Nuclear Instruments and Methods in Physics Research A 778 (2015) 92–101

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

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

Densities mixture unfolding for data obtained from detectors with finite resolution and limited acceptance

N.D. Gagunashvili *,[1]

University of Akureyri, Borgir, v/Nordurslód, IS-600 Akureyri, Iceland

ARTICLE INFO

Article history:
Received 9 October 2014
Received in revised form
29 December 2014
Accepted 5 January 2015
Available online 12 January 2015

Keywords:
Deconvolution
Mixture densities
Adaptive algorithm
Inverse problem
Single sided strongly varying spectra
Regularization

ABSTRACT

A procedure based on a Mixture Density Model for correcting experimental data for distortions due to finite resolution and limited detector acceptance is presented. Addressing the case that the solution is known to be non-negative, in the approach presented here, the true distribution is estimated by a weighted sum of probability density functions with positive weights and with the width of the densities acting as a regularization parameter responsible for the smoothness of the result. To obtain better smoothing in less populated regions, the width parameter is chosen inversely proportional to the square root of the estimated density. Furthermore, the non-negative garrote method is used to find the most economic representation of the solution. Cross-validation is employed to determine the optimal values of the resolution and garrote parameters. The proposed approach is directly applicable to multi-dimensional problems. Numerical examples in one and two dimensions are presented to illustrate the procedure.

1. Introduction

The probability density function (PDF) $P(x')$ of an experimentally measured characteristic x', in general, differs from the true physical PDF $p(x)$ because of the limited acceptance (probability) $A(x)$ to register an event with true characteristic x, finite resolution and bias in the response function $R(x'|x)$, which describes the probability to observe x' for a given true value x. Formally the relation between $P(x')$ and $p(x)$ is given by

$$P(x') \propto \int_{\Omega} p(x)A(x)R(x'|x)\,dx. \tag{1}$$

The integration in (1) is carried out over the domain Ω of the variable x. In practical applications the experimental distribution is usually discretised by using a histogram representation, obtained by integrating $P(x')$ over n finite sized bins

$$P_j = \int_{c_{j-1}}^{c_j} P(x')\,dx' \quad j = 1,\ldots,n \tag{2}$$

with c_{j-1}, c_j the limits of bin j.

If a parametric (theoretical) model $p(x, a_1, a_2, \ldots, a_l)$ for the true PDF is known, then the unfolding can be done by determining the parameters. For example, by a least squares fit to the binned data [1–3]. Here the model, which allows to describe the true distribution by a finite number of parameter values, constitutes a priori information which is needed to correct for the distortions by the experimental setup,

In contrast, model independent unfolding, as considered, e.g. in [4–14], is an ill-posed problem, and every approach to solve it requires a priori information about the solution. Methods differ, directly or indirectly, in the way a priori information is incorporated in the result.

2. Description of the unfolding method

To solve the unfolding problem (1), a representation of the true distribution has to be chosen. This representation should be as flexible as possible and allow introducing a priori information. Classical kernel statistics is an example that approximates the true distribution by putting a $1/N$-weighed copy of a kernel PDF at the location of each of N observed data points and adding them up (see e.g. [15]). With enough data, this comes arbitrarily close to any PDF. There exist methods that use a kernel representation of the true distribution to solve also the inverse problem [16]. One drawback of this approach is that one has to store all the data points, another is that the known kernel based algorithms expect the response function of a set-up in analytical form, i.e. computer modeling cannot be used.

* Tel.: +354 4608505; fax: +354 4608998.
E-mail address: nikolai@unak.is
[1] Present address: Max-Planck-Institut für Kernphysik, P.O. Box 103980, 69029 Heidelberg, Germany.

http://dx.doi.org/10.1016/j.nima.2015.01.014
0168-9002/

In this paper the use of a Mixture Density Model (MDM) [17,19] to describe the true distribution $p(x)$ is proposed,

$$p(x) = \sum_{i=1}^{s} w_i K_i(x; a_{1i}, \ldots, a_{li}), \tag{3}$$

where the $K_i(x; a_{1i}, \ldots, a_{li})$ is the ith Probability Density Function in Mixture (PDFM) with parameters $a_{1i}, \ldots, a_{li}$ and the weight w_i the fraction of the ith PDFM.

