Estimation of combinatoric background in seaquest using an event-mixing method

Published by IOP Publishing for Sissa Medialab

Received: *February 9, 2023*
Revised: *June 29, 2023*
Accepted: *September 19, 2023*
Published: *October 26, 2023*

Estimation of combinatoric background in seaquest using an event-mixing method

S.F. Pate,[a,*] **A. Pun,**[a] **M.F. Hossain,**[a] **K. Nagai,**[b] **C.A. Aidala,**[c] **C. Ayuso,**[c] **L. El Fassi,**[d,e] **D.F. Geesaman,**[f] **T.J. Hague,**[g,1] **E.R. Kinney,**[h] **W. Lorenzon,**[c] **K. Nakano,**[i,j] **P.E. Reimer,**[f] **M.B.C. Scott**[c] **and R.S. Towell**[g]

[a] *Department of Physics, New Mexico State University, Las Cruces, NM, 88003, U.S.A.*
[b] *Physics Division, Los Alamos National Laboratory, Los Alamos, NM, 87545, U.S.A.*
[c] *Department of Physics, University of Michigan, Ann Arbor, MI, 48109, U.S.A.*
[d] *Department of Physics and Astronomy, Rutgers, The State University of New Jersey, Piscataway, NJ, 08854, U.S.A.*
[e] *Department of Physics and Astronomy, Mississippi State University, Mississippi State, MS, 39762, U.S.A.*
[f] *Physics Division, Argonne National Laboratory, Lemont, IL, 60439, U.S.A.*
[g] *Department of Engineering and Physics, Abilene Christian University, Abilene, TX, 79601, U.S.A.*
[h] *Department of Physics, University of Colorado, Boulder, CO, 80309, U.S.A.*
[i] *Department of Physics, University of Virginia, Charlottesville, VA, 22904, U.S.A.*
[j] *RIKEN Nishina Center for Accelerator-Based Science, Wako, Saitama 351-0198, Japan*

E-mail: spate@nmsu.edu

Abstract: All experiments observing dilepton pairs (e.g. e^+e^-, $\mu^+\mu^-$) must confront the existence of a *combinatoric* background caused by the combining of tracks not arising from the same physics vertex. Some method must be devised to calculate and remove this background. In this document we describe a particular event-mixing method relying on many of the unique aspects of the SeaQuest spectrometer and data. The method described here calculates the combinatoric background with correct normalization; i.e., there is no need to assign a floating normalization factor that is then determined in a subsequent fitting procedure. Numerous tests are applied to demonstrate the reliability of the method.

Keywords: Analysis and statistical methods; Data processing methods; Data reduction methods

ArXiv ePrint: 2302.04152

[1] Currently at Lawrence Berkeley National Laboratory, Berkeley, CA, 94720, U.S.A.
[*] Corresponding author.

https://doi.org/10.1088/1748-0221/18/10/P10032

Contents

1 Introduction

The SeaQuest experiment looks for dimuon signals coming from either the Drell-Yan process or from the decay of the J/ψ or ψ' mesons [1, 2]. The track pairs we reconstruct from the data not only contain muon pairs from the aforementioned processes but also random combinations of single muons from uncorrelated processes. These background pairs arise not only from multiple physics interactions in the same beam bunch, but also from complex single events like open-charm production, where each charmed meson decays into a muonic channel. Such random combinations of muons, called the *combinatoric* background, need to be subtracted from the data to extract the true signal yields. We discuss an event-mixing method to estimate the combinatoric background from the SeaQuest data, and we demonstrate that this method has the correct absolute normalization.[1]

[1]This manuscript does *not* describe the method used in ref. [2] to estimate the combinatoric background. The method described here has been developed later.

2023 JINST 18 P10032

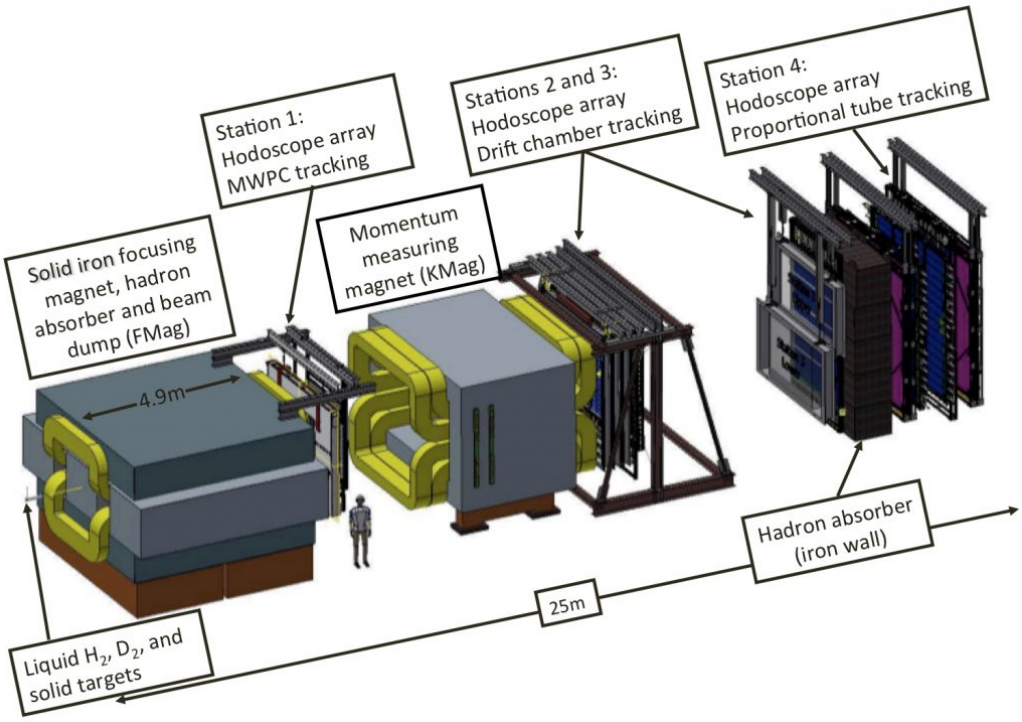

Figure 1. Perspective view of the SeaQuest spectrometer.

Crochet and Braun-Munzinger [3] emphasize two important characteristics of a successful event-mixing process for estimating the combinatoric background. Firstly, the signal density in the data stream must be low. This is the case in SeaQuest, where only approximately 5% of events in the data stream contain a candidate dimuon pair from a photon or vector meson decay; the vast majority of triggered events do not contain tracks from a physics signal. (Additional analysis requirements on these candidate tracks will reduce this raw 5% value.) Secondly, the tracks must be mixed from events that are "similar" to each other. This "similarity" is to guarantee that the tracks are drawn from events with similar track distributions. In ref. [3] the discussion is concentrated around heavy-ion collision data, so it is suggested to divide the events up into centrality and flow classes; centrality strongly affects track multiplicity and flow introduces momentum correlations. In SeaQuest, we will see that the relevant quantity is the station-1 drift chamber occupancy, which is largely driven by the tremendous variation in proton beam bunch sizes delivered to the target; therefore we will sort events by the chamber occupancy.

2 Characteristics of the SeaQuest data sample

The SeaQuest spectrometer is fully described in ref. [1] and a schematic diagram is shown in figure 1. A 120 GeV proton beam from the Fermilab Main Injector was incident upon liquid hydrogen and deuterium targets, and also on variety of solid targets. Particles produced by interactions in the target passed into a 5m-thick iron beam dump, which served to absorb all particles except for highly energetic muons. The beam dump was also a magnet which served to focus the muons into the spectrometer. Following the beam dump were the "station 1" detectors, comprising x- and

2023 JINST 18 P10032

y-measuring hodoscopes and six planes of drift chambers. Following station 1 was an open magnet which served to measure the momentum of the muons. Following this magnet came stations 2 and 3, each composed of hodoscopes and wire chambers. After station 3 was an additional iron absorber for muon identification purposes. Lastly came station 4 comprising hodoscopes and proportional tubes. The hodoscopes were used for triggering purposes, while the wire chambers and proportional tubes were used for track reconstruction. SeaQuest took physics data during the years 2014–2017.

The event trigger in SeaQuest is determined by a field-programmable gate array (FPGA) which looks at the x-measuring (bend-plane) hodoscopes at the four stations of the spectrometer. The FPGA is programmed to look for likely opposite-sign track pairs, based on a simulation of such pairs passing through the hodoscope stations. Even though the FPGA trigger rejects a tremendous amount of useless particle tracks, it is still programmed to be a "loose" trigger, in the sense that it rejects very few valid events. In this study, we will only use one of the FPGA triggers, the one named "top/bottom", which looked for pairs where one track passed through the upper half of the spectrometer and the second track passed through the lower half.

For the purposes of studying the proposed event-mixing method, we used ten one-hour data runs, from the calendar year 2015 beam period. The beam quality was typical for SeaQuest runs; the distribution of the number of protons per 1-ns bucket had a mean of about 25,000 accompanied by a long tail extending to approximately 80,000 protons-per-bucket. The large variation in protons-per-bucket resulted in a large variation in the drift chamber occupancy per event.

The occupancy of the drift chambers at station 1, ω, is defined as the number of hits in the station-1 drift chambers, collectively called D1. The number of wires in each plane in D1 is shown in table 1. The D1 occupancy distribution for top/bottom-triggered events is shown in figure 2.

Table 1. Number of wires on each plane in the station-1 drift chambers.

