Inverse Problems in Experimental Particle Physics

Sean Gilligan

Abstract

This report provides a survey of some of the common methods used by the high energy physics community to understand and solve ill-posed inverse problems as they pertain to signal distortions that result from imperfect measuring devices and processes. These methods are in general collectively referred to as unfolding. The specifics of data and data collection methods are generalized. Common features are discussed insofar as they contribute to the necessary understanding of the data and implementation of any covered unfolding methods. In order to construct a slightly more wholistic picture some additional topics are briefly touched upon if they relate to other common aspects of data analysis in particle physics, but only during parts of relevant discussions where they would otherwise normally appear.

1 Introduction

A common problem faced in the quantitative sciences and their associated technologies is the introduction of errors during the data collection process. While the possible sources of these errors are as varied as the possible events which the data might describe, significant work has been done to develop methods the can help would-be analysts reconcile them. The requisite understanding of a scenario's underlying systematic and stochastic processes might not allow researchers to truly reverse entropy or make up for the finite resolution of a detector, but it can approximate them with a quantifiable degree of certainty. The applied mathematics that this involves falls within the general category of **inverse problems**, and there are a variety of labels used to refer to the procedures in its arsenal. Within the applications described here there is the colloquially vague **unsmearing**, but there are also names that reference specific applications and methods, such as those characterized in this report.

For the sake of simplicity, the manner of inverse problems addressed here will only involve linear operations that map from one Hilbert space[1] to another. Symbolically this can be

[1]The definition of a Hilbert space is provided in Appendix A.1 for convenience.

expressed by the equation

$$Az = u,$$

where A is a linear operator acting on an element $z \in Z$, the sought solution, to produce an element $u \in U$, the observed data. Within the context of the methods described herein z and u take the form of continuous or discrete distributions that when integrated or summed over the domain of their arguments result in finite real quantities.

The difficulty of solving for z can be nominally classified into one of two camps. The easiest cases involve conditions that create a **well-posed** problem, which requires that [14]

1. a solution exists $\forall u \in U$,
2. the solution is unique,
3. and if $u_n \longrightarrow u$, $Az_n \longrightarrow u_n$, and $Az \longrightarrow u$, then $z_n \longrightarrow z$.

Conditions 1 and 2 work together to imply that the inverse operator A^{-1} exists, and Condition 3 is often worded to describe the inverse as continuous, which means that small deviations in u should correspond to similarly small deviations in z. When one or more of these conditions are not meant, the problem is said to be **ill-posed**, and some of the consequences of assuming otherwise should hopefully become clear in the coming pages. Entire books have been written on this subject that do not begin to cover the full scope of the methods developed to deal with ill-posed problems. The hope of this paper is for it to serve as an introduction to this content and provide some degree of direction for those who would like to know more. In the next section convolutions and deconvolutions are discussed in some detail for the continuous case before generalizing and shifting over to discrete approaches that allow for the better use of computational methods.

1.1 The Deconvolution

One way to characterize a basic example of a situation suitable for being treated as a convolution would be one that should be very familiar to anyone who has ever taken statistics course. Assume that data collected regarding n statistical events represent the measurement of n independent and identically distributed (i.i.d.) random variables $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ from a distribution of possible values represented by the probability density function (PDF) $f_X(x)$, such that the probability of a random variable X_i having a value between x_a and x_b is

$$P(\, x_a < X_i < x_b \,) = \int_{x_a}^{x_b} f_X(x)\, dx$$

and

$$\int_{\mathcal{X}} f_X(x)\,dx = 1,$$

where $\mathcal{X}$ represents the domain of x. The error introduced during the measurement process is similarly represented by a set of i.i.d. random variables $\boldsymbol{\varepsilon} = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n\}$ with a PDF $f_\varepsilon(\varepsilon)$, where the sets $\boldsymbol{\varepsilon}$ and $\boldsymbol{X}$ are typically assumed to be independent of each other. The set of measured/reconstructed values $\boldsymbol{Y} = \{Y_1, Y_2, \ldots, Y_n\}$ then are also i.i.d. and can be defined in terms of the preceding sets of variables such that for event $i \in \{1, \ldots, n\}$,

$$\begin{aligned} Y_i &= g(X_i, \varepsilon_i) \\ &= X_i + \varepsilon_i. \end{aligned} \tag{1}$$

In light of this relationship, the corresponding PDF $f_Y(y)$ can be found explicitly through an operation on $f_X(x)$ and $f_\varepsilon(\varepsilon)$ using the mathematics of functional analysis. Stated in more general terms, the empirical density function f_Y is formed from the **convolution** of the true density function f_X and the error density function f_ε, and is defined by [11]

$$f_Y \equiv f_X * f_\varepsilon \tag{2}$$

$$\begin{aligned} f_Y(y) &\equiv \int_{\mathcal{X}} f_X(x) f_\varepsilon(\varepsilon)\,dx \\ &= \int_{\mathcal{X}} f_X(x) f_\varepsilon\left(g_x^{-1}(y)\right) \left|J_{g_x^{-1}}(y)\right| dx \\ &= \int_{\mathcal{X}} f_X(x) f_\varepsilon(y - x)\,dx, \end{aligned} \tag{3}$$

where J represents the Jacobian of the transformation involved in performing the change of basis on f_ε from ε to x, which is necessary for the evaluation of the integral for a given y. The magnitude of the Jacobian for transformation of ε to $y - x$ through the manipulation of Equation (1) happens to be 1.

As the collection of measured values $\boldsymbol{Y}$ accumulates an estimate of empirical density $\hat{f}_Y$ can readily be formed. However, a major goal in an analysis of data like this is typically to develop an accurate estimate of the true density $\hat{f}_X$. Using the information contained in $\hat{f}_Y$ to accomplish this necessarily requires some attempt at finding an inverse process to the convolution, i.e. the **deconvolution**.

For cases in the form of this particular example there are a variety approaches, but they commonly involve the Fourier transform of the density functions $\{f_X, f_\varepsilon, f_Y\}$ into their corresponding characteristic functions $\{\phi_X, \phi_\varepsilon, \phi_Y\}$ [10][11]. Minor aspects of the definition for the Fourier transform can vary slightly between applications, resulting primarily from the

use of different scale factors and sign conventions. Here it will be defined for some random variable $U \in \mathbb{R}$ with density function $f_U(u)$ and random variable $T \in \mathbb{R}$ as

$$\phi_T(t) = \int_{-\infty}^{\infty} f_U(u)\, e^{itu}\, du. \tag{4}$$

When conditions permit the inverse Fourier transform can be found via

$$f_U(u) = \int_{-\infty}^{\infty} \phi_T(t)\, e^{-itu}\, dt. \tag{5}$$

The Fourier transform is important in deconvolution methods because when you apply it to the convolution of two density functions the link between their respective characteristic functions becomes purely multiplicative, i.e.

$$f_Y = f_X * f_\varepsilon \implies \phi_Y = \phi_X \phi_\varepsilon.$$

An instructional proof of this result is provided on page 447 of [4]. The steps so far characterize a typical deconvolution scheme, with later steps consisting of various ways to perform density estimation and addressing issues similar to those that will be seen ahead[10].

1.2 Generalizing

The remainder of this paper is dedicated to a more generalized study of these type of problems. With the understanding that even experts can be fairly loose and inconsistent with their vocabulary, this paper will do its best to provide clear demarcations to distinguish between similar processes and utilize consistent notation throughout. To begin, while most literature on deconvolution methods do use the word "convolution", this operation is also referred to by the German word *faltung* [13]. The latter's English translation, **folding**, is featured prominently in the particle physics community, but refers to a more generalized process than what is described by Equation (3) [8][1][2]. In general, the terms folding and **unfolding** are used to describe two supersets of processes that respectively include convolution and deconvolution.

One way to arrive at the intended generalization is with the help of conditional probability. Thinking of $\{X, Y\}$ as a continuous bivariate random vector with joint PDF $f(x, y)$ and marginal PDFs $f_X(x)$ and $f_Y(y)$, we can define the conditional PDF of Y given that $X = x$ as function of y, $f(y\,|x)$ [6]. The relationship between these PDFs is sufficient to define any one of them in terms of operations involving one or more of the others. As such, for $f_Y(y)$ it

can be shown

$$\begin{aligned} f_Y(y) &= \int_{\mathcal{X}} f(x,y)\,dx \\ &= \int_{\mathcal{X}} f(y\,|x) f_X(x)\,dx \\ &= \int_{\mathcal{X}} K(x,y) f_X(x)\,dx. \end{aligned} \tag{6}$$

While integrating over x, $f(y\,|x)$ is implicitly treated as a function of both x and y. Acknowledging this allows for understanding Equation (6) as a Fredholm integral of the first kind with a Kernel function $K(x,y)$ that reflects the physical measurement process [3]. The relationship between x and y in $K(x,y)$ is not defined, but when the kernel is a function of the difference of its arguments, such that $K(x,y) = K(y-x)$, Equation (6) becomes the convolution described in Equation (3).

