Nuclear Instruments and Methods in Physics Research A 729 (2013) 410–416

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

journal homepage: www.elsevier.com/locate/nima

An algorithm for automatic unfolding of one-dimensional data distributions

Hans P. Dembinski *, Markus Roth

IKP, Karlsruhe Institute of Technology (KIT), Postfach 3640, D-76021 Karlsruhe, Germany

ARTICLE INFO

Article history:
Received 19 April 2012
Received in revised form
16 July 2013
Accepted 19 July 2013
Available online 26 July 2013

Keywords:
Deconvolution
Unfolding
Resolution correction
Statistics
Regularisation
Non-parameteric

ABSTRACT

We discuss a non-parametric algorithm to unfold detector effects from one-dimensional data distributions. Unfolding is performed by fitting a flexible spline model to the data using an unbinned maximum-likelihood method while employing a smooth regularisation that maximises the relative entropy of the solution with respect to an *a priori* guess. A regularisation weight is picked automatically such that it minimises the mean integrated squared error of the fit. The algorithm scales to large data sets by employing an adaptive binning scheme in regions of high density. An estimate of the uncertainty of the solution is provided and shown to be accurate by studying the frequentist properties of the algorithm in Monte-Carlo simulations. The simulations show that the regularisation bias decreases as the sample size increases.

1. Introduction

Experimentalists want to publish measurements that are independent of the detection device. If the measurement is the distribution of an observable – for example, the energies of cosmic rays arriving at Earth – the finite resolution of the detector can be an issue since it will smear out this distribution. Mathematically, the true distribution $b(x)$ of the variable x is convolved with a kernel function $K_{\mathrm{det}}(y,x)$ which describes the fluctuations and losses introduced by the detector

$$f(y) = \int_{-\infty}^{\infty} \mathrm{d}x\, K_{\mathrm{det}}(y,x) b(x) \tag{1}$$

so that we obtain a Fredholm integral equation of the first kind. Only a particular realisation of the distorted distribution $f(y)$ is accessible in an experiment. The kernel $K_{\mathrm{det}}(y,x)$ can be further splitted into a conditional probability density $K(y,x)$ and a probability $\epsilon(y,x)$.

The conditional probability density $K(y,x)$ describes the frequency of converting a quantity x into an observable y. It models random fluctuations that occur in the detection process and may include a non-linear response. The efficiency of the detector is described by the detection probability $\epsilon(y,x)$. It summarises effects that cause event loss, such as events missing the sensitive range of the detector or events that are in range, but produce a signal below the detection threshold.

If $K(y,x)$ and $\epsilon(y,x)$ are known one can in principle solve Eq. (1) for the distribution $b(x)$. The task is straightforward if a parametric model of $b(x)$ exists, defined up to few free coefficients. In that case Eq. (1) can be fitted to the data directly with a maximum-likelihood method (this is sometimes called the forward folding method [1]). If nothing is known about $b(x)$, the task becomes very difficult since we lack information about the complexity and variability of the source distribution $b(x)$. In that sense it is similar to the challenging problem of non-parametric density estimation [2]. Model-independent unfolding can be approximated by choosing a parametric model with a very large number of parameters so that $b(x)$ can accommodate an arbitrary structure. However, since it is impossible to obtain all information about $b(x)$ from the finite data, a regularisation needs to be added which enforces a degree of smoothness on $b(x)$. This effectively reduces the number of free parameters to a tractable amount. Due to the regularisation, the statistical estimate $\hat{b}(x)$ of the true $b(x)$ in general has a non-zero bias: $E[\hat{b}(x)-b(x)]\neq 0$.

The art of unfolding is to find a regularisation that works well for real-life problems and that achieves an optimal balance between bias and statistical variance of the solution. A variety of algorithms are used in high energy physics [3–7]. We add ARU[1] to this set, an algorithm with some conceptual improvements that may improve the unfolding performance [8]. ARU was developed

* Corresponding author. Tel.: +49 72160829183.
E-mail address: hans.dembinski@kit.edu (H.P. Dembinski).

[1] http://projects.hepforge.org/aru

0168-9002/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.nima.2013.07.060

after studying the existing algorithms and uses central concepts of the RUN algorithm [4] and the maximum-entropy unfolding [1,5].

Our goal was to create an algorithm that can be used as a black box without requiring the user to tweak the algorithm for a specific problem and that uses the maximum amount of information in the data. One step into that direction is to avoid histogramming the data. We fit Eq. (1) directly to the raw event distribution $\{y_i\}$ after expanding the unknown function $b(x)$ into a sum of B-splines (like in RUN) and performing the numerical integration of Eq. (1). We show how an adaptive binning approach speeds up the solution of problems with huge statistics. The details are given in Section 2.

The fit of $b(x)$ is based on maximizing a regularized log-likelihood. The regularisation is up to a term that ensures proper normalisation of the relative entropy between the unfolded solution $b(x)$ and a reference function $g(x)$. The relative entropy is a well-defined measure of the similarity of two probability density functions and approaches zero as $b(x)$ approaches $g(x)$. By choosing $g(x)$ close to the correct solution we reduce the systematic bias. We argue that a fit to the original data distribution corrected by the detector efficiency is the best *a priori* guess of the final solution.

This scheme is a generalisation of the maximum-entropy regularisation [5] in which the reference distribution $g(x)$ is the uniform distribution. Like the latter our regularisation is invariant to monotonic transformations $x \to x'$ that do not change the statistical information of the problem. This is an advantage over regularisations of the Tikhonov-type [4,7,9] which do not share this property.