Plane Name	Number of wires
D1U, D1Up, D1V, D1Vp	201 for each plane
D1X, D1Xp	160 for each plane
Total	1124

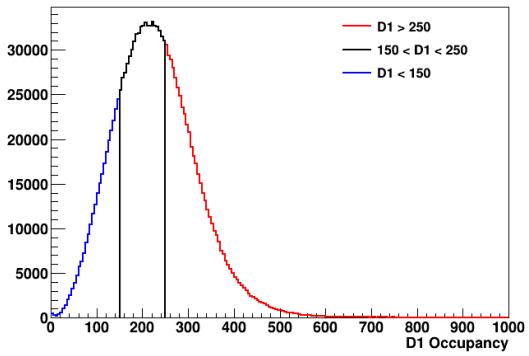

Figure 2. Distribution of the D1 occupancy (ω) for top/bottom-triggered events. The vertical lines show the separation between the regions of low ($\omega < 150$), middle ($150 < \omega < 250$) and high ($\omega > 250$) occupancy.

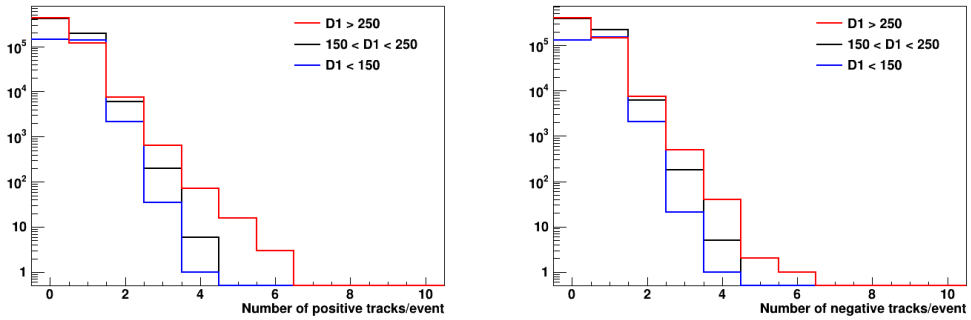

Figure 3. Track multiplicities for positive (left) and negative (right) tracks at different D1 occupancies. The blue, black, and red lines correspond to low, middle and high occupancy values, respectively. There are in general more positive tracks than negative tracks because more positive pions than negative pions are produced in collisions of protons with nuclei.

The D1 occupancy is divided into three different regions; low ($\omega < 150$), middle ($150 < \omega < 250$) and high ($\omega > 250$). Track multiplicities and momentum (P_z and p_T) for different D1 occupancy bins are compared. These histograms are produced using the raw reconstructed tracks; no additional cuts have been applied. The track (positive and negative) multiplicity for different occupancy regions are shown in figure 3. The number of tracks per event increases with increasing occupancy. The P_z momentum distributions for both positive and negative tracks for different occupancy regions are shown in figure 4. The distribution becomes flatter and wider for higher occupancies. The p_T distributions for both positive and negative tracks for different occupancy regions are shown in figure 5. The p_T distribution changes greatly for higher occupancies.

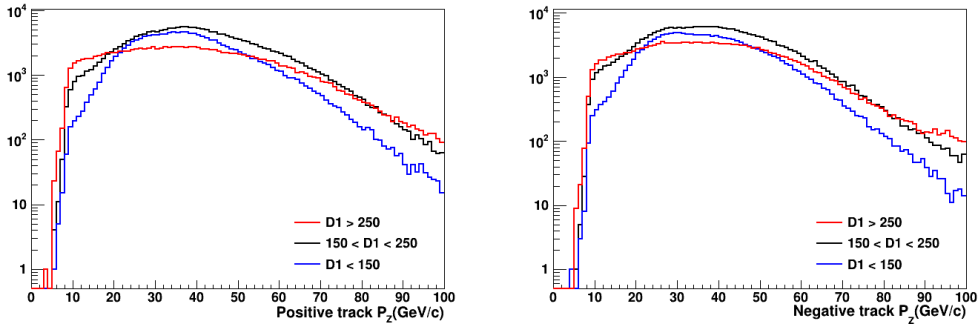

Figure 4. Longitudinal momentum P_z for positive (left) and negative (right) tracks at different D1 occupancies. The blue, black, and red lines correspond to low, middle and high occupancy values, respectively.

In summary, the SeaQuest data stream has a number of important features:

- The majority of top/bottom-triggered events contain zero reconstructed tracks. A similarly large fraction contain only one track.
- Due to the low density of signal events in the data stream, even in events with two tracks there is only a 30% probability that they are a signal pair.

2023 JINST 18 P10032

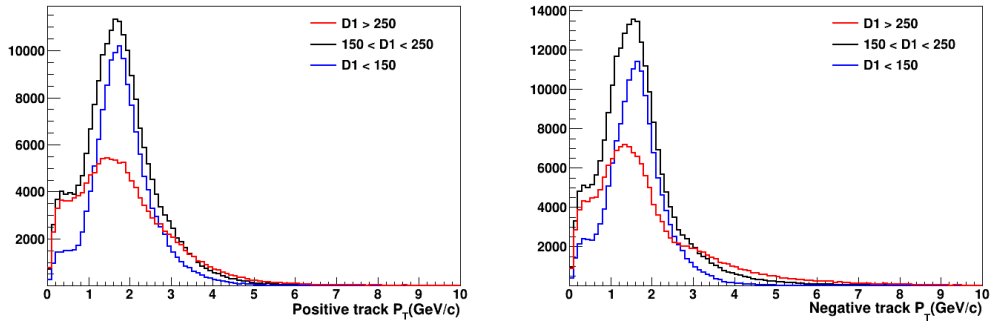

Figure 5. Transverse momentum p_T for positive (left) and negative (right) tracks at different D1 occupancies. The blue, black, and red lines correspond to low, middle and high occupancy values, respectively.

- The number and momentum distribution of the tracks depends strongly on the D1 occupancy; if we plan to mix tracks from different events, we must make sure those events have similar occupancy.

3 Combining tracks to form a spectrum

As we have seen, an event may have 0, 1, 2, or more reconstructed tracks within it. We can separate those tracks into two groups: *signal tracks* which have been produced by a muon from a J/ψ, ψ', or Drell-Yan decay; and *background tracks* that have arisen from any other mechanism. The distinguishing characteristic of signal tracks is that they come in pairs comprising one positive and one negative track. The background tracks do not share a physics vertex with any other track. N.B.: an individual track from a J/ψ, ψ', or Drell-Yan decay will be a background track if the other track in the pair was not reconstructed.

Within each event, we need to create all possible pairs of positive and negative tracks in search of the dimuon pairs from the signal sources. In the process of doing so, we will combine signal tracks with background tracks, and also combine background tracks with other background tracks, thus forming the *combinatoric* background in the spectrum. It is important to keep in mind that the combinatoric background contains a contribution from signal tracks as well as background tracks.

4 Mimicking the track-pairing process to estimate the combinatoric background

To estimate the combinatoric background correctly, we need to combine tracks from the same populations of positive and negative tracks as found in the top/bottom data stream, but without the possibility of making any signal pairs. We have seen that the wire chamber occupancy of an event strongly influences the number and momentum distributions of tracks, so we need to combine tracks from events with similar occupancy. We also want to make sure that each track is only combined with tracks from one event, since that is what we do when forming a spectrum using the original data.

These considerations lead to the following algorithm for mixing tracks from different events.

1. Choose a single normal data run, lasting about 1 hour.

2. Select top/bottom-triggered events.

2023 JINST 18 P10032

3. Order the events according to occupancy (low to high, for example).

4. Put all positive tracks from event i and all negative tracks from event $(i + 1)$ into a single new *mixed* event. This is implemented for all events, including those with no reconstructed tracks. There might be zero positive tracks in a given event, for example. These "mixed events" are placed into a file structure called a "mixed run".

5. Subsequent processing for mixed events occurs with identical conditions as for normal events. Especially, we make sure that the two tracks forming a dimuon from both normal and mixed events must satisfy the top/bottom trigger condition. This last requirement is important so that we preserve the bias of the top/bottom trigger in the mixed events.

In the discussion that follows, a "normal run" will contain "normal events" that came from the original data stream, while a "mixed run" will contain "mixed events" created according to the algorithm described above.

5 Normalization of the estimated combinatoric background

In this mixing method, we obtain the correct absolute normalization for the mixed distribution. This statement depends on four conditions: (1) The density of events with signal pairs in the data stream is very small, as required in ref. [3]; (2) Sorting the events according to D1 occupancy before mixing ensures the similarity of two events being mixed; (3) A given track is only combined with tracks from within one event. It is either combined with tracks within its own event (a normal run), or with tracks from one similar event (a mixed run); (4) The normal events and the mixed events are subject to exactly the same cuts in the subsequent analysis. In particular, all track pairs must satisfy the original top/bottom trigger condition.