In particle physics experiments, analysts make use of Monte-Carlo (MC) simulations to estimate detector response to randoms samples from some true distribution $f_X(x)^{\text{MC}}$, which is itself estimated by way of MC simulations using models that typically contain theory being tested by the experiment in question. The resulting measured distribution $f_Y(y)^{\text{MC}}$ grants implicit knowledge of $K(x,y)$ by way of Equation (6) [2]. Finding the inverse of this Kernel is then the goal, as it should in theory allow for the mapping of experimental observations $\boldsymbol{Y}$, as randomly sampled from $f_Y(y)$, back to their true values $\boldsymbol{X}$.

1.3 Discretization

In practice researchers are only ever dealing with estimates $\hat{f}_X$, $\hat{f}_Y$, $\hat{f}_X^{\text{MC}}$, and $\hat{f}_Y^{\text{MC}}$, and the sets of data that contribute to these estimates are organized by bin into histograms that form unnormalized granular approximations of their true distributions. Thinking in terms of these histograms allows for the reformulation of Equation (6) into the linear matrix equation:

$$\boldsymbol{\nu} = \boldsymbol{R}\boldsymbol{\mu}. \tag{7}$$

The vectors $\boldsymbol{\nu}$, $\boldsymbol{\mu}$ and matrix $\boldsymbol{R}$ relate to their continuous counterparts by [2]:

$$\begin{aligned} \text{true distribution } f_X(x) &\longrightarrow \boldsymbol{\mu} \in \{\mathcal{U} \equiv \mathbb{R}_+^M \cup \mathbf{0}\} \text{ the unknown true bin counts,} \\ \text{measured distribution } f_Y(y) &\longrightarrow \boldsymbol{\nu} \in \mathcal{V} \equiv \{\mathbb{R}_+^N \cup \mathbf{0}\} \text{ the measured bin counts,} \\ \text{Kernel } K(x,y) &\longrightarrow \boldsymbol{R} \text{ the rectangular } N\text{-by-}M \textbf{ response matrix}. \end{aligned}$$

The components of vectors $\boldsymbol{\nu}$ and $\boldsymbol{\mu}$ represent the number of events that have occurred within the regions of x and y that define the components' corresponding bins. For $i = 1, \ldots, N$ and $j = 1, \ldots, M$ the components of matrix $\boldsymbol{R}$ are defined by the conditional probability [7]

$$\begin{aligned} R_{ij} &= P(\text{measured value in bin } i | \text{true value in bin } j) \\ &= \frac{P(\text{measured value in bin } i \text{ and true value in bin } j)}{P(\text{true value in bin } j)} \\ &= \frac{\int_{\text{bin } i} \int_{\text{bin } j} K(x,y) f_X(x) dx\, dy}{\int_{\text{bin } j} dx\, f_X(x)} \\ &\equiv P(\nu_i | \mu_j). \end{aligned} \tag{8}$$

In terms of $P(\nu_i|\mu_j)$ the full response matrix then has the form

$$\boldsymbol{R} = \begin{pmatrix} P(\nu_1|\mu_1) & P(\nu_1|\mu_2) & \ldots & P(\nu_1|\mu_N) \\ P(\nu_2|\mu_1) & P(\nu_2|\mu_2) & \cdots & P(\nu_2|\mu_N) \\ \vdots & \vdots & \ddots & \vdots \\ P(\nu_M|\mu_1) & P(\nu_M|\mu_2) & \ldots & P(\nu_M|\mu_N) \end{pmatrix}. \tag{9}$$

With these definitions Equation (7) tells us that an event produced in bin μ_j has some probability ≥ 0 of being measured in each of the N bins of $\boldsymbol{\nu}$, and that each bin count ν_i receives potential contributions from each of the M bins in $\boldsymbol{\mu}$, i.e.

$$\nu_i = \sum_{j=1}^{M} R_{ij} \mu_j. \tag{10}$$

The number of bins are typically set such that $M \leq N$, with the convention $N = M + 1$ being common. A higher number of bins in the measured distribution reflects that the measuring process is expected to map some events in $\boldsymbol{X}$ to values of $\boldsymbol{Y}$ that are outside the region of values that define the initial M bins. These one or more extra bins are intended to account for all the possible values that a particular event could be mapped to, such that for a given event starting in bin j one might expect the probabilities of it being measured in each of the N final bins to sum to 1.

However, in practice there are a variety of constraints on events that can either result in them not being included for analysis or even prevent them from being detected at all. For example, an analyst might cut events observed in regions of a detector that result in insufficient data collection, or maybe some event information carriers miss the detector entirely, resulting in such events going unseen. In either case the effect of these missing events is described

using the detector **efficiency**, and represented mathematically by the N-vector $\boldsymbol{\epsilon}$, where component ϵ_j is the efficiency of the jth true bin defined[2] by [7]:

$$\sum_{i=1}^{N} P(\nu_i|\mu_j) = \sum_{i=1}^{N} R_{ij} = \epsilon_j \leq 1. \tag{11}$$

In contrast to this are contributions to measured counts from **background** processes. Just as events produced in a region of interest can be smeared out of it, events produced out of it can be smeared into it. The crossed barrier could correspond to the variable of interest, but it can also include events excluded from analysis due to assigned constraints on other variables that describe the event. Background processes are often studied and dealt with prior to the unfolding procedures described in the paper. It is briefly mentioned here to provide a slightly more holistic picture of particle physics analysese. Mathematically, background would be included by modifying Equation (10) to read

$$\nu_i = \sum_{j=1}^{M} R_{ij}\mu_j + \beta_i, \tag{12}$$

where β_i is the ith component of the N-vector $\boldsymbol{\beta}$, which represents the binned background counts. This leads to equations like $\nu_i^{\text{sig}} = \nu_i - \beta_i$ in order to specify the expected number of measured counts that are from the signal of interest. Going forward background will be assumed to already have been accounted for, and ν_i will refer to the expected signal counts of bin i.

As all these variables so far have been derived from the exact continuous distributions $f_X(x)$ and $f_Y(y)$, they correspond to the expectation values that researchers are estimating during data collection and analysis. As this is a counting process the components of the observed number of signal events $\boldsymbol{n}$, an N-vector, are often related to the components of the expected number of observed counts $\boldsymbol{\nu}$ as a collection of N separate and independent Poisson processes. That is to say the observed counts n_i in bin i are treated as i.i.d. random variables with the probability mass function

$$P(n_i|\nu_i) = \frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}. \tag{13}$$

[2]In the continuous case it is typically written as $\epsilon(x)$, and understood to be the conditional probability of an event producing any measured value given it has a true value of x. It is typically absorbed into $K(x, y)$ where it goes on to manifest within $\boldsymbol{R}$ in the manner shown in Equation (11) [2].

The counts n_i would in theory then form the estimate $\hat{\nu}_i$ of the expected counts ν_i by

$$\begin{aligned}\nu_i &= \mathrm{E}[\hat{\nu}_i] = \mathrm{E}[n_i] \\ &= \mathrm{Var}[\hat{\nu}_i] = \mathrm{Var}[n_i].\end{aligned}$$

Understanding the probability distribution of $\boldsymbol{n}$ allows for unfolding methods that involve the use of maximum likelihood estimation. For methods based on least squares it becomes necessary to find the covariance matrix $\boldsymbol{\Sigma}^{\nu}$ of the observations, which for independent Poisson processes has components of the form

$$\begin{aligned}\Sigma^{\nu}_{ij} &= \mathrm{Cov}[n_i, n_j] \\ &= \delta_{ij}\nu_i,\end{aligned} \tag{14}$$

where δ_{ij} is the Kronecker delta[3]. The path to the covariance matrix of the estimated true distribution $\hat{\boldsymbol{\mu}}$ can be considered briefly by considering the maximum log-likelihood, where it can be shown

$$\begin{aligned}\log L(\boldsymbol{\mu}) &= \sum_{i=1}^{N} \log\left(\frac{\nu_i^{n_i} e^{-\nu_i}}{n_i!}\right) \\ &= \sum_{i=1}^{N} \left(n_i \log \nu_i - \nu_i - \log n_i!\right) \\ \frac{\partial \log L}{\partial \mu_k} &= \sum_{i=1}^{N} \frac{\partial \log L}{\partial \nu_i}\frac{\partial \nu_i}{\partial \mu_k} \\ &= \sum_{i=1}^{N} \left(\frac{n_i}{\nu_i} - 1\right) R_{ik} = 0.\end{aligned}$$

Some minor algebra here reproduces the estimate $\hat{\boldsymbol{\nu}} = \boldsymbol{n}$, which has been assumed up until now. Continuing with an additional derivative shows

$$\begin{aligned}\frac{\partial^2 \log L}{\partial \mu_k \partial \mu_l} &= -\sum_{i=1}^{N} \left(\frac{n_i}{\nu_i^2}\frac{\partial \nu_i}{\partial \mu_l}\right) R_{ik} \\ &= -\sum_{i=1}^{N} \frac{n_i R_{il} R_{ik}}{\nu_i^2},\end{aligned}$$

the expectation value of which is the Fisher information.

