PROCEEDINGS OF SCIENCE

A First Look at Improved Probabilistic Event Weighting via Covariance-Corrected Q-Factors for Signal Isolation

Zachary Baldwin[a,*] **and Nathaniel Dene Hoffman**[a]

[a] *Carnegie Mellon University,*
Pittsburgh PA, United States of America

E-mail: zbaldwin@cmu.edu

In complex particle-physics analyses, where signal and background overlap across multidimensional phase space, statistically consistent event-by-event weights are often needed to extract signal observables without bias from the background. A common strategy is to fit a mixture model for the contributing components in a discriminating variable, typically an invariant mass, and then use the fitted component shapes and normalizations to assign a signal weight to each event. In practice, widely used weighting methods can perform poorly when their assumptions are only approximately valid, in particular when the discriminating variable is correlated with the observables of interest, or when the calculation of the weights does not account for correlations among the fitted signal and background yields.

We assess the limitations of these standard techniques with particular focus on the Q-factor method, a local fitting method based on k-nearest neighbors. Although Q-factors offer greater flexibility than a single global fit, their local probability weights can bias spectator observables, variables not included in the discriminating-variable fit but whose signal distributions are to be determined, even when the underlying mixture fits appear well behaved. We introduce a corrected formalism, the ${}_s Q$-factor method, which retains the Q-factor neighborhood construction but replaces the local probability weight with an ${}_s\mathcal{P}$lot-style covariance-corrected weight computed from each neighborhood fit. Using toy Monte Carlo studies, we demonstrate that ${}_s Q$-factors improve the recovery of non-discriminatory signal distributions and reduce the bias in fitted physics parameters in comparison to standard probability-weight methods.

The European Physical Society Conference on High Energy Physics (EPS-HEP 2025)
6-11 July, 2025
Palais Du Pharo, Marseille, France

*Speaker

PoS(EPS-HEP2025)578

1. Introduction

Separating signal from background remains a recurring challenge in many particle-physics measurements where the observables of interest span high-dimensional phase space, while the separation power resides in only one or a few discriminating variables, typically an invariant mass. Analyses introduce a model for the signal and background shapes in the discriminating variable and obtain their relative contributions from a fit to the measured distribution. Event-by-event weighting methods use these fits in the discriminating variable to reconstruct the distributions of *control* variables, observables not included in the fit, and to extract physics parameters. However, the reliability of these methods depends on whether the data satisfy the underlying assumptions and on how the fit uncertainties are propagated into the weighted observables.

We introduce the ${}_sQ$-factor method, a covariance-corrected local event-weighting approach. It uses k-nearest neighbor (kNN) fits to obtain local signal and background yields and their covariance matrices, and assigns ${}_s\mathcal{P}$lot-style weights computed from those local fits. We validate the method with toy Monte Carlo studies and compare it to common alternatives.

2. Weighting Methods

2.1 ${}_s\mathcal{P}$lot Method

The ${}_s\mathcal{P}$lot method [1] starts from an extended unbinned maximum-likelihood fit in a discriminating variable x_r. The distribution of the discriminating variable is modeled as a sum of components $\gamma \in \{S, B\}$ (signal S and background B), each described by a probability-density function (PDF) $f_\gamma(x_r)$. The variables of interest x (angles, kinematic quantities, etc.) are not included in this fit and are treated as the control variables. A key assumption is that within each component, x_r is independent of x, so that the joint PDF can be written as $F_\gamma(x, x_r) = G_\gamma(x)\, f_\gamma(x_r)$.

