Nuclear Instruments and Methods in Physics Research A 771 (2015) 39–48

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

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

Interpolation between multi-dimensional histograms using a new non-linear moment morphing method

M. Baak [a,1], S. Gadatsch [b,2], R. Harrington [c], W. Verkerke [b]

[a] CERN, CH-1211 Geneva 23, Switzerland
[b] Nikhef, PO Box 41882, 1009 DB Amsterdam, The Netherlands
[c] School of Physics and Astronomy, University of Edinburgh, Mayfield Road, Edinburgh, EH9 3JZ, Scotland

ARTICLE INFO

Article history:
Received 24 June 2014
Received in revised form
2 October 2014
Accepted 17 October 2014
Available online 28 October 2014

Keywords:
Analysis
Distribution
Histogram
Interpolation
Simulation

ABSTRACT

A prescription is presented for the interpolation between multi-dimensional distribution templates based on one or multiple model parameters. The technique uses a linear combination of templates, each created using fixed values of the model's parameters and transformed according to a specific procedure, to model a non-linear dependency on model parameters and the dependency between them. By construction the technique scales well with the number of input templates used, which is a useful feature in modern day particle physics, where a large number of templates are often required to model the impact of systematic uncertainties.

1. Introduction

In particle physics experiments, data analyses generally use shapes of kinematical distributions of reconstructed particles to interpret the observed data. These shapes are simulated using Standard Model or other theoretical predictions, and are determined separately for signal and background processes. Simulations of known fundamental physics processes are run through a detailed detector simulation, and are subsequently reconstructed with the same algorithms as the observed data. These simulated samples may depend on one or multiple model parameters, for example the simulated Higgs particle mass, and a set of such samples may be required to scan over the various parameter values. Since Monte Carlo simulation can be time-consuming, there is often a need to interpolate between the limited number of available Monte Carlo simulation templates.

In particular, the statistical tests widely used in particle physics, *e.g.* for the construction of confidence intervals on model parameters or the discovery of new phenomena, rely strongly on continuous and smooth parametric models that describe the signal and background processes in the data.[3] These parametric models describe parameters of interest, such as a shifting mass parameter or the rate of a signal process, and the so-called nuisance parameters that parametrize the impact of systematic uncertainties. As such, the models are often constructed in terms of those parameters by interpolating between simulated Monte Carlo templates, thereby ensuring continuity in those parameters.

Several algorithms exist that can be used to interpolate between Monte Carlo sample distributions [1,2]. Interpolation techniques have been used on multiple occasions in particle physics, for example to predict kinematic distributions for intermediate values of a model parameter, *e.g.* the simulated Higgs boson, *W* boson or top quark mass, or to describe the impact of systematic uncertainties, which are often modeled as shape or rate variations about a nominal template of a kinematic distribution.

This work describes a new morphing technique, *moment morphing*, which has the advantage over existing methods in that it is fast, numerically stable, allows for both binned histogram and continuous templates, has proper vertical as well as horizontal morphing (explained in Section 2), and is not restricted in the

E-mail addresses: max.baak@cern.ch (M. Baak), stefan.gadatsch@nikhef.nl (S. Gadatsch).
[1] Tel.: +41 22 7671255; fax: +41 22 7678350.
[2] Tel.: +31 20 5925140; fax: +31 20 592 5155.

[3] Typically, a statistical test involves the maximization of a likelihood function, which has been built from both the parametric model and the observed data. The maximization procedure relies on the derivatives of the likelihood with respect to the model's parameters.

http://dx.doi.org/10.1016/j.nima.2014.10.033
0168-9002/

number of input templates, the number of model parameters or the number of input observables. In particular, the latter feature allows the moment morphing technique to model the impact of a non-factorizable response between different model parameters, where varying one model parameter at a time is insufficient to capture the full response function.

The paper is organized as follows: Section 2 describes in detail how the moment morphing function, used to interpolate between histograms, is constructed using one or more morphing parameters. Section 3 describes how the moment morphing technique can be used to properly take into account systematic uncertainties in a high energy physics analysis, giving an example of a typical application. A comparison in terms of accuracy of moment morphing with alternative morphing algorithms is provided in Section 4. Section 5 describes the implementation of the moment morphing algorithm in publicly available C++ code, including benchmarking of its performance.

2. Construction of the morphing p.d.f.

This section details the construction of the moment morphing probability density function (p.d.f.). The method proposed here is based on the linear combination of input templates. The dependency on the morphing parameter(s) can be non-linear, and is captured in multiplicative coefficients and a transformation of the template observables. Interpolation using a single morphing parameter is described in Section 2.1. Section 2.2 describes interpolation using multiple morphing parameters and shows that dependencies between morphing parameters can be readily modeled. Other choices of basis functions for the construction of the morphing p.d.f. are considered in Section 2.3.

2.1. Interpolation with a single morphing parameter

Consider an arbitrary p.d.f. $f(\mathbf{x}|m)$, where f depends on the single morphing parameter m and describes the observables $\mathbf{x}$. The true dependency on m is not known or difficult to obtain. Instead, the p.d.f. f has been sampled at n different values of m, with each $f(\mathbf{x}|m_i)$ representing a known input template shape for a single value of the morphing parameter, labeled as m_i. In the following the goal is to construct a parametric approximation of $f(\mathbf{x}|m)$ for arbitrary m, which is continuous and smooth in the model parameter, as required for example by the statistical tests used in particle physics alluded to in Section 1. There are two steps to this.