[3]The Kronecker delta δ_{ij} is a piecewise function of variables i and j defined by $\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$.

1.4 A Simulated Example

Consider an example of the form described by Equation (1), i.e. $Y_i = \epsilon_i (X_i + \varepsilon_i)$. Let X_i be a i.i.d. random variable from a bimodal distribution of the form $X_i = Z_i X_{1,i} + (1 - Z_i) X_{2,i}$, where

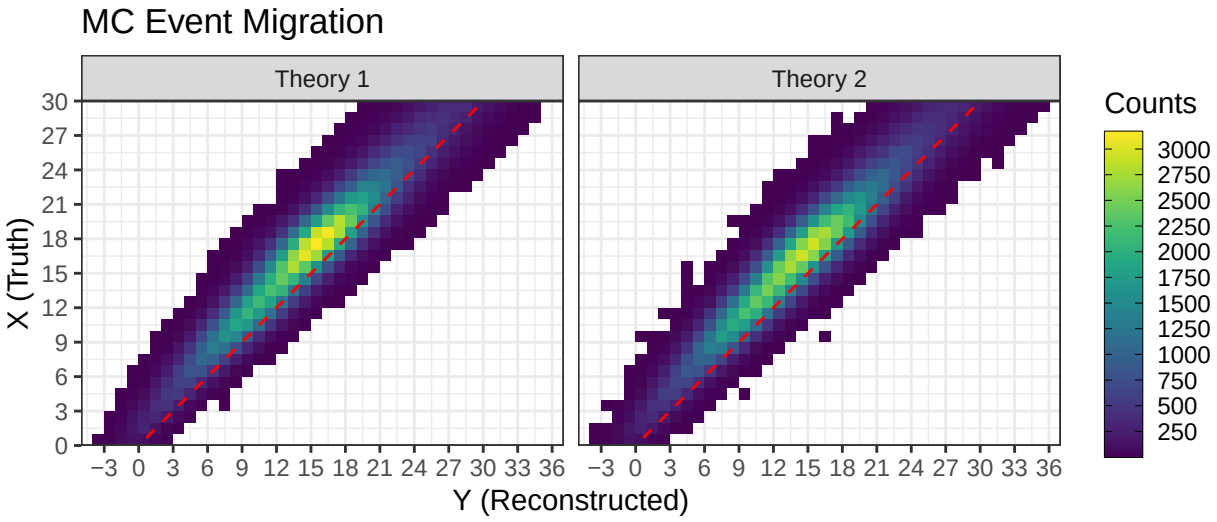
$$\begin{aligned} X_{1,i} &\sim \text{Gamma}(24, 0.4), \\ X_{2,i} &\sim \text{Gamma}(42, 0.4), \\ \text{and } Z_i &\sim \text{Bernoulli}(2/7), \end{aligned}$$

and let the effects of detector smearing be represented by i.i.d random variables generated by the conditional Gaussian process $\varepsilon_i \sim N\left(\mu(X_i), \sigma(X_i)^2\right)$, the mean and variance of which are functions defined by

$$\begin{aligned} \mu(X_i = x) &= -x^{1/4} \text{ and} \\ \sigma(X_i = x) &= \log\left(\frac{x + 10}{4}\right). \end{aligned}$$

The efficiency is similarly conditional on X_i, and is modeled here as a Bernoulli process with i.i.d random variables $\epsilon_i \sim \text{Bernoulli}\Big(p(X_i)\Big)$, where the average detection rate (when $\epsilon_i = 1$) is a function of the form

$$p(X_i = x) = 1 - e^{-\sqrt{x}/4}.$$



2 Unfolding in particle physics

2.1 Bin-by-bin

In this approach a multiplicative **correction factor** C_i is applied to the observed number of signal events n_i for each bin to produce the estimator of μ_i [7],

$$\hat{\mu}_i = C_i n_i. \tag{15}$$

The correction factors are determined by taking the respective ratios of a bin's MC simulated truth signal event counts μ_i^{MC} to its MC simulated reconstructed signal event counts ν_i^{MC},

$$C_i = \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}}. \tag{16}$$

The covariance matrix $\mathbf{\Sigma}_\mu$ of this estimator derives naturally from Equations (14) and (15), with components

$$\begin{aligned} \Sigma_{ij}^{\mu} &= \text{Cov}[\hat{\mu}_i, \hat{\mu}_j] \\ &= C_i C_j \text{Cov}[n_i, n_j] \\ &= C_i^2 \delta_{ij} \nu_i. \end{aligned} \tag{17}$$