From the likelihood fit, both the fitted yields N_γ and their covariance matrix $V_{\gamma\omega}$ are provided, which encodes the statistical uncertainties and correlations among the fitted yields. The corresponding sWeights for component γ and event i are defined as

$$w_\gamma^{(s\mathcal{P})}(x_r^{(i)}) = \frac{\sum_{\omega\in\{S,B\}} V_{\gamma\omega}\, f_\omega(x_r^{(i)})}{\sum_{\chi\in\{S,B\}} N_\chi f_\chi(x_r^{(i)})} \,. \tag{1}$$

Under the independence assumption stated above, histogramming the control variable x with weights $w_\gamma^{(s\mathcal{P})}$ yields, on average over repeated samples, the distribution $N_\gamma G_\gamma(x)$. The defining quantity of the ${}_s\mathcal{P}$lot weights is the covariance matrix $V_{\gamma\omega}$ in Eq. (1), which corrects for correlations among the fitted yields and therefore permits negative weights when pairs of fitted component yields are strongly anti-correlated.

2.2 Q-factor Method

The Q-factor method [2] extends one-dimensional sideband subtraction to multivariate analyses by assigning probabilistic event weights without binning the data. It does so by fitting the reference coordinate x_r (discriminating variable) locally in phase space. Each event is represented by $\vec{x} = (x_r, \vec{z})$, where x_r is a reference coordinate whose signal and background shapes can be parameterized, while $\vec{z}$ collects the remaining observables of interest, corresponding to the control variables x in Sec. 2.1. No parameterization is introduced for the signal or background dependence on $\vec{z}$. The method assumes that, for both signal and background, x_r is uncorrelated with $\vec{z}$ and that signal and background do not interfere. It additionally relies on neighborhoods chosen for each event so that the local model remains a good approximation.

In contrast to the global fit used in ${}_s\mathcal{P}$lot, Q-factors determine weights locally in the non-reference variables $\vec{z}$. A metric is defined in the non-reference coordinates $X \subset \vec{z}$ (e.g. a normalized Euclidean distance) and for each event i the n_c nearest neighbors are selected. The reference coordinate distribution x_r of the selected neighbors is then fitted with the sum of a signal $f_S(x_r)$ and background $f_B(x_r)$ PDF to obtain the local yields $N_S^{(i)}$ and $N_B^{(i)}$ (and any local shape parameters)

for event i. The Q-factor, i.e. the weight for event i, is then

$$Q_i = \frac{N_S^{(i)}\, f_S(x_r^{(i)})}{N_S^{(i)}\, f_S(x_r^{(i)}) + N_B^{(i)}\, f_B(x_r^{(i)})} \,, \tag{2}$$

which is interpreted as a local signal probability. Filling histograms with weights Q_i (or $1 - Q_i$) approximates the signal (or background) distributions in $\vec{z}$. The same weights can also be used in unbinned fits to extract physics observables.

Compared with ${}_s\mathcal{P}$lot, the Q-factor method differs in two important ways. (i) The event weights are obtained from many local fits rather than a single global fit, which provides adaptivity in case the signal fraction or the shape parameters of $f_{S,B}(x_r)$ vary across phase space but introduces sensitivity to the choice of metric and neighbor number n_c. (ii) Since Q-factors use local probability weights rather than covariance-corrected weights, their bias and reliability of uncertainties are typically checked with dedicated Monte Carlo studies [2]. The ${}_s\mathcal{P}$lot method instead yields unbiased control variable distributions provided the discriminating and control variables are uncorrelated within each component.

2.3 ${}_sQ$-factor Method

The ${}_sQ$-factor method leverages the strengths of the Q-factor method in multidimensional analyses and, highlighted in toy Monte Carlo studies shown in Sec. 3, reduces bias in the extracted control variable distributions and improves the reliability of the associated uncertainties. It retains the local adaptivity of Q-factors, but replaces the usual neighborhood probability weight with a covariance-corrected weight that incorporates the yield covariance from each local fit when extracting signal distributions in control variables. This extension is motivated by the distinction made in Ref. [1] between naive probability weights and covariance-based sWeights. ${}_s\mathcal{P}$lot uses the yield covariance matrix from the global fit to obtain statistically consistent control variable distributions when, within each component, the reference distribution is independent of the control variables. The standard Q-factor method computes an analogous probability weight from each local neighborhood fit [2]. The weight is a local posterior in x_r rather than a covariance-corrected sWeight. Agreement of the neighborhood fit in x_r therefore tests only the description of the reference coordinate. It does not ensure that the same weight yields a clean extraction of the signal distribution in the remaining coordinates, so residual background structure can persist even when the local x_r fits are described well.