We show that this method provides an estimate of the combinatoric background that is statistically consistent with the actual background. A "run" is a set of data collected with a specific trigger with the experimental conditions unchanged; in SeaQuest, a run lasted about one hour. A particular run may have N_E events. Each event i has zero or more reconstructed tracks, which are broken into four groups:

- s_i^+ is the number of positive tracks from a signal (J/ψ, ψ', or Drell-Yan) $= 0$ or 1.
- s_i^- is the number of negative tracks from a signal (J/ψ, ψ', or Drell-Yan) $= 0$ or 1.
- b_i^+ is the number of positive tracks from a background $= 0, 1, 2, \ldots$
- b_i^- is the number of negative tracks from a background $= 0, 1, 2, \ldots$

The signal tracks (positive and negative) come from a correlated source and only appear in pairs in the same event; we always have $s_i^+ = s_i^-$. The background tracks come from uncorrelated sources. If only one of a pair of signal tracks is reconstructed, it falls into the background category. Then the total number of unlike-sign track pairs, N_P, in a normal run is

$$N_P = \sum_{i=1}^{N_E} \left(s_i^+ s_i^- + s_i^+ b_i^- + b_i^+ s_i^- + b_i^+ b_i^- \right) .$$

2023 JINST 18 P10032

The first term in the sum is special, because s_i^+ and s_i^- come from a correlated source. The sum over this term is the total number of signal dimuon pairs in this run, N_S.

$$N_S = \sum_{i=1}^{N_E} s_i^+ s_i^-$$

The other three terms generate the combinatoric background, N_C.

$$N_C = \sum_{i=1}^{N_E} \left(s_i^+ b_i^- + b_i^+ s_i^- + b_i^+ b_i^- \right)$$

At this point it is appropriate to sort the events into similar groups; in the case of SeaQuest, this means sorting them according to the D1 chamber occupancy, ω, from low to high. Then the sum can be broken down into sub-sums where all events have the same occupancy. The number of events at a given occupancy ω is N_ω.

$$N_C = \sum_{\omega=0}^{\omega_{\max}} \sum_{i=1}^{N_\omega} \left(s_i^+ b_i^- + b_i^+ s_i^- + b_i^+ b_i^- \right)$$

The numbers s_i^+, b_i^-, and so on are all small integers (typically no larger than 7, see figure 3) drawn from a distribution depending on the occupancy. On the other hand, N_ω will tend to be large, certainly a few hundreds or thousands for the most popular occupancies in a run. The sum over events with the same occupancy will sample all possible values of $s_i^+ b_i^-$ (e.g.) many times. Therefore, we can replace the sum with averages:

$$N_C = \sum_{\omega=0}^{\omega_{\max}} N_\omega \left(\left\langle s^+ b^- \right\rangle_\omega + \left\langle b^+ s^- \right\rangle_\omega + \left\langle b^+ b^- \right\rangle_\omega \right),$$

where $\langle s^+ b^- \rangle_\omega$ is the average value of the product $s_i^+ b_i^-$ at the given occupancy ω, etc. Then the total number of pairs in the run is

$$N_P = N_S + N_C.$$

Now consider the total number of unlike-sign track pairs, N'_P, in a *mixed run*, where we have combined the positive tracks from event i with the negative tracks from event $i+1$ sourced from a normal run. We have sorted the events by occupancy, so that adjacent events contain tracks drawn from the same distributions.

$$N'_P = \sum_{i=1}^{N_E} \left(s_i^+ s_{i+1}^- + s_i^+ b_{i+1}^- + b_i^+ s_{i+1}^- + b_i^+ b_{i+1}^- \right)$$

The sum over the first term $s_i^+ s_{i+1}^-$ is non-zero but may be negligible; tracks from signals are rare and are only found in pairs in the same event. The probability that adjacent events in our mixed run will both have signal tracks is very small. We call this the "adjacent signals" term, N_{AS}, and we ignore this term for the moment; this is the most important quantitative requirement for the proposed mixing method to work properly.

$$N_{\mathrm{AS}} = \sum_{i=1}^{N_E} s_i^+ s_{i+1}^- \approx 0$$

2023 JINST 18 P10032

The remaining three terms may be treated in the same way as in the normal run.

$$N'_C = \sum_{i=1}^{N_E} \left(s_i^+ b_{i+1}^- + b_i^+ s_{i+1}^- + b_i^+ b_{i+1}^-\right) = \sum_{\omega=0}^{\omega_{\max}} N_\omega \left(\left\langle s^+ b^- \right\rangle_\omega + \left\langle b^+ s^- \right\rangle_\omega + \left\langle b^+ b^- \right\rangle_\omega\right)$$

The sums N_C (from the normal run) and N'_C (from the mixed run) are equal in the limit of large statistics. In the case of limited statistics, they will be equal within statistical uncertainties. Then to estimate the number of signal pairs, we need only subtract the mixed run from the normal run.

$$N_P - N'_P = (N_S + N_C) - N'_C \approx N_S$$

as $N_C \approx N'_C$ within uncertainties.

The question of whether the adjacent signals term N_{AS} can be ignored depends on two considerations: the probability f that a given event will contain a signal pair, and the overall statistical significance of the experiment. A numerical example is useful. Suppose $f = 0.01$ (1% of events have signal pairs) and in total there are 10^4 events. The number of signal pairs is approximately 100. The probability that two adjacent events have signal pairs is f^2, so the number of adjacent signal pairs is just 1. Compared to the statistical uncertainty in the number of signals (10), the adjacent signals term may be ignored. On the other hand, if there are 10^6 events, then the number of signal pairs is 10^4 and the number of adjacent signal pairs is 100, which is the same order as the uncertainty in the number of signal pairs. In this second case the adjacent signals term needs to be taken into account. We will show that the effect of the adjacent signals can be quantified via simulation and embedding, and so may be corrected for.

6 Simple models illustrating the proposed event-mixing method

To demonstrate the abilities and limitations of this mixing method, we create simple statistical models of the event stream. We will use the simple shape $\exp(-An)$ for the parent distribution for signals and backgrounds; n is the number of signal pairs or background tracks, and A is the shape constant of the curve; increasing the value of A means the signal or background becomes more rare.

In our first simple model, we have chosen the values of A so that the signal pairs are very rare compared to the background tracks.

$P_S \propto \exp(-7n_s)$ — probability of the number of signal pairs in the event

$P_P \propto \exp(-4n_p)$ — probability of the number of positive background tracks in the event

$P_N \propto \exp(-5n_n)$ — probability of the number of negative background tracks in the event

In each event, integer values of n_s, n_p, and n_n are chosen based on these parent probability distributions, determining the number of signal and background tracks in the event. For example, a selection of $(n_s, n_p, n_n) = (1, 1, 2)$ would mean $(s^+, s^-, b^+, b^-) = (1, 1, 1, 2)$. Then the following process is followed.

1. Create a list of 100,000 events using these probability distributions.

2023 JINST 18 P10032

2. Calculate the total number of track pairs, the true number of signal pairs, the true number of combinatoric pairs, the estimated number of combinatoric pairs using the proposed method, and the difference in the number of combinatoric pairs (true − estimated).

3. Repeat steps 1 and 2 10,000 times, and histogram the above quantities.

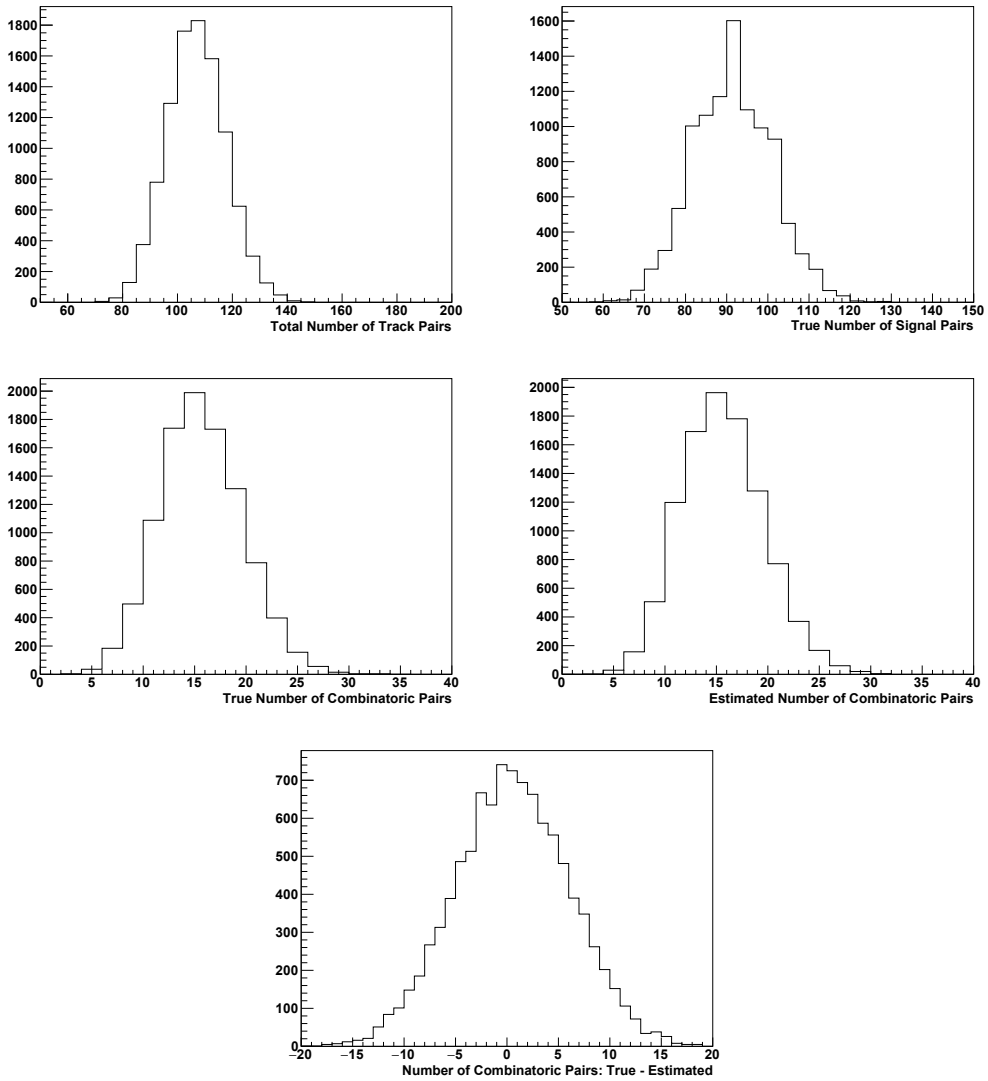

Figure 6. Results from first simple model. Top Left: histogram of the total number of track pairs. Top Right: true number of signal pairs. Middle Left: true number of combinatoric background pairs. Middle Right: estimated number of combinatoric background pairs. Bottom: difference between true and estimated number of combinatoric background pairs.

