Computing *sWeights* as described in Section 2 is equivalent to using $I_{\mathrm{II}}(m)$, which is asymptotically optimal in this case.

3.1. Mismodelled signal density

A common task is to extract a single signal component that factorises in m and t, contaminated with a background that is not factorising. To construct the signal-extracting COW, w_0', one requires a model for the signal density, $g_0(m)$ according to Eq. (32), and a set of background p.d.f.s $g_k(m)$ (where $k = 1, \dots, n$) and a variance function, $I(m)$. Often, the signal shape $g_0(m)$ is a non-trivial function containing a number of nuisance parameters. We now investigate what happens if $g_0(m)$ is mismodelled and does not match the true signal density.

We distinguish between the true p.d.f.s, $g_k(m)$, and their mismodelled proxy p.d.f.s, $G_k(m)$, which are used in the calculation of COWs, $w_k(m)$. If we review the mathematical steps in the previous section, we find that $G_k(m) = g_k(m)$ is not required for the construction of COWs. The construction steps and the properties of the A and W matrices remain the same if $G_k(m) \neq g_k(m)$. We only get a different result if we integrate over the product of the total density $\rho(m,t)$ and the COWs.

We can write the expected signal weight in an infinitesimal bin of width dt in the control variable as

$$\begin{aligned}\frac{\mathrm{dE}[w_0']}{\mathrm{d}t} &= \int \rho(m,t) D \frac{w_0(m)}{\epsilon(m,t)} \,\mathrm{d}m \\ &= \sum_{k=0}^{n} z_k \, h_k(t) \int g_k(m) \, w_0(m) \mathrm{d}m \\ &= \sum_{k=0}^{n} z_k \, h_k(t) \sum_{\ell=0}^{n} A_{0\ell} \int \frac{g_k(m) \, G_\ell(m)}{I(m)} \mathrm{d}m\end{aligned} \tag{45}$$

Since we only consider the signal p.d.f. to be mismodelled, we have $G_k(m) = g_k(m)$ for $k \geq 1$, and

$$\int \frac{g_k(m) \, G_\ell(m)}{I(m)} \mathrm{d}m = \int \frac{G_k(m) \, G_\ell(m)}{I(m)} \mathrm{d}m = W_{k\ell} \text{ for } k \geq 1.$$

We use this and the symmetry of W_{kl} to find

$$\begin{aligned}\frac{\mathrm{dE}[w_0']}{\mathrm{d}t} &= z_0 \, h_0(t) \left[\sum_{\ell=0}^{n} A_{0\ell} \int \frac{g_0(m) \, G_\ell(m)}{I(m)} \mathrm{d}m \right] \\ &\quad + \sum_{k=1}^{n} z_k \, h_k(t) \underbrace{\sum_{\ell=0}^{n} A_{0\ell} W_{\ell k}}_{\delta_{0k}} \\ &\propto h_0(t) \,,\end{aligned} \tag{46}$$

since the first sum in the square bracket is a constant independent of t, and the second sum is zero.

Therefore, a mismodelling of the signal component $G_0(m)$ in the discriminatory variable introduces no bias for the estimation of $h_0(t)$, although the method is then no longer optimal in the previous sense. This offers an intriguing possibility in practice. If one is only interested in projecting out the normalised signal p.d.f., $h_0(t)$, there is great freedom in the construction of the p.d.f. $G_0(m)$ and the variance function $I(m)$ in the discriminating variable. Any p.d.f. works for $G_0(m)$ which is not a linear combination of background p.d.f.s $g_k(m)$ with $k \geq 1$. The best results are nevertheless obtained with a $G_0(m)$ that is as close to $g_0(m)$ as possible, and as few basis polynomials for the background as possible. Poorly chosen functions $I(m)$ and $G_0(m)$ will increase the variance of the estimate $\hat{h}_0(t)$.

3.2. COWs in the wild

In this section, we remark on points that are important for the practical applications of COWs, which arose in discussions on this method.

Firstly, we emphasise the important consequence for analyses in particle physics that follows from the previous section. If the signal factorises in m and t, a COW for the signal can be computed without a fit. We refer to this variant as *COWs lite*, to contrast it with *full COWs* where both signal and background are non-factorising. A histogram of the simulated signal distribution in m obtained from simulation can be used for $G_0(m)$, the proxy p.d.f. for the signal density, and a histogram of the signal and background distribution $g(m)$ in the simulation can be used for the variance function $I(m)$. Even if the simulation does not describe the real experiment perfectly, using these proxies does not create a bias and the substitutes will usually be close to optimal. The optimal variance function $q(m)$ can be estimated from data with a $1/\epsilon^2(m,t)$-weighted histogram of the m-distribution. The details of the binning are not important provided bins are not empty. A non-optimal binning will increase the variance of the signal COW above the minimum possible, but the influence is very weak and therefore it is not necessary to carefully optimise the binning. The extreme case of a single-bin histogram is equivalent to $I(m) = 1$ (the scale of $I(m)$ is irrelevant) and also a valid choice.

A good description of the background under the signal is however crucial to avoid bias. The background can be expanded generally into Bernstein or B-spline polynomials, which can approximate any p.d.f. with enough terms. A systematic bias related to the choice of the background model, *i.e.* incurred by not having sufficient terms in the background description, can be probed by adding more terms and verifying with goodness-of-fit tests (see *e.g.* Ref. [11]) that the model is sufficient to describe the m distribution. For a chi-square test statistic in a binned fit, the Fisher F-test indicates whether adding further terms improves the model significantly. When two models are tested on a histogram with n bins, where model 1 has k_1 terms and model 2 has $k_2 > k_1$ terms, the statistic

$$F = \frac{(\chi_1^2 - \chi_2^2)/(k_2 - k_1)}{\chi_2^2/(n - k_2)} \tag{47}$$

is asymptotically F distributed. Cross-validation [12] is another option if the fit is unbinned.

Secondly, we want to give some insight into why it is beneficial to include the efficiency function $\epsilon(m,t)$ into the *sWeights* estimation, instead of trying to separate signal and background first and then correct for efficiency in a separate step. If the efficiency function is not directly included in the analysis, both signal and background are non-factorising because of the effect of the efficiency function. In the COWs framework, we can still extract *sWeights* by setting $\epsilon(m,t) = 1$, but now sufficiently many terms in the signal and background part of the data model are required to account for factorisation-breaking effects. This reduces the statistical power of the method and an additional complication arises. Once the signal density in the observed sample, $\tilde{h}_s(t)$, has been determined, the efficiency correction must be done with the signal efficiency projected into the control variable, $\bar{\epsilon}(t)$, to obtain the true density $h_s(t) \propto \tilde{h}_s(t)/\bar{\epsilon}(t)$. In weighted unbinned fits or when generating a histogram estimate of $h_s(t)$, one has to use $w_s(m_i)/\bar{\epsilon}(t_i)$, where i indexes the data points in the sample. Using $w_s(m_i)/\epsilon(m_i, t_i)$ as an event-by-event weight instead is wrong, since the m-dependence in the efficiency factor destroys the orthogonality relations for the COW, and the signal estimate in t becomes polluted by background. The projected efficiency is given by

$$\bar{\epsilon}(t) = \int \mathrm{d}m\, \epsilon(m,t)\, f_s(m,t)\;, \tag{48}$$

where $f_s(m,t)$ denotes the signal component of the true p.d.f., and is not straight-forward to estimate from the sample. One could take $\bar{\epsilon}(t)$ from simulation, with the caveat that the simulation may differ from the real experiment. If the efficiency function factorises, $\epsilon(m,t) = \epsilon_m(m)\,\epsilon_t(t)$, we get

$$\bar{\epsilon}(t) = \epsilon_t(t) \left(\int \mathrm{d}m\, \frac{\rho_s(m)}{\epsilon_m(m)} \right)^{-1} \propto \epsilon_t(t)\;, \tag{49}$$

where $\rho_s(m) = \int \mathrm{d}t\, \rho_s(m,t)$ is the observed signal p.d.f. in m. This marginally simplifies the matter, since $w_s(m_i)/\epsilon_t(t_i)$ can be used instead of $w_s(m_i)/\bar{\epsilon}(t_i)$ in this case, if only the shape of $h_s(t)$ is of interest. Nevertheless, these additional complications make the two-step approach unfavourable.

4. Variance of estimates from weighted data

This section discusses how to correctly perform parameter uncertainty estimation in an unbinned fit of weighted data, when using *sWeights*. The more complex binned fit for this case is described in Ref. [4]. We will give explicit formulas for classic *sWeights* computed with variant B, as introduced in Section 2.2. The general approach we discuss here also applies to fits of weighted data obtained from variant A or the COWs method, but we will not give explicit formulas for the other variants for the sake of brevity, as each variant has its own varied uncertainty estimate. We will point out how the formulas have to be adapted, but leave the explicit calculations to the reader.

Parameter estimation using weighted unbinned data sets can be performed by maximising the weighted likelihood [13], which is equivalent to solving the weighted score functions

$$\sum_i w_i \frac{\partial \ln h_s(t_i;\boldsymbol{\phi})}{\partial \phi_k} \overset{!}{=} 0, \tag{50}$$

with $w_i = w_s(m_i)$ in case of *sWeights* or $w_i = w_s'(m_i, t_i)$ in case of COWs and shape parameters $\boldsymbol{\phi}$ of the signal p.d.f. $h_s(t;\boldsymbol{\phi})$. The weighted likelihood is not a true likelihood (product of probabilities) and so the inverse of the Hessian matrix [13] of the weighted likelihood does not asymptotically provide an estimate of the covariance matrix of the parameters. Eq. (50) is an example of an *M-estimator* [14]. A complete derivation of the asymptotic covariance matrix for the parameters $\boldsymbol{\phi}$ can be found in the appendix of Ref. [4], here we only summarise the main findings.

A complication arises due to the fact that the *sWeights* depend, via Eq. (17), on the elements of the W matrix, which is determined via Eq. (21). The estimates $\widehat{W}_{cd}$ in turn depend on the estimates of the signal and background yields, $\hat{N}_s$ and $\hat{N}_b$, usually determined from an extended maximum likelihood fit. Problems of this type are described as *two-step M-estimation* in the statistical literature [15,16]. To account for the fact that the parameters are estimated from the same data sample and are not independent, one has to combine the estimating equations for the parameters of interest with those of the yields and the inverse covariance matrix elements in a single vector.

We construct the quasi-score function $\boldsymbol{S}(\lambda)$, where $\lambda = \{N_s, N_b, \boldsymbol{\theta}, W_{ss}, W_{sb}, W_{bb}, \boldsymbol{\phi}\}$ is the vector of all such parameters; $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are also vectors for the shape parameters in m and t, respectively. The elements of $\boldsymbol{S}$ are given by

$$\boldsymbol{S}(\lambda) = \begin{pmatrix} \partial \ln\mathcal{L}(N_s, N_b, \boldsymbol{\theta})/\partial N_s \\ \partial \ln\mathcal{L}(N_s, N_b, \boldsymbol{\theta})/\partial N_b \\ \partial \ln\mathcal{L}(N_s, N_b, \boldsymbol{\theta})/\partial \theta_1 \\ \vdots \\ \partial \ln\mathcal{L}(N_s, N_b, \boldsymbol{\theta})/\partial \theta_n \\ \psi_{ss}(N_s, N_b, \boldsymbol{\theta}, W_{ss}) \\ \psi_{sb}(N_s, N_b, \boldsymbol{\theta}, W_{sb}) \\ \psi_{bb}(N_s, N_b, \boldsymbol{\theta}, W_{bb}) \\ \xi_1(\boldsymbol{\theta}, W_{ss}, W_{sb}, W_{bb}, \boldsymbol{\phi}) \\ \vdots \\ \xi_p(\boldsymbol{\theta}, W_{ss}, W_{sb}, W_{bb}, \boldsymbol{\phi}) \end{pmatrix}, \tag{51}$$

where,

$$\begin{aligned} \frac{\partial \ln \mathcal{L}}{\partial N_c} &= \sum_i \left[\frac{g_c(m_i, \boldsymbol{\theta})}{N_s\, g_s(m_i, \boldsymbol{\theta}) + N_b\, g_b(m_i, \boldsymbol{\theta})} - \frac{1}{N} \right], \\ \frac{\partial \ln \mathcal{L}}{\partial \theta_k} &= \sum_i \frac{N_s\, \partial g_s(m_i, \boldsymbol{\theta})/\partial\theta_k + N_b\, \partial g_b(m_i, \boldsymbol{\theta})/\partial\theta_k}{N_s\, g_s(m_i, \boldsymbol{\theta}) + N_b\, g_b(m_i, \boldsymbol{\theta})}, \\ \psi_{(cd)} &= \sum_i \left[\frac{g_c(m_i, \boldsymbol{\theta})\, g_d(m_i, \boldsymbol{\theta})}{\left(N_s\, g_s(m_i, \boldsymbol{\theta}) + N_b\, g_b(m_i, \boldsymbol{\theta})\right)^2} - \frac{W_{cd}}{N} \right], \\ \xi_k &= \sum_i w_s(m_i; \boldsymbol{\theta}, W_{ss}, W_{sb}, W_{bb}) \frac{\partial \ln h_s(t_i;\boldsymbol{\phi})}{\partial \phi_k} \end{aligned}$$

with $c, d \in \{s, b\}$, (cd) iterating over the three unique combinations $\{ss, sb, bb\}$, and the shape parameters of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ running between $\{1 \dots n\}$ and $\{1 \dots p\}$, respectively. For reference these can be compared to the equivalent expressions in Eq. (21) and Eq. (24). One can show that $\mathrm{E}[\boldsymbol{S}(\lambda_0)] = \boldsymbol{0}$, if λ_0 is the vector of true parameter values [4]. Therefore, a consistent estimate $\hat{\lambda}$ can be constructed as the solution to $\boldsymbol{S}(\lambda) \overset{!}{=} \boldsymbol{0}$. We note that the elements of $\boldsymbol{S}(\lambda)$ can be multiplied by arbitrary non-zero constants without changing these results.

