Carnegie Mellon University

Improved Probabilistic Event Weighting via Covariance-Corrected Q-Factors for Signal Isolation

Zachary Baldwin*
and Nathaniel Dene Hoffman

ABSTRACT

In complex particle physics analyses where signal and background events are intertwined across multidimensional phase space, statistically consistent event-by-event weighting is indispensable for unbiased extraction of signal observables. However, many widely used methods can often fail to correctly estimate this separation, particularly in the presence of statistically independent variables or when key model assumptions fail. We assess the limitations of these standard techniques with particular focus on Q-factors -- an adaptive local fitting method based on k-nearest neighbors. Although Q-factors offer enhanced flexibility over global fits, it inherently assumes statistical independence between discriminating and weighted variables, leading to a bias when this condition is violated. To address this, we introduce a corrected formalism, ${}_sQ$-factors (pronounced /skju:/, as in "skew"), which integrates the local adaptivity of Q-factors with the covariance-based corrections from ${}_s\mathscr{P}$lot. This hybrid approach restores statistical consistency across dimensions while still preserving local sensitivity, enabling unbiased signal extraction in complex, multidimensional analyses. Through Monte Carlo simulations, we demonstrate that ${}_sQ$-factors consistently outperform traditional methods in both signal recovery and physics parameter estimation. These studies highlight the robustness and accuracy of the method in high-dimensional analyses.

Traditional Methods

- *Cut based*
- *Sideband Subtraction*

$$w^{SB}(y_e) = \begin{cases} 1 & y_e \in U_S \\ -\dfrac{\int_{U_S} f_B(y)\,dy}{\int_{U_L} f_B(y)\,dy + \int_{U_R} f_B(y)\,dy} & y_e \in U_L \cup U_R \end{cases}$$

Signal: $U_S = [y^S_{min}, y^S_{max}]$

Sidebands: $U_L = [y^B_{min}, y^B_{max}]$ $U_R = [y^S_{min}, y^B_{max}]$

- in$\mathscr{P}$lot

$$w^{PR}_\gamma(y_i) = \frac{N_\gamma f_\gamma(y_i)}{\sum_{\chi\in\{S,B\}} N_\chi f_\chi(y_i)}$$

PDF $f_\gamma(y_i, \vec{\alpha})$ Total Likelihood

Advanced Methods

GOAL: Properly handle signal-to-background separation

${}_s\mathscr{P}$lot

⇒ allows for the unbinned, model-independent separation of signal and background

⇒ based on control variables that are not used in the fit via maximum likelihood estimation

$$w^{SW}_\gamma(y_i) = \frac{\sum_{\omega\in\{S,B\}} V_{\gamma\omega} f_\omega(y_i)}{\sum_{\chi\in\{S,B\}} N_\chi f_\chi(y_i)}$$

$$V^{-1}_{\gamma\omega} \equiv \sum_{j=1}^{N_{tot}} \frac{f_\gamma(y_j) f_\omega(y_j)}{\left(\sum_{\chi\in\{S,B\}} N_\chi f_\chi(y_j)\right)^2}$$

Q-factor

⇒ provides local, per-event signal weights obtained from a k-Nearest Neighbor (kNN) fit in phase-space

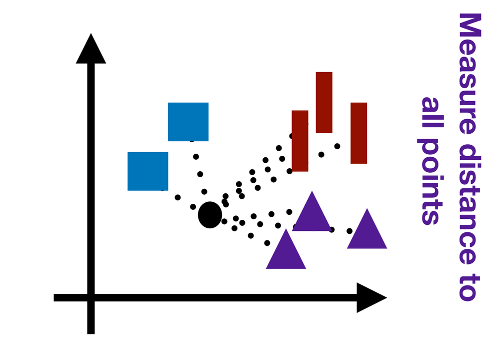
⇒ generalizes side-band subtraction to high dimensions w/out an explicit spectator-PDF only kNN distances used

$$Q_{x_i} = \frac{N_S\, f_S(x_i)}{\sum_{\chi\in\{S,B\}} N_\chi f_\chi(x_i)}$$

Taking the limit $k \to N_{Tot}$, reduces to in$\mathscr{P}$lot

$$d^2_{ij} = \sum_{x\in X\subseteq\Theta} \left[\frac{x_i - x_j}{\sigma_x}\right]^2$$

Normalized Euclidean Distance



Build on the foundation of Q-factors by incorporating corrections from $s\mathscr{P}$lot