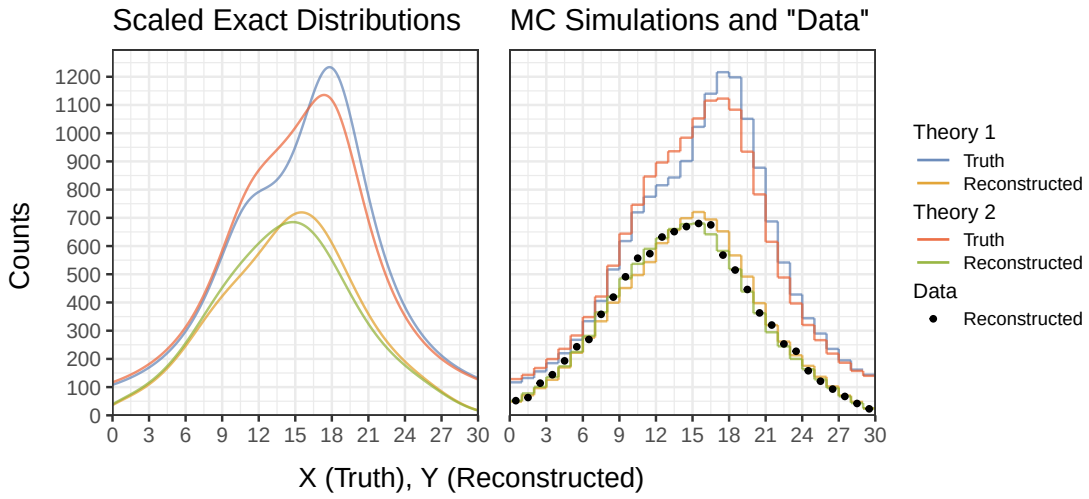
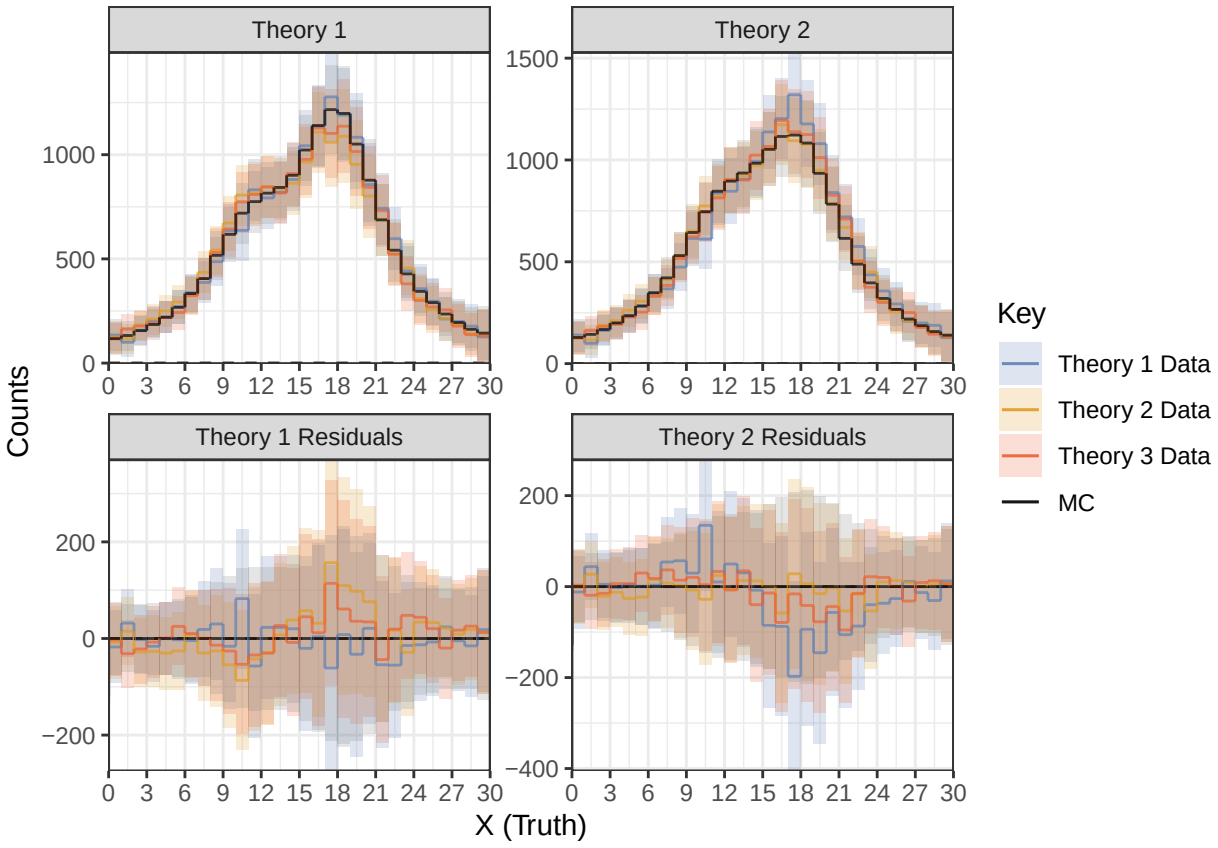
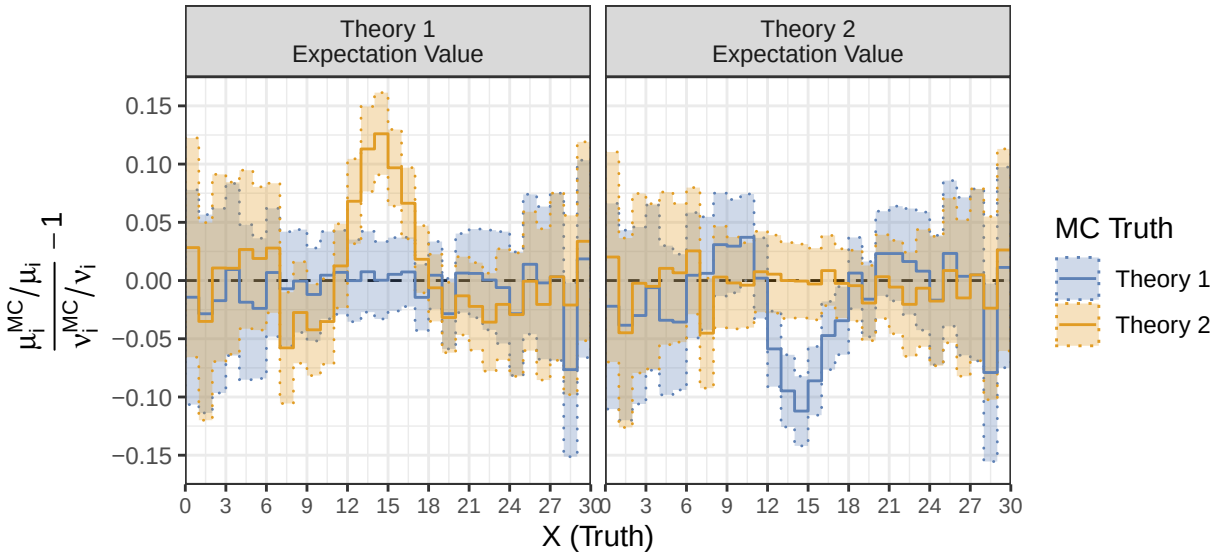


Figure 1: *The above plots feature two bimodal gamma distributions depicting events before and after after detector effects. (Top Left) The two continuous distributions correspond to the theoretical PDFs of the two distributions rescaled to correspond with the counts from 20,000 events. The histograms are calculated from the PDFs and correspond to the expected event counts from 20,000 simulated events. (Bottom Left) These four histograms consist of the same expected event counts as well as one instance of actual counts resulting from 20,000 simulated events. (Right) A visual study of simulated detector efficiency is provided by a side-by-side comparison of two heat maps that demonstrate the skewness and dispersion added by a simulated measurement process for detected and undetected events. This study is not intended to meaningfully represent a hypothetical distribution of undetected events in any real detector. Actual detector efficiencies are almost certainly governed by much more complicated collections of parameters.*

The expectation value of the estimate can be calculated easily enough as well, and with it the bias

$$\begin{aligned}
\text{Bias}[\hat{\mu}_i] &= E_i[\hat{\mu}_i] - \mu_i \\
&= C_i E[n_i] - \mu_i \\
&= \frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} \nu_i - \mu_i \\
&= \left(\frac{\mu_i^{\text{MC}}}{\nu_i^{\text{MC}}} - \frac{\mu_i}{\nu_i} \right) \nu_i.
\end{aligned} \tag{18}$$





2.2 Inverting the Response Matrix

In the event of Equation (7) being well-posed the obvious approach would be to construct the unique inverse of the response matrix $\boldsymbol{R}^{-1}$ and map the reconstructed counts back to an

estimate of the true counts via

$$\hat{\boldsymbol{\mu}} = \boldsymbol{R}^{-1}\boldsymbol{n}. \tag{19}$$

A statistical justification for this comes from performing generalized least squares [9] fit to estimate $\boldsymbol{\mu}$, which relies on approximating bin count n_i as normally distributed with mean ν_i and variance $1/\nu_i$. Minimizing the sums of squares yields

$$\begin{aligned}
\min_{\boldsymbol{\mu}} \nabla_{\mu}\boldsymbol{\chi}^2(\boldsymbol{\mu}) &= \nabla_{\mu}(\boldsymbol{R\mu} - \boldsymbol{n})^T\boldsymbol{\Sigma}_{\nu}^{-1}(\boldsymbol{R\mu} - \boldsymbol{n}) \\
&= \nabla_{\mu}(\boldsymbol{\mu}^T\boldsymbol{R}^T - \boldsymbol{n}^T)\boldsymbol{\Sigma}_{\nu}^{-1}(\boldsymbol{R\mu} - \boldsymbol{n}) \\
&= \nabla_{\mu}(\boldsymbol{\mu}^T\boldsymbol{R}^T\boldsymbol{\Sigma}_n^{-1}\boldsymbol{R\mu} - \boldsymbol{\mu}^T\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{n} - \boldsymbol{n}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{R\mu} + \boldsymbol{n}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{n}) \\
&= \nabla_{\mu}(\boldsymbol{\mu}^T\boldsymbol{R}^T\boldsymbol{\Sigma}_n^{-1}\boldsymbol{R\mu} - 2\boldsymbol{\mu}^T\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{n} + \boldsymbol{n}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{n}) \\
&= 2\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{R\mu} - 2\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{n} \\
&= 0 \\
\implies \boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{R\mu} &= \boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{n} \\
\implies \hat{\boldsymbol{\mu}} &= (\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{R\mu})^{-1}\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{n} = \boldsymbol{R}^{+}\boldsymbol{n} \qquad (20) \\
\implies \boldsymbol{R}^{+} &= (\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{R\mu})^{-1}\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}, \qquad (21)
\end{aligned}$$

where $\boldsymbol{R}^{+}$ is the Moore-Penrose generalized inverse (or pseudo-inverse) [2], and it is assumed that $\boldsymbol{R}^T\boldsymbol{\Sigma}_{\nu}^{-1}\boldsymbol{R\mu}$ is not singular. The corresponding covariance matrix is more quickly calculated by way of

$$\boldsymbol{\Sigma}_{\mu} = \mathrm{Cov}[\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\mu}}] = \mathrm{Cov}[\boldsymbol{R}^{+}\boldsymbol{n}, \boldsymbol{R}^{+}\boldsymbol{n}] = \boldsymbol{R}^{+}\mathrm{Cov}[\boldsymbol{n}, \boldsymbol{n}]\boldsymbol{R}^{+^T} = \boldsymbol{R}^{+}\boldsymbol{\Sigma}_{\nu}\boldsymbol{R}^{+^T}.$$

One such inverse matrix was calculated from each of the MC Theory simulations and both were applied to every set of the Reconstructed counts. The residuals of these unfolding calculations are paired with a 1σ uncertainty from the estimated Truth counts. They are then scaled by the expected Truth for the assumed theory and plotted on a log10 scale in Figure 2.