The asymptotic covariance of λ, which includes the parameters of interest $\boldsymbol{\phi}$, is then given by [17–19]

$$\boldsymbol{C}_\lambda = \mathrm{E}\left[\frac{\partial \boldsymbol{S}}{\partial \lambda^T}\right]^{-1} \times \boldsymbol{C}_{\boldsymbol{S}} \times \mathrm{E}\left[\frac{\partial \boldsymbol{S}}{\partial \lambda^T}\right]^{-T}, \tag{52}$$

where $\partial\boldsymbol{S}/\partial\lambda^T$ is defined as the Jacobian matrix built from the derivatives $\partial S_k/\partial\lambda_\ell$ and $\boldsymbol{C}_{\boldsymbol{S}} = \mathrm{E}\left[\boldsymbol{S}\boldsymbol{S}^T\right]$. We note that the inverse of the Jacobian $\partial\boldsymbol{S}/\partial\lambda^T$ introduces correlations between the parameter uncertainties. In a finite sample, the expectation values in Eq. (52) can be estimated from the sample. The estimate for $\mathrm{E}[\partial\boldsymbol{S}/\partial\lambda^T]$ is $\partial\boldsymbol{S}/\partial\lambda^T|_{\hat{\lambda}}$, while the elements of the matrix $\widehat{\boldsymbol{C}}_{\boldsymbol{S}}$ are provided in Appendix H. In the literature, Eq. (52) is often referred to as the *sandwich estimator*, but in this case the variance of the score is modified because we consider fluctuations in the sample size.

This general pattern repeats for *sWeights* obtained with variant A and for COWs. Eq. (52) is general and holds for all variants, and the ξ_k in the quasi-score vector in Eq. (51) always remain the same, but the other parts of the quasi-score vector change. The vector has to include a quasi-score for each sample estimate that is used in the calculation of the weights.

- If *sWeights* are computed with variant A, the estimates of the W matrix given by Eq. (18) are not computed from the sample; they are a function of other estimates, $\hat{z} = \hat{N}_s/N$ and $\hat{\boldsymbol{\theta}}$. The $\psi_{(cd)}$ drop out of the quasi-score vector, since the weights $w_s(m; \boldsymbol{\theta}, N_s, N_b)$ in ξ_k can now be expressed directly as a function of these parameters by inserting Eq. (18).
- If *COWs lite* are computed as described in Section 3.2 with a fixed signal model $G_0(m)$ or full COWs with a generic expansion for the signal, a fixed variance function $I(m) = 1$ (or any other fixed function), and a fixed efficiency function, the quasi-score vector reduces to the ξ_k. Only in this case, the weights are independent of the sample and the sum of weights squared is an asymptotically correct estimate of the weight variance, as described in

Appendix B, and the signal p.d.f. $h_s(t)$ can be estimated with a simple weighted histogram, as described in Appendix C.

- If *COWs* are computed and the optimal signal p.d.f. $g_0(m)$ is estimated from the sample, the estimating equations $\partial \ln \mathcal{L}/\partial \theta_k$ for the shape parameters $\boldsymbol{\theta}$ of $g_0(m;\boldsymbol{\theta})$ have to be included in the score vector as in case of classic *sWeights*. If the estimated variance function $\hat{I}(m) = \hat{g}(m)$ is used, one needs in addition $\partial \ln \mathcal{L}/\partial N_c$, where N_c is the amplitude of component c, also analogue to classic *sWeights*. If the optimal variance function $\hat{I}(m) = \hat{q}(m)$ is estimated from the sample as a histogram, as described in Section 3.2, one has to include a quasi-score function for each bin of the histogram, because the value in each bin is a sample estimate. If the efficiency function is estimated from the sample as well, its score functions also need to be included.

We close with a discussion of an explicit result obtained for classic *sWeights* (signal and background are each independent in discriminatory and control variable, and there is no factorisation-breaking efficiency correction) computed with variant B, when the shapes of $g_s(m)$ and $g_b(m)$ are fixed. Then, some simplifications in Eq. (52) are possible, as detailed in Ref. [4]. They result in the following covariance matrix

$$\hat{\boldsymbol{C}}_{\boldsymbol{\phi}} = \boldsymbol{H}^{-1}\boldsymbol{H}'\boldsymbol{H}^{-T} - \boldsymbol{H}^{-1}\boldsymbol{E}\boldsymbol{C}'\boldsymbol{E}^{T}\boldsymbol{H}^{-T}, \tag{53}$$

for the parameters of interest $\boldsymbol{\phi}$, with

$$\begin{aligned}
H_{k\ell} &= \sum_i \hat{w}_s(m_i) \frac{\partial^2 \ln h_s(t_i;\boldsymbol{\phi})}{\partial \phi_k \, \partial \phi_\ell}\bigg|_{\hat{\boldsymbol{\phi}}}, \\
H'_{k\ell} &= \sum_i \hat{w}_s^2(m_i) \left(\frac{\partial \ln h_s(t_i;\boldsymbol{\phi})}{\partial \phi_k} \frac{\partial \ln h_s(t_i;\boldsymbol{\phi})}{\partial \phi_\ell} \right)\bigg|_{\hat{\boldsymbol{\phi}}}, \\
E_{k(cd)} &= \sum_i \frac{\partial w_s(m_i)}{\partial W_{cd}}\bigg|_{\widehat{W}_{ss},\widehat{W}_{sb},\widehat{W}_{bb}} \frac{\partial \ln h_s(t_i;\boldsymbol{\phi})}{\partial \phi_k}\bigg|_{\hat{\boldsymbol{\phi}}}, \\
C'_{(cd)(uv)} &= \sum_i \frac{g_c(m_i)\, g_d(m_i)\, g_u(m_i)\, g_v(m_i)}{\left(\hat{N}_s\, g_s(m_i) + \hat{N}_b\, g_b(m_i)\right)^4},
\end{aligned}$$

where (cd) and (uv) iterate over $\{ss, sb, bb\}$, and $\hat{w}_s(m_i) = w_s(m_i; \hat{W}_{ss}, \hat{W}_{sb}, \hat{W}_{bb})$. The asymptotically correct expression for the binned approach is also derived in Ref. [4]. The first term of Eq. (53) is the covariance for a weighted score function as described by Eq. (50) with independent weights w_i. The second term is specific to *sWeights* computed with variant B. Since the term itself is always positive, it reduces the estimated covariance of $\hat{\phi}$. This reduction is caused by the fact that *sWeights* are estimated from the same data sample. If the shapes of $g_s(m)$ and $g_b(m)$ are also estimated from the data sample, Eq. (53) has to be extended with further terms, see Appendix H.

5. Practical applications of COWs and sweights

In this section we investigate the performance of *sWeights* and the new COW methods in various test-case scenarios. The studies presented in this section make use of the `sweights` Python package [20], developed by the authors. The software includes generic implementations of extracting classic *sWeights* (Section 2) and COWs (Section 3) with the variants detailed in this document, as well as a class which performs a correction to the covariance matrix when fitting unbinned weighted data (Section 4). The Python interface offers support for probability distribution functions defined in either `scipy` [21], `ROOT` (via `TTrees`) [22] or `RooFit` [23]. Classic *sWeights* computed with variant B are also implemented in the `RooStats` [24] package, but there is no implementation of the other variants or COWs.

We emphasise again that computing *sWeights* only requires sensible estimates for the signal and background shapes, $\hat{g}_s(m)$, $\hat{g}_b(m)$, and the signal fraction $\hat{z}$. It does not require any special refitting or yield-only fitting which has been commonly recommended in other *sWeights* discussions, and is enforced in the `RooStats` software implementation. Using the description for *sWeights* outlined in this article, one only needs to fit the discriminant variable(s) (usually a candidate invariant mass) once to obtain $\hat{g}_s(m)$, $\hat{g}_b(m)$, and the signal fraction $\hat{z}$, with the freedom to float, fix or constrain (by means of penalty terms in the likelihood) any of the shape or yield parameters.

More generally, the *sWeights* formalism described in this article extracts the weight function $w_k(m)$ for each component k, which is generally valid for any control variable that is independent of the discriminant variable. There can be several control variables or the control variable can be multi-dimensional. There are other beneficial consequences. The range used to estimate the p.d.f.s and yields can be different from the range used to extract the weights. It is also possible to use a binned fit to obtain estimates of the p.d.f.s and still extract per-event *sWeights*. This is helpful if the sample size is very large, when an unbinned fit can take much more computation time than a binned fit.

In the case of extracting COWs for a factorising signal component (we call this case *COWs lite*), a fit never even needs to be performed. One needs an approximation $G_s(m)$, which does not have to agree with the true p.d.f. $g_s(m)$ (although it should be close to minimise the variance of *sWeights*), a generic expansion for the background, and the variance function $q(m)$ that can be estimated as a histogram from the data, as described in Section 3.2.

We demonstrate in the examples below that there are some pitfalls to be wary of. Generally, we recommend that each non-trivial use-case follows our approach here: produce ensembles of simulated events to check that biases are small and variances are as expected. We start off by explaining how a statistical test of independence can be helpful in deciding whether *sWeights* can be applied.

5.1. Statistical test of independence

A prerequisite for the extraction of *sWeights*, described in Section 2, is that the signal and background samples are each independent in the discriminant and control variables, so that the p.d.f. of the respective component factorises for the discriminant and control variables. If this is not the case then the extracted *sWeights* are biased in general. The COWs method, described in Section 3, overcomes this, but it is useful to know when the *sWeights* method can be applied and is sufficient. A statistical test of independence is useful here.

It is important to truly test for independence. In practice, it is common to compute the correlation coefficient of the discriminant and control variable and test whether it is compatible with zero, but this cannot detect a non-linear dependence that has no linear component, for example, the functional relationship $y = x^2$ will erroneously pass this test, if x is distributed symmetrically around zero. We recommend the USP test of independence [25], which is applied to a two-dimensional histogram constructed from pairs of the discriminant and control variable. The test has proven optimal properties and high statistical power [25].

Implementations of the USP test are available in R [26] and Python [27]. The implementations use a histogram as input and compute a p-value. The binning of the histogram should be fine enough to capture the essential variation in both variables. Bins with a small number of entries or even zero entries are not an issue for this test. If the p-value is very small, evidence for a dependence was detected. The test needs to be applied to the signal and background samples separately, which requires that the test is applied to the simulation where signal and background can be disentangled. If the test passes for the signal, but not the background, one can use the COWs lite approach described in Section 3.2. If both components show dependence, one has to use the full COWs method described in Section 3.

5.2. A simple example comparing sWeights variants

We apply the *sWeights* method on a simple toy example and illustrate the small differences between the variants to compute the W matrix described in Section 2.2. A common application of *sWeights*

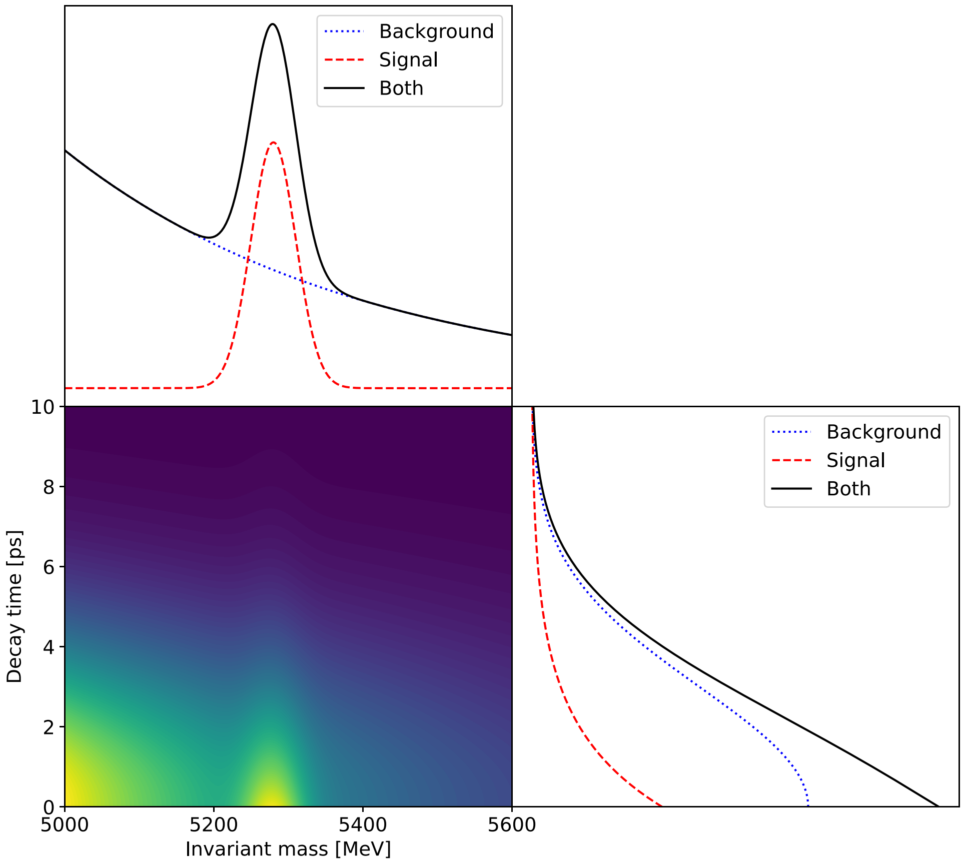

Fig. 1. The p.d.f. used to generate the pseudo-experiments studied in Section 5.2, shown in two-dimensions (bottom left) along with the one-dimensional projections in invariant mass, m, (top left) and decay time, t, (bottom right).

in particle physics is to extract the lifetime of a candidate using its invariant mass to isolate it from the background. We consider two independent variables; the invariant mass, m, and decay time, t, of a B-meson candidate. The simulated dataset contains a mixture of signal and background events. The signal is normally (exponentially) distributed in m (t), whilst the background is exponentially (normally) distributed in m (t). The p.d.f. used to generate events, which is the $f(m,t)$ of Eq. (1), is shown in Fig. 1.

For each simulated dataset, the estimates $\hat{g}_s(m)$ and $\hat{g}_b(m)$ are obtained from an unbinned maximum likelihood fit to the generated invariant mass distribution. The $\widehat{W}$ matrix, with variants A to C, is computed with Eqs. (18), (21), and (26), respectively. Within variant C, we compute the $\widehat{W}$ matrices from the covariance matrix obtained from the HESSE routine in the `iminuit` package [6,28] using both of the methods described in Section 2.3: (i) by inverting the covariance matrix obtained from the full fit and extracting the relevant components, and (ii) by fitting once to obtain maximum-likelihood estimates for all parameters, and then again with only the event yields as free parameters, while all other parameters are fixed to their previous values, and inverting the resulting 2 × 2 covariance matrix. Finally, the weight functions, $\hat{w}_s(m)$ and $\hat{w}_b(m)$, are computed for each variant using Eq. (17).

