# A multidimensional, event-by-event, statistical weighting procedure for signal to background separation

*Zachary* Baldwin[1,*]

[1]Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

**Abstract.** Numerous analyses performed in nuclear and particle physics are in search of signals that are contaminated by irreducible background that cannot be suppressed using event-selection criteria. These background events can lead to unphysical or biased results when extracting physical observables and need to be taken into account. Exploring a data set across multiple dimensions allows us to characterize the phase space of a desired reaction through a set of coordinates. For a subset of these coordinates, known as reference coordinates, signal and background follow different distributions with known functional forms with potential unknown parameters. The Quality Factor approach uses the space defined by the remaining non-reference phase space coordinates to determine the $k$-nearest neighbors of an event. The distribution of these neighbors in the reference coordinates undergoes a fit with the sum of the signal and background model functions, employing techniques like the unbinned maximum likelihood method, to extract the signal fraction, or Q-factor. This quality factor, which is defined for each event, is equal to the probability that it originates from the signal of interest. In this document, we will give a brief overview of this procedure and illustrate examples using Monte Carlo simulations and data from the GlueX experiment at Jefferson Lab.

## 1 Introduction

Practically all experimental data are expected to contain some undesired background components. Rather than ignoring these backgrounds, which could lead to inaccurate and biased results, it is necessary to devise algorithms which can reduce their impact. Therefore, developing a sophisticated method to separate out meaningful regions of interest, impacts numerous science domains, especially in high-energy and nuclear physics.

Extensively based on Ref. [1], where a complete description of the method can be found, this document will offer a brief summary of the characteristics of the Quality Factor method. This approach is a multidimensional, event-by-event, statistical weighting procedure that is capable of separating signal regions of interest from background regions. Unlike other popular methods, including machine learning algorithms that rely on training predetermined data sets, the Quality Factor method is distinct due to not requiring *a priori* information of the signal or background spectra across diverse coordinates, only in the defined reference coordinates. To showcase the advantages and accuracy of the Quality Factor method, we analyze Monte Carlo simulations and photo-production data obtained from the GlueX experiment at Jefferson National Lab. The presented results validate the method's ability to reliably distinguish signal regions from background.

*e-mail: zbaldwin@cmu.edu

## 2 Signal to Background Separation Approaches

Due to the nature of certain particle reactions, measurements may lead to data that are contaminated with *irreducible* backgrounds. This form of background cannot be suppressed by any basic selection or rejection techniques, making it impossible to be certain whether an event is purely signal or background.

Take for instance the reaction $\gamma p \to p\eta$ where $\eta \to \pi^0\pi^+\pi^-$. Backgrounds from the reaction $\gamma p \to p\omega$ where also $\omega \to \pi^0\pi^+\pi^-$ could be irreducible for the $\eta$ reaction. Since the cross section for $\omega \to \pi^0\pi^+\pi^-$ is large , the natural width of the $\omega$ resonance, in combination with finite detector resolution, will make the low mass tail of the $\omega$ in the $\pi^0\pi^+\pi^-$ spectrum leak into the $\eta$ region and hence contaminating the $\eta$ peak. These and other comparable background events will lead to a significant amount of non-$\eta$ event contamination and become indistinguishable to the $\eta$ events of interest.

While other methods, were developed for signal to background separation (e.g. the ${}_s\mathcal{P}lot$ technique, see Ref. [10]), this document will specifically compare the commonly utilized sideband subtraction technique and the Quality Factor method.

### 2.1 Sideband Subtraction Description

The simplest method to assist in dealing with irreducible background is the *sideband subtraction* technique. This technique estimates the number of background events in a desired signal region of interest, based on a pure background sample extracted in one or more background regions (i.e. sidebands), and then statistically subtracts them. The underlying assumption is that the background distributions in the non-discriminating kinematic variables are sufficiently similar in the sidebands and in the signal region. For example, close to threshold, sidebands could behave notably different from the signal region creating issues. The technique is extremely useful due to the fact that it is straightforward, has no formal definition for the relative size of each region, and can incorporate appropriate weights that can be used in further studies (e.g. partial wave analysis). However, like most methods, there are limiting factors. Apart from differing background distributions in the sidebands and the signal region, there is a risk of under or over subtraction, leading to issues with systematic effects. This is compounded by the fact that the scalability of the method diminishes in higher dimensions, as sideband regions quickly become sparsely populated.