Our algorithm adjusts the strength of the regularisation automatically by minimising the mean integrated squared error (MISE) of the fit which determines the optimal balance between bias and variance of the estimate $\hat{b}(x)$ for the given problem, following the typical approach used in non-parametric density estimation [2]. The approach is explained in Section 3.

An analytical calculation of the uncertainty of the solution completes the algorithm, shown in Section 4. We compare the uncertainty estimate with the true frequentistic variance of the solution using Monte-Carlo simulations in Section 5. We close with a discussion and remarks on the use of the algorithm in Section 6.

2. Regularised maximum-likelihood fit

We construct the objective function for our fit initially from the maximum-likelihood principle. According to it, the best parameters $\hat{\boldsymbol{c}}$ of a model $f(y)$ maximise the joint probability P of all observations, the likelihood. As a matter of convention, we minimise $l_1 = -\ln P$ instead, which is equivalent.

In order to formulate the unfolding problem as a fit, we need a flexible parametrisation of $f(y)$. A good idea [4] is to expand the unfolded distribution $b(x)$ into a series of basis functions $b(x) = \sum_k c_k b_k(x)$, exploiting the linearity of Eq. (1):

$$f(y) = \int \mathrm{d}x\, \epsilon(y,x) K(y,x) b(x) = \sum_k c_k \int \mathrm{d}x\, \epsilon(y,x) K(y,x) b_k(x) = \sum_k c_k f_k(y). \quad (2)$$

The problem is thus reduced to a fit of the folded basis functions $f_k(y)$.

A convenient choice for the basis functions $b_k(x)$ of the unfolded distribution is B-splines [4,10], although other choices are possible. B-splines are bell-shaped functions with finite support. They are non-negative, which allows us to impose the mathematical condition $b(x) > 0$ by the constraint $c_k > 0$ that is supported by most numerical optimisation algorithms. A basis spline $b_k(x)$ on a grid with m knots is defined by the recursion formula

$$b_{k,0}(x) = \begin{cases} 1 & \text{if } x_k \le x < x_{k+1} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$b_{k,n}(x) = \frac{x - x_k}{x_{k+n} - x_j} b_{k,n-1}(x) + \frac{x_{k+n+1} - x}{x_{k+n+1} - x_{k+1}} b_{k+1,n-1}(x), \quad k = 0, \ldots, m+n-2. \quad (4)$$

We use $n=3$ and omit the index n in the following, so that $b_k(x) := b_{k,3}(x)$. The grid of knots is extended by three virtual knots to the left and the right of the interval in order to define the basis splines at the borders properly. A grid of m knots defines $m+2$ basis splines $b_k(x)$ so that the spline model has $m+2$ coefficients. No boundary conditions are enforced to reduce the number of coefficients. An example of a spline curve is shown in Fig. 1.

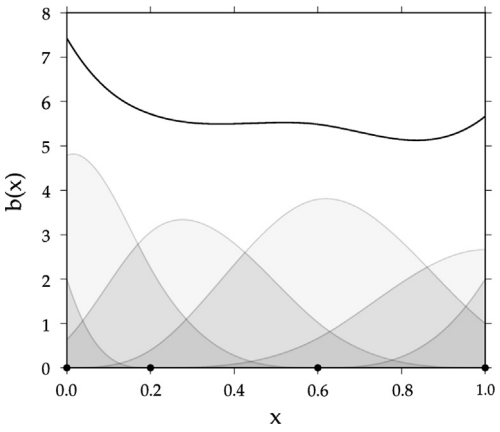

Fig. 1. Example of a spline model $b(x)$ (black line) over a vector of four knots (black dots). The shaded areas indicate the six scaled basis splines $c_k b_k(x)$ that constitute the curve.

The coefficient vector $\boldsymbol{c}$ needs to be very large in order to make the model sufficiently adjustable to approximate the limit of unbounded variance. Therefore, the fit will be under-constrained and the minimum of the negative log-likelihood l_1 will be a long thin valley, leading to huge variance and strong anti-correlations of the coefficients $\hat{\boldsymbol{c}}$. The valley is constrained by adding a weighted regularisation term wl_2, constructed to have a minimum at an *a priori* guess of the true solution. The solution $\hat{\boldsymbol{c}}$ to the unfolding problem is then found by minimising the combination

$$l(\boldsymbol{c}) = l_1(\boldsymbol{c}) + w l_2(\boldsymbol{c}) \quad (5)$$

with standard non-linear optimisation algorithms [11–13]. The price to pay for this approach is a statistical bias $E[\hat{\boldsymbol{c}} - \boldsymbol{c}] \neq 0$, since the guess will differ from the correct solution in general.

From the frequentist point of view we constructed a biased estimator with the intention of reducing the variance. From the Bayesian point of view we added prior information to our estimation.

2.1. First part l_1 of the regularised log-likelihood

We want to fit the shape and the normalisation of $b(x)$ and therefore construct the so-called extended likelihood function [1]. Let $f^*(y) = f(y)/\nu$ be the normalized probability density function (p.d.f.) that describes a set of data points y_i, with $\nu = \int \mathrm{d}y\, f(y)$ being the total expected number of observations. We assume Poisson fluctuations for the realised number N of events and obtain for the joint probability P of the data

$$P(\boldsymbol{c}) = \frac{\nu^N}{N!} e^{-\nu} \prod_i \int_{y_i}^{y_i + \Delta y_i} \mathrm{d}y\, f^*(y) \underset{\Delta y_i \to 0}{\longrightarrow} \frac{\nu^N}{N!} e^{-\nu} \prod_i f^*(y_i) \Delta y_i$$