This degree of failure is disastrous. What is evinced here is that this problem does not satisfy Condition 3 of a well-posed problem as described in Section 1. That the inverse calculated here maps $\boldsymbol{n}$ to negative counts indicates that the estimate $\hat{\boldsymbol{\mu}} \notin \mathcal{U}$, the space of allowable solutions defined at the beginning of Section 1.3. Recall $\boldsymbol{\nu} \in \mathcal{V}$ back to some $\boldsymbol{\mu} \in \mathcal{U}$, as were defined to contain vectors with elements ≥ 0 .

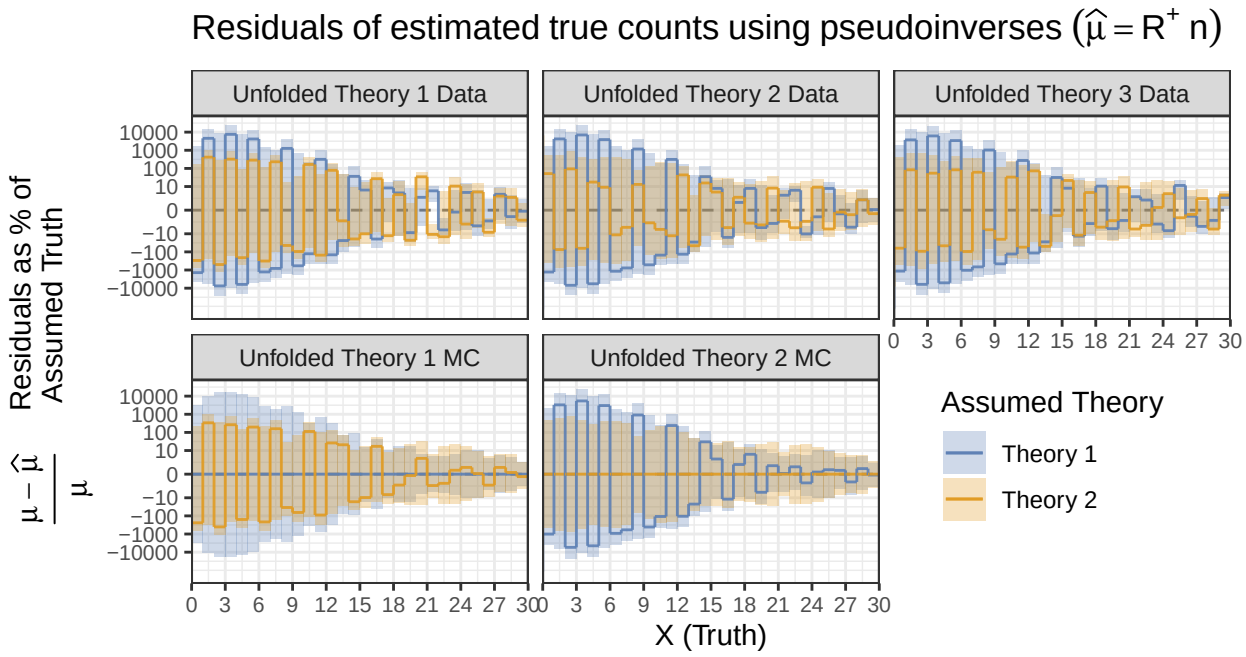
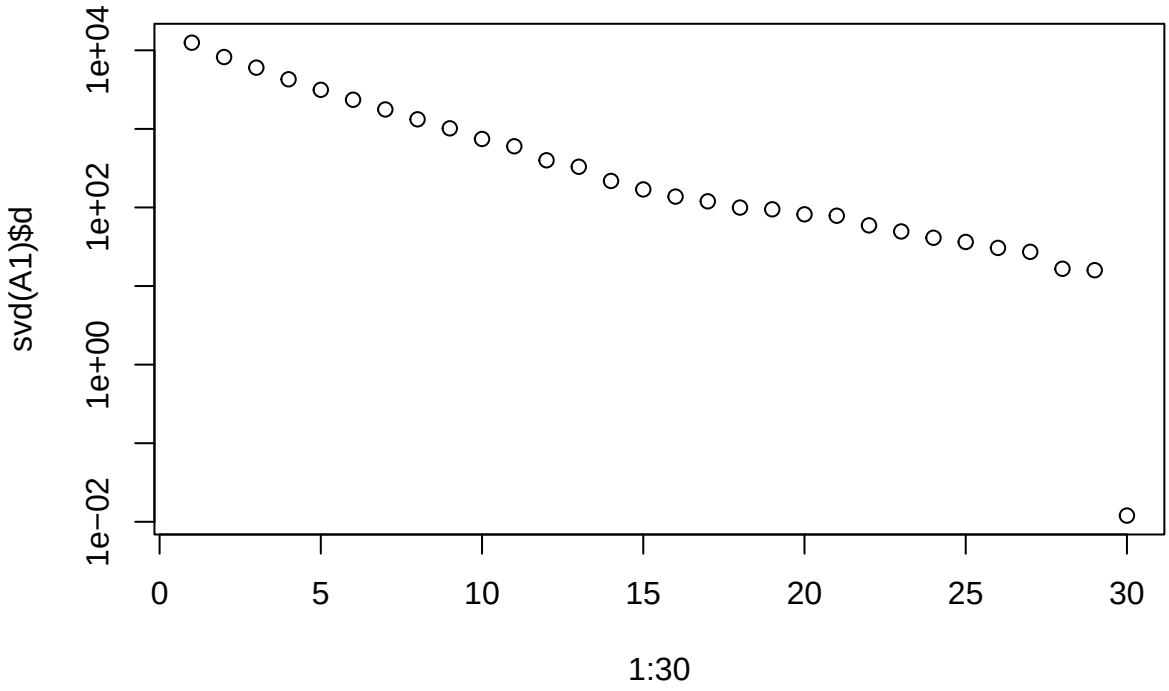
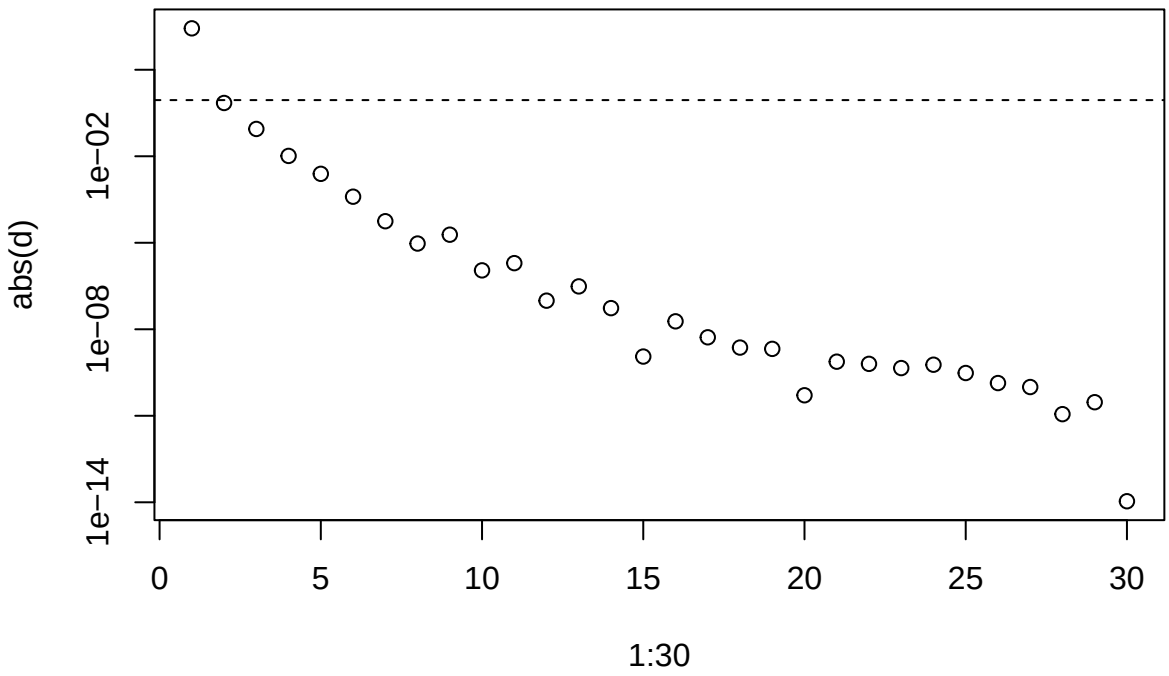
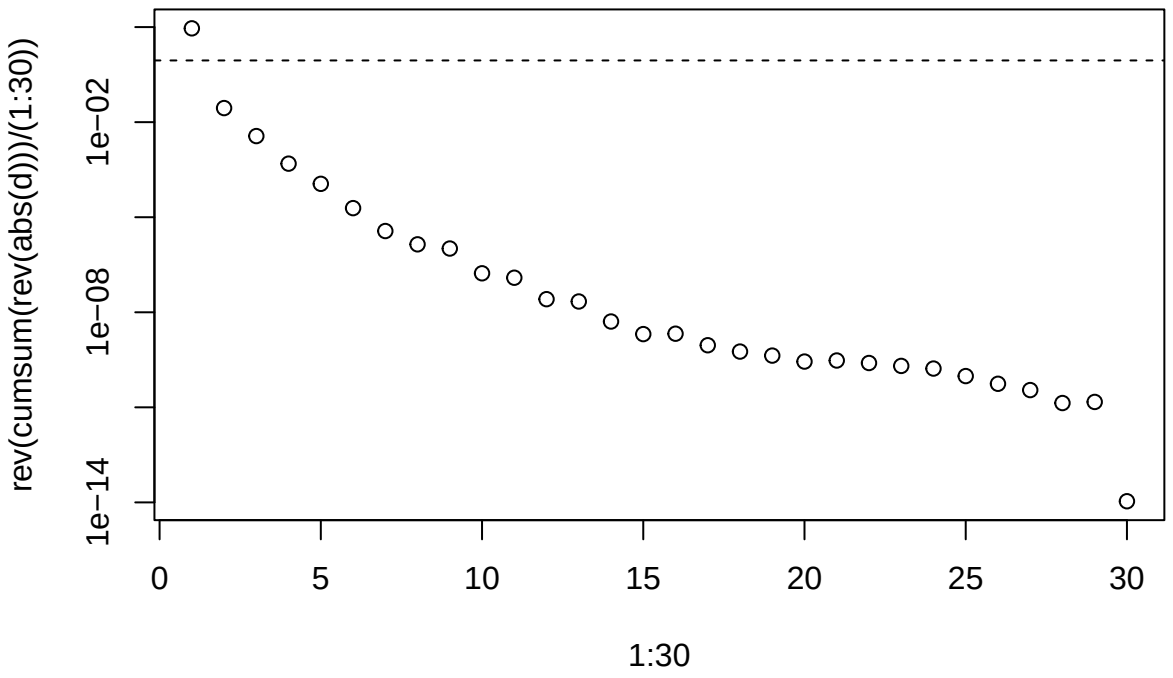