### 2.2 Quality Factor Method Description

To mitigate some of the limitations that are viable while performing the sideband subtraction method, Mike Williams, et al. developed the *Quality Factor* method. It specifically generalizes the one (and two) dimensional sideband subtraction method to higher dimensions. This enhanced technique alleviates the need for data to be binned and reduces concerns about background variations in kinematically distinct regions during analyses.

A probabilistic weight is determined on an event-by-event basis (i.e. a quality factor $Q$) such that the event of interest is a signal event ($Q^i$) or a background event ($1 - Q^i$) [1]. The Quality Factor method relies on knowledge that the phase space of a particle reaction is described by a set of coordinates ($\vec{\xi} = (\xi_1, \ldots, \xi_m)^T$). For a subset of these coordinates (i.e. the reference coordinates), the signal and background distributions are sufficiently different, and their functional form is known up to unknown parameters. In the analyses presented here, the subspace of the reference coordinates is one-dimensional with the reference coordinate $\xi_r$, but the method can be applied also to higher-dimensional reference-coordinate spaces. There are several assumptions that need to be made when utilizing the Quality Factor algorithm. These include: the distributions of signal and background being known in a subset of coordinates, signal and background must not interfere, and signal and background kinematics not varying rapidly in the non-reference coordinates.

The normalized Euclidean distance is chosen as the metric where,

$$d_{ij}^2 = \sum_{k\neq r}^{m} \left[ \frac{(\xi_k^i) - (\xi_k^j)}{\sigma_k} \right]^2 \tag{1}$$

with $\sigma_k$ being the root mean square of the $k^{th}$ value between any pair of events $\xi_k$ with indices $i$ and $j$ for the non-reference coordinates. This metric allows us to compute the distance between any combination of events within phase space, determining how close two events are in the non-reference coordinates. For each event, we gather the $n_c$ nearest neighbor events, including the event itself.

Once the $n_c$ nearest neighbor events are obtained, their distribution in the references coordinates are fit by the sum of the number densities $F_s(\xi_r, \hat{\alpha}_s)$ for the signal events and $F_b(\xi_r, \hat{\alpha}_b)$ for the background events using, for instance, the unbinned maximum likelihood method (MLE). Here, $\hat{\alpha}_{s,b}$ represents the unknown parameters of the distribution functions. From the resulting fits, a Q-factor value can be determined for each event $i$ in the data, analogous to the Q-factor calculation in the original work [1]:

$$Q^i = \frac{F_s(\xi_r^i, \hat{\alpha}_s^i)}{F_s(\xi_r^i, \hat{\alpha}_s^i) + F_b(\xi_r^i, \hat{\alpha}_b^i)} \tag{2}$$

where $\xi_r^i$ represents the value of the reference coordinate for the given event and $\hat{\alpha}_{s,b}^i$ are the estimates for the parameters obtained from the individual fit for the event $i$. Since $F_s$ denotes the signal function dependent on $\xi_r$ (similar for $F_b$), then each Q-factor value is the signal fraction. Therefore, by weighting each event in the data set by the calculated Q-factor, the background is subtracted.

One caveat of the Quality Factor method is that it is a very computationally expensive technique [2]. For every event in the data, the method requires calculating the distance, determining the $n_c$ nearest neighbors, and then performing an unbinned maximum likelihood fit, making each application resource-intensive as well as time consuming. An important parameter of the method is the number $n_c$ of nearest neighbors. For the best results, we want only events close to the event of interest [3]. If $n_c$ is chosen too large in comparison to the number of data events, then the advantages of the Quality Factor method are lost and are analogous to the sideband subtraction technique. While too small of an $n_c$ value and fit instability takes control such that the errors will be significant. Since this method is the generalized sideband subtraction, it is satisfying to see that as the value of the closest nearest neighbors $n_c$ approaches the total number of events $n$, this method becomes just a sideband subtraction in the limit of large clusters [1].