First, given the sampling points, $f(\mathbf{x}|m)$ can be expanded in a Taylor series up to order $n-1$ around reference value m_0:

$$f(\mathbf{x}|m) \approx \sum_{j=0}^{n-1} \frac{\mathrm{d}^{(j)} f(\mathbf{x}|m_0)}{\mathrm{d}m^{(j)}} \frac{(m-m_0)^j}{j!} = \sum_{j=0}^{n-1} f'_j(\mathbf{x}|m_0)(m-m_0)^j \tag{1}$$

where the second equality defines $f'(\mathbf{x}|m)$. For the n given values of m follows the vector equation:

$$f(\mathbf{x}|m_i) \approx \sum_{j=0}^{n-1} (m_i-m_0)^j f'_j(\mathbf{x}|m_0) = \sum_{j=0}^{n-1} M_{ij} f'_j(\mathbf{x}|m_0) \tag{2}$$

where $M_{ij} = (m_i - m_0)^j$ defines a $n \times n$ transformation matrix. Inverting Eq. (2) gives

$$f'_j(\mathbf{x}|m_0) = \sum_{i=0}^{n-1} (M^{-1})_{ji} f(\mathbf{x}|m_i) \tag{3}$$

which allows us to determine the n values $f'_j(\mathbf{x}|m_0)$. Substituting this in Eq. (1), $f(\mathbf{x}|m)$ reads

$$f(\mathbf{x}|m) \approx \sum_{i,j=0}^{n-1} (m-m_0)^j (M^{-1})_{ji} f(\mathbf{x}|m_i) \tag{4}$$

which can be used to predict the template shape at any new value of the morphing parameter given by m':

$$f_{\mathrm{pred}}(\mathbf{x}|m') = \sum_{i=0}^{n-1} c_i(m') f(\mathbf{x}|m_i) \tag{5}$$

which is a linear combination of the input templates $f(\mathbf{x}|m_i)$, each multiplied by a coefficient

$$c_i(m') = \sum_{j=0}^{n-1} (m'-m_0)^j (M^{-1})_{ji} \tag{6}$$

which themselves are non-linear and depend only on the distance to the reference points. This approach of weighting the input templates is also known as vertical morphing. Note that the coefficients c_i are independent of the derivatives of $\mathbf{f}$ with respect to morphing parameters or to the observable set $\mathbf{x}$, making their computation easy.

The coefficient for a point included in the set of input templates is one, *i.e.*

$$c_i(m_j) = \delta_{ij} \tag{7}$$

and by construction the sum of all coefficients c_i equals one:

$$\sum_i c_i(m) = 1. \tag{8}$$

This turns out to be a useful normalization, as will be seen below.

To illustrate, one can consider a morphed p.d.f. using only input templates at two values of the morphing parameter, $m_{\min}$ and $m_{\max}$. The coefficients $c_i(m)$ become linear in m and reduce to the simple fractions:

$$c_{i_{\min}} = 1 - m_{\mathrm{frac}} \tag{9}$$

$$c_{i_{\max}} = m_{\mathrm{frac}} \tag{10}$$

where $m_{\mathrm{frac}} = (m - m_{\min})/(m_{\max} - m_{\min})$, $c_{i_{\min}}$ and $c_{i_{\max}}$ sum up to one, and all other coefficients are zero.

Second, it may be that the sampled input p.d.f.s f_i describe distributions in $\mathbf{x}$ that vary strongly as a function of m in shape and location. This is equivalent to the first and second moments (*i.e.* the means and variances) of the input distributions having a dependence on the morphing parameter m.

Since the input p.d.f.s in Eq. (5) are summed linearly, it is imperative to translate all input distributions $f_i(\mathbf{x})$ in the sum before combining in the morphed p.d.f. such that their locations match up. The process of translating the input observables (but not scaling; see below) is also called the horizontal morphing. In addition it is necessary to take into account the change in the width of the input distributions as a function of the morphing parameter.

To achieve this, the mean μ_{ij} and width σ_{ij} of each input distribution i and observable x_j are shifted to the common values of $\mu'_j(m)$ and $\sigma'_j(m)$. These are obtained by multiplying the underlying means and widths with the coefficients $c_i(m)$ of Eq. (6):

$$\mu'_j(m) = \sum_i c_i(m) \cdot \mu_{ij} \tag{11}$$

$$\sigma'_j(m) = \sum_i c_i(m) \cdot \sigma_{ij} \tag{12}$$

In order to shift the input p.d.f.s a linear transformation of each observable is applied. For each p.d.f. i and observable j define

$$x'_{ij} = a_{ij} x_j + b_{ij}, \tag{13}$$

with slope

$$a_{ij} = \frac{\sigma_{ij}}{\sigma'_j} \tag{14}$$

and offset

$$b_{ij} = \mu_{ij} - \mu'_j a_{ij}. \tag{15}$$