The distribution of the weight functions, $\hat{w}_s(m)$ and $\hat{w}_b(m)$, from one pseudo-experiment containing 50 000 (200 000) signal (background) events, are shown for the nominal variant B method in Fig. 2 (left). The other variants give very similar distributions. As expected, tiny differences can be seen when inspecting their relative differences as shown in Fig. 2 (right), which vary from pseudo-experiment to pseudo-experiment. It is explained in Section 2.3 why the results obtained with variant Ci and Cii differ from B, although they should be mathematically identical.

Table 1
A comparison of the fitted component yields and the errors from the fit with the extracted sum of weights and sum of weights squared. This numerically demonstrates one reason why we recommend variant B as the best choice, because the weights precisely reproduce both the central value and uncertainty of the fitted yield in a yield only fit.

Fit methods	N_s	$\sigma(N_s)$	N_b	$\sigma(N_b)$
EML Fit (all pars.)	49591.22	351.23	200409.16	523.61
EML Fit (yields only)	49591.22	311.25	200409.16	497.69
sWeight methods	$\sum w_s$	$\sqrt{\sum w_s^2}$	$\sum w_b$	$\sqrt{\sum w_b^2}$
Variant A	49591.01	311.26	200408.99	497.70
Variant B	49591.22	311.25	200409.16	497.69
Variant Ci	49595.97	311.24	200408.98	497.67
Variant Cii	49596.17	311.24	200410.08	497.67

We evaluate the sum of weights and sum of squared weights for all four methods in order to make a comparison with the yield estimates and uncertainties extracted from the discriminant variable fit. The sum of squared weights is an approximate estimate for the variance of the sum of *sWeights*. As shown in Appendix B, it is an exact estimate only when the weights are independent of the sample, but *sWeights* are not independent since the component p.d.f.s have nuisance parameters that are estimated from the same sample. The deviation in this example is small, but that is not always the case. The results are shown in Table 1, along with those from maximum-likelihood fits to the distribution in m, in which firstly all parameters are estimated, and secondly only the event yields are estimated, while the other parameters are set to their respective true values. As expected, we find that the sum of squared weights is comparable to the variance obtained from a fit in which only the yields are estimated and all shape parameters are fixed. The sum of squared weights alone underestimates the variance of the

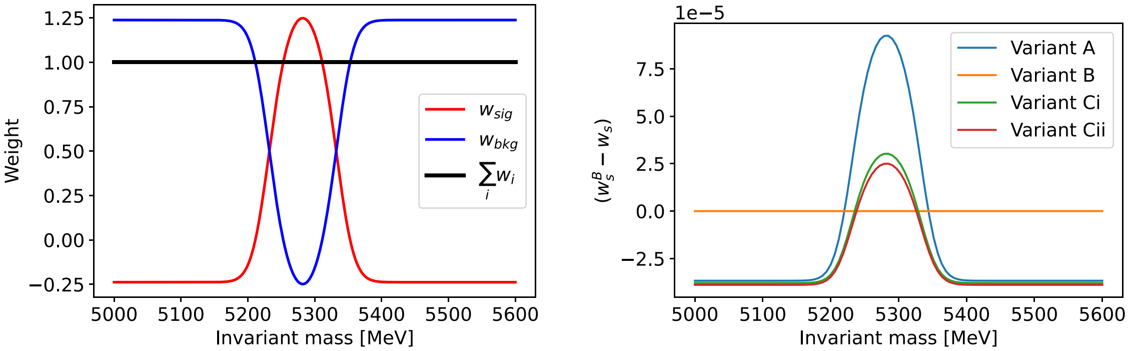

Fig. 2. Left: Distribution of the weight functions, $w_s(m)$ (red) and $w_b(m)$ (blue), as well as their sum (black), extracted using the variant B method. The other variants give very similar results. Right: Difference between the extracted signal weights, $w_s(m)$, from each variant with variant B as the reference.

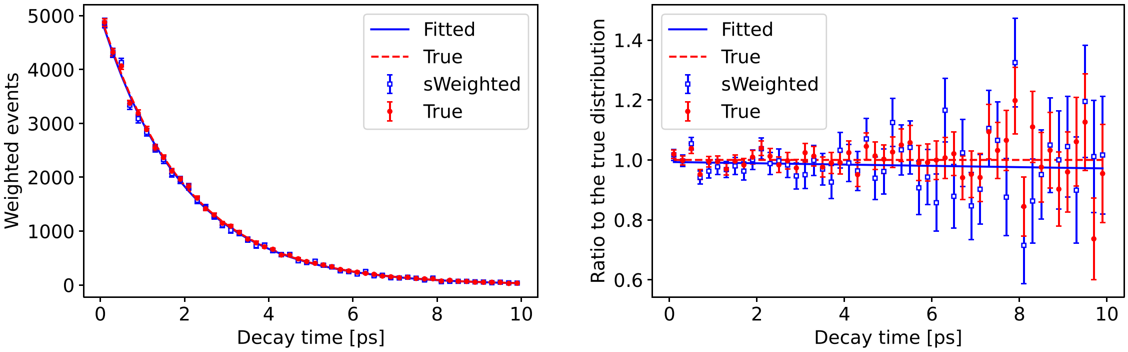

Fig. 3. Left: The decay-time distribution of true signal candidates (red points) and the total signal and background dataset, weighted with $w_s(m)$ (blue points), extracted using variant B. The solid blue line shows the result of an exponential fit to the weighted distribution. The dashed red line shows the true underlying decay-time distribution used to generate the sample. Weights extracted using the other variants give very similar results. Right: The ratio of the fitted function, the weighted distribution and the true distribution with respect to the true function.

Table 2
A comparison of the values and uncertainties extracted from a fit to the exponential slope of the weighted control variable distribution with the outcome of a two-dimensional fit.

Method	Fit result
2D Fit	2.0025 ± 0.0137
Variant A	2.0067 ± 0.0138
Variant B	2.0067 ± 0.0138
Variant Ci	2.0068 ± 0.0138
Variant Cii	2.0068 ± 0.0138

fitted yields when also the shape parameters are estimated from the sample. This is important to keep in mind and why one has to use the sandwich estimator described in Section 4 in general to fully propagate the uncertainty, which also takes into account that the weights are not independent of the sample. The validity of Eq. (22) for variant A, B, and C is also demonstrated, *i.e.* that the fitted yield is exactly reproduced by the sum of weights. In case of variant A and C, the small deviations originate from round-off errors and the comparably low accuracy of a numerically computed second derivative.

Finally, the *sWeights* are applied to the distribution in the control variable, t, which is then fitted with an exponential distribution. This accurately reproduces the true shape, $h_s(t)$, and gives an estimate of the exponential slope parameter, λ, consistent with that obtained from a two-dimensional unbinned maximum likelihood fit, as shown in Fig. 3 and Table 2. The estimated exponential slopes for each variant of *sWeights* and for the full two-dimensional fit are given in Table 2. In the case of the fits to weighted data, the uncertainties on the slopes are calculated with Eq. (53). The precision when fitting the weighted data is comparable to the full two-dimensional fit in this particular case, but in general this will not always be true.

This study is repeated on ensembles containing 500 pseudo-experiments in order to ensure that any of the behaviour seen is not just a fluke of the specific dataset shown in this example. We further perform the same study on ensembles with smaller sample sizes and with different signal to background ratios. The results are shown in Figs. 4 and 5. We find that the four *sWeights* variants give very similar results and can accurately reproduce the correct decay-time slope, and with a precision comparable to a two-dimensional fit.

We further see in Fig. 4 that the sum of weights (left two panels) for variant B exactly reproduces the fitted yield within numerical precision, as expected. Variant A also produces an unbiased estimate, but has a slightly larger spread (note the very small range on the y-axis). Variants Ci and Cii give a very small bias and tend to overestimate the yield by about 0.1 per mill. This is an artefact of computing the second derivative of the log-likelihood numerically and shows why we recommend variant B over C. When inspecting the variance properties, using the sum of squared weights (right two panels of Fig. 4), we find that the sum of squared weights slightly overestimate the true variance of the yield, which is due to the aforementioned fact that *sWeights* are not independent from the sample.

The bias of the fitted slope parameters and their average variance estimates computed over the ensembles are shown for different scenarios in Fig. 5. We find that the variance estimates for the parameter obtained from the *sWeights* variants are comparable to that from a full two-dimensional fit when the sample size is large. For very small amounts of signal, either small overall sample size or small values of the signal to background ratio, we see slight biases and a smaller variance estimate when using the weights method, compared to the full two-dimensional fit. Inspection of the studentised residual distributions suggest a small level (∼ 10%) of under-coverage in these cases, which arises from the asymptotic assumptions made when computing the

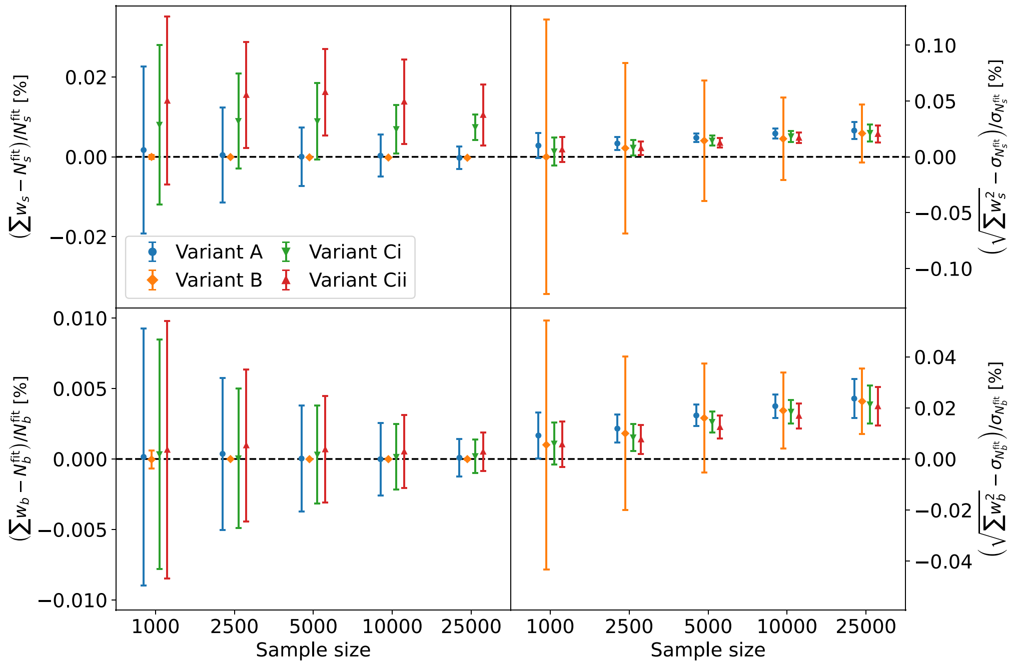

Fig. 4. The upper (lower) left plots show the percentage difference between the sum of weights and the fitted yield, from a fit to the discriminant mass variable in which only the yields float, for the signal (background) components. The right plots show the percentage difference between the square root of the sum of squared weights and the error on the fitted yield, obtained from the second derivative of the likelihood. The points (error bars) show the mean (width) of the distribution across the ensemble of pseudo-experiments.

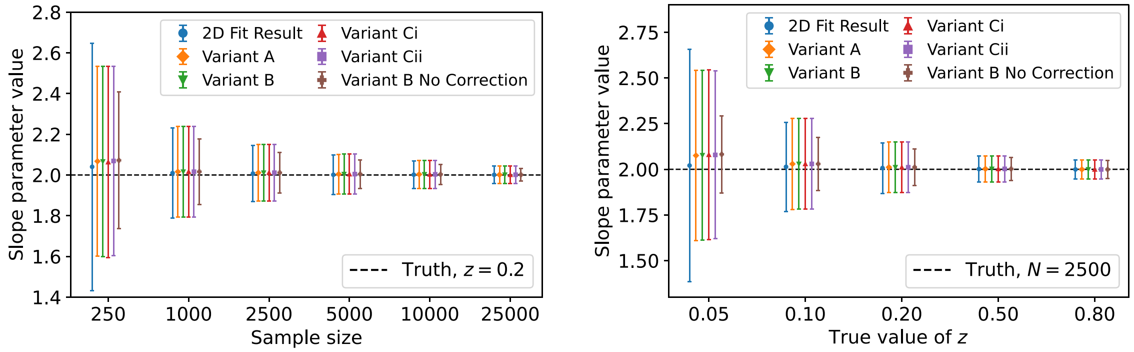

Fig. 5. A comparison of the performance of each *sWeight* variant with a full two-dimensional fit as a function of the sample size (left) and signal to background ratio, z, (right). In the left figure $z = 0.2$ and in the right figure the sample size is $N = 2500$. The points show the mean of the distribution of fitted exponential slope values across the ensemble of pseudo-experiments. The error bars shows the square root of the mean of the variances of the fitted slope extracted across the ensemble. The variance of the brown crosses is computed, in the naive and incorrect way, directly from the double differentials of the weighted likelihood given in Eq. (50), whereas the orange diamonds, green downward-triangles, red upward-triangles and purple squares use the full sandwich estimator of Eq. (52). Note, neither of the x-axes are on a linear scale.

variance with the sandwich estimator. The importance of using the full sandwich estimator, which takes all sources of variance into account, is demonstrated by the points labelled as *Variant B No Correction*, which were computed with the first term of Eq. (53) only (which is sufficient for independent weights) and produce a variance estimate that is too small.

5.3. *sWeights applied to a more complex example*

In this section we show a more complex example with multiple factorising components within $f(m,t)$, of which some may be signal and some may be backgrounds. Because each component factorises in m and t we can still use the classic *sWeights* approach described in Section 2. In this example, we use the invariant mass of a reconstructed B-meson candidate as the discriminant variable once more, but now have six different components. Some are even peaking under or near the signal in a similar way to the signal, as shown in Fig. 6 (left). We label the components with a discrete integer, $c \in [1,6]$, where,

- $c = 1$ represents the signal, which peaks in invariant mass, and is labelled "Signal",
- $c = 2$ and $c = 3$ represent backgrounds where one of the final state particles is mis-reconstructed, *e.g.* a neutral pion is reconstructed as a photon or a charged kaon is reconstructed as a charged pion. These are still peaking but are broader than the signal and have their peak position shifted with respect to the signal. They are labelled "MisRec 1" and "MisRec 2",
- $c = 4$ and $c = 5$ represent backgrounds which have been partially reconstructed, *i.e.* one final state particle has been missed (shifting the mass shape down) or one final state particle from another source is erroneously added (shifting the mass shape up). These are labelled "PartRec 1" and "PartRec 2",
- $c = 6$ represents the continuum background containing random combinations of particles not from the same decay. This is labelled "Background".

