SLAC-PUB-1929
April 1977
(E)

Reconstruction of Missing Data in Principal Component Analysis*

Paul E. Condon

Physics Department
University of California Irvine
Irvine, California 92664

and

Stanford Linear Accelerator Center
Stanford University
Stanford, California 94305

Abstract

A simple technique is described for reconstructing missing data for use in principal component analysis. The technique is described in the context of track finding and track reconstruction in wire chamber spectrometers, but should have a much wider range or applicability.

Submitted for Publication

* Work supported by Energy Research and Development Administration

Reconstruction of Missing Data in Principal Component Analysis

The use of principal componenet analysis in track finding in wire chambers has been suggested by H. Wind[1][2] in several papers. It is commonly thought that the method fails or at least becomes quite intractible if one or more of the wire chambers is inefficient and the corresponding data missing. This note is written to point out an efficient technique for constructing the missing data and handling the consequent reduction in the redundancy in the data. We first review the basics of principal component analysis, in order to establish our notation.

In a wire spark chamber or proportional wire chamber spectrometer each plane or layer of wires measures a coordinate on a particle track. For a single particle passing through a spectrometer with M layers, M coordinates are measured. The M-tuple of coordinate values may be treated as a vector or a point in an M-dimensional space. Since we expect to have many tracks in an experiment we label tracks with a subscript α

$$\vec{X}_\alpha = (X_{\alpha 1}, X_{\alpha 2}, X_{\alpha 3}, \ldots, X_{\alpha M}) \qquad \alpha = 1, 2, 3, \ldots, N$$

The points $\vec{X}_\alpha$ should cluster somehow in the M-dimensional space of all possible coordinate values. In fact, we know that a real track is fully describable by five independent parameters (e.g. two transverse positions, two direction angles, and reciprocal momentum). As a consequence, all the points corresponding to real tracks must lie in a 5-dimension sub-space of our original M-dimension data space. Wind has suggested that for a wide variety of experimental situations, this subspace may be very close to a linear subspace or hyperplane of 5 dimensions. If one forms the dispersion matrix, $\vec{\vec{A}}$, such that

$$A_{ij} = \frac{1}{N}\sum_{\alpha} X_{\alpha i} X_{\alpha j} - \left(\frac{1}{N}\sum_{\alpha} X_{\alpha i}\right)\left(\frac{1}{N}\sum_{\alpha} X_{\alpha j}\right)$$

and solves the eigen value equation.

$$\vec{\vec{A}}\,\vec{W} = \lambda\,\vec{W}$$

The resulting eigen vectors, $\vec{W}$, have some very useful properties which are set forth in the work of Wind. For a system with M dimension there are in general M eigen values and eigen vectors. The matrix, $\vec{\vec{W}}$, formed from the full set of eigen vectors is an orthonormal matrix which rotates the track coordinates $\vec{X}_{\alpha}$ so that the