The MDM lies between the cases of the parametric representation of the true density on one hand, i.e. the case when there is only one distribution in the sum (3), and the kernel statistics approach where the number of terms in the sum (3) is equal to the number of observations N. The MDM has a limited number of parameters for representing a PDF and computer modeling can be used to calculate the response of the system. The MDM is also convenient for taking into account different type of a priori information, such as knowledge about the type of distributions, constraints on parameters and smoothness of the distributions. Ideas and achievements of regression analysis as well as classical kernel statistics can be used in applications of a MDM for estimating the densities.

Using Eq. (3) to parameterize the solution $p(x)$ reduces the unfolding problem from finding a solution in the infinite-dimensional space of all functions to finding a solution in a finite dimensional space. This way an approximation of the true density is performed which, in contrast to e.g. a discretization by a histogram, has the advantage to introduce negligible quantization errors for sufficiently smooth distributions.

Without loss of generality two-parametric PDFMs will be used throughout the paper. The first parameter, x_i, defines the mean value (location) of term i and the second one, λ_i, represents the standard deviation. Different smooth PDFMs commonly employed by kernel statistics, such as biweight, triweight, tricube, cosine, Cauchy, B-spline and other kernels can be used. Rather popular is the Gaussian Mixture Model (GMM) [18] with PDFMs

$$K_i(x; x_i, \lambda_i) = \frac{1}{\lambda_i \sqrt{2\pi}} \exp\left(-\frac{(x-x_i)^2}{2\lambda_i^2} \right), \tag{4}$$

which provides a rather flexible model in the approximation of a wide class of statistical distributions. The standard deviation λ_i acts as a regularization parameter, which allows to adjust the smoothness of the result. Weights, positions x_i and standard deviations λ_i are determined by the unfolding procedure described below.

Substituting $p(x)$ as represented by Eq. (3) into the basic Eq. (1) yields

$$P(x') = \sum_{i=1}^{s} w_i \int_{\Omega} K_i(x; x_i, \lambda_i) A(x) R(x' \mid x)\, dx, \tag{5}$$

and taking statistical fluctuations into account, the relation between the weights w_i and the histogram of the observed distribution becomes a set of linear equations

$$\boldsymbol{P} = \mathbf{Q}\boldsymbol{w} + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{P}$ is the n-component column vector of the experimentally measured histogram, $\boldsymbol{w} = (w_1, w_2, \ldots, w_s)^t$ is the s-component vector of weights and $\mathbf{Q}$ is an $n \times s$ matrix with elements

$$Q_{ji} = \int_{c_{j-1}}^{c_j} K_i(x; x_i, \lambda_i) A(x) R(x' \mid x)\, dx \quad j = 1, \ldots, n; \quad i = 1, \ldots, s. \tag{7}$$

The vector $\boldsymbol{\epsilon}$ is an n-component vector of random deviates with expectation value $E[\boldsymbol{\epsilon}] = \mathbf{0}$ and covariance matrix $\mathbf{C}$, the diagonal elements of which being $\mathrm{Var}[\boldsymbol{\epsilon}] = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$, where σ_j is the statistical error of the measured distribution for the jth bin. Each column of the matrix $\mathbf{Q}$ is the response of the system to one of the PDFM in the mixture model for the true distribution. Numerically the calculation of the column vectors can be done by weighting events of a Monte Carlo sample such that they follow the corresponding PDFM, see Ref. [20], and taking the histogram of the observed distribution obtained with the weighted entries.