The control variables are two so-called Dalitz variables. We have assumed that the discriminant invariant mass variable is constructed from

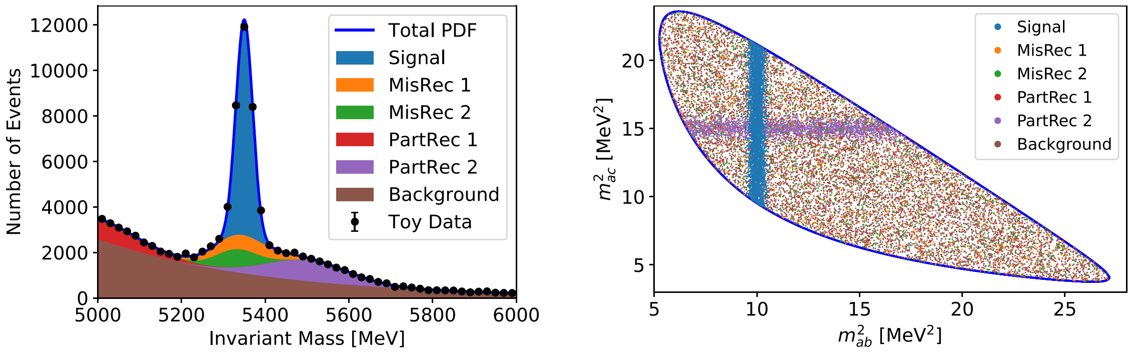

Fig. 6. Left: the probability distribution functions in the discriminant variable for the more complex example. Right: The true distribution of the Dalitz variables in the more complex example, with events coloured by their true event type. The Dalitz variables are uniformly distributed for all components apart from the signal (blue) and one of the backgrounds (purple) which appear as the vertical and horizontal bands in the Dalitz plot, respectively.

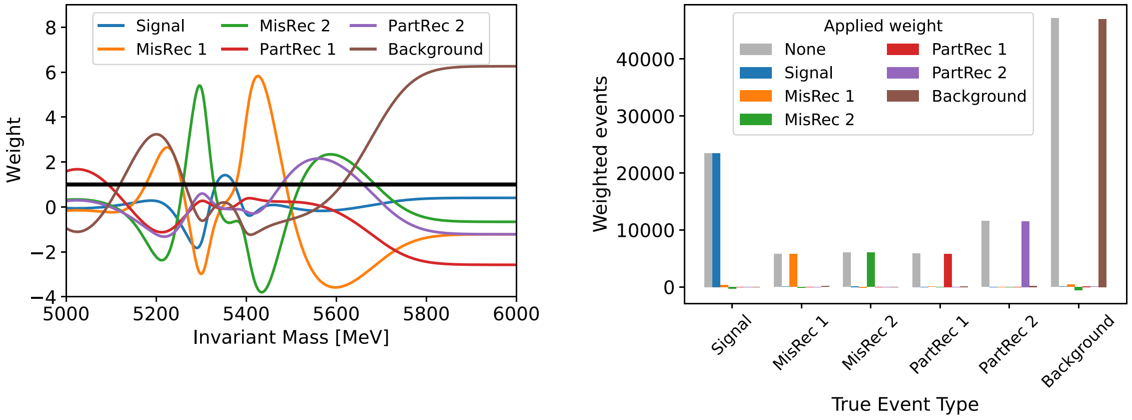

Fig. 7. Left: The distributions of the weight functions, $w_c(m)$, for each of the components in the invariant mass fit. Their sum (equal to 1) is shown by the black line. Right: The sum of weights for each component and its true amount in grey. The colour of each bar (bars have their x position offset to aid the visualisation) represents the respective component weight that has been applied. One can see that each sum of weights, $\sum_i w_c(m_i)$, projects out its own component and nullifies the other components.

a three-body decay of the form $X \to ABC$ and in this case the Dalitz variables are the invariant mass squared of the AB and AC combinations. We generate a pseudo-experiment from the true underlying model, in which the Dalitz variables are uniformly distributed across the phase space for all components, apart from the signal which has a resonance in the AB invariant mass, and one of the backgrounds which has a resonance in the AC invariant mass. These appear as horizontal and vertical bands in the Dalitz plot, shown in Fig. 6 (right).

The generated dataset is fitted in the invariant mass, by maximising the extended unbinned likelihood with the shape parameters of the p.d.f.s, $g_c(m)$, fixed to their true values, in order to obtain estimates for the event yields. The result of this fit is shown in Fig. 6 (left). We then use variant B to obtain the (in this case 6 × 6) $\widehat{W}$ matrix and extract the corresponding weight functions, $w_c(m)$. The distributions of these weight functions are shown in Fig. 7 (left). As shown in Fig. 7 (right), each sum of weights, $\sum_i w_c(m_i)$, projects out its component c and not the others.

In contrast to the previous example, this more complex example exhibits rapidly oscillating weight functions. They oscillate much more quickly than the actual variation of the relevant component shapes themselves. This is because the weight is related to how the shapes overlap *as well as* how they vary themselves with mass. One can also see that the weight functions for competing (*i.e.* similar) shapes have anti-correlated weights, which is what we would expect as their yields are anti-correlated. The sum of all component weights for any value of the discriminant variable is still unity.

The weighted Dalitz variables for each component are shown in Fig. 8. The weights project out the respective components, also the peaking ones. It is worth highlighting that the components with the smallest yields have the largest amplitudes of the weight function, as can be seen in Fig. 7 (left). This is a general feature of *sWeights*: a component with a small yield has larger uncertainties and correspondingly more oscillations in the weight function and a larger weight variance. This can then lead to sizeable fluctuations when weights are visualised like in Fig. 8. When inspecting these plots, it is possible to mistake fluctuations as features in a distribution; for example, a band might seem to appear in a Dalitz distribution when in reality it is just large fluctuations around zero. Minimising the size of these fluctuations is prudent as it is generally undesirable to have few events with large weights. Since the oscillations of the weight functions cannot be reduced (they are already minimal by construction), we recommend, for display purposes, to increase the bin size accordingly in plots like Fig. 8 for components that have a very small yield. Generally, we recommend to proceed with caution when trying to use *sWeights* for a component which is considerably smaller than the others.

5.4. COWs applied to an example with non-factorising background and efficiency

In the final example, we consider a case similar to the first example in Section 5.2, but which now contains non-factorising background and a non-factorising efficiency. The signal is still factorising and is normally (exponentially) distributed in m (t). The background is exponentially (normally) distributed in t (m) but with the shape parameters in m (t) containing a dependence on t (m). The background p.d.f. can be written as

$$f_b(m,t) = Ne^{-(\lambda - s_\lambda t)m} \times \frac{e^{-(t-\mu-s_\mu m)^2}}{2(\sigma + s_\sigma m)^2}, \tag{54}$$

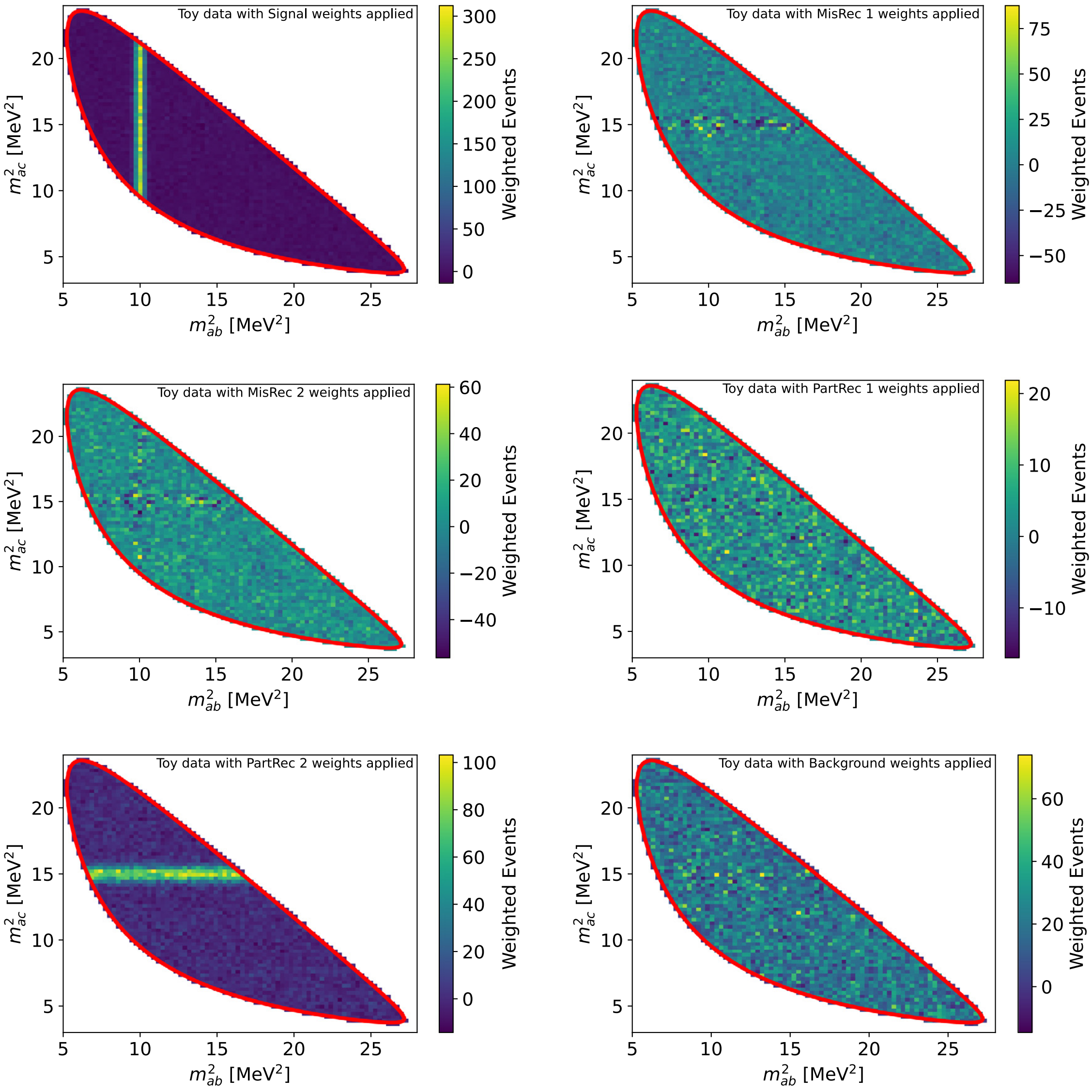

Fig. 8. The weighted Dalitz distribution when applying each of the six components weights: signal (top left), mis-reconstructed 1 (top right), mis-reconstructed 2 (middle left), partially reconstructed 1 (middle right), partially reconstructed 2 (bottom left) and background (bottom right). One can see that the true distributions are appropriately recovered, with some fluctuations.

where N is a normalisation constant and s_λ (s_μ, s_σ) encode the strength of the dependence in m (t) on t (m). This emulates the more realistic use case in which the signal model is straightforward, but the background model is not, and where the efficiency loss is an additional complication. The nature of the true model used to generate ensembles of experiments is shown in Fig. 9, in which the non-factorising nature of the background is manifest in that the exponential slope of the background in mass varies with decay time, and both the mean and width of the normal distribution describing the background in decay time vary with mass. Projections of the integrated distributions along with the projection of the efficiency model used are also shown in Fig. 9.

To demonstrate the failure of classic *sWeights* in contrast to COWs in this case, we apply both methods to a single high-statistics pseudo-experiment containing 500 000 events. For classic *sWeights*, a fit to the invariant mass, shown in Fig. 10 (left), is used to estimate the signal and background p.d.f.s required to extract the *sWeights*. When constructing COWs for this case, no fit is performed as described in Section 3.2. Therefore, only the signal function, the background functions (in this case polynomials up to order 4 are chosen) and the variance function, $I(m) = 1$, used to build the COWs, are shown in Fig. 10 (right). The choice of $I(m)$ is not optimal, but it is an allowed simple choice.

The extracted weight functions for both methods are shown in Fig. 11. A comparison of the weighted events and true distributions in the control variable, t, is shown in Fig. 12. The *sWeights* method cannot handle the non-factorising nature of the background and the efficiency, and therefore the correct distributions are not recovered. The COWs method correctly reproduces the desired distribution and can be applied even without fitting the data.

We further perform an analysis on ensembles of simulated datasets using variant B of the *sWeights* procedure along with various implementations of the COW formalism presented in Section 3. The simulated sample size is 2 000 with equal amounts of signal and background. For the *sWeights* implementation the signal, $\hat{g}_s(m)$, and background, $\hat{g}_b(m)$ distributions are estimated by fitting the simulated sample. To compute COWs, the same signal model is used, while the background expansion and the variance function $I(m)$ are varied. We use these variants for the background:

- A single background component, the same as the fitted estimate of the background from the *sWeights* calculation, $\hat{g}_b(m)$.