## 3 Toy-Model Monte Carlo Example

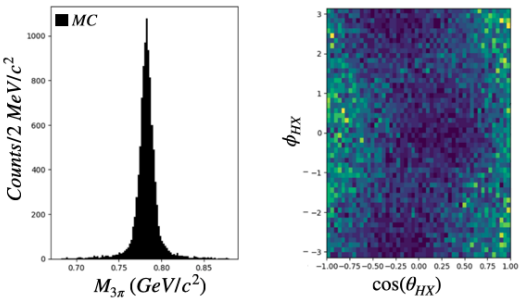



**Figure 1:** Generated signal Monte Carlo with $M_{3\pi}$ invariant mass alongside the decay angular distributions, $\phi_{HX}$ and $\cos(\theta_{HX})$. These plots accurately depict the true signal distribution.

To demonstrate the capability of the Quality Factor method in comparison to the sideband subtraction technique, a toy-model Monte Carlo sample was generated, similar to the procedure in the original Quality Factor paper.

The simulation depicted in Fig. 1 was generated for the reaction $\gamma p \rightarrow p\omega$ where the $\omega$ decays to $\pi^0\pi^+\pi^-$ approximately 90% of the time. The $\omega$ signal events were generated according to 3-body phase space weighted by a Voigtian in the mass of the $3\pi$ system, while the non-$\omega$ background was generated according to 3-body phase space weighted by a linear function in the $3\pi$ mass. Opting to work in the helicity system, which is consistent with the original paper, the angular components (as depicted in the rightmost plot of Fig. 1) represent the polar and azimuthal angles of vector $(\vec{p}_{\pi^+} \times \vec{p}_{\pi^-})$ in the decay plane of the $\omega$ at rest [1]. One of the main assumptions in this Monte Carlo generation is that signal and background do not interfere with one another.

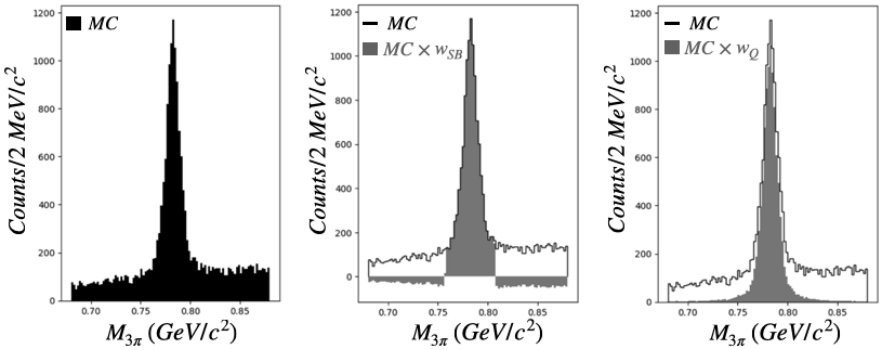



**Figure 2:** The invariant mass of simulated $\pi^0\pi^+\pi^-$ for $\omega$ signal events + background events (left) is displayed. Additionally, the generated Monte Carlo sample is displayed showing the weighted distribution for the signal region and sidebands (middle), along with the same Monte Carlo sample, now weighted by the Q-factors.

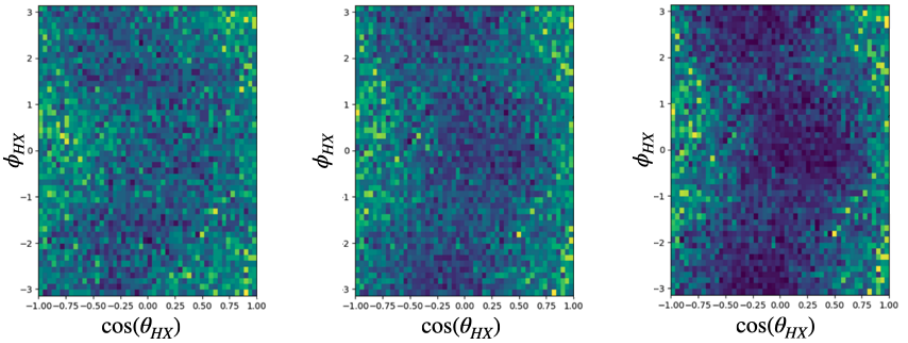



**Figure 3:** Similar to Ref. [1], $\phi_{HX}$ (radians) vs $\cos(\theta_{HX})$ is plotted for simulated events. By using the signal+background Monte Carlo (left), a vast difference in signal separation power can be observed when applying the sideband subtraction technique (middle) versus the Quality Factor method (right) when compared to Fig. 1.