By a non-negative least-squares fit, the weight vector $\boldsymbol{w}$ in Eq. (6) for a given set of PDFMs is determined such that it minimizes

$$X^2 = (\boldsymbol{P} - \mathbf{Q}\hat{\boldsymbol{w}})^t \mathbf{C}^{-1} (\boldsymbol{P} - \mathbf{Q}\hat{\boldsymbol{w}}) \tag{8}$$

under the constraints

$$w_i \geq 0 \quad i = 1, \ldots, s. \tag{9}$$

Following Ref. [21], if an unconstrained solution satisfies Eq. (9) then $\hat{\boldsymbol{w}}$ solves the constrained problem. Otherwise, the solution to the constrained problem must be a boundary point of $[0, +\infty)^s$ and therefore at least one $w_i = 0$. It follows that after performing all possible regressions with one or more w_i in Eq. (9) set to zero, the non-negative problem is solved by picking the subset of w_i satisfying Eq. (9) such that X^2 as defined in Eq. (8) is smallest. The numerical algorithm and computer program for solving this minimization problem has been developed in references [21,22]. Here, first the subset of components equal to zero is determined iteratively, and the vector of the remaining indices $\hat{\boldsymbol{w}}$ is found by simple linear regression

$$\hat{\boldsymbol{w}} = (\mathcal{Q}^t \mathbf{C}^{-1} \mathcal{Q})^{-1} (\mathcal{Q}^t \mathbf{C}^{-1}) \boldsymbol{P}, \tag{10}$$

where $\mathcal{Q}$ is the submatrix of $\mathbf{Q}$ that corresponds to the subset of indices of positive components of the solution. The result of the fit is an estimate of the unfolded distribution $\hat{p}(x)$, defined by a subset of parameters x_i, λ_i, $i = 1, \ldots, k$ which are summed with positive weights $\hat{w}_i, i = 1, \ldots, k$ to yield

$$\hat{p}(x) = \sum_{i=1}^{k} \hat{w}_i K_i(x; x_i, \lambda_i). \tag{11}$$

The choices of the optimal type of PDFMs and the values of parameters (mean values and the standard deviations for the GMM model) are driven by the accuracy and the complexity of the model. The goal is a simple, and at the same time, accurate solution of the problem. A figure of merit for the accuracy is the Prediction Error (PE) [23], defined as the expectation value of the average squared normalized residual when using the predictor $\mathcal{Q}\hat{\boldsymbol{w}}$ to describe an independent experimentally measured histogram $\boldsymbol{P}^{new}$ drawn from the same parent distribution as the original,

$$PE(\mathcal{Q}\hat{\boldsymbol{w}}) = E\left[\frac{1}{n} (\boldsymbol{P}^{new} - \mathcal{Q}\hat{\boldsymbol{w}})^t \mathbf{C}^{-1} (\boldsymbol{P}^{new} - \mathcal{Q}\hat{\boldsymbol{w}}) \right]. \tag{12}$$

The expectation is taken over $\boldsymbol{P}^{new}$. In the following we will denote the predictor $\mathcal{Q}\hat{\boldsymbol{w}}$ as $\hat{\boldsymbol{P}}$ and call it the fitting histogram.

Following Ref. [23], V-fold Cross-Validation allows us to estimate $PE(\mathcal{Q}\hat{\boldsymbol{w}})$. Here the given data set $\mathcal{U}$ is split into V subsets $\mathcal{U}_1, \ldots, \mathcal{U}_V$ with equal number of events. The complementary sets are denoted by $\mathcal{U}^{(v)} = \mathcal{U} - \mathcal{U}_v$. Applying the minimization procedure to $\mathcal{U}^{(v)}$ and forming the predictors $\mathcal{Q}\hat{\boldsymbol{w}}^{(v)}$, the Cross-Validation error (CV) is defined by

$$CV = \frac{1}{n} \sum_{v=1}^{V} (\boldsymbol{P}_v - \mathcal{Q}\hat{\boldsymbol{w}}^{(v)})^t \mathbf{C}^{-1} (\boldsymbol{P}_v - \mathcal{Q}\hat{\boldsymbol{w}}^{(v)}), \tag{13}$$

where $\boldsymbol{P}_v$ is the vector of histogram contents for the subset of the data $\mathcal{U}_v$. The Cross-Validation error is the estimate of the Prediction Error

$$CV = \widehat{PE}(\mathcal{Q}\hat{\boldsymbol{w}}). \tag{14}$$

In order to have sufficient sampling of the configuration space, the number of folders used in the Cross-Validation procedure should