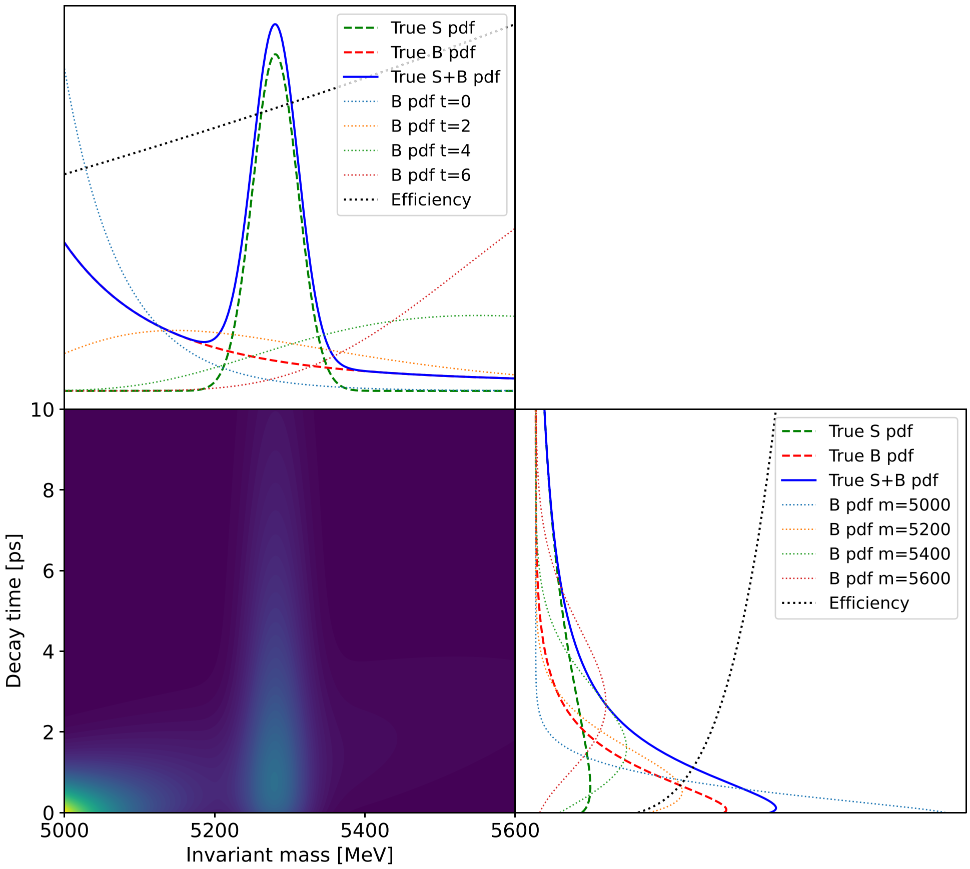

Fig. 9. The p.d.f. used to generate the pseudo-experiments studied in Section 5.4, shown in two-dimensions (bottom left) along with the one-dimensional projections in invariant mass, m, (top left) and decay time, t, (bottom right). The projections show the true signal (dashed green line) and background (dashed red line) p.d.f.s, along with their sum (solid blue line). Projections of the efficiency model (dotted black line) are also shown. The different coloured dotted lines show the background p.d.f. conditional on values of the other parameter, and demonstrate how dependent the shape in m (t) is on the value of t (m).

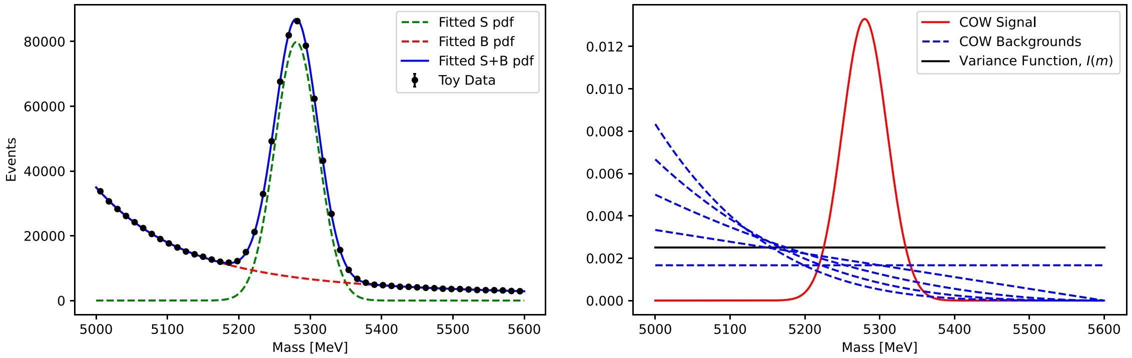

Fig. 10. Left: a fit to the invariant mass to determine the signal and background functions used to compute the *sWeights*. Right: the signal (red), background (blue) and variance (black) functions used when constructing the COW.

- Several background components, given by polynomial powers of m, up to 1st, 3rd and 5th order.

We do not expect COWs to perform well with a single background component, because the background is non-factorising, which requires several components. We use the following variance functions:

- Unity, $I(m) = 1$.
- The true p.d.f. in the mass variable as in Eq. (43), $I(m) = g(m)$. This is the COWs equivalent of *sWeights* computed with variant A.
- Estimates computed from the data sample itself using a histogram of the $1/\epsilon^2(m,t)$ weighted m-distribution, as in Eq. (42), $I(m) = \hat{q}_b(m)$, where b is the number of bins in the histogram from 10 to 50. This estimate approaches the optimal variance function (general advice on an appropriate binning is given in Section 3.2).

After the weights are constructed, we calculate an estimate of the slope parameter, λ, with an unbinned weighted maximum-likelihood fit to the signal-weighted decay-time distribution. For reference, we also include in the comparison a 2D maximum-likelihood fit of the two-dimensional parametric model in which all parameters apart from the slope parameter are known, and a fit of the distributions in the discriminant variable m in 25 slices of the control variable t. The slope parameter of the signal in the control variable is then estimated from the signal yield per slice.

Whether a method produces consistent results is indicated by the distribution of the so-called pull of the slope estimate computed over

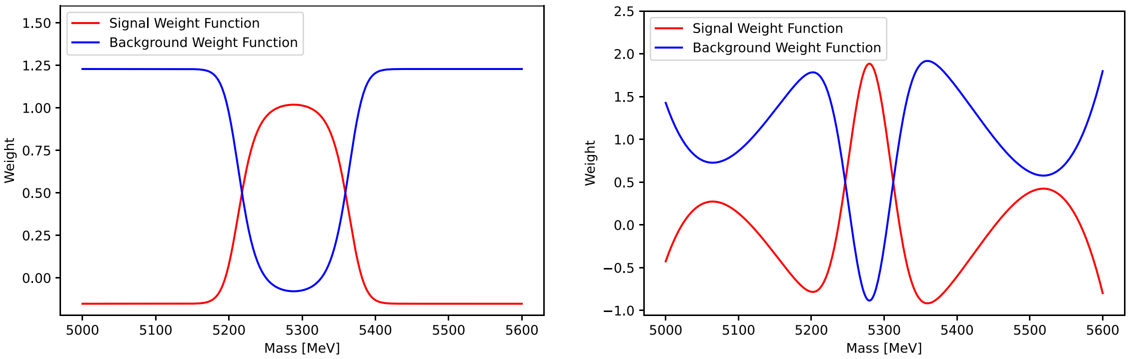

Fig. 11. The extracted weight functions for the signal (red) and background (blue) when using classic *sWeights* (left) and COWs (right). For the COWs, the background weight function is taken as the sum of weight functions for each polynomial component (which is equal to 1).

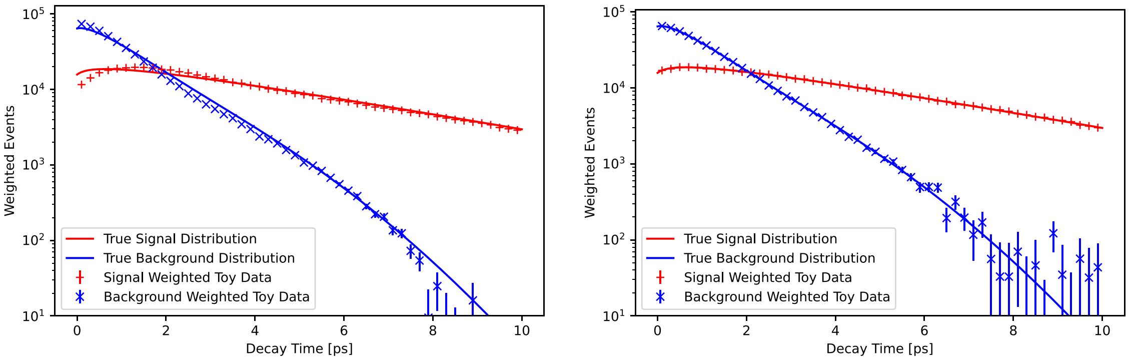

Fig. 12. Signal weighted (blue points) and background weighted (red points) distributions in decay-time along with the true p.d.f.s (blue and red lines) when using classic *sWeights* (left) and a COW (right).

the ensemble. The pull is defined as the difference of the estimated and true values of λ divided by the estimated uncertainty of λ. For an unbiased estimate with the correct variance estimate, the pull distribution has a mean consistent with zero and a width consistent with one. As described in Section 4, the sandwich estimator is needed to obtain an asymptotically correct estimate of the slope uncertainty in the weighted unbinned maximum-likelihood fit. The standard HESSE routine in MINUIT [6] applied to a weighted likelihood computes the inverse matrix of second derivatives of the weighted likelihood function, but this does not give correct uncertainty estimates since the weighted likelihood is not a true likelihood. We further assess the statistical power of all methods by computing the variance of the slope parameter relative to the ideal case where each signal event can be correctly identified and the uncertainty is just Poissonian.

The results of this study are shown in Fig. 13. As expected, classic *sWeights* and COWs without sufficiently many background components yield a biased estimate in this factorisation-breaking example; this is indicated by the pull that deviates from zero. COWs with sufficiently many background components produce an unbiased result. Adding more background components than necessary reduces the statistical power of COWs. The optimal trade-off is case-specific, but can be found empirically using simulations or with goodness-of-fit tests on the data, as described in Section 3.2.

The variance function $\hat{q}_b(m)$ performs better than $I(m) = f(m)$ or $I(m) = 1$, this is also expected as $\hat{q}_b(m)$ approaches the optimal choice $q(m)$ as b increases. The 2D fit has better statistical power than COWs, but it requires a parametric model of the background, which is often not known. Fitting in slices still does not produce an unbiased estimate of the slope because the factorisation breaking is severe enough that the p.d.f.s do not factorise within each slice.

We have further tested cases with different signal-to-background ratios and with different sample sizes and the conclusions are similar, although fewer orders of background polynomial are required for the COWs to produce a minimal bias when the sample size is smaller. In small samples, the residual bias from using a finite number of orders gets masked by the statistical variance. So when using COWs in general there is a trade-off between systematic bias and statistical precision. It is also worth noting that for small samples (< 100 events) there are small biases due to the fact that the covariance computed with the sandwich estimator is only asymptotically valid.

6. Conclusions

This article provides a new perspective on the classic *sWeights* method, by re-deriving the results in the context of orthonormal functions. This approach provides many new insights and allowed us to generalise the method to what we dub *Custom Orthogonal Weight functions* (COWs). COWs produce correct results in cases when *sWeights* fail, namely when the background is not factorising in the discriminant and control variable, or when the detection is affected by finite efficiency that does not factorise. Both of these ailments are not uncommon in practice. We show how to obtain optimal (minimum variance) COWs under these conditions. When only the shape of the signal distribution in the control variable is of interest, COWs can be constructed without performing a fit to the distribution in the discriminant variable. We further summarise the work presented elsewhere [4] on the correct application of the sandwich estimator to weighted maximum-likelihood fits that use *sWeights*.

For the specific toy examples that we studied in this paper, fits to *sWeighted* data had statistical precision comparable to a fully parametric two-dimensional maximum-likelihood fit of the distribution of the discriminant and control variables. This is remarkable, but we also note that weighted fits are in general less efficient than fully parametric maximum-likelihood fits [4]. In a toy example with non-factorising background and efficiency, the statistical power for COWs is lower than

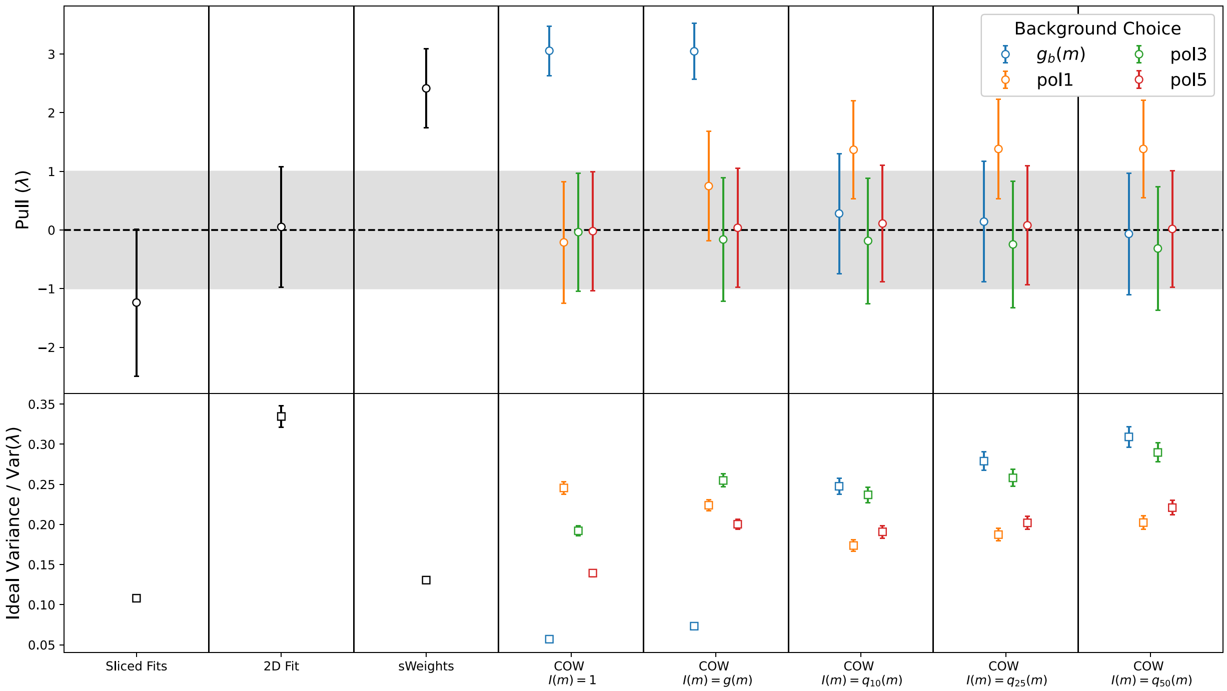

Fig. 13. Results of the ensemble study on an example with non-factorising efficiency and non-factorising background. The top panel shows the pull of the fitted exponential slope parameter over the ensemble. The bottom panel shows the variance of the fitted exponential slope parameter values over the ensemble with respect to the ideal variance that would be obtained if all signal events could be correctly identified. The eight panels, from left-to-right, show the performance under various different scenarios. The different colours represent different choices for the modelling of the background function, $g_b(m)$, in the COWs. The two benchmark scenarios and classic *sWeights* are shown in black.