The left-hand plot in Fig. 2 illustrates the features of the simulated signal+background Monte Carlo sample, displaying a prominent peak at the nominal $\omega$ resonance mass of 0.782 GeV/$c^2$ that is formed against a linear background. This shows the successful generation of the desired $\omega$ signal signature amidst the non-$\omega$ background interference. The middle plot shows the signal region of interest and sidebands with applied negative weights that are used

for statistical subtraction. Lastly, on the right, the Quality Factor method demonstrates its separation power by clearly reproducing the generated signal distribution displayed in Fig. 1. As observed, the sideband subtraction technique is not able to adequately display it's true ability in the reference coordinate, as the Quality Factor method can.

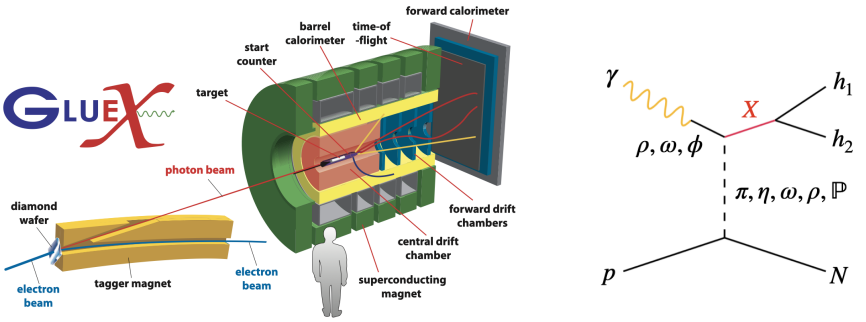
More interestingly, the stark contrast between the strength of these two methods is evident in Fig. 3. By observing the decay angular distributions, $\phi_{HX}$ and $\cos(\theta_{HX})$, with each of the method's respected weights applied to the Monte Carlo, it is clear that the Q-factor weighted distribution is closer to the true distribution as seen in Fig. 1.

## 4 Experimental Results

The Quality Factor method was originally developed to study $\omega$ photo-production, in CLAS data [4]. Since then, it has contributed to various analyses across different experiments [5]– [6] etc., and shows promise in benefiting the *Gluonic Excitation* (GlueX) experiment as well.

### 4.1 GlueX Experiment

The GlueX experiment has the main goal of mapping the spectrum of light hadrons in addition to understanding the fundamental nature of confinement within Quantum Chromodynamics (QCD). To accomplish this, an emphasis is placed on searching for the evidence of quantum numbers a quark-antiquark ($q\bar{q}$) system would be forbidden to have in the current constituent quark model. Observing a state with these non-$q\bar{q}$ quantum numbers would be ideal evidence for new exotic forms of matter [7]. Lattice QCD provides predictions that feature the possibility of such evidence, a light exotic meson with the forbidden quantum numbers $J^{PC} = 1^{-+}$, which can be found in the $\eta\pi^0$ system if the system is observed in an odd orbital angular momentum state [8]. Here, $J$ represents the total angular momentum of the $q\bar{q}$ system, $P$ is parity, and $C$ is charge conjugation.




(a) (b)

**Figure 4:** Provided in Fig. 4a is a diagram of the unique experimental design of GlueX apparatus which entails each detector subsystem. Fig. 4b presents a concise Feynman diagram depicting the typical photo-production processes investigated at GlueX.

The GlueX detector apparatus was designed to achieve precise measurements of both neutral and charged final states with high-energy and momentum resolution. A distinctive feature of this experiment is the utilization of linearly polarized photons. This type of photo-production produces resonances with many different $J^{PC}$ quantum numbers, which leads to a greater potential to discover exotic mesons.