Figure 2: *Plots featuring the ratio of the residuals to the assumed theory's expected truth for the five sets of reconstructed counts when unfolded using response matrices derived from MC simulations for both Theory 1 and Theory 2. Note the log10 scale of the vertical axes, the large uncertainties, and negative estimated counts. The MC results are exact when unfolded under the correct theory for tautological reasons.*




abs(d)
1e−02
1e−08
1e−14
0
5
10
15
20
25
30
1:30



3 Algorithm Construction

Unfolding is ultimately concerned with finding a reliable inverse to the process by which an event occurring in some true bin i maps to an observed bin j. In generalizing this a bit we can think in terms of *causes*, C_i $(i = 1, \ldots, n_C)$ and *effects*, E_j $(j = 1, \ldots, n_E)$, representing the true and observed bins respectively. In regard to a single event we are then interested in conditional probabilistic view for causation, $P(C_i|E_j, I) \equiv \theta_{ij}$, the probability that we can attribute some cause C_i to an observed effect E_j. Using Bayes' theorem we can define this in terms of other probabilities that can be estimated more directly,

$$\begin{aligned} P(C_i|E_j, I) &= \frac{P(E_j|C_i, I) \cdot P(C_i|I)}{\sum_{i=1}^{n_C} P(E_j|C_i, I) \cdot P(C_i|I)} \\ \theta_{ij} &= \frac{\lambda_{ji} \cdot P(C_i|I)}{\sum_{i=1}^{n_C} \lambda_{ji} \cdot P(C_i|I)}, \end{aligned} \tag{22}$$

where the conditional probability regarding inference (effect), $P(E_j|C_i, I) \equiv \lambda_{ji}$, is the probability that some effect E_j will result with some cause C_i and $P_o(C_i|I)$ is the true probability of an event occurring from cause C_i. The term I represents any implicit conditional

information regarding the analysis, such as the choice of prior, and is usually apparent when the probabilities are written out as density functions.

At the analysis level we care less about individual events and more about mapping the total number events per effect, $\boldsymbol{x}_E = \{x(E_1), \ldots, x(E_{n_E})\}$, to the total number of events per cause, $\boldsymbol{x}_C = \{x(C_1), \ldots, x(C_{n_C})\}$. However, so far this regards only observed events as categorized into the n_E effects, as we cannot expect to observe or select for all effects resulting from some arbitrary cause C_i. In light of this, while we can add causes to account for any independent background sources to assume the normalization of $P_o(C_i|E_j, I)$ and $P_o(C_i|I)$, such that $\sum_{i=1}^{n_C} P_o(C_i|E_j, I) = 1$ and $\sum_{i=1}^{n_C} P(C_i|I) = 1$, we cannot say the same for $P(E_j|C_i, I)$. A necessarily imperfect effect selection capability results in

$$0 \leq \sum_{j=1}^{n_E} P(E_j|C_i, I) = \sum_{j=1}^{n_E} \lambda_{ji} \equiv \epsilon_i \leq 1,$$

the exact value of which provides for us a definition for ϵ_i, the *efficiency* at which we detect cause C_i from all accounted for observed effects, being also defined and useable as the ratio of observed events resulting from cause C_i to the true number of events resulting from C_i, $x^{obs}(C_i)/x(C_i)$.

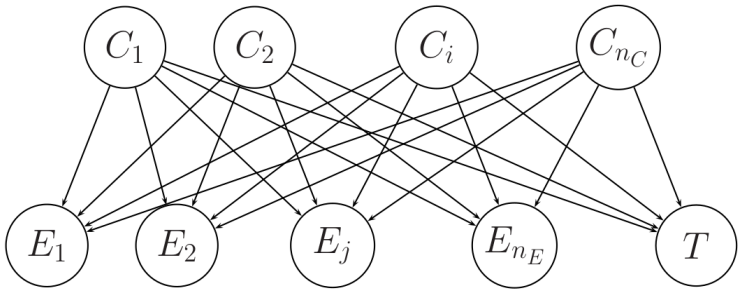


Figure 3: *Thinking of causes and effects as distinct subsets within one or more dimensional cause and effect phase spaces, this figure shows how events are probabilistically mapped from the subsets used to define our causes to the subsets used to define our effects. The node indicated by T ('trash') represents the event selection inefficiency, and can be thought of as an additional effect E_{n_E+1} that contains an unobserved number of events. Unlike the possibility of alloting independent sources of background to different causes, E_{n_E+1} (T) can consist of any number of potentially distinguishable effects depending on how the lost events are distributed across the complement of $\cup(E_1, E_2, \ldots, E_{n_E})$ in our effect phase space.*

A visualization of these lost events can be seen in Figure [3], where some collection of undocumented effects resulting from our collection of causes are lumped into a composite effect E_{n_E+1}, which should relate to our efficiency regarding cause C_i by $P(E_{n_E+1}|C_i, I) = \lambda_{n_E+1,i} = 1 - \epsilon_i$. Including this new effect with the others results in $\sum_{j=1}^{n_E} \lambda_{ji} = 1$, creating

normalized basis vectors to define the columns of a *smearing matrix* Λ,

$$\begin{aligned}\Lambda &= \begin{pmatrix} P(E_1|C_1,I) & P(E_1|C_2,I) & \dots & P(E_1|C_{n_C},I) \\ P(E_2|C_1,I) & P(E_2|C_2,I) & \cdots & P(E_2|C_{n_C},I) \\ \vdots & \vdots & \ddots & \vdots \\ P(E_{n_E}|C_1,I) & P(E_{n_E}|C_2,I) & \dots & P(E_{n_E}|C_{n_C},I) \\ P(E_{n_E+1}|C_1,I) & P(E_{n_E+1}|C_2,I) & \dots & P(E_{n_E+1}|C_{n_C},I) \end{pmatrix} \\ &= (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \dots, \boldsymbol{\lambda}_{n_c}),\end{aligned} \tag{23}$$

where $\boldsymbol{\lambda}_i$ refers to the i-th column consisting of $\{\lambda_{1,i}, \lambda_{2,i}, \dots, \lambda_{n_E+1,i}\}$.