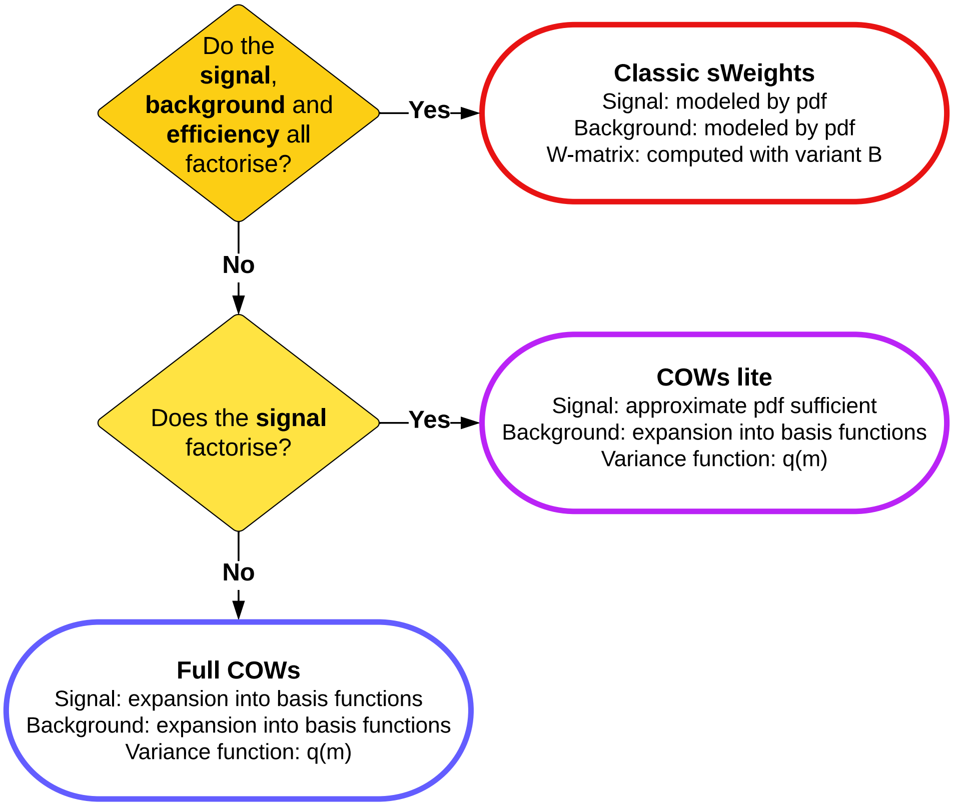
Do the signal, background and efficiency all factorise?
Yes
Classic sWeights
Signal: modeled by pdf
Background: modeled by pdf
W-matrix: computed with variant B
No
Does the signal factorise?
Yes
COWs lite
Signal: approximate pdf sufficient
Background: expansion into basis functions
Variance function: q(m)
No
Full COWs
Signal: expansion into basis functions
Background: expansion into basis functions
Variance function: q(m)

Fig. 14. Guide for the practical application of *sWeights* or COWs. A description of this chart is given in the conclusions.

what is obtained in a fully parametric fit, but better than using multiple fits in slices of the control variable.

A summary of our recommendations for the optimal use of *sWeights* and COWs is given in Fig. 14. When the signal, background, and efficiency all factorise in the discriminant and control variables, one should use classic *sWeights* computed with variant B, described in Section 2.2. This method minimises the variance of the weights, and is the one previously described in the literature [1,2]. When only the signal factorises, one should use what we dub *COWs lite* described in Section 3.2. In this case, an approximate signal p.d.f. is sufficient. The result is unbiased for any signal p.d.f., but the statistical power is maximised when the signal p.d.f. is close to the true signal. It can be obtained as a histogram from simulation. The non-factorising background should be expanded into basis functions, we recommend Bernstein polynomials or basis splines. The optimal variance function, $q(m)$, for this case can be estimated from the data sample with a histogram too, as described in Section 3.2. Finally, if the signal is also non-factorising, then one has use the *full COWs* approach, which is like COWs lite but now the signal is also expanded into basis functions, which must be linear-independent from the basis functions used in the expansion of the background.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The authors wish to thank their colleagues in the LHCb experiment, and members of the wider community of particle physicists with an interest in statistics, for the fruitful and enjoyable discussions which inspired this study. In particular they would like to thank Roger Barlow, Arnau Brossa Gonazalo, Vladimir Gligorov, Louis Lyons, Richard Samworth and Tom Berrett. We would also like to thank the journal reviewers whose constructive feedback helped to considerably improve the discussion and details in the paper. MK is supported by the Science and Technology Facilities Council (STFC), UK, under grant #ST/R004536/2. HD acknowledges funding from the Deutsche Forschungsgemeinschaft (DFG – German Research Foundation) under project number 449728698. CL gratefully acknowledges support by the Emmy Noether programme of the Deutsche Forschungsgemeinschaft (DFG), grant identifier LA 3937/1-1. The authors would like to thank UK Research and Innovation (UKRI) for providing funds to allow open access.

Appendix A. Constrained minimisation problem

We use Lagrange multipliers to find the function $w_s(m)$ which minimises Eq. (6) under the constraints in Eq. (3). We need to find the extremum of

$$\begin{aligned} L(w_s(m), \alpha_s, \alpha_b) \;=\; & \int w_s(m)^2\, g(m)\, \mathrm{d}m \;-\; z^2 \\ & - 2\alpha_s \left(\int \mathrm{d}m \; w_s(m)\, g_s(m) - 1 \right) \\ & - 2\alpha_b \int \mathrm{d}m \; w_s(m)\, g_b(m). \end{aligned} \tag{A.1}$$

The Lagrange multipliers $\alpha_{s,b}$ in L were scaled by a factor of two without loss of generality. Since L is a functional of $w_s(m)$, we need to use variational calculus. With

$$\delta \int \mathrm{d}m \; w_s(m)\, \phi(m) = \int \mathrm{d}m \; \delta w_s(m)\, \phi(m)$$

$$\delta \int \mathrm{d}m \; w_s(m)^2\, \phi(m) = \int \mathrm{d}m \; 2 w_s(m)\, \delta w_s(m)\, \phi(m)$$

the variational score function is

$$\delta L = 2 \int \mathrm{d}m \; \delta w_s(m) \Big[w_s(m)\, g(m) - \alpha_s\, g_s(m) - \alpha_b\, g_b(m) \Big] \overset{!}{=} 0. \tag{A.2}$$

According to the fundamental lemma of calculus of variations, the equation is satisfied for any continuous $\delta w_s(m)$ only if the integrand inside the square brackets is zero. So we obtain

$$w_s(m) = \frac{\alpha_s\, g_s(m) + \alpha_b\, g_b(m)}{g(m)}. \tag{A.3}$$

Appendix B. Variance of a sum of weights

We compute the variance of a sum of independently and identically distributed weights, $T = \sum_i^n w_i$, where the sample size n is a Poisson-distributed number. The latter changes the computation of the variance of T. *sWeights* and COWs in general are not independently distributed, so that the simple formula derived here only applies in rare special cases (as pointed out in Appendix C).

We follow the derivation in Ref. [29]; the key insight is that the sampling of n is independent of the sampling of the w_i. The variance of T is $\mathrm{Var}(T) = \mathrm{E}[T^2] - \mathrm{E}[T]^2$, so we need the respective expectations. The expectation of T is

$$\mathrm{E}[T] = \mathrm{E}_n[\mathrm{E}_w[T]] = \mathrm{E}_n \left[\sum_i^n \mathrm{E}[w] \right] = \mathrm{E}[n]\; \mathrm{E}[w], \tag{B.1}$$

where E_n is an expectation taken with respect to n only, likewise for E_w. The expectation of T^2 is

$$\begin{aligned} \mathrm{E}[T^2] = \mathrm{E}_n[\mathrm{E}_w[T^2]] \;&=\; \mathrm{E}_n \left[\mathrm{Var}_w(T) + \mathrm{E}_w[T]^2 \right] \\ &= \mathrm{E}_n \left[\sum_i^n \mathrm{Var}(w) + n^2\; \mathrm{E}[w]^2 \right] \\ &= \mathrm{E}[n]\; \mathrm{Var}(w) + \mathrm{E}[n^2]\; \mathrm{E}[w]^2. \end{aligned} \tag{B.2}$$

Here we used that the variance of a sum of independent random variables is equal to the sum of their variances. The variance of T then is

$$\begin{aligned} \mathrm{Var}(T) \;&=\; \mathrm{E}[n]\; \mathrm{Var}(w) + \mathrm{E}[n^2]\; \mathrm{E}[w]^2 - \mathrm{E}[n]^2\; \mathrm{E}[w]^2 \\ &= \mathrm{E}[n]\; \mathrm{Var}(w) + \mathrm{Var}(n)\; \mathrm{E}[w]^2. \end{aligned} \tag{B.3}$$

With $\mathrm{Var}(n) = \mathrm{E}[n]$ for a Poisson distribution, the variance reduces to

$$\mathrm{Var}(T) = \mathrm{E}[n]\, (\mathrm{Var}(w) + \mathrm{E}[w]^2) = \mathrm{E}[n]\; \mathrm{E}[w^2]. \tag{B.4}$$

An unbiased estimate of this is given by

$$\widehat{\mathrm{Var}}(T) = n \times \frac{1}{n} \sum_i w_i^2 = \sum_i w_i^2. \tag{B.5}$$

Appendix C. Covariance matrix of a weighted histogram

For *sWeights* and COWs, the sums of weights in bins (a weighted histogram) of the control variable are asymptotically unbiased estimates of the respective component p.d.f.. The asymptotically correct covariance matrix of the bin contents is obtained with the sandwich estimator generally described in Section 4, which properly accounts for correlations between all estimates extracted from the same sample.

In special cases, the bins of the weighted histograms are uncorrelated and the variance of each bin is correctly estimated by the sum of the squares of the weights, which simplifies the analysis of the weighted histograms. This is the case if the efficiency is constant and the component p.d.f.s $g_c(m)$ are fixed *a priori*, no component $g_c(m)$ is misspecified, and one of the following conditions applies to the variance function $I(m)$:

A. The variance function $I(m)$ is fixed *a priori*.
B. The variance function is the estimated density $\hat{g}(m) = \sum_c \hat{z}_c\, g_c(m)$.
C. The variance function $I(m)$ is a histogram of the observed distribution in m, a binned estimate of $\hat{g}(m)$.

These cases are not only of academic interest. A fixed variance function $I(m) = 1$ is a valid (albeit not optimal, in the sense of achieving minimum variance for *sWeights*) variance function. The component p.d.f.s $g_c(m)$ are fixed when a general expansion into basis functions is used to construct COWs. Case B corresponds to classic *sWeights* with fixed component p.d.f.s (not common in practice) computed with variant A. Cases B and C correspond to the application of the full COWs where both the signal and background components are expanded into fixed basis functions and when the efficiency is constant. The choices for $I(m)$ are then optimal or nearly optimal in case of the histogram.

We will now prove these claims. Since the component p.d.f.s $g_c(m)$ are complete and not misspecified, the true density is a linear combination of the components, $g(m) = \sum_c z_c\, g_c(m)$. The efficiency is constant, which is equivalent to setting $\epsilon(m,t) = 1$. For a variance function $I(m)$, the weight function $w_c(m)$ that projects out a single component c is given by Eq. (32). We will omit the index c in the following, $w(m) := w_c(m)$. The content of the kth bin of a weighted histogram for component c in t is given by

$$\hat{B}_k = \sum_{i=1}^{\hat{N}_k} w(m_i;\boldsymbol{\alpha})\,, \tag{C.1}$$

where the sum is taken over samples with $t_k \le t < t_{k+1}$ and $w(m;\boldsymbol{\alpha})$ potentially depends on parameters $\boldsymbol{\alpha}$ that are estimated from the sample. The number of entries $\hat{N}_k$ in bin k is a Poisson distributed random variable with expectation value N_k. If $\mathrm{E}\left[w(m;\boldsymbol{\alpha})\right] = z$, the true fraction of component c, $\hat{B}_k$ is an unbiased estimate of the expected bin content B_k,

$$\mathrm{E}\left[\hat{B}_k\right] = \mathrm{E}\left[\sum_{i=1}^{\hat{N}_k} w(m_i;\boldsymbol{\alpha})\right] = \mathrm{E}\left[\hat{N}_k\right]\mathrm{E}\left[w(m;\boldsymbol{\alpha})\right] = N_k\, z = B_k\,. \tag{C.2}$$

We will now compute the covariance matrix C of the histogram $\boldsymbol{B}$. We introduce unspecified score functions $\boldsymbol{P}$ whose roots shall determine all nuisance parameters $\boldsymbol{\alpha}$ in the calculation of the weights $w(m;\boldsymbol{\alpha})$, which may include, for example, the component fractions z_c and shape parameters $\boldsymbol{\theta}$ of the component p.d.f.s $g_c(m;\boldsymbol{\theta})$. The bin-contents in t are defined by the roots of the score functions $\boldsymbol{Q}$,

$$Q_k = B_k - \sum_{i=1}^{\hat{N}_k} w(m_i;\boldsymbol{\alpha})\,. \tag{C.3}$$

The combined estimate of all parameters is obtained from the roots of the joint vector $(\boldsymbol{P},\boldsymbol{Q})$. Asymptotically, the covariance matrix of nuisance parameters and bin contents is given by the sandwich estimator

$$\begin{pmatrix} \mathrm{E}\left[\dfrac{\partial \boldsymbol{P}}{\partial \boldsymbol{\alpha}}\right] & \mathrm{E}\left[\dfrac{\partial \boldsymbol{P}}{\partial \boldsymbol{B}}\right] \\ \mathrm{E}\left[\dfrac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\alpha}}\right] & \mathrm{E}\left[\dfrac{\partial \boldsymbol{Q}}{\partial \boldsymbol{B}}\right] \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{E}\left[\boldsymbol{P}\boldsymbol{P}^T\right] & \mathrm{E}\left[\boldsymbol{P}\boldsymbol{Q}^T\right] \\ \mathrm{E}\left[\boldsymbol{Q}\boldsymbol{P}^T\right] & \mathrm{E}\left[\boldsymbol{Q}\boldsymbol{Q}^T\right] \end{pmatrix} \begin{pmatrix} \mathrm{E}\left[\dfrac{\partial \boldsymbol{P}}{\partial \boldsymbol{\alpha}}\right] & \mathrm{E}\left[\dfrac{\partial \boldsymbol{P}}{\partial \boldsymbol{B}}\right] \\ \mathrm{E}\left[\dfrac{\partial \boldsymbol{Q}}{\partial \boldsymbol{\alpha}}\right] & \mathrm{E}\left[\dfrac{\partial \boldsymbol{Q}}{\partial \boldsymbol{B}}\right] \end{pmatrix}^{-T}. \tag{C.4}$$

The partial derivatives represent Jacobian matrices of the derivatives of the score-function parts $\boldsymbol{P}$ and $\boldsymbol{Q}$ with respect to the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{B}$. We are only interested in the latter; the covariance matrix C is the lower-right block matrix of the matrix product. By construction, we have $\mathrm{E}\left[\partial\boldsymbol{P}/\partial\boldsymbol{B}\right] = 0$ and $\mathrm{E}\left[\partial\boldsymbol{Q}/\partial\boldsymbol{B}\right] = 1$. If the derivatives $\mathrm{E}\left[\partial\boldsymbol{Q}/\partial\boldsymbol{\alpha}\right]$ are all zero, C reduces to $\mathrm{E}\left[\boldsymbol{Q}\boldsymbol{Q}^T\right]$, which has a simple structure as we will see. For the trivial case A of no nuisance parameters, $\mathrm{E}\left[\partial\boldsymbol{Q}/\partial\boldsymbol{\alpha}\right]$ is trivially zero and Eq. (C.2) trivially holds.