The results are illustrated in the histograms in figure 6. The average number of track pairs per 100,000 events is 105, of which on average there are 91 true signal pairs. The histograms for the true and estimated numbers of combinatoric background pairs are extremely similar, with the same average of 15. The difference between true and estimated combinatoric background pairs is centered

2023 JINST 18 P10032

about zero, with a root-mean-square deviation of 5.6 consistent with the difference of the averages; the statistical error on 15 − 15 would be $\sqrt{15+15} = 5.5$. We see that the proposed method works very well in this scenario.

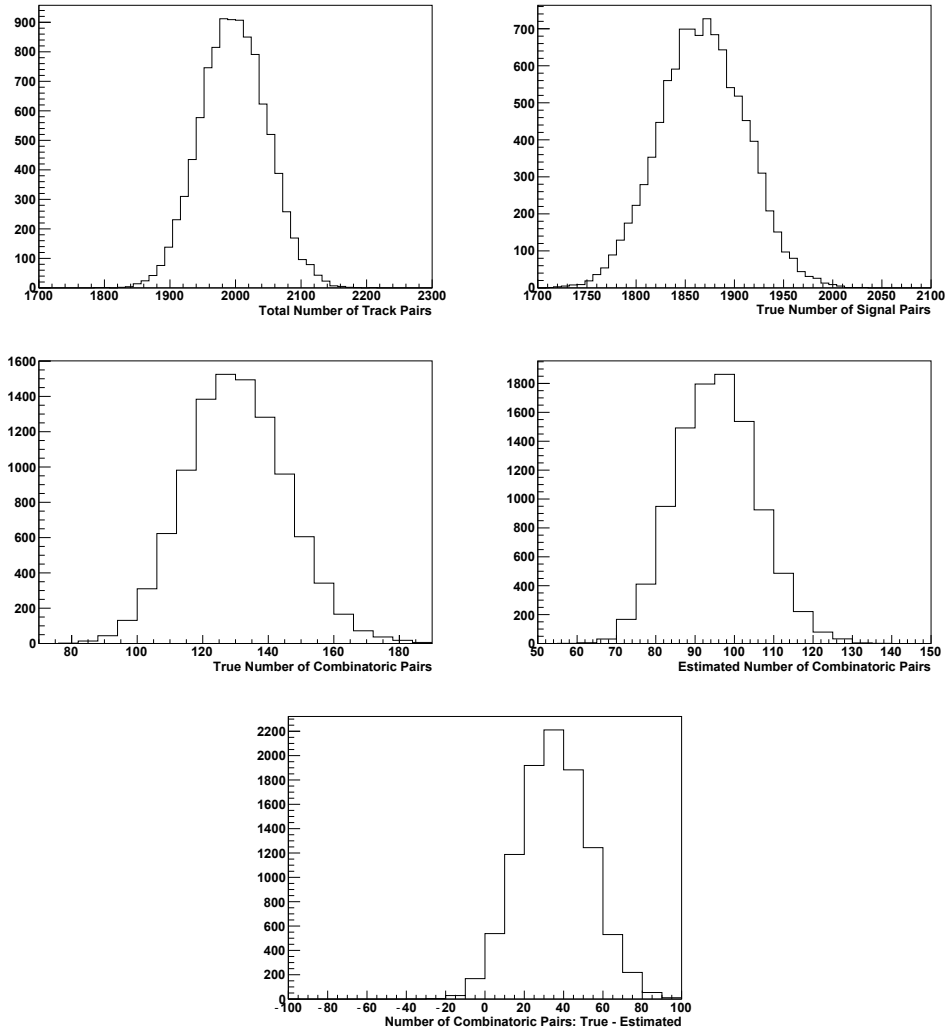

Figure 7. Results from second simple model. Top Left: histogram of the total number of track pairs. Top Right: true number of signal pairs. Middle Left: true number of combinatoric background pairs. Middle Right: estimated number of combinatoric background pairs. Bottom: difference between true and estimated number of combinatoric background pairs.

In our second simple model, we increase the rate of signal pairs so that it is comparable to the rate of background tracks; this violates one of the assumptions of the proposed mixing method and so we expect this to fail. The parent distributions are now:

$P_S \propto \exp(-4n_s)$ — compare to $\exp(-7n_s)$ in the first model

$P_P \propto \exp(-4n_p)$ — same as in first model

$P_N \propto \exp(-5n_n)$ — same as in first model

2023 JINST 18 P10032

The same procedure is followed, and the results are shown in figure 7. With a much greater density of signal pairs, the average total number of track pairs is increased to approximately 2,000, of which on average there are approximately 1,870 signal pairs. The number of true combinatoric pairs therefore averages around 130, but the proposed method *underestimates* this to be about 95. The reason for the underestimation is we now can have more than one signal pair per event, and the proposed method does not reproduce the extra combinatoric background created by the signal pairs among themselves. The method fails if the signal-to-background ratio is too high.

In our third simple model, we make the signal pairs rare again, but we make the background track distributions alternate between even-numbered and odd-numbered events.

$$\begin{aligned} P_S &\propto \exp(-7n_s) && \text{in all events} \\ P_P &\propto \exp(-4n_p) && \text{in even-numbered events} \\ &\propto \exp(-6n_p) && \text{in odd-numbered events} \\ P_N &\propto \exp(-5n_n) && \text{in even-numbered events} \\ &\propto \exp(-4n_n) && \text{in odd-numbered events} \end{aligned}$$

In this scenario, when we mix tracks from adjacent events in the list of events, we will be mixing tracks from events with different track distributions, which violates one of the assumptions of the proposed mixing method, and so we expect this to fail.

In figure 8, we see that the proposed method vastly overestimates the actual number of combinatoric tracks, because of the mis-match in track distributions that occurs during the mixing. The proposed method fails if you mix tracks from different distributions.

7 Tests of the event-mixing method using SeaQuest data

We performed various tests, using actual SeaQuest track data, to check the validity of the mixing method. There were two distinct types of tests. Both types employ simulated track data embedded into the actual data stream.

Type 1 Test

- We start with a *mixed run* as described above. This is a set of events containing tracks from a real run, sorted by occupancy, and then the positive tracks from one event mixed with the negative tracks from the next event. There is no physics signal in a mixed run; all possible unlike-sign track pairs within mixed events are uncorrelated pairs.
- Into these events, we embed reconstructed track pairs from a simulated signal. (The simulated signal is called "GMC" for "generated Monte Carlo.") A track pair is embedded into every n^{th} event; the number n is chosen so that the embedded signal is sparse (on order of a few percent) in the same way that real signal pairs are sparse (about 5%) in the real data.
- We analyze this set of events containing embedded tracks like it was a normal run. First, we loop over the events and form unlike-sign track pairs to make a spectrum; then we follow our procedure to mix tracks from different events, and produce a combinatoric background spectrum; we subtract the second spectrum from the first; the result should be the signal that we embedded.