Searching for an exotic signal within the mass spectrum of any system is inherently challenging due to the substantial background and the presence of overlapping states within the relevant mass range for a potential exotic signature [7]. It's crucial to note that "background"

here encompasses both conventional resonances and exotic resonances, which arise from a superposition of quantum amplitudes. Traditional statistical subtraction alone cannot eliminate this background, as it stems from the intertwined nature of quantum amplitudes associated with both exotic and conventional resonances.

It becomes necessary to incorporate a partial-wave analysis (PWA) for distinguishing exotic resonances from conventional ones by exploiting the distinct angular distributions of each state's decay products. For an effective PWA, a high-purity sample is crucial, minimizing contaminations from other processes. This is exactly where the Quality Factor method becomes significant. This algorithm effectively reduces unwanted background, aiding the extraction of potential exotic signatures by discerning differences in angular distributions, mass, and other relevant parameters between signal and background events [3].

### 4.2 Application to GlueX data on photo-production of $\eta\pi^0$

In order to discuss the application of the Quality Factor to actual GlueX data, an understanding of the different reference frames and their respective angles is essential.

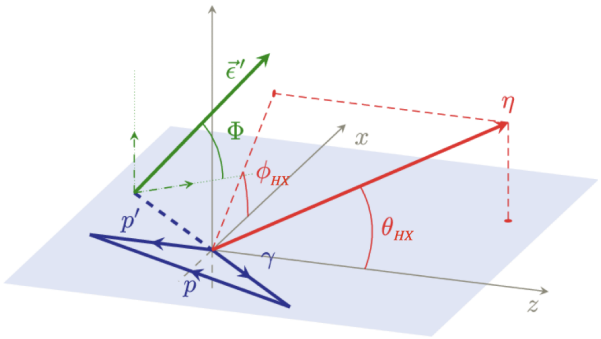


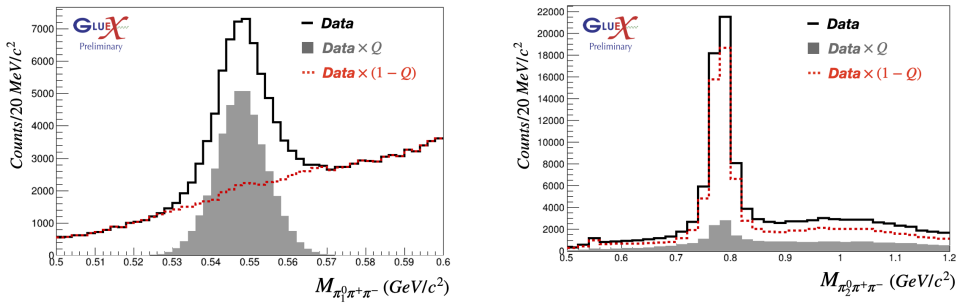

**Figure 5:** Illustration of the helicity frame used to describe the photo-production of $\eta\pi^0$ system at GlueX as seen in Ref. [9]. To acquire the Gottfried-Jackson angles, a simple rotation is necessary where the direction of the beam momenta is aligned with the $z$-axis.

For this experiment, the reaction plane is defined by the incoming photon ($\gamma$), and the recoiling proton ($p'$) in the $\eta\pi^0$ center-of-mass frame. With respect to this reaction plane, the linearly polarized photon beam produces an angle $\Phi$ from its beam polarization angle $\epsilon$ as shown in Fig. 5. The angles $\phi_{HX}$ and $\theta_{HX}$ are the azimuthal and polar angles of the $\eta$ [9].