We compute the matrix C under this condition,

$$\begin{aligned} C_{k\ell} &= \mathrm{E}\left[Q_k Q_\ell\right] = \mathrm{E}\left[\left(B_k - \sum_{i=1}^{\hat{N}_k} w(m_i)\right)\left(B_\ell - \sum_{j=1}^{\hat{N}_\ell} w(m_j)\right)\right] \\ &= -B_k B_\ell + \mathrm{E}\left[\sum_{i=1}^{\hat{N}_k}\sum_{j=1}^{\hat{N}_\ell} w(m_i)\, w(m_j)\right]. \end{aligned} \tag{C.5}$$

Since the bins are non-overlapping, the sums over i and j are disjoint and the expectation value for $k \neq l$ factorises. The off-diagonal elements are zero,

$$C_{k\ell} = -B_k B_\ell + N_k N_\ell z^2 = 0\,. \tag{C.6}$$

We calculate the diagonal elements by splitting the double sum into independent pairs and identical pairs, and we use $\mathrm{E}[\hat{N}_k^2] = N_k^2 + N_k$ for a Poisson-distributed variable,

$$\begin{aligned} C_{kk} &= -B_k^2 + \mathrm{E}\left[\sum_{i=1}^{\hat{N}_k}\sum_{j=1}^{\hat{N}_k} w(m_i)\, w(m_j)\right] \\ &= -B_k^2 + \mathrm{E}\left[\hat{N}_k(\hat{N}_k - 1)\right]\mathrm{E}\left[w\right]^2 + \mathrm{E}\left[\hat{N}_k\right]\mathrm{E}\left[w^2\right] \\ &= -B_k^2 + N_k^2 z^2 + N_k\,\mathrm{E}\left[w^2\right] = N_k\,\mathrm{E}\left[w^2\right] = \mathrm{E}\left[\sum_{i=1}^{\hat{N}_k} w^2(m_i)\right]. \end{aligned} \tag{C.7}$$

The variance of the bin is estimated by the sum of weights squared and different bins are uncorrelated. The same result can be derived from the fact that the weights in each bin are independently sampled. It also follows from independence that different bins are uncorrelated and the variance of the sum of weights per bin is given by Appendix B in this case.

We now show that Eq. (C.2) is true and $\mathrm{E}\left[\partial\boldsymbol{Q}/\partial\boldsymbol{\alpha}\right]$ is zero for the non-trivial cases B and C. More generally, $\mathrm{E}\left[\partial\boldsymbol{Q}/\partial\boldsymbol{\alpha}\right]$ is zero, if only the variance function $I(m;\boldsymbol{\alpha})$ depends on nuisance parameters $\boldsymbol{\alpha}$, but not the component p.d.f.s $g_c(m)$. In case B, the nuisance parameters are the fractions $\boldsymbol{z}$ in $I(m;\boldsymbol{z}) = \sum_c z_c\, g_c(m)$, which are estimated from the sample. In case C, the expected variance function is

$$I(m;\boldsymbol{\beta}) = \sum_k \beta_k\, \eta_k(m), \tag{C.8}$$

with $\eta_k(m) := H(m - m_k)H(m_{k+1} - m)$ where $H(x)$ is the Heaviside step function, m_k is the lower edge of bin k, and the nuisance parameters are the bin fractions $\boldsymbol{\beta}$, with expectation values $\beta_k = \int_{m_k}^{m_{k+1}} \mathrm{d}m\, g(m)$.

For Eq. (C.2) to hold, we need to show that $\mathrm{E}\left[w_0(m;\boldsymbol{\alpha})\right] = z_0$ for any $\boldsymbol{\alpha}$. We compute the expectation for weights computed with a variance function $I(m;\boldsymbol{\alpha})$,

$$\begin{aligned} \mathrm{E}\left[w_0(m;\boldsymbol{\alpha})\right] &= \sum_\ell A_{0\ell} \int \mathrm{d}m\, \frac{g_\ell(m)}{I(m;\boldsymbol{\alpha})} g(m) \\ &= \sum_{\ell,k} z_k\, A_{0\ell} \underbrace{\int \mathrm{d}m\, \frac{g_\ell(m) g_k(m)}{I(m;\boldsymbol{\alpha})}}_{W_{\ell k}} \\ &= \sum_k z_k\, \delta_{0k} = z_0\,, \end{aligned} \tag{C.9}$$

where we used that A is the inverse of W.

Regarding $\mathrm{E}\left[\partial\boldsymbol{Q}/\partial\boldsymbol{\alpha}\right]$, we find for component c,

$$\begin{aligned} \mathrm{E}\left[\frac{\partial Q_k}{\partial \alpha_\ell}\right] &= \mathrm{E}\left[\frac{\partial}{\partial \alpha_\ell}\left(B_k - \sum_{j=1}^{\hat{N}_k} w_c(m_j;\boldsymbol{\alpha})\right)\right] \\ &= -\mathrm{E}\left[\sum_{j=1}^{\hat{N}_k} \frac{\partial w_c(m_j;\boldsymbol{\alpha})}{\partial \alpha_\ell}\right] = -N_k\,\mathrm{E}\left[\frac{\partial w_c}{\partial \alpha_\ell}\right]. \end{aligned} \tag{C.10}$$

So it remains to be shown that $\mathrm{E}\left[\partial w_c/\partial\alpha_\ell\right]$ is zero. We insert Eq. (32) and $g(m) = \sum_c z_c\, g_c(m)$,

$$\begin{aligned} \mathrm{E}\left[\frac{\partial w_c}{\partial \alpha_k}\right] &= \sum_\ell \left(\frac{\partial A_{c\ell}}{\partial \alpha_k}\int \mathrm{d}m\, \frac{g_\ell(m)}{I(m;\boldsymbol{\alpha})} g(m)\right. \\ &\qquad \left. +\ A_{c\ell}\int \mathrm{d}m\, \frac{\partial}{\partial \alpha_k}\left(\frac{g_\ell(m)}{I(m;\boldsymbol{\alpha})}\right) g(m)\right) \\ &= \sum_{\ell,d} z_d \left(\frac{\partial A_{c\ell}}{\partial \alpha_k}\int \mathrm{d}m\, \frac{g_\ell(m) g_d(m)}{I(m;\boldsymbol{\alpha})}\right. \\ &\qquad \left. +\ A_{c\ell}\frac{\partial}{\partial \alpha_k}\int \mathrm{d}m\, \frac{g_\ell(m) g_d(m)}{I(m;\boldsymbol{\alpha})}\right) \end{aligned}$$

$$= \sum_{\ell,d} z_d \left(\frac{\partial A_{c\ell}}{\partial \alpha_k} W_{\ell d} + A_{c\ell} \frac{\partial W_{\ell d}}{\partial \alpha_k} \right) = \sum_{\ell,d} z_d \frac{\partial}{\partial \alpha_k} (A_{c\ell} W_{\ell d})$$

$$= \sum_{d} z_d \frac{\partial}{\partial \alpha_k} \underbrace{\left(\sum_{\ell} A_{c\ell} W_{\ell d} \right)}_{\delta_{cd}} = 0\,, \tag{C.11}$$

where we used that the A and W matrices are inverses of each other. Since $\mathrm{E}\left[\partial \boldsymbol{Q}/\partial \boldsymbol{\alpha}\right]$ is zero, the covariance matrix C also takes the same simple form as in the trivial case A.

We emphasise that this only holds for COWs and *sWeights* computed with variant A. If classic *sWeights* are computed with variant B, extra terms appear in the calculation of the covariance matrix, as discussed in Ref. [4].

Appendix D. Self-consistency of *sWeights*

Here, we prove Eq. (22) for *sWeights* computed with variant A or B. The sum of *sWeights* is given by Eq. (7), if true p.d.f.s are replaced by estimates,

$$T = \sum_i \hat{w}_s(m_i) = \sum_i \frac{\hat{A}_{ss}\, \hat{g}_s(m_i) + \hat{A}_{sb}\, \hat{g}_b(m_i)}{\hat{g}(m_i)}\,, \tag{D.1}$$

where $\hat{A}$ is the inverse of the $\widehat{W}$ matrix, which can be computed with Eq. (18) (variant A) or Eq. (21) (variant B).

The proof for variant A requires that the component fractions N_c with $c \in \{s, b\}$ are estimated with the extended maximum-likelihood (EML) method. The EML estimates $\hat{N}_s$ and $\hat{N}_b$ are solutions to the score functions (compare to Eq. (24)),

$$\sum_i \frac{\hat{g}_s(m_i)}{N\, \hat{g}(m_i)} = 1 \text{ and } \sum_i \frac{\hat{g}_b(m_i)}{N\, \hat{g}(m_i)} = 1\,, \tag{D.2}$$

with $N\, \hat{g}(m) = \hat{N}_s\, \hat{g}_s(m) + \hat{N}_b\, \hat{g}_b(m)$ and estimated p.d.f.s $\hat{g}_c(m)$, whose nuisance parameters are obtained from the EML fit. With Eq. (D.2), we obtain

$$\frac{T}{N} = \left(\hat{A}_{ss} \sum_i \frac{\hat{g}_s(m_i)}{N\, \hat{g}(m_i)} + \hat{A}_{sb} \sum_i \frac{\hat{g}_b(m_i)}{N\, \hat{g}(m_i)} \right) = \hat{A}_{ss} + \hat{A}_{sb} = \sum_k \hat{A}_{sk}\,. \tag{D.3}$$

We now use Eq. (18),

$$1 = \int \mathrm{d}m\, \hat{g}_k(m) = \int \mathrm{d}m\, \hat{g}_k(m) \frac{\hat{g}(m)}{\hat{g}(m)} = \sum_\ell \frac{\hat{N}_\ell}{N} \widehat{W}_{k\ell} \tag{D.4}$$

Since $\hat{A}$ is the inverse of $\widehat{W}$, we finally get

$$T = N \sum_{k,\ell} \hat{A}_{sk} \frac{\hat{N}_\ell}{N} \widehat{W}_{k\ell} = \sum_\ell \delta_{s\ell}\, \hat{N}_\ell = \hat{N}_s\,. \tag{D.5}$$

To prove the result for variant B, we go back to Eq. (D.1) and use Eq. (21),

$$\begin{aligned} T &= \sum_{i,k} \frac{\hat{A}_{sk}\, \hat{g}_k(m_i)}{\hat{g}(m_i)} \\ &= \frac{1}{N} \sum_{i,k} \frac{\hat{A}_{sk}\, \hat{g}_k(m_i) \left(\hat{N}_s\, \hat{g}_s(m_i) + \hat{N}_b\, \hat{g}_b(m_i) \right)}{[\hat{g}(m_i)]^2} \\ &= \frac{1}{N} \sum_{i,k,\ell} \frac{\hat{N}_\ell\, \hat{A}_{sk}\, \hat{g}_k(m_i)\, \hat{g}_\ell(m_i)}{[\hat{g}(m_i)]^2} \\ &= \sum_{k,\ell} N_\ell\, \hat{A}_{sk}\, \widehat{W}_{k\ell} = \sum_\ell N_\ell\, \delta_{s\ell} = \hat{N}_s \end{aligned} \tag{D.6}$$

This proof does not make use of Eq. (D.2) and therefore holds more generally.