2023 JINST 18 P10032

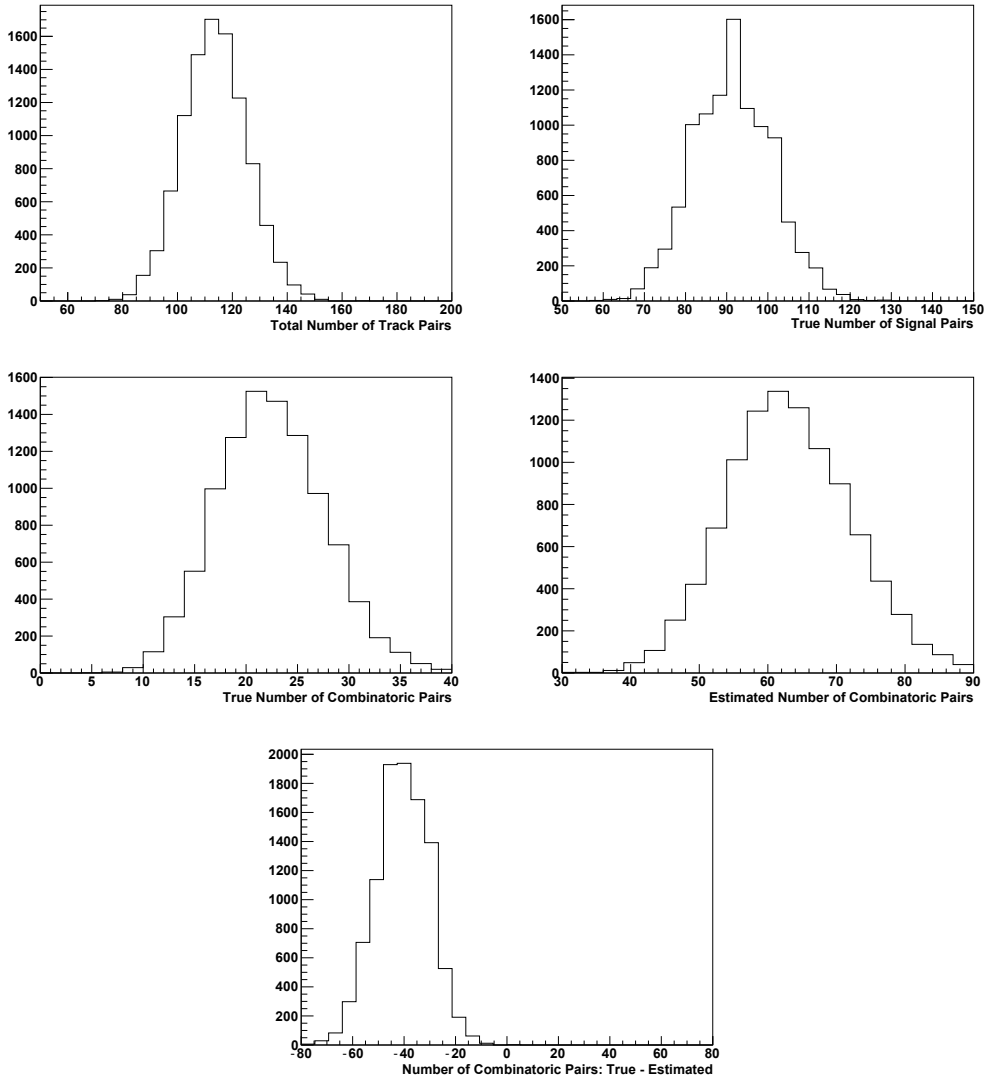

Figure 8. Results from third simple model. Top Left: histogram of the total number of track pairs. Top Right: true number of signal pairs. Middle Left: true number of combinatoric background pairs. Middle Right: estimated number of combinatoric background pairs. Bottom: difference between true and estimated number of combinatoric background pairs.

That type of test was done with two different simulated signals; in one case with a broad set of generated Drell-Yan events, and in a second case with a "resonance" at an invariant mass of 6 GeV/c^2.

Type 2 Test

- We start with a *normal* run, as described above. This is a set of events containing tracks from a real run, containing physics signal tracks as well as background tracks.
- Into these events, we embed reconstructed track pairs from a simulated (GMC) signal. A track pair is embedded into every n^{th} event, in the same manner as in the Type 1 Test.

2023 JINST 18 P10032

– Now we perform two analyses.

1. One is done with the original normal run (without embedded tracks), using the event mixing method. We calculate a spectrum from unlike-sign pairs (Yield1) and also the corresponding combinatoric background (Comb1) and then subtract the combinatoric background: Signal1 = Yield1 − Comb1.
2. The second analysis is done with the normal run containing the embedded GMC tracks. We calculate a spectrum from unlike-sign pairs (Yield2) and also the corresponding combinatoric background (Comb2) and then subtract the combinatoric background: Signal2 = Yield2 − Comb2.

– Now we subtract Signal1 from Signal2 and the result should be the embedded GMC signal.

This more elaborate test was performed with three different simulated signals: a resonance at 3.14 GeV/c^2, a resonance at 6 GeV/c^2, and a broad spectrum of Drell-Yan events.

An important feature available to us in these embedding schemes is that we can turn off the adjacent signals term N_{AS} by embedding the simulated tracks pairs at fixed intervals in the sequence of events; this means there are certainly no adjacent events with embedded signals. Alternatively, we can embed randomly in the event stream, and thus turn on the adjacent signals term. Looking at both kinds of embedding (fixed vs. random) enables us to quantify the effect of adjacent signals.

The results of one of these tests are given in the following subsection. The balance of the test results are given in the appendix.

7.1 Type 2 test with 6 GeV/c^2 resonance embedded signal

Here we explain in detail a Type 2 test with a simulated 6 GeV/c^2 resonance signal. In this test, the embedded signals are placed at fixed intervals in the event stream; this means the effect of the adjacent signals term N_{AS} is turned off, and we are demonstrating that the N'_C term generated in the mixed data correctly estimates the combinatoric background N_C generated from the normal data. We used 10 normal runs of SeaQuest data, using top/bottom-triggered events. The reconstructed tracks of the simulated signal were obtained by running generated dimuon pairs with an invariant mass of 6 GeV/c^2 through the full detector simulation and reconstruction. For this analysis, no further cuts and conditions are applied other than requiring the dimuons to satisfy the top/bottom trigger and to form a proper dimuon vertex. The following are the details for this test.

- We took 10 normal data runs.
- The mixing method was applied to each individual run, and then the results combined together. The left side of figure 9 shows the total dimuon mass distribution for real (black line, Yield1) and mixed events (green line, Comb1) from the data.
- The total signal (Signal1) is obtained by subtracting the mixed distribution (Comb1) from the real data distribution (Yield1); this is shown in the right side of figure 9. This spectrum contains a strong J/ψ peak near 3 GeV/c^2, a shoulder (barely visible) from ψ' production just above 3 GeV/c^2, and a continuum of Drell-Yan events. The spectrum falls to zero beyond 1 GeV/c^2 and 9 GeV/c^2 due to the acceptance of the spectrometer.

2023 JINST 18 P10032

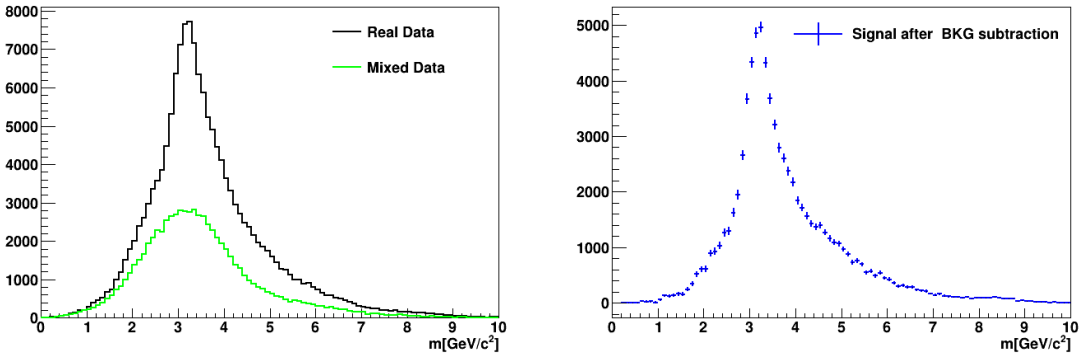

Figure 9. Left: invariant mass distribution from the 10 runs of normal data (black) and corresponding mixed (green) distribution. Right: signal obtained by subtracting mixed distribution (BKG) from data. Blue = Black – Green.

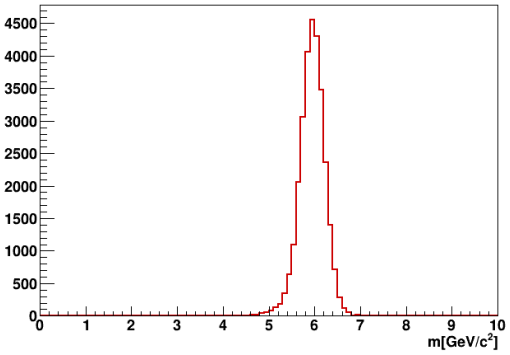

Figure 10. The simulated 6-GeV/c^2 resonance signal that will be embedded into the data displayed in figure 9.

- The reconstructed simulated dimuon signal with mass of 6 GeV/c^2 is shown in figure 10. The tracks for the simulated signal are embedded into every 50th event in the 10 runs.
- The mixing method is applied to individual runs containing the embedded tracks. The total dimuon distribution (Yield2) after embedding as mentioned above is shown in the black line in the left side of figure 11. In the same figure, the green line shows the mixed distribution (Comb2) after applying the mixing procedure.
- The total signal (Signal2) obtained from subtracting the mixed distribution from the embedded data distribution is shown in the right side of figure 11.

The left hand side of figure 12 shows the various signals as we discussed above. The magenta points are the signal we obtained from the embedded data, the blue points show the signal from the real data, and the red histogram shows the total simulated signal we embedded. So, if the normalization from the mixing method is unity then the difference between the signal from embedded data (Signal2) and that from real data (Signal1) should be same as the embedded data. That difference is shown as the blue points in the right hand side of the figure 12 together with the red signal histogram. The figure shows that the two distributions are in good agreement.

2023 JINST 18 P10032

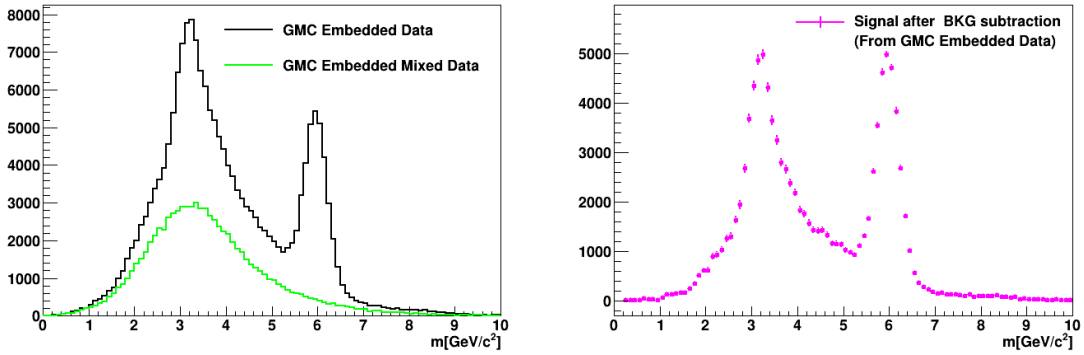

Figure 11. Left: dimuon distribution (black) after the simulated signal from figure 10 is embedded into the data of figure 9, and the corresponding mixed distribution (green). Right: signal obtained by subtracting mixed distribution from data. Magenta = Black − Green.