We analyze the $\eta\pi^0$ system where $\eta\pi^0 \to \pi^0_1\pi^0_2\pi^+\pi^- \to 4\gamma\pi^+\pi^-$. The branching fraction for $\eta \to \pi^0_{1,2}\pi^+\pi^-$ and both $\pi^0_{1,2}$ decays are approximately 39% and 99%, respectively. This channel is complicated due to the indistinguishable neutral pions, which can form a $\pi^0\pi^+\pi^-$ subsystem using either $\pi^0_1$ or $\pi^0_2$. To implement the procedure outlined in Sec. 2.2, the chosen reference coordinate $\xi_r$ in the Quality Factor method corresponds to the invariant mass of the $\pi^0_1\pi^+\pi^-$ subsystem, focusing in the range of $0.5\ \mathrm{GeV}/c^2 < M_{\pi^0_1\pi^+\pi^-} < 0.6\ \mathrm{GeV}/c^2$. This decision is motivated by the presence of a distinct $\eta$ signal peak in the GlueX data, situated atop a background that may include a mixture of combinatorial or other resonance backgrounds, and proves challenging to reduce further through selection criteria alone.

The non-reference coordinates $\xi_k$ in the analysis encompass: $\Phi$, $\cos(\theta_{HX})$ (as illustrated in Fig. 5), $\cos(\vartheta_{HX}^{(\pi^0_1\pi^+\pi^-)})$ and $\phi_{HX}^{(\pi^0_1\pi^+\pi^-)}$ (in the helicity $\pi^0_1\pi^+\pi^-$ decay reference frame), as well as $\cos(\vartheta_{HX}^{(\pi^0_2\pi^+\pi^-)})$ and $\phi_{HX}^{(\pi^0_2\pi^+\pi^-)}$ (in the helicity $\pi^0_2\pi^+\pi^-$ decay reference frame). Considering the orientation of the normal to the decay plane $\eta \to (\pi^0)_{1,2}\pi^+\pi^-$ subsequent to boosting from the rest frame of the $\eta\pi^0$ system, parameters such as $\cos(\vartheta_{HX}^{(\pi^0_1\pi^+\pi^-)})$, $\phi_{HX}^{(\pi^0_1\pi^+\pi^-)}$, $\cos(\vartheta_{HX}^{(\pi^0_2\pi^+\pi^-)})$ and

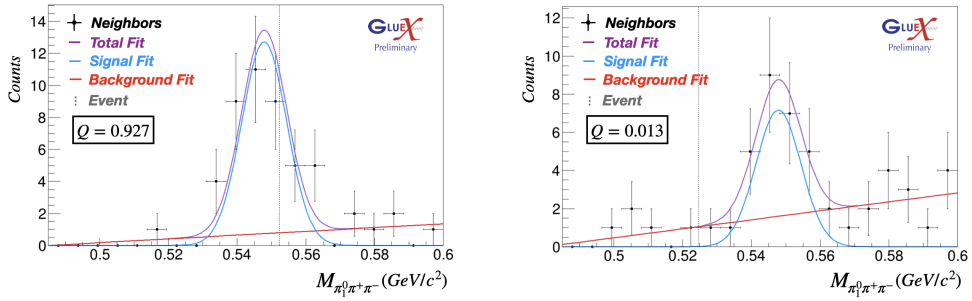
$\phi_{HX}^{(\pi_2^0\pi^+\pi^-)}$ played a crucial role in further enhancing the separation of signal and background in the $\eta\pi^0$ system.




(a) (b)

**Figure 6:** Both plots showcase the data (black line), the signal distribution (data weighted by the Q-factor weights; shaded grey), and the background distribution (data weighted by $(1-Q)$; red dashed line). Specifically, Fig. 6b displays the *other* $\pi^0\pi^+\pi^-$ system not used for the reference coordinate. Here, challenges and subtleties within the method's application can be observed.

A distinct peak at the nominal $\eta$ resonance mass of 0.547 GeV$/c^2$ is showcased in Fig. 6a along with the background clearly described after the Q-factors are applied. Since the $\eta\pi^0$ system contains 4 photons that can be reconstructed back to 2 indistinguishable neutral pions, an observation of Fig. 6b provides the *other* $\pi^0\pi^+\pi^-$ system where the neutral pion is not the same (i.e. $M_{\pi_2^0\pi^+\pi^-}$) as in Fig. 6a. The method excellently handles most of the background encompassing the desired $\eta$ signal but is unable to remove all $\omega$ background throughout the whole system. Individual fits in Fig. 7 are shown for a particular event to display what the algorithm is executing behind the curtains to calculate the probabilistic Q-factors.