The ${}_sQ$-factor method therefore combines the local neighborhood construction of Q-factors with the covariance-corrected weighting of ${}_s\mathcal{P}$lot. As in Sec. 2.2, we construct a local neighborhood of n_c nearest neighbors in the chosen non-reference space around each event i and perform a signal-plus-background fit of the reference distribution of the selected neighbors. This provides the local yields $N_S^{(i)}$ and $N_B^{(i)}$ together with their yield covariance matrix $V_{\gamma\omega}^{(i)}$, with $\gamma, \omega \in \{S, B\}$, as estimated by the extended maximum-likelihood fit.

The only change relative to the standard Q-factor method is the weight definition. Instead of assigning the local probability weight of Eq. (2), the ${}_sQ$-factor method assigns an ${}_s\mathcal{P}$lot-style weight computed from the *local* fit outputs,

$$w_\gamma^{(sQ)}(x_r^{(i)}) = \frac{V_{\gamma S}^{(i)}\, f_S(x_r^{(i)}) + V_{\gamma B}^{(i)}\, f_B(x_r^{(i)})}{N_S^{(i)}\, f_S(x_r^{(i)}) + N_B^{(i)}\, f_B(x_r^{(i)})} \,, \tag{3}$$

where $f_S(x_r)$ and $f_B(x_r)$ are the same PDFs used in the neighborhood fits of the reference distributions. When these PDFs are approximately independent of the extracted variables over the selected neighborhoods, the covariance terms in Eq. (3) correct the local probability-weight construction for yield correlations. This correction should reduce residual background structure and bias in extracted control variable distributions while preserving the phase space dependent adaptivity of Q-factors.

In the limit $n_c \to N_{tot}$, where N_{tot} is the total number of events in the dataset, the neighborhood becomes the full dataset and the construction reduces to the ${}_s\mathcal{P}$lot method. For $n_c < N_{tot}$, the weights are computed from local neighborhood fits rather than a single global fit, so the method interpolates between Q-factors and ${}_s\mathcal{P}$lot as n_c varies. As with the Q-factor method, performance depends on the choice of distance metric and neighborhood size, which must keep the local fit well-conditioned while remaining sufficiently local. In sparse or poorly separated neighborhoods, the local yield covariance matrix can become ill-conditioned and the weights can fluctuate strongly. Stability is assessed with dedicated closure tests below.

3. Monte Carlo Closure Tests

3.1 Simulated dataset

Pseudo-data are generated following the original Q-factor toy model construction, using a single toy model that isolates a reference coordinate from the remaining phase space variables used for the physics extraction. Each event contains the reference coordinate $x_r = M_r$, taken to be a generic mass-like discriminant and the angular variables $\cos\theta$ and ϕ, which enter the model and define the target signal angular distributions and fit parameters reported below. We extend the setup by introducing two auxiliary spectator coordinates, t and g, chosen to have distinct signal and background shapes while remaining external to the reference-variable fit. Specifically, t is drawn from exponential distributions with different slopes for signal and background, and g is drawn from zero-mean normal distributions with different widths for signal and background as well. These features make the sample a nontrivial test of event-by-event weighting beyond the reference coordinate. The stacked signal and background contributions in Fig. 1 illustrate a narrow signal peak on a smooth background in the reference mass M_r and distinct signal and background shapes in t and g, together with structured angular dependence in $\cos\theta$ and ϕ.