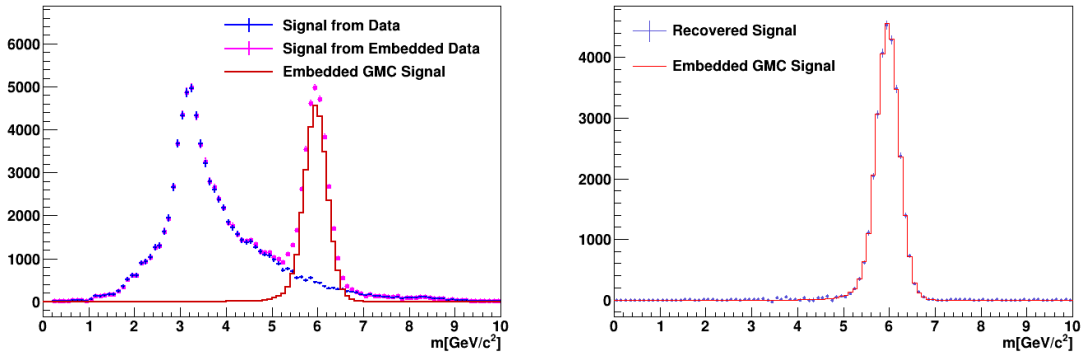

Figure 12. Left: comparison of signals from different stages of analysis. The blue is total signal (Signal1) from the 10 normal runs, magenta is signal from simulated embedded data (Signal2); both are obtained after subtracting the combinatoric background. The red histogram is the embedded simulated signal, same as figure 10. Right: the blue histogram is the difference (Signal2 − Signal1) between the blue and magenta histograms in the left-hand panel. The red is the embedded signal, same as figure 10.

Figure 13 shows the difference between the embedded simulated signal and signal recovered from the mixing method. The distribution has statistical fluctuations centered around zero; there is no residual signal.

As an additional test of our claim of correct normalization, we attach a common normalization factor (NM) to the combinatoric backgrounds and see if the residual spectrum (like figure 13) is affected. This means we will calculate

$$\text{Signal1} = \text{Yield1} - \text{NM} \cdot \text{Comb1} \qquad \text{and} \qquad \text{Signal2} = \text{Yield2} - \text{NM} \cdot \text{Comb2}.$$

The result is shown in figure 14. We see that any choice of NM other than 1 leaves a residual signal. This is numerically demonstrated in table 2, where we have integrated the area of each histogram in figures 13 and 14.

A number of other tests, of Type 1 and Type 2, are shown in the appendix. In all of these tests, the embedded events have been placed at fixed intervals, so that the adjacent signals term N_{AS} has

2023 JINST 18 P10032

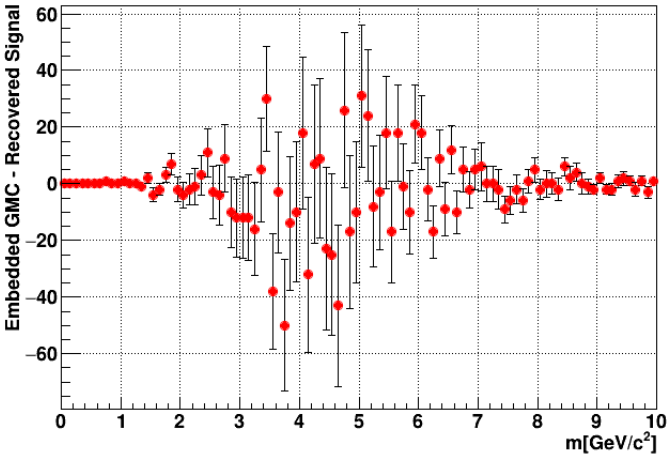

Figure 13. The difference between signal recovered and signal embedded; this is the difference between the two histograms in figure 12 right, above.

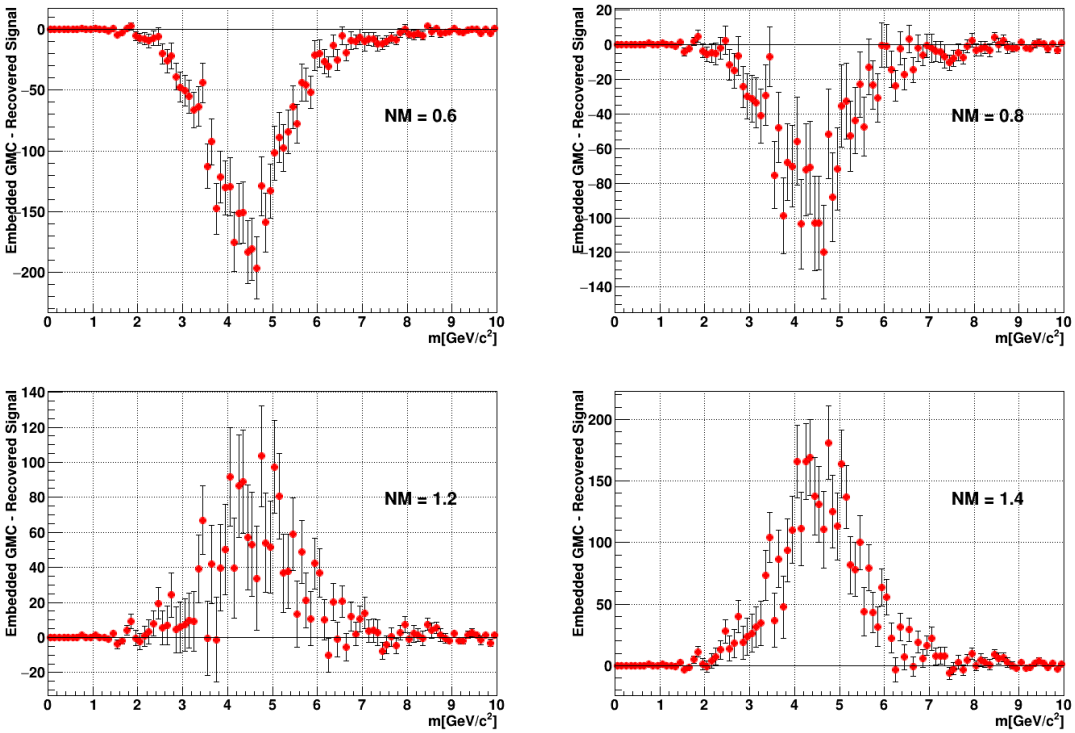

Figure 14. The difference between signal recovered and signal embedded when both mixed distributions as described in section 7 are scaled by a common normalization factor. Compare to figure 13 where NM = 1.

been turned off. We show in all these cases that the N'_C term generated in the mixed data correctly estimates the combinatoric background N_C. In the next section we turn to the issue of the effect of adjacent signals in the event stream.

2023 JINST 18 P10032

Table 2. Area of each histogram in figures 13 and 14 as a function of the normalization factor NM.

Normalization	0.6	0.8	1.0	1.2	1.4
Area	-3610 ± 118	-1876 ± 125	-143 ± 132	1590 ± 138	3324 ± 145

8 Quantification of the effect of the adjacent signals term

Previously, in section 5, we mentioned that the adjacent signals term might not be negligible under some circumstances. This does not invalidate the use of the proposed mixing method, because the size and effect of this term can be quantified using the simulation and embedding methods described in section 7. In this section, we begin to embed the simulated signals randomly into the event stream, thus introducing the possibility of signals being in adjacent events.

We use a Type 1 test as described above in section 7. We show the results for each of the 10 runs, and we show results both for embedding the simulated signals at fixed intervals and at random locations in the event stream. The simulated signals were embedded into 5% of the events. In table 3 we show the residual signal after subtracting the recovered signal from the generated signal. The residual signal is the integral of the difference. In each run, when embedding at fixed intervals, the residual signal statistically fluctuates around zero. On the other hand, when embedding randomly, the effect of the adjacent signals term is seen as a net positive residual signal.

In each of the ten runs, there are about 150,000 events. In this test, we embedded simulated signal pairs into 5% of those events, that is about 7500 embedded signal pairs. When we embed randomly in the event stream, then about 0.25% of events will have embedded signals in adjacent events, that is about 375 such events; this would be the value of the N_{AS} term in the mixed run. However, not all of those 375 mixed pairs will pass the top/bottom-trigger condition, nor will all of them form a proper dimuon vertex, because they are a pair of unrelated tracks. So, we should see fewer than N_{AS} extra counts in our residual signal. Table 3 confirms this; the residual signal is 192.6 ± 41.5 and not 375. The N_{AS} term is an upper limit on the size of the final residual signal produced by adjacent signal pairs; additional analysis requirements ("cuts") will reduce the effect of these pairs.

9 Conclusion

We have developed a method to estimate the combinatoric background valid for dilepton experiments where (1) the population density of signal pairs in the data stream is sufficiently low, and (2) the events can be sorted into classes containing the same track distributions. The method has the correct normalization and the computed distribution can be directly subtracted from the total yields to recover the signal yields. In the case of experiments with sufficiently high statistical significance, the effect of signal pairs that occur in adjacent events can perturb the results, and we have demonstrated a technique for quantifying this effect and correcting for it.

In an experiment with low statistics for both the signal and the background, it can be desirable to improve the statistical significance of the estimate of the background. In principle, one could double the statistics of the estimated combinatoric background by combining the positive tracks from event i with the negative tracks from both event $i+1$ and $i+2$ and retain the correct normalization by dividing by 2. (And one could imagine extending this to events $i+3$, $i+4$, etc.) We have not

2023 JINST 18 P10032

Table 3. List of residual signals in each run, for embedding done at fixed intervals in the event stream and for embedding done randomly, using a Type 1 test.