(a) (b)

**Figure 7:** A single unbinned maximum likelihood fit from the computation of Q-factors in the GlueX data. The gray-dashed line is an evaluation of the of the fitted signal (blue) and background (red) distribution at the current events reference coordinate value.

The distributions in the reference coordinate are modelled by utilizing a Gaussian function for the $\eta$ signal (blue) and a Bernstein polynomial of the $2^{nd}$ degree for the non-$\eta$ background (red). The fit of the $M_{\pi_1^0\pi^+\pi^-}$ distribution in Fig. 7a resulted in a calculation of $Q = 0.927$ for that candidate. Therefore, this event has a 92.7% probability that it originated from the signal. Similarly, the fit in Fig. 7b gave a calculation of $Q = 0.013$, which is 1.30%. Therefore, this event most likely is not signal and is indeed background.

## 5 Conclusion

In this study, we applied the Quality Factor method from Ref. [1] to photo-production data from the GlueX experiment and Monte Carlo simulations to showcase its effectiveness in

separating signal events from non-interfering backgrounds. Performing input-output Monte Carlo studies revealed that the Quality Factor method removed more background compared to the sideband subtraction technique. A short list of the advantages and disadvantages for the method is provided below.

| Advantages | Disadvantages |
|---|---|
| • Event weights applied to log-likelihood for background subtraction | • Potential issues with correlated coordinates |
| • Scales to higher dimensions without binning | • Sensitivity to outliers may impact signal extraction |
| • No *a priori* information regarding the signal or background over various coordinates | • Dependence on the quality of the reference coordinates selection |
| • Offers interpretability through the probabilistic nature | • Computationally expensive |

### 5.1 Outlook

While the Quality Factor method has demonstated promise in several analyses, further studies still need to be performed to comprehensively understand its performance. Those potentially include: using a different tactic of data classification instead of the nearest neighbors approach, applying machine learning procedures to assist with the separation of differing events, comparing other separation methods besides just the sideband subtraction technique, investigating the impact of correlations in data, and leveraging High-Performance Computing resources for computational speed enhancement. These endeavors are expected to contribute to the continued success in the application of the Quality Factor method in experimental analyses.

## References


[1] M. Williams, M. Bellis, C. Meyer, *Multivariate side-band subtraction using probabilistic event weights*, JINST **4**, P10003 (2009)

[2] L. Ng (2021), *Q-Factors* (Source code), URL: https://github.com/lan13005/Q-Factors

[3] C. Meyer (2020), *Notes on Q-Factor Analyses* (GlueX-doc-5526), URL: https://halldweb.jlab.org/DocDB/0055/005526/004/qfactor.pdf

[4] M. Williams et al., *Partial wave analysis of the reaction $\gamma p \to p\omega$ and the search for nucleon resonances*, Phys. Rev. C **80**, 065209 (2009)

[5] M. Albrecht et al. *Coupled channel analysis of $\bar{p}p \to \pi^0\pi^0\eta$, $\pi^0\eta\eta$ and $K^+K^-\pi^0$ at 900 MeV/c and of $\pi\pi$-scattering data*, Eur. Phys. J. C **80.5**, 453 (2020)

[6] P. Roy et al., *Measurement of the beam asymmetry $\Sigma$ and the target asymmetry $T$ in the photo-production of $\omega$ mesons off the proton using CLAS at Jefferson Laboratory*, Phys. Rev. C **97.5**, 055202 (2018)

[7] C.A. Meyer and E.S. Swanson, *Hybrid mesons* Progress in Particle and Nucl. Phys. **82**, 21–58 (2015)

[8] J. Dudek et al., *Isoscalar meson spectroscopy from lattice QCD*, Phys. Rev. D **83**, 5 (2011)

[9] V. Mathieu et al., *Moments of angular distribution and beam asymmetries in $\eta\pi^0$ photoproduction at GlueX*, Phys. Rev. D **100**, 16 (2019)

[10] M. Pivk, F.R. Le Diberder, *sPlot: A statistical tool to unfold data distributions*, Nucl. Instrum. and Meth. in Phys. Research **A 555**, 356 (2005)