Now that Eq. (2) accounts for lost events we can begin to construct a conditional probability for $\boldsymbol{x}_C$ similar to that for Eq. (1) using Bayes' theorem,

$$P(\boldsymbol{x}_C|\boldsymbol{x}_E, \Lambda, I) \propto P(\boldsymbol{x}_E|\boldsymbol{x}_C, \Lambda, I) \cdot P(\boldsymbol{x}_C|I), \tag{24}$$

and account for uncertainties in Λ with

$$P(\boldsymbol{x}_C|\boldsymbol{x}_E, I) = \int P(\boldsymbol{x}_C|\boldsymbol{x}_E, \Lambda, I)\ f(\Lambda|I)\ \mathrm{d}\Lambda.$$

At this point one should recognize in Eq. (3) $P(\boldsymbol{x}_E|\boldsymbol{x}_C, \Lambda, I)$ as the likelihood and $P(\boldsymbol{x}_C|I)$ as the prior in the formal calculation of a posterior. Ignoring the prior for now and focusing on the likelihood, for a given cause C_i we can model this expression as a multinomial distribution such that

$$P(\boldsymbol{x}_E|x(C_i), \Lambda, I) = \frac{x(C_i)!}{\prod_j^{n_E+1} x(E_j)!} \prod_j^{n_E+1} \lambda_{ji}^{x(E_j)}, \tag{25}$$

which leads us finally to asking how we should estimate our λ_{ji}'s. Once again, using Bayes' theorem we can see that for a fixed i,

$$f(\boldsymbol{\lambda}_i|\boldsymbol{x}_E, x(C_i), I) \propto P(\boldsymbol{x}_E|x(C_i), \boldsymbol{\lambda}_i, I) \cdot f(\boldsymbol{\lambda}_i|I). \tag{26}$$

Previously mentioned properties about $\boldsymbol{\lambda}_i$ make a Dirichlet($\boldsymbol{\alpha}_{prior_i}$) prior appropriate. A flat prior in which $\boldsymbol{\alpha}_{prior_i} = \{1, \dots, 1\}$ is often chosen, and is done so here. Regardless of the

choice for $\boldsymbol{\alpha}_{prior_i}$, multiplying by a multinomial distribution results in the posterior

$$\begin{aligned} f(\boldsymbol{\lambda}_i|\boldsymbol{x}_E, x(C_i), I) &\propto \left[\frac{x(C_i)!}{\prod_j^{n_E+1} x(E_j)!}\prod_j^{n_E+1}\lambda_{ji}^{x(E_j)}\right]\cdot\left[\frac{1}{\mathrm{B}(\boldsymbol{\alpha}_{prior_i})}\prod_{j=1}^{n_E+1}\lambda_{ji}^{\alpha_{prior_{ji}}-1}\right] \\ &\propto \prod_j^{n_E+1}\lambda_{ji}^{\left(\alpha_{prior_{ji}}+x(E_j)\right)-1} \\ &= \mathrm{Dirichlet}(\boldsymbol{\alpha}_{prior_i}+\boldsymbol{x}_E). \end{aligned} \tag{27}$$

Samples from this distribution informs much of the the uncertainty resulting from the unfolding process, creating distributions for objects fully or partially calculated from them, such as the smearing matrix, efficiency, and inverse probabilities θ_{ij} once a prior for $P(C_i|I)$ is made. These will be necessary, as $P(\boldsymbol{x}_C|\boldsymbol{x}_E, I)$ becomes the sum of independent multinomial distributions, which does not have a closed solution that we can analytically maximize the likelihood of. We have to make the rest of our progress starting from Eq. (1) where the choice around a prior for $P(C_i|I)$ is due. The choice of $P(C_i|I) = \textit{constant}$ is considered here, which D'Agostini acknowledges is a strong prior that produces biases that will require iterations to be resolved.

Defining $\boldsymbol{\theta}_j = \{\theta_{1,j}, \theta_{2,j}, \ldots, \theta_{n_C,j}\}$, we can model how the $x(E_j)$ observed events with the effect E_j are likely distributed from potential causes by the multinomial distribution

$$\boldsymbol{x}_C|_{x(E_j)} \sim \mathrm{Mult}(x(E_j), \boldsymbol{\theta}_j).$$

Before summing over the effects to get the total observed causes we should acknowledge that each $x(E_j)$ is the result of a Poisson process with an unknown rate parameter μ_j. Using the conjugate prior $\mu_j \sim \mathrm{Gamma}(c_j, r_j)$, with $c_j = 1$ and very small r_j to create a flat prior, we arrive at

$$\mu_j|_{x(E_j)} \sim \mathrm{Gamma}(c_j + x(E_j), r_j + 1),$$

which tells us not to use $x(E_j)$, but μ_j. In dealing with fractional values of μ_j D'Agostini suggests:

1. Rounding μ_j to its nearest positive integer m_j,
2. Sampling from $\boldsymbol{x}_C|_{m_j} \sim \mathrm{Mult}(m_j, \boldsymbol{\theta}_j)$,
3. Rescaling by $\boldsymbol{x}_C|_{\mu_j} = \frac{\mu_j}{m_j}\boldsymbol{x}_C|_{m_j}$,
4. Summing over each effect with $\boldsymbol{x}_C|_{\boldsymbol{x}_E} = \sum_{j=1}^{n_E}\boldsymbol{x}_C|_{\mu_j}$,
5. And applying the inefficiency correction with $\boldsymbol{x}_C = \frac{\boldsymbol{x}_C|_{\boldsymbol{x}_E}}{\epsilon_i}$.

The drawing of multiple samples from the posteriors is used to form an ensemble of values of $\boldsymbol{x}_C$ and estimate credible intervals. The performance of second and later iterations is accomplished by using the previous iteration's posteriors as the new iteration's priors.

4 A Basic Example

In the following example 400,000 random samples are drawn from a Cauchy distribution and then subject to some processes that disperse, bias, and reduce event selection efficiency. The results of these simulations are shown in Figure [**??**]. The migration matrix does a decent job of demonstrating how the true data was smeared. For unaffected data one would see just a diagonal line from the bottom left to the top right. It was with this in mind that instead of choosing the earlier mentioned flat prior for $\boldsymbol{\alpha}_{prior_i}$ I decided to go with

$$\alpha_{ij} = e^{-|x_{truth} - x_{smeared}|},$$

the values of which are represented in Figure [**??**].

5 Discussion of Results and Conclusion

The results of my unfolding are shown below in Figure [**??**]. The iterations get off to an okay start, but instead of converging they begin to behave erratically. The tails blowing up clearly comes from using random samples from a posterior distribution as basis for new priors, embedding events in places where there were zero before. I attempted to remedy this with a prior that disfavored events occurring far from the diagonal, as mentioned previously in the context of Figure [**??**]. It kept these tails from blowing up for at least the first iteration. In the future I will look more into options relating to this issue.

There is almost certainly an error in the code governing the 3rd iteration, resulting in a vastly larger confidence band. Instead of trying to fix this issue I think it would be better use of my time to study some similar working examples [5].

A Supplementary Mathematical Definitions and Derivations

A.1 Hilbert Spaces

Hilbert spaces are a major structural component of the field of functional analysis. They see significant application in partial differential equations, quantum mechanics, and signal processing, where they are commonly implemented in the performance of Fourier analysis. Mathematically they represent an extension beyond the real and complex geometric-like vector spaces developed by earlier generalizations of Euclidean spaces in the 19th century. Developments in real analysis at the beginning of the 20th century lead the spaces of functions and sequences to being conceptualized as linear spaces in their own right.

As extensions of previously understood spaces they necessarily exist at the intersection of several other important spaces that aught to be understood beforehand. With that said, the following definitions come from Rudin in [12]. To start, a **vector space**, as defined here, consists of a set X of vectors for which addition and scalar multiplication are defined such that for all $x, y, z \in X$ and any complex number $\alpha \in \mathbb{C}$

1. there exists a vector in X such that
   (a) addition is commutative: $x + y = y + x$,
   (b) addition is associative: $x + (y + z) = (x + y) + z$,
2. αx exists in X such that $1x = x$, $0x = 0$ (the zero vector), and multiplication is distributive:
   (a) $\alpha(\beta x) = (\alpha\beta x)$,
   (b) $\alpha(x + y) = \alpha x + \alpha y$, and
   (c) $(\alpha + \beta)x = \alpha x + \beta x$.

The range of α above describes a complex vector space. If α is restricted to the reals $\mathbb{R}$, then X is considered a real vector space. Note that vector spaces include more than just traditional coordinate-style vectors, but also include function spaces such as the vector space of all polynomials with degree of at most n, which has the basis $\{1, x, x^2, \ldots, x^{n-1}, x^n\}$.