Run Number	Fixed Embedding	Random Embedding
12525	143 ± 137.7	90 ± 138.1
12527	−43 ± 146.4	206 ± 147.3
12528	1 ± 143.3	225 ± 144.4
12529	−14 ± 109.6	148 ± 109.9
12530	84 ± 139.4	188 ± 140.4
12531	−22 ± 147.5	42 ± 148.1
12532	37 ± 91.3	160 ± 92.3
12533	−174 ± 146.8	119 ± 147.6
12534	150 ± 149.7	487 ± 150.1
12535	182 ± 145.0	364 ± 145.2
Weighted Average	33.1 ± 41.3	192.6 ± 41.5

explored this. The effect of the adjacent signals would need to be investigated for this case; doubling the size of the event pool will double the number of "adjacent signals".

Other groups attempting to use this method should perform the same sorts of tests, as we have shown here, to make sure this method is applicable to their situation.

Acknowledgments

This work was partially supported under grants DE-FG02-94ER40847 (New Mexico State), DE-AC02-06CH11357 (Argonne), DE-FG02-07ER41528 (Mississippi State), DE-FG02-96ER40950 (Virginia) from the US Department of Energy, Office of Nuclear Physics, as well as by the National Science Foundation under grants 2110229 & 2012926 (Michigan) and 2013002 (Colorado).

A Consistency checks performed by embedding various simulated distributions into data

Here we show additional tests we have done using the embedding of simulated tracks into real track data. The embedding here is not done randomly, but instead at fixed intervals in the event stream, thus turning off the effect of the adjacent signals term N_{AS}.

A.1 Type 2 test with 3.14 GeV/c^2 dimuon embedded signal

Instead of a 6-GeV/c^2 resonance, we used a 3.14-GeV/c^2 resonance, because this places the embedded signal in the middle of the region with the largest yield. Here we embed the simulated signals in every 200$^{\mathrm{th}}$ event. We also included the variable normalization factor NM described above, and a table of the residual signals showing that the best choice is NM=1. Please read the captions for figures 15-19 and see table 4.

2023 JINST 18 P10032

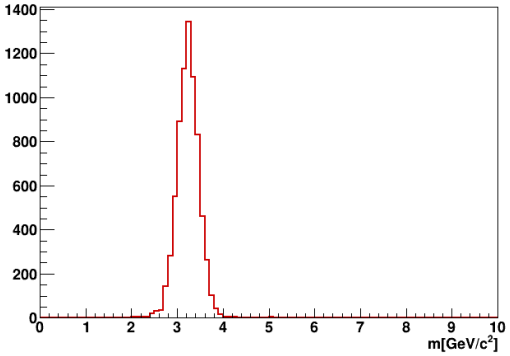

Figure 15. The simulated 3.14-GeV/c^2 resonance signal that will be embedded into the data displayed in figure 9.

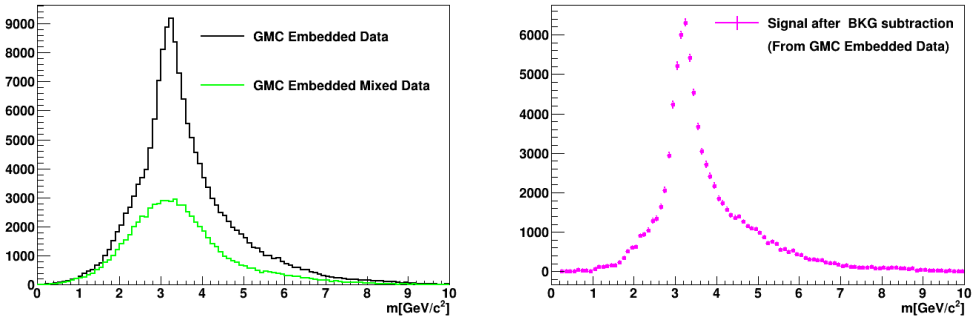

Figure 16. Left: dimuon distribution (black) after simulated signal from figure 15 is embedded into the data from figure 9, and the corresponding mixed distribution (green). Right: signal obtained by subtracting mixed distribution from data. Magenta = Black − Green.

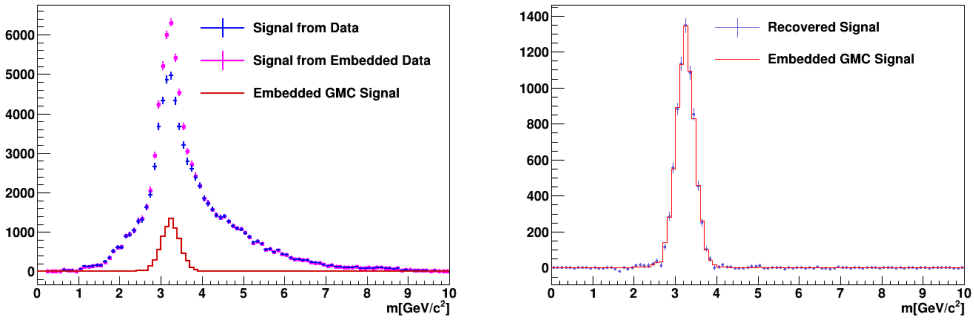

Figure 17. Left: comparison of signals from different stages of analysis. The blue points are the total signal from a single normal run; same as figure 9 Right. The magenta points are the signal from data with the embedded simulated signal; same as figure 16 Right. Red is the embedded simulated signal, the same as figure 15. Right: signal recovered (blue) vs. signal embedded (red).

2023 JINST 18 P10032

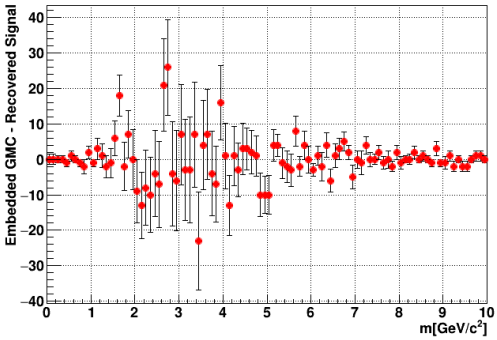

Figure 18. The difference between signal recovered and signal embedded; this is the difference between the two histograms in the right hand side of figure 17 above. It is seen there is no residual signal.

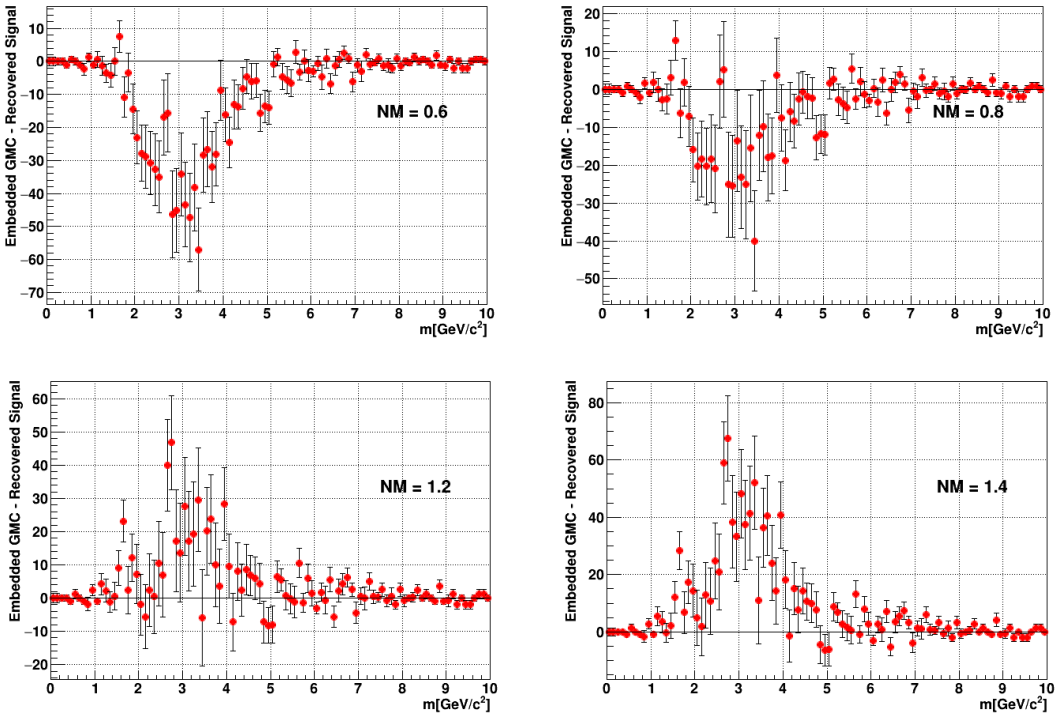

Figure 19. The difference between signal recovered and signal embedded when both mixed distributions as described in section 7 are scaled by different normalization factors. Compare to figure 18 where NM=1.

Table 4. Area of each histogram in figures 18 and 19 as a function of the normalization factor NM.

Normalization	0.6	0.8	1.0	1.2	1.4
Area	-858 ± 58	-430 ± 62	-1 ± 65	428 ± 68	856 ± 72

A.2 Type 2 test embedding simulated Drell-Yan events into real data

For this consistency check, we took one normal data run. We embedded reconstructed dimuon tracks from simulated Drell-Yan events in every 25^{th} event. The Drell-Yan events were generated

2023 JINST 18 P10032

uniformly in the mass range 0-10 GeV/c^2, and then passed through the detector simulation and reconstruction package. After embedding, we implemented the mixing procedure to get the mixed distribution. Finally, the mixed distribution is subtracted from the embedded data distribution. The signal distribution thus recovered is consistent with the embedded signal. Please see figure 20.