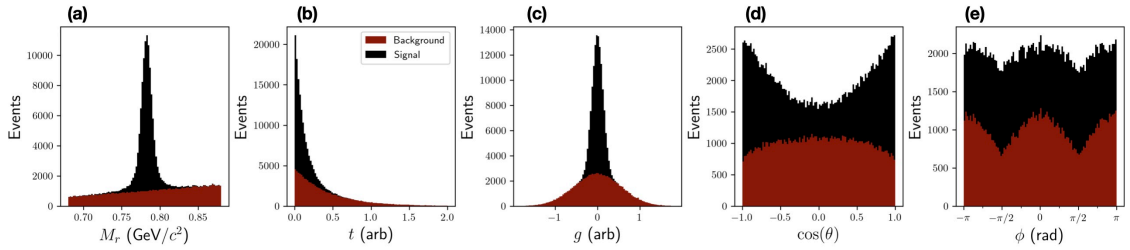

Figure 1: Distributions of generated signal (black) and background (red) events in the toy Monte Carlo sample used to test ${}_sQ$-factor weighting. Panels show (a) the reference mass M_r, (b) and (c) the spectator variables t and g respectively, and the angular distributions in $\cos\theta$ (d) and ϕ (e).

3.2 Results

Figure 2 shows that the ${}_sQ$-factor weighted spectra closely follow the true signal distributions for the spectator variables t and g as well as for the angular variables. The M_r projection is shown for completeness and carries limited diagnostic power for methods whose event weights are constructed from a fit in x_r. The non-reference variables therefore provide the more stringent closure test because they are not directly constrained by the reference-variable fit.

Table 1 summarizes the results obtained from performing 1000 pseudo-experiments generated with fixed truth parameters and analyzed independently with each weighting method. For each pseudo-experiment and each fitted parameter θ, a weighted unbinned maximum-likelihood fit returns an estimate $\hat{\theta}$ and an estimated fit uncertainty $\sigma_{\hat{\theta}}$. The table reports the mean absolute pull $\langle|(\hat{\theta} - \theta_{true})/\sigma_{\hat{\theta}}|\rangle$ over the set of pseudo-experiments. For an unbiased estimator with correctly estimated uncertainties, pulls are approximately Gaussian with zero mean and unit width, implying an expected mean absolute pull of $\sqrt{2/\pi} \approx 0.8$. With 1000 pseudo-experiments, statistical fluctuations of order ≈ 0.02 around this value are expected, while larger values indicate biases or non-Gaussian tails. A

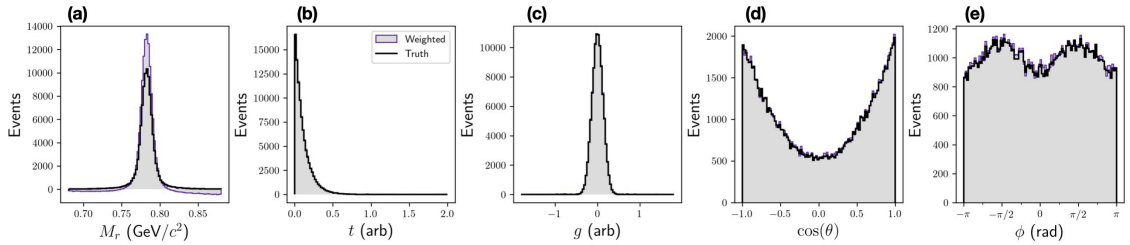

Figure 2: Comparison of ${}_sQ$-factor weighted distributions (shaded) and the corresponding true signal distributions (black) in the generated Monte Carlo data. Panels show (a) the reference mass M_r, (b) and (c) the spectator variables t and g respectively, and the angular distributions in cos θ (d) and ϕ (e).

useful next step would be a comparison with orthogonal-weight methods such as COWs [3] under the same pseudo-experiment closure tests.