Typically associated in applications, metric spaces form a another relevant set of spaces that has some significant overlap with the vector spaces. A space X is said to be a **metric space** if for all $x, y \in X$ there exists an operator $d(x, y)$ that maps them to a nonnegative real

number that defines their distance from each other within X. The properties of this operator are

1. $0 \leq d(x, y) < \infty$ for all x and $y \in X$,
2. $d(x, y) = 0$ iff $x = y$,
3. $d(x, y) = d(y, x)$ for all x and $y \in X$,
4. $d(x, z) \leq d(x, y) + d(y, z)$ for all x, y, $z \in X$.

For a metric space X, the distance operator d is referred to as the metric on X. The intersection of the vector and metric spaces form the set of normed spaces. As an extension of the conditions thus far, a space X is a **normed space** if $\forall x \in X$ there exists a nonnegative real number $||x||$, called the **norm** of x such that

1. $||x + y|| \leq ||x|| + ||y|| \; \forall x, y \in X$,
2. $||\alpha x|| = |\alpha| \, ||x||$ if $x \in X$ and α is a scalar,
3. $||x|| > 0$ if $x \neq 0$.

Such a set is said to be **complete** if every **Cauchy sequence** in X converges to a point in X. A Cauchy sequence in a metric space X is any sequence $\{x_n\}$ that $\forall \varepsilon > 0$ there exists an integer N such that $d(x_m, x_n) < \varepsilon$ when $m > N$ and $n > N$. A quick example of this is the sequence defined by $x_n = \sqrt{n}$. For some starting x_m and x_n where $m - n = \delta$, we have

$$\begin{aligned} d(x_m, x_n) &= \sqrt{m} - \sqrt{n} \\ &= \sqrt{n + \delta} - \sqrt{n} \\ &= (\sqrt{n + \delta} - \sqrt{n}) \frac{\sqrt{n + \delta} + \sqrt{n}}{\sqrt{n + \delta} + \sqrt{n}} \\ &= \frac{n + \delta - n}{\sqrt{n + \delta} + \sqrt{n}} \\ &= \frac{\delta}{\sqrt{n}(\sqrt{1 + \delta/n} + \sqrt{1})} \\ &< \frac{1}{\sqrt{n}} \left(\frac{\delta}{2} \right) < \varepsilon \\ \implies n &> \left(\frac{\delta}{2\varepsilon} \right)^2. \end{aligned}$$

Noting that for constant δ the limit of $\frac{1}{\sqrt{n}} \left(\frac{\delta}{2} \right)$ as $n \longrightarrow \infty$ is the zero vector (the point of convergence) would also be sufficient to show that $x_n = \sqrt{n}$ is a Cauchy sequence.

Incidentally, a normed vector space that is complete as defined here meets the definition of a **Banach space**. An additional subset of the normed vector spaces consists of those spaces in which for all $x, y \in X$ there exists a real or complex number $\langle x, y\rangle$ defined by an operator called the **inner product**. For all $x, y, z \in X$ this operation must satisfy

1. $\langle x, y\rangle = \langle y, x\rangle^*$ (where the * represents the complex conjugate),
2. $\langle x + y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$,
3. $\langle \alpha x, y\rangle = \alpha\langle x, y\rangle$ (for $\alpha \in \mathbb{C}$),
4. $\langle x, x\rangle \geq 0$, and
5. $\langle x, x\rangle = 0$ iff $x = 0$.

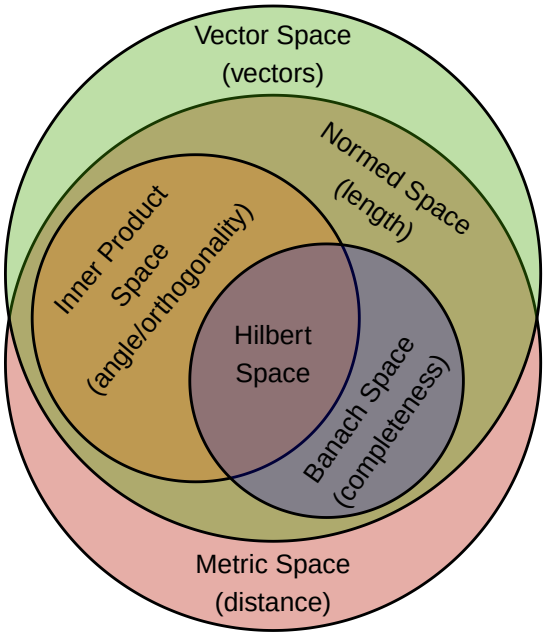


Figure 4: *A Venn diagram representing the intersection and nesting of the spaces described in Appendix A.1.*

A space that satisfies these requirements forms an **inner product space**, and the inner product defined in such a space relates to the form of its norm, such that $||x|| = \langle x, x\rangle^{1/2}$. Finally, at the intersection of Banach spaces and inner product spaces are the Hilbert spaces. I.e. a **Hilbert space** is a complete vector space with an inner product defined by its norm.

A commonly presented example is the L^2 function space, which consists of functions that are square integrable, i.e. if $f(x) \in L^2 \implies ||f(x)||^2 = \int_\chi |f(x)|^2 dx < \infty$, where χ is the domain of x. The subset $L^2[-\pi, \pi]$, where $\chi = [-\pi, \pi]$, has the well known Fourier series as a basis, which is commonly written such that for $f(x) \in L^2[-\pi, \pi]$

$$f(x) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos(nx) + b_n \sin(nx)\right],$$

where

$$a_n = \tfrac{1}{\pi} \textstyle\int_{-\pi}^{\pi} f(x) \cos(nx) dx$$
$$\text{and}$$
$$b_n = \tfrac{1}{\pi} \textstyle\int_{-\pi}^{\pi} f(x) \sin(nx) dx.$$

Verification that this basis meets all the requirements laid out so far is beyond the scope of

this paper.

B R Code

References

[1] Tim Adye. "Unfolding algorithms and tests using RooUnfold". In: *PHYSTAT 2011*. Geneva: CERN, 2011, pp. 313–318. DOI: 10.5170/CERN-2011-006.313. eprint: 1105.1160.

[2] Volker Blobel. "Unfolding". In: *Data analysis in high energy physics: A practical guide to statistical methods*. Ed. by Olaf Behnke et al. Weinheim, Germany: Wiley-VCH, 2013. Chap. 6, pp. 187–226.

[3] Volker Blobel. "Unfolding Methods in Particle Physics". In: *PHYSTAT 2011*. Geneva: CERN, 2011, pp. 240–251. DOI: 10.5170/CERN-2011-006.252. eprint: 1105.1160.

[4] Mary L. Boas. *Mathematical Methods in the Physical Sciences*. Third. Wiley, 2005. ISBN: 9780471198260.

[5] Carsten Daniel Burgard. *RooUnfold*. https://gitlab.cern.ch/RooUnfold/RooUnfold. 2021.

[6] G. Casella and R.L. Berger. *Statistical Inference*. Second. Cengage Learning, 2001. ISBN: 9780534243128.

[7] G. Cowan. *Statistical Data Analysis*. Oxford University Press, USA, 1998. ISBN: 9780198501558.

[8] G. D'Agostini. "A Multidimensional unfolding method based on Bayes' theorem". In: *Nucl. Instrum. Meth. A* 362 (1995), pp. 487–498. DOI: 10.1016/0168-9002(95)00274-X.

[9] R. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Sixth. Pearson, 2007. ISBN: 9780131877153.

[10] Alexander Meister. *Deconvolution Problems in Nonparametric Statistics*. Vol. Lecture Notes in Statistics. Springer, 2009. ISBN: 9783540875567.

[11] Victor M. Panaretos. "A Statistician's View on Deconvolution and Unfolding". In: *PHYSTAT 2011*. Geneva: CERN, 2011, pp. 252–259. DOI: 10.5170/CERN-2011-006.252. eprint: 1105.1160.

[12] W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 1991. ISBN: 9780070542365.

[13] Eric W. Weisstein. *Convolution. From MathWorld—A Wolfram Web Resource*. Last visited on 15/2/2022. URL: https://mathworld.wolfram.com/Convolution.html.

[14] Anatoly G. Yagola. “Ill-Posed Problems and Methods for Their Numerical Solution”. In: *Optimization and Regularization for Computational Inverse Problems and Applications*. Ed. by Yanfei Wang, A.G. Yagola, and Changchun Yang. Springer, Berlin, Heidelberg, 2011. Chap. 2, pp. 17–34. ISBN: 9783642137419. DOI: https://doi.org/10.1007/978-3-642-13742-6.