Appendix E. Sum of component weights obtained from COWs

When $I(m)$ is a linear combination of the p.d.f.s,

$$I(m) = \sum_{k=0}^{n} a_k g_k(m), \tag{E.1}$$

then the normalisation of the $g_k(m)$ implies that

$$\begin{aligned} 1 = \int \mathrm{d}m\, g_k(m) &= \int \mathrm{d}m\, g_k(m) \frac{I(m)}{I(m)} \\ &= \sum_{l=0}^{n} a_l \int \mathrm{d}m\, \frac{g_k(m) g_l(m)}{I(m)} = \sum_{l=0}^{n} a_l W_{kl}. \end{aligned} \tag{E.2}$$

Inverting this matrix equation, it follows that $a_l = \sum_{k=0}^{n} A_{kl}$ and thus,

$$\sum_{k=0}^{n} w_k(m) = \sum_{k=0}^{n} \sum_{l=0}^{n} \frac{A_{kl}\, g_l(m)}{I(m)} = \frac{1}{I(m)} \sum_{l=0}^{n} a_l g_l(m) = 1. \tag{E.3}$$

Appendix F. Variance function for COWs which minimises the variances of $\hat{z}_k$

The variance of $\hat{z}_k$ from Eq. (37) is

$$\begin{aligned} \mathrm{Var}(\hat{z}_k) &= \mathrm{E}[\hat{z}_k^2] - \mathrm{E}[\hat{z}_k]^2 \\ &= \frac{1}{N^2} \mathrm{E}\left[\hat{D}^2 \sum_{i,j=1}^{N} \frac{w_k(m_i) w_k(m_j)}{\epsilon(m_i, t_i)\epsilon(m_j, t_j)} \right] - z_k^2 \end{aligned} \tag{F.1}$$

$$= \frac{1}{N^2} \left(\sum_{i \neq j}^{N} \mathrm{E}\left[\hat{D}^2 \frac{w_k(m_i) w_k(m_j)}{\epsilon(m_i, t_i)\epsilon(m_j, t_j)} \right] + \sum_{i=j}^{N} \mathrm{E}\left[\hat{D}^2 \frac{w_k^2(m_i)}{\epsilon^2(m_i, t_i)} \right] \right) - z_k^2 \tag{F.2}$$

$$= \frac{1}{N^2} \left(N(N-1) z_k^2 + N\, \mathrm{E}\left[\hat{D}^2 \frac{w_k^2(m)}{\epsilon^2(m, t)} \right] \right) - z_k^2 \tag{F.3}$$

$$= \frac{1}{N} \left(\mathrm{E}\left[\hat{D}^2 \frac{w_k^2(m)}{\epsilon^2(m, t)} \right] - z_k^2 \right). \tag{F.4}$$

If the weight $I(m)$ is to be such that the variance of $\hat{z}_k$ is minimal it then follows that the expectation value $\mathrm{E}[w_k^2(m)/\epsilon^2(m, t)]$ is minimal, since $\hat{D}$ does not depend on $I(m)$. The minimisation has to incorporate the constraints that the integrals of $w_k(m) g_l(m)$ are either zero or one, which is done by Lagrange multipliers, $2\lambda_l$. The extremum condition becomes

$$\int \mathrm{d}m\, \mathrm{d}t\, \rho(m, t) \left[\frac{w_k^2(m)}{\epsilon^2(m, t)} - \sum_{l=0}^{n} 2\lambda_l w_k(m) g_l(m) \right] \overset{!}{=} \min. \tag{F.5}$$

Only $\rho(m, t)$ and $\epsilon(m, t)$ depend on t. Encompassing the t-integral by introducing

$$q(m) = \int \mathrm{d}t\, \frac{\rho(m, t)}{\epsilon^2(m, t)} \tag{F.6}$$

and using the extremum condition, which requires that any variations $\delta w_k(m)$, with $\delta w_k^2(m) = 2w(m)\delta w(m)$, lead to zero variation of the remaining m integral, one finds

$$\int \mathrm{d}m\; 2\,\delta w_k(m) \left[w_k(m) q(m) - \sum_{l=0}^{n} \lambda_l g_l(m) \right] = 0. \tag{F.7}$$

This is true under any variations $\delta w_k(m)$ provided the term in square brackets zero. This implies that the functional form of the weight functions is

$$w_k(m) = \sum_{l=0}^{n} \frac{\lambda_l g_l(m)}{q(m)}, \tag{F.8}$$

which in turn means that the optimal variance weight function is given by

$$I(m) = q(m) = \int \mathrm{d}t\, \frac{\rho(m, t)}{\epsilon^2(m, t)}. \tag{F.9}$$

Appendix G. Variance function for COWs for which $\hat{z}_k$ are maximum likelihood estimates

Consider an Extended Maximum Likelihood fit of the yields, N_k, for each component of the data model. The Maximum Likelihood (ML) estimates, $\hat{N}_k$ are obtained by minimising

$$\mathcal{L} = \sum_{l=0}^{n} N_k - \sum_{i=1}^{N} \frac{1}{\epsilon(m_i, t_i)} \ln \left[\sum_{l=0}^{n} N_l g_l(m) \right]. \tag{G.1}$$

The requirement of a stationary point $\partial \mathcal{L} / \partial \hat{N}_k = 0$ leads to

$$1 = \sum_{i=1}^{N} \frac{1}{\epsilon(m_i, t_i)} \frac{g_k(m_i)}{\sum_l \hat{N}_l g_l(m_i)}. \tag{G.2}$$

Inserting the estimates $\hat{z}_k = \hat{N}_k D / N$ means that

$$\frac{N}{D} = \sum_{i=1}^{N} \frac{1}{\epsilon(m_i, t_i)} \frac{g_k(m_i)}{\sum_l \hat{z}_l g_l(m_i)} \; \forall \, k. \tag{G.3}$$

The solution for this system of non-linear equations requires that the right-hand-side is the same for all k, namely N/D. Noticing here the similarity with Eq. (37), one can choose $I(m)$ such that the sum in Eq. (G.3) becomes N/D. In this case one finds that

$$\hat{z}_k = \sum_{l=0}^{n} A_{kl} = a_k \tag{G.4}$$

and therefore

$$I(m) = \sum_{l=0}^{n} \hat{z}_l g_l(m). \tag{G.5}$$

Appendix H. Sample estimate for variance of the quasi-score vector

Below we give the sample estimate for $\boldsymbol{C_S} = \mathrm{E}\left[\boldsymbol{S}\boldsymbol{S}^T\right]$ in Eq. (52). We obtain

$$\hat{\mathrm{E}}\left[\frac{\partial \ln\mathcal{L}}{\partial N_c}\frac{\partial \ln\mathcal{L}}{\partial N_d}\right] = \sum_i \frac{\hat{g}_c(m_i)\hat{g}_d(m_i)}{\left(\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)\right)^2} \tag{H.1}$$

$$\hat{\mathrm{E}}\left[\frac{\partial \ln\mathcal{L}}{\partial N_c}\frac{\partial \ln\mathcal{L}}{\partial \theta_k}\right] = \sum_i \frac{\hat{g}_c(m_i)\left(\hat{N}_s\frac{\partial g_s(m_i)}{\partial \theta_k} + \hat{N}_b\frac{\partial g_b(m_i)}{\partial \theta_k}\right)|_{\hat{\boldsymbol{\theta}}}}{\left(\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)\right)^2} \tag{H.2}$$

$$\hat{\mathrm{E}}\left[\frac{\partial \ln\mathcal{L}}{\partial N_c}\psi_{(uv)}\right] = \sum_i \frac{\hat{g}_c(m_i)\hat{g}_u(m_i)\hat{g}_v(m_i)}{\left(\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)\right)^3} \tag{H.3}$$

$$\hat{\mathrm{E}}\left[\frac{\partial \ln\mathcal{L}}{\partial N_c}\xi_k\right] = \sum_i \frac{\hat{w}_s(m_i)\hat{g}_c(m_i)}{\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)} \frac{\partial \ln h_s(t_i)}{\partial \phi_k}\bigg|_{\hat{\boldsymbol{\phi}}} \tag{H.4}$$

$$\hat{\mathrm{E}}\left[\frac{\partial \ln\mathcal{L}}{\partial \theta_k}\frac{\partial \ln\mathcal{L}}{\partial \theta_\ell}\right] = \sum_i \frac{(\hat{N}_s\frac{\partial g_s(m_i)}{\partial \theta_k} + \hat{N}_b\frac{\partial g_b(m_i)}{\partial \theta_k})|_{\hat{\boldsymbol{\theta}}}}{\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)} \tag{H.5}$$

$$\times \frac{(\hat{N}_s\frac{\partial g_s(m_i)}{\partial \theta_\ell} + \hat{N}_b\frac{\partial g_b(m_i)}{\partial \theta_\ell})|_{\hat{\boldsymbol{\theta}}}}{\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)} \tag{H.6}$$

$$\hat{\mathrm{E}}\left[\frac{\partial \ln\mathcal{L}}{\partial \theta_k}\psi_{(cd)}\right] = \sum_i \frac{\hat{g}_c(m_i)\hat{g}_d(m_i)}{\left(\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)\right)^3} \tag{H.7}$$

$$\times \left(\hat{N}_s\frac{\partial g_s(m_i)}{\partial \theta_k} + \hat{N}_b\frac{\partial g_b(m_i)}{\partial \theta_k}\right)\bigg|_{\hat{\boldsymbol{\theta}}} \tag{H.8}$$

$$\hat{\mathrm{E}}\left[\frac{\partial \ln\mathcal{L}}{\partial \theta_k}\xi_\ell\right] = \sum_i \frac{(\hat{N}_s\frac{\partial g_s(m_i)}{\partial \theta_k} + \hat{N}_b\frac{\partial g_b(m_i)}{\partial \theta_k})|_{\hat{\boldsymbol{\theta}}}}{\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)} \tag{H.9}$$

$$\times \hat{w}_s(m_i)\frac{\partial \ln h_s(t_i)}{\partial \phi_\ell}\bigg|_{\hat{\boldsymbol{\phi}}} \tag{H.10}$$

$$\hat{\mathrm{E}}\left[\psi_{(cd)}\psi_{(uv)}\right] = \sum_i \frac{\hat{g}_c(m_i)\hat{g}_d(m_i)\hat{g}_u(m_i)\hat{g}_v(m_i)}{\left(\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)\right)^4} \tag{H.11}$$

$$\hat{\mathrm{E}}\left[\psi_{(cd)}\xi_k\right] = \sum_i \frac{\hat{w}_s(m_i)\hat{g}_c(m_i)\hat{g}_d(m_i)}{\left(\hat{N}_s\hat{g}_s(m_i) + \hat{N}_b\hat{g}_b(m_i)\right)^2} \tag{H.12}$$

$$\times \frac{\partial \ln h_s(t_i)}{\partial \phi_k}\bigg|_{\hat{\boldsymbol{\phi}}} \tag{H.13}$$

$$\hat{\mathrm{E}}\left[\xi_k\xi_\ell\right] = \sum_i \hat{w}_s^2(m_i)\left(\frac{\partial \ln h_s(t_i)}{\partial \phi_k}\frac{\partial \ln h_s(t_i)}{\partial \phi_\ell}\right)\bigg|_{\hat{\boldsymbol{\phi}}}, \tag{H.14}$$

where $\hat{g}_c(m_i) = g_c(m_i; \hat{\boldsymbol{\theta}})$, $h_s(t_i) = h_s(t_i, \boldsymbol{\phi})$, $\hat{w}_s(m_i) = w_s(m_i; \hat{g}_s, \hat{g}_b, \widehat{W}_{ss}, \widehat{W}_{sb}, \widehat{W}_{bb})$, $c, d \in \{s, b\}$, (cd) and (uv) each iterate over $\{ss, sb, bb\}$, and k, l index the shape parameters of $\boldsymbol{\theta}$ or $\boldsymbol{\phi}$.

References

[1] R.J. Barlow, Event classification using weighting methods, J. Comput. Phys. 72 (1987) 202–219.

[2] M. Pivk, F.R. Le Diberder, SPlot: A Statistical tool to unfold data distributions, Nucl. Instrum. Methods A 555 (2005) 356–369.

[3] B. Efron, R. Tibshirani, An introduction to the bootstrap, Statist. Sci. 57 (1) (1986) 54–75.

[4] C. Langenbruch, Parameter uncertainties in weighted unbinned maximum likelihood fits, Eur. Phys. J. C 82 (2022) 393.

[5] R.J. Barlow, Extended maximum likelihood, Nucl. Instrum. Methods A 297 (1990) 496–506.

[6] F. James, M. Roos, Minuit: a system for function minimization and analysis of the parameter errors and correlations, Comput. Phys. Comm. 10 (1975) 343–367.

[7] A.N. Kolmogorov, On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition, Dokl. Akad. Nauk SSSR 114 (1957) 953–956.

[8] A.B. Givental, B.A. Khesin, J.E. Marsden, A.N. Varchenko, V.A. Vassiliev, O.Y. Viro, V.M. Zakalyukin, On the Representation of Functions of Several Variables as a Superposition of Functions of a Smaller Number of Variables, Springer, Berlin, 2009.

[9] S. Bernstein, Proof of the theorem of weierstrass based on the calculus of probabilities, Comm. Kharkov Math. Soc. 13 (1912) 1–2.

[10] C. de Boor, A Practical Guide to Splines, Springer, New York, NY, 1978.

[11] S. Baker, R.D. Cousins, Clarification of the use of chi square and likelihood functions in fits to histograms, Nucl. Instrum. Methods 221 (1984) 437–442.

[12] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Ser. B Stat. Methodol. 36 (2) (1974) 111–133.

[13] F. James, Statistical Methods in Experimental Physics, second ed., World Scientific, Singapore, 2006.

[14] P.J. Huber, Robust Statistics, second ed., Wiley, New York, 1981.

[15] J. Wooldridge, Econometric Analysis of Cross Section and Panel Data, Econometric Analysis of Cross Section and Panel Data, MIT Press, 2010.

[16] W.K. Newey, D. McFadden, Large sample estimation and hypothesis testing, in: Handbook of Econometrics, Vol. 4, Elsevier, 1994.

[17] H. White, Maximum likelihood estimation of misspecified models, Econometrica 50 (1) (1982) 1–25.

[18] A. van der Vaart, Asymptotic Statistics, Cambridge University Press, 2000.

[19] A. Davison, Statistical Models, in: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2003.

[20] M. Kenzie, COWs and sWeights source code, 2021, available at https://github.com/matthewkenzie/sweights.

[21] P. Virtanen, et al., SciPy 1.0: fundamental algorithms for scientific computing in python, Nature Methods (2020).

[22] R. Brun, F. Rademakers, ROOT: AN object oriented data analysis framework, Nucl. Instrum. Methods A 389 (1997) 81–86.

[23] W. Verkerke, D.P. Kirkby, The RooFit toolkit for data modeling, eConf C0303241, 2003, MOLT007.

[24] L. Moneta, K. Belasco, K.S. Cranmer, S. Kreiss, A. Lazzaro, D. Piparo, G. Schott, W. Verkerke, M. Wolf, The RooStats project, PoS ACAT2010, 2010, 057.

[25] T.B. Berrett, R.J. Samworth, USP: an independence test that improves on pearson's chi-squared and the G-test, Proc. R. Soc. A 477 (2256) (2021) 20210549.

[26] T. Berrett, I. Kontoyiannis, R. Samworth, USP: U-statistic permutation tests of independence for all data types, 2020, available at https://cran.r-project.org/web/packages/USP/index.html.

[27] H. Dembinski, D. Saxton, resample: Randomisation-based inference in Python based on data resampling and permutation, 2022, Available at https://github.com/scikit-hep/resample.

[28] H. Dembinski, P. Ongmongkolkul, C. Deil, H. Schreiner, M. Feickert, et al., scikit-hep/iminuit: v2.11.2, 2022, http://dx.doi.org/10.5281/zenodo.6389982.

[29] C.A. Benjamin, J.R. Cornell, Probability, Statistics, and Decisions for Civil Engineers, Dover Publications, New York, 2014.