Nuclear Instruments and Methods in Physics Research A 780 (2015) 81–90

Contents lists available at ScienceDirect

Nuclear Instruments and Methods in Physics Research A

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

Handling missing data for the identification of charged particles in a multilayer detector: A comparison between different imputation methods

S. Riggi [a,*], D. Riggi [b], F. Riggi [c,d]

[a] INAF - Osservatorio Astrofisico di Catania, Italy
[b] Keras Strategy - Milano, Italy
[c] Dipartimento di Fisica e Astronomia - Università di Catania, Italy
[d] INFN, Sezione di Catania, Italy

ARTICLE INFO

Article history:
Received 1 September 2014
Received in revised form
28 December 2014
Accepted 17 January 2015
Available online 28 January 2015

Keywords:
Particle identification
Neural networks
Missing data imputation
Skew-normal mixture
EM algorithm

ABSTRACT

Identification of charged particles in a multilayer detector by the energy loss technique may also be achieved by the use of a neural network. The performance of the network becomes worse when a large fraction of information is missing, for instance due to detector inefficiencies. Algorithms which provide a way to impute missing information have been developed over the past years. Among the various approaches, we focused on normal mixtures' models in comparison with standard mean imputation and multiple imputation methods. Further, to account for the intrinsic asymmetry of the energy loss data, we considered skew-normal mixture models and provided a closed form implementation in the Expectation-Maximization (EM) algorithm framework to handle missing patterns. The method has been applied to a test case where the energy losses of pions, kaons and protons in a six-layers' Silicon detector are considered as input neurons to a neural network. Results are given in terms of reconstruction efficiency and purity of the various species in different momentum bins.

1. Introduction

The treatment of missing data in different areas of science, statistics, economics, or pattern recognition has been and still is a wide subject of interest, since a variety of situations may lead to incomplete data in a set of information. In particle and nuclear physics there are several situations where a certain fraction of data may be missing. Typical cases are the set of space points which contribute to the tracking of charged particles, or the different values of the energy loss of particles in a multilayer detector. There are several reasons why data may be missing. In the simplest case, they may be missing completely at random (MCAR), e.g. the probability that a data is missing does not depend on the value of the variable. In most cases however such probability either depends on the other variables in the data set or on the value of the variable under consideration. In the former situation the data are said to be missing at random (MAR), while in the latter situation the missing data mechanism is denoted as non-ignorable or missing not at random (MNAR). Examples of the MCAR case are the passage of minimum ionizing charged particles through a multilayer detector, in the limit of very small thickness (thus limiting the amount of energy loss and multiple scattering in each layer), with detection efficiency smaller than 100%, due to dead areas or other inefficiencies. Examples of the MNAR case are given by a detector which has a finite energy threshold, modelled for instance as a sigmoid function.

The simplest approach in the case of a certain fraction of missing information from a set of variables is to disregard the event where at least one variable is missing. Such an approach, although retaining only complete events, may lead to a substantial loss of events either when the elementary fraction of missing events is large or when the number of variables is large. As an example, in a process with $d=100$ variables (such as the number of space points in a large tracking detector), even an elementary fraction $\eta=0.1\%$ in each variable leads to a net loss of $1-(1-\eta)^d = 1-(0.999)^{100} \sim 10\%$. Fig. 1 shows the contour lines corresponding to different overall fractions of missing events (10%, 20%, 30% and 40%), as a function of the dimensionality d of the problem and of the elementary fraction of missing events η in each variable. The importance of using all collected events is of special concern in the case of rare events, where disregarding the event due to its incompleteness may lead to a substantial fraction of potentially interesting events being lost. This is the reason why different methods have been employed to impute the missing values of a variable according to the statistical properties of the variables of interest

* Corresponding author.
E-mail address: sriggi@oact.inaf.it (S. Riggi).

http://dx.doi.org/10.1016/j.nima.2015.01.063
0168-9002/

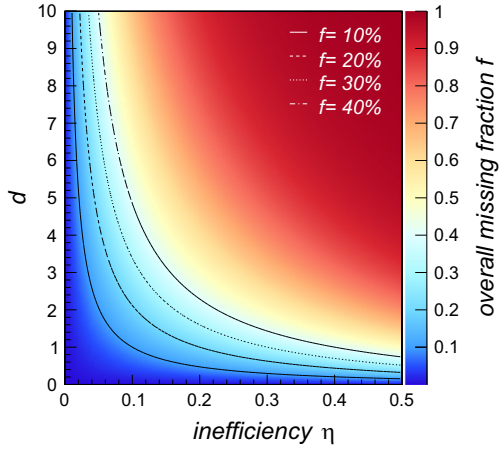

Fig. 1. Overall missing fraction f as a function of the variable detection inefficiency η and the number of observables d in the analysis. The contour lines are relative to $f=10\%$, 20%, 30%, 40%.

which define the event [1–3]. Most of such methods rely on normal distributions of the variables, which however is not a good representation in many problems. The energy loss of a charged particle in a thin detector is a typical example of a variable whose distribution has an asymmetric shape (Landau tail). For such reason, it is of interest to develop and test methods which are not biased by the assumption on normal distributions.