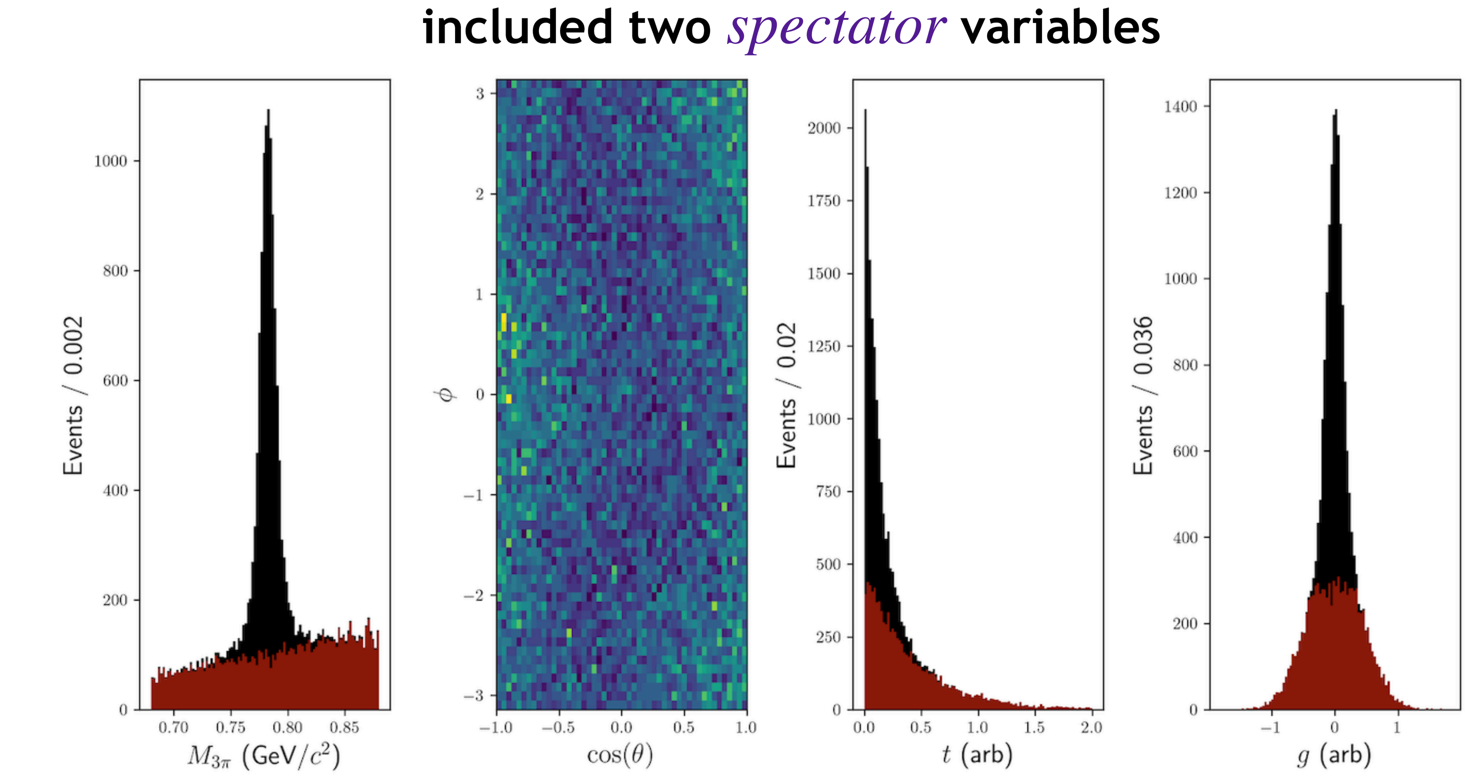
$$Q_{x_i} = \frac{N_S\, f_S(x_i)}{\sum_{\chi\in\{S,B\}} N_\chi f_\chi(x_i)} \longrightarrow {}_sQ_{\gamma x_i} = \frac{V_{\gamma S}\, f_S(x_i) + V_{\gamma B}\, f_B(x_i)}{N_s f_S(x_i) + N_B f_B(x_i)}$$