Weighting Method	θ, φ			t	g
	ρ^0_{00}	$\rho^0_{1,-1}$	$\mathrm{Re}\left[\rho^0_{10}\right]$	τ	σ
No Weights	36.38	17.62	15.88	73.71	100.80
Sideband Subtraction	0.80	0.79	0.80	0.79	0.79
Q-Factor	0.80	2.26	0.91	34.73	47.60
Q-Factor (with t)	0.83	3.61	1.27	4.86	40.13
Q-Factor (with g)	1.14	3.88	1.49	30.24	17.06
Q-Factor (with t and g)	1.47	4.14	1.84	7.17	21.11
${}_s\mathcal{P}$lot	0.81	0.80	0.79	0.94	0.92
${}_sQ$-Factor	0.82	0.80	0.79	0.85	0.84
${}_sQ$-Factor (with t)	0.80	0.80	0.79	1.44	0.82
${}_sQ$-Factor (with g)	0.80	0.79	0.79	0.81	1.12
${}_sQ$-Factor (with t and g)	0.79	0.80	0.79	1.36	1.03

Table 1: Performance comparison of event-weighting methods using 1000 pseudo-experiments based on the angular parameters ρ^0_{00}, $\rho^0_{1,-1}$, Re$[\rho^0_{10}]$ and spectator parameters τ, and σ. Color indicates the mean absolute pull used for the comparison (blue better, red worse).

In this benchmark, the Q-factor method gives acceptable results for the angular parameters but performs poorly for spectator parameters τ and σ, where the mean absolute pull becomes very large. This behavior persists even when the neighborhood definition is augmented with the spectator coordinates t and/or g, indicating that this issue is not resolvable by simply adding t and/or g to the neighborhood definition. Replacing the local probability weights with covariance-corrected ${}_sQ$-factor weights substantially improves the agreement with the true parameter values and brings the metric back toward the ~ 0.8 scale in the best configurations, while keeping the angular results competitive and for some parameters, slightly better than the global ${}_s\mathcal{P}$lot in this benchmark. Sideband subtraction also performs very well because the signal peak is narrow and the background under the peak is smooth, so sidebands closely approximate the true background in this simplified toy model. This is a property of the model chosen to generate the Monte Carlo data rather than evidence that sidebands are generally optimal in high-dimensional analyses.

4. Conclusions

Methods that infer signal content from a single discriminating variable, whether through probability weights from a fitted mixture model or through sideband subtraction, provide no guarantee without additional assumptions of correct signal extraction in other observables, so apparent agreement in the discriminant alone is not a reliable validation when correlations or phase space dependence are present. The ${}_sQ$-factor method addresses this by retaining the kNN locality of Q-factors while replacing the neighborhood probability weight with a covariance-corrected weight built from the local yield covariance, following the ${}_s\mathcal{P}$lot logic. In toy Monte Carlo studies, this improves the recovery of non-reference signal distributions and reduces bias in extracted parameters relative to standard probability weight methods. The ${}_sQ$-factor method is most useful when phase space variation necessitates local fits and when the reference variables remain only weakly correlated with the observables being extracted at the neighborhood scale.

Acknowledgments

The work leading to this article was partially supported by the U.S. Department of Energy, Office of Science, Office of Nuclear Physics under contract DOE Grant No. DE-FG02-87ER40315.

References

[1] M. Pivk and F. R. Le Diberder, "${}_s\mathcal{P}lot$: A statistical tool to unfold data distributions," *Nucl. Instrum. Methods Phys. Res., Sect. A*, **555**, 356–369 (2005), doi:10.1016/j.nima.2005.08.106.

[2] M. Williams, M. Bellis, and C. A. Meyer, "Multivariate side-band subtraction using probabilistic event weights," *J. Instrum.*, **4**, P10003 (2009), doi:10.1088/1748-0221/4/10/P10003.

[3] H. Dembinski, M. Kenzie, C. Langenbruch, and M. Schmelling, "Custom Orthogonal Weight functions (COWs) for event classification," *Nucl. Instrum. Methods Phys. Res., Sect. A*, **1040**, 167270 (2022), doi:10.1016/j.nima.2022.167270.