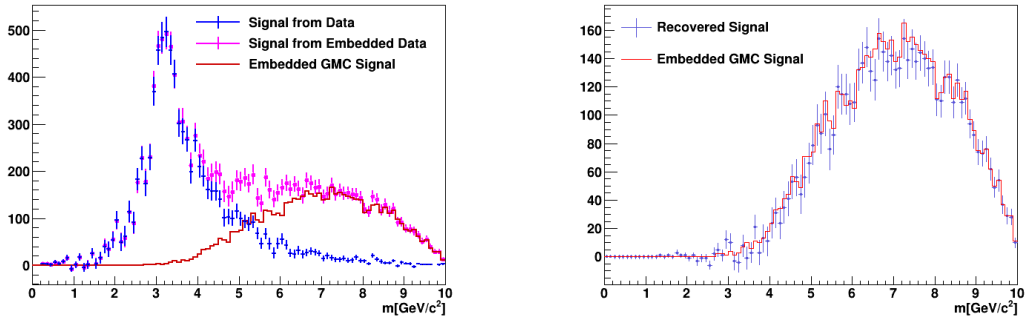

Figure 20. Left: comparison of signals from different stages of analysis. The blue points are the total signal from a single normal data run, while the magenta points are the signal from data with a simulated embedded signal. Both are obtained after subtracting the combinatoric spectrum calculated from the event-mixing method. The red histogram is the embedded simulated signal. Right: signal recovered (blue) vs. signal embedded (red).

A.3 Type 1 test with 6-GeV/c^2 dimuon embedded signal

In this case, a simulated 6-GeV/c^2 resonance (figure 21 right) is embedded into a mixed run (figure 21 left), using every 25$^{\text{th}}$ event. The resulting total spectrum and combinatoric background (figure 22 left) and the difference between them (figure 22 right) are shown; the embedded signal and the extracted signal are compared directly in figure 23.

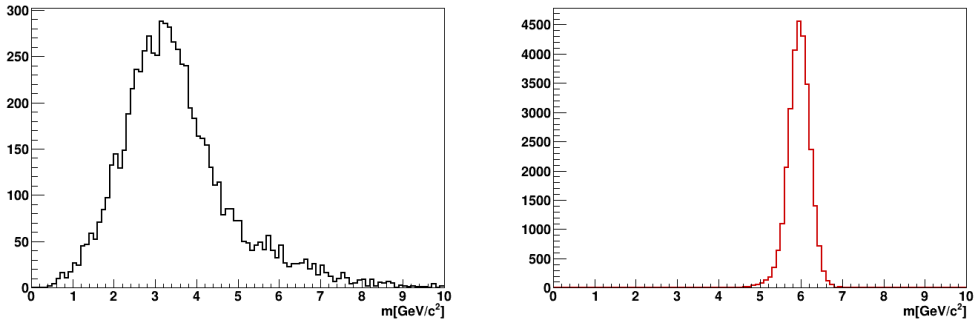

Figure 21. Left: dimuon mass distribution from uncorrelated data (obtained from applying mixing method in real data). Right: simulated signals to be embedded in uncorrelated data.

A.4 Type 1 test with simulated Drell-Yan dimuon distribution

In this case, a spectrum of simulated Drell-Yan events (figure 24 right) is embedded into a mixed run (figure 24 left), in every 25$^{\text{th}}$ event. This is the same spectrum of Drell-Yan events used in

2023 JINST 18 P10032

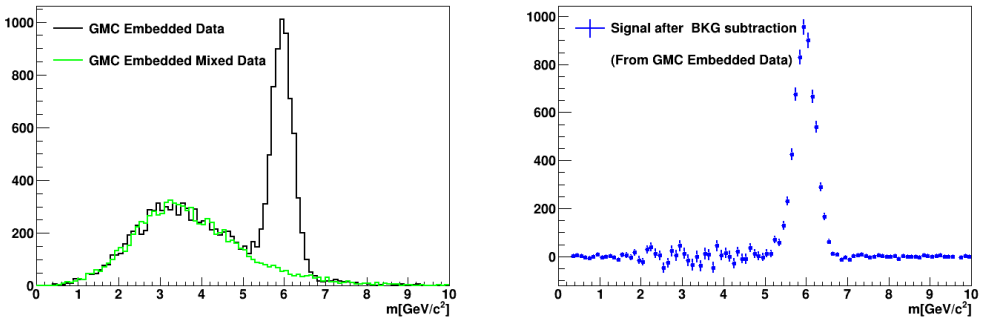

Figure 22. Left: simulated dimuon mass distribution embedded in uncorrelated data (black), and distribution from corresponding mixed events (green). Right: signal obtained from subtracting. Blue = Black − Green.

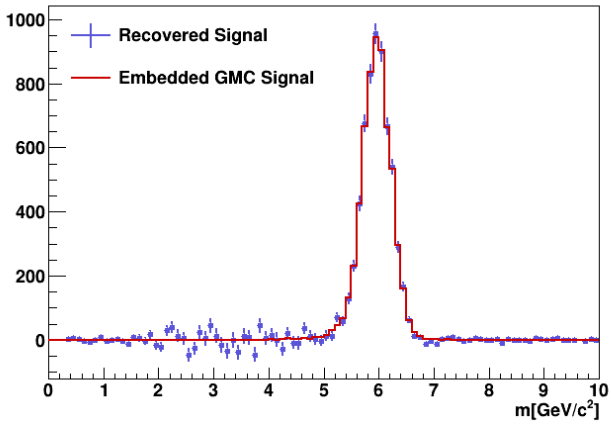

Figure 23. Signal recovered from mixing method (blue points) and embedded simulated events (red line) plotted together.

subsection A.2 above. The resulting total spectrum and combinatoric background (figure 25 left) and the difference between them (figure 25 right) are shown; the embedded signal and the extracted signal are compared directly in figure 26.

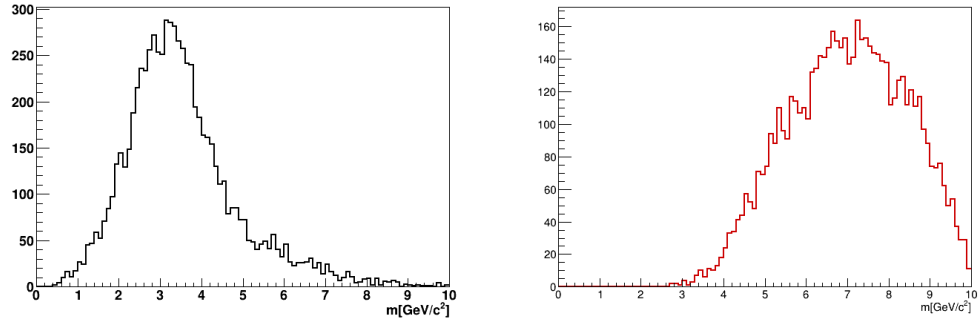

Figure 24. Left: dimuon mass distribution from uncorrelated data (obtained from applying mixing method in real data), same as figure 21. Right: simulated signals to be embedded in uncorrelated data.

2023 JINST 18 P10032

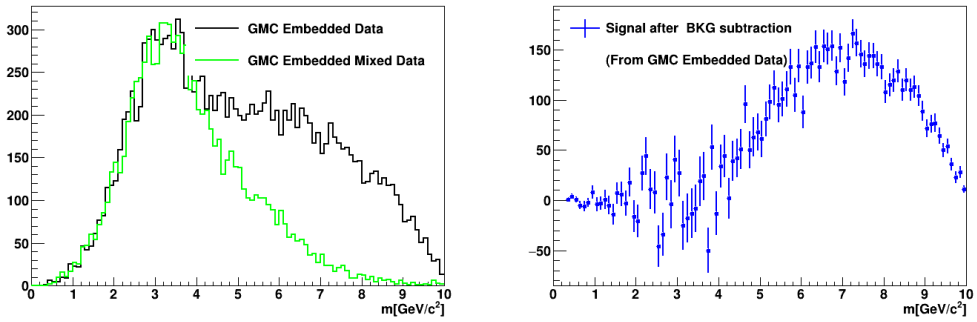

Figure 25. Left: simulated dimuon mass distribution embedded into uncorrelated data (black) and distribution from corresponding mixed events (green). Right: signal obtained from subtracting. Blue = Black − Green.

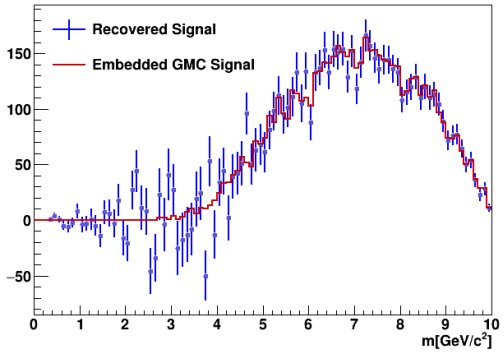

Figure 26. Signal recovered from mixing method (blue points, figure 25 right), and the embedded simulated events (red line, figure 24 right), plotted together.

References

[1] SeaQuest collaboration, *The SeaQuest Spectrometer at Fermilab*, *Nucl. Instrum. Meth. A* **930** (2019) 49 [arXiv:1706.09990].

[2] SeaQuest collaboration, *The asymmetry of antimatter in the proton*, *Nature* **590** (2021) 561 [*Erratum ibid.* **604** (2022) E26] [arXiv:2103.04024].

[3] P. Crochet and P. Braun-Munzinger, *Investigation of background subtraction techniques for high mass dilepton physics*, *Nucl. Instrum. Meth. A* **484** (2002) 564 [nucl-ex/0106008].

2023 JINST 18 P10032