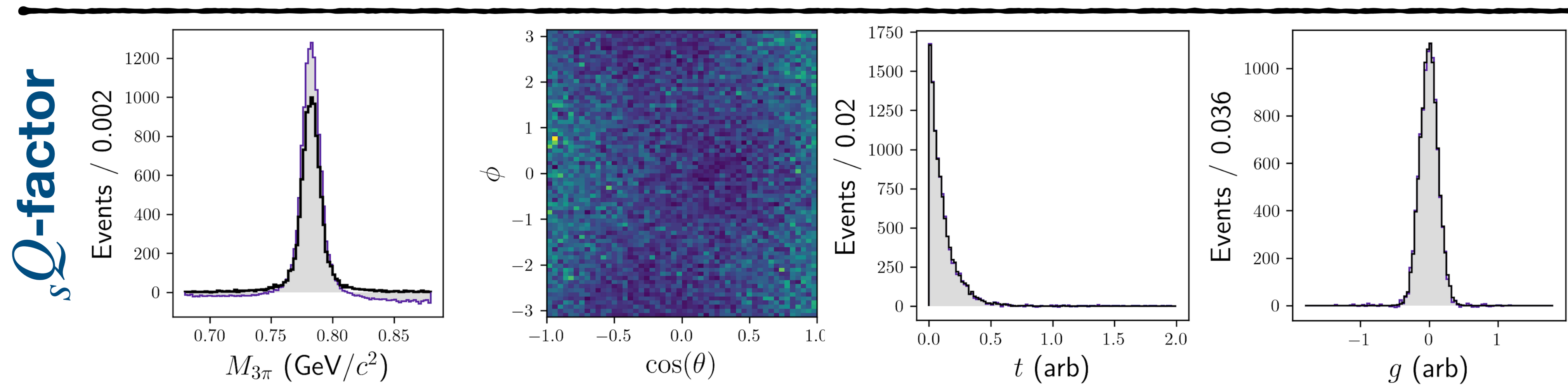
Correction

$$V^{-1}_{\gamma\omega} = \sum_{j=1}^{N_{tot}} \frac{f_\gamma(x_j) f_\omega(x_j)}{(N_S f_S(x_j) + N_B f_B(x_j))^2} = \frac{\partial^2(-ln\mathscr{L})}{\partial N_S \partial N_B}$$

Covariance Matrix Element w.r.t. N_s/N_b

Generated similar dataset to original Q-factors paper, but included two *spectator* variables

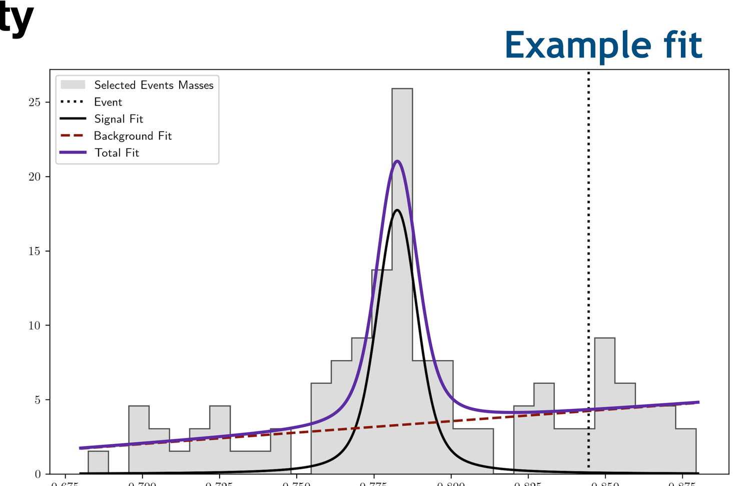




- ${}_sQ$-factor recovers the true signal while honoring kNN locality and embedding the yield-covariance correction
- Q-factors either need ${}_s\mathscr{P}$lot or Monte Carlo closure checks
  → ${}_sQ$-factors carry that covariance term by design

When spectator–mass correlations are strong or data are background-dominated,
→ ${}_sQ$-Factors could be the preferred method

Example fit



Conclusion & Next Steps

Publishing Soon!

${}_sQ$-factors routinely out performs other methods

Multiple iteration fits

Weighting Method	θ,φ: ρ^0_{00}	θ,φ: $\rho^0_{1,-1}$	θ,φ: Re$[\rho^0_{10}]$	t: τ	g: σ
No Weights	36.378	17.620	15.877	73.712	100.801
Sideband Subtraction	3. 0.797	2. 0.793	0.798	1. 0.790	1. 0.789
InPlot	12.899	6.505	5.827	35.825	49.336
Q-Factor	5. 0.802	2.258	0.908	34.732	47.597
Q-Factor (with t)	0.830	3.611	1.274	4.857	40.134
Q-Factor (with g)	1.135	3.884	1.494	30.243	17.060
Q-Factor (with t and g)	1.473	4.141	1.840	7.169	21.111
sPlot	0.813	0.796	5. 0.794	4. 0.943	4. 0.918
sQ-Factor	0.818	4. 0.795	3. 0.792	3. 0.853	3. 0.837
sQ-Factor (with t)	2. 0.795	5. 0.796	2. 0.790	1.442	2. 0.820
sQ-Factor (with g)	4. 0.802	1. 0.792	1. 0.789	2. 0.809	1.121
sQ-Factor (with t and g)	1. 0.791	3. 0.795	4. 0.792	5. 1.356	5. 1.031

Absolute mean pull of each parameter over 1000 independent simulations (*smaller numbers are better*)

- limitations to Q-factor (${}_s\mathscr{P}$lot) method → ${}_sQ$-factor introduced to account for these
- maintains robustness across a variety of test conditions and scenarios
- implementation on GlueX data provides accurate results compared to traditional methods

References

M. Pivk and F. R. Le Diberder, ${}_s\mathscr{P}$lot: A statistical tool to unfold data distributions, Nucl. Instrum. Meth. A 555 (2005) 356–369, https://doi.org/10.1016/j.nima.2005.08.106.

${}_s\mathscr{P}$lot | Q-Factor

M. Williams, M. Bellis, and C. A. Meyer, Multivariate side-band subtraction using probabilistic event weights, JINST 4(2009) P10003, https://doi.org/10.1088/1748-0221/4/10/P10003.