In this paper an approach based on multivariate skew-normal (MSN) distribution was developed. The method was then applied to a test case, where simulated energy losses of pions, kaons and protons in a multilayer Silicon detector were considered. The performance of a neural network was then evaluated under missing data and after recovering them with the approach described here. Section 2 describes in some more detail the problem of missing data reconstruction and the methods adopted throughout the paper, namely Multiple Imputation (MI) and Maximum Likelihood (ML) methods, while the application to the test case is discussed in Section 3. Various methods of imputation of the missing data were then considered and applied to the same set of simulated data, comparing their performance in terms of identification efficiency. The results of such comparison are reported in Section 4. Some details of the algorithms being employed are described in Appendix A.

2. Missing data reconstruction

The most common strategy of dealing with missing data is list-wise deletion (LD), that consists in rejecting all events with missing observables. Such an approach decreases the effective sample size, especially with a large number of observables in analysis, and correspondingly the power of any statistical tests to be performed with that sample, which in the presence of rare events is not desirable. If the data are not missing at random, such a procedure may also lead to a selection bias. In cases where the missing data mechanism is MCAR or MAR list-wise deletion does not add any bias and if the event sample size is not a critical issue it should be the preferred approach.

The alternative approach than removing events with missing data is to fill in or impute the missing values, with the aim of recovering the relevant features of the data set as a whole (mean, variance or any additional parameter), rather than obtaining precise estimates of single missing values. Indeed, imputed values should not be trusted nor directly used to draw inferences.

The simplest way is to replace each missing value with the mean of the observed values for that variable. This strategy significantly alters the distribution for that observable and all derived summary indicators, for example the variance which is typically underestimated. If more than one group is present in the data sample, the mean estimation tends to be pulled towards the most abundant group. For that reason such a method is not recommended.

A wide number of modern imputation methods exist in the literature, also used in the context of neural network analysis, among them regression-based single imputation, multiple imputation (MI), maximum likelihood (ML) methods, methods based on K-Nearest Neighbor (KNN) or K-means clustering, Singular Value Decomposition (SVD), Support Vector Machines (SVM). A review of some of these methods is available in [1–4].

Neural networks themselves could be also used to impute missing data [4,5]. This is typically achieved by designing a set of neural network classifiers, each specialized to learn given missing values from observed data. However when missing data are present in many variables, the method requires to design a large number of networks, one per each combination of missing attributes.

In this work we will adopt maximum likelihood methods based on multivariate normal (MN) mixtures as compared to the popular multiple imputation approach. Both methods assume a normal model, which is not appropriate to the energy deposit data under consideration. For that reason we designed in this work a ML method based on multivariate skew-normal (MSN) mixture models, presented in Section 2.2.2.

2.1. Multiple imputation

Single imputation methods, for example those based on linear regression, are intrinsically limited. They proceed by calculating the regression of the incomplete variable on the other complete variables and substituting the predicted mean for each missing observation. Since imputed values always lie on the regression line, the actual dispersion of the data is ignored and therefore the variance is underestimated, leading to a bias in the parameter estimates.

The multiple imputation approaches proceed instead by introducing a random variation in the process, i.e. a random normal error into the regression equation, and generating several data sets, each with different imputed values, to partially restore the lost variance. To account for the fact that only a single draw from the data population is taken, multiple random draws from the posterior distribution of the population, each imputed several times, are also introduced to completely restore the variance of the data. Multiple imputation algorithms available in the literature typically differs for how this latest step is performed. Some uses bootstrap procedures to generate random draws, others adopt Data Augmentation (DA) techniques or the Markov chain Monte Carlo (MCMC) algorithm and so forth. If the MAR assumption holds the method that leads to almost unbiased estimators.

2.2. Maximum likelihood imputation

In the likelihood approach a model is assumed to describe the observed data. If p variables are considered for the analysis, we denote a sample of N data observations with $\mathbf{x}=(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, being $\mathbf{x}_i$ a p-dimensional data vector. For a given observation i we may have missing patterns. We indicate with $\mathbf{x}_{i,o}$ being the observed patterns and with $\mathbf{x}_{i,m}$ being the missing patterns. A widely used approach consists in assuming a mixture model $F(\mathbf{x};\boldsymbol{\Theta})$ with K components with probability density functions pdf $f(\mathbf{x};\boldsymbol{\theta}_k)$ and parameters $\boldsymbol{\theta}_k$:

$$F(\mathbf{x};\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{x};\boldsymbol{\theta}_k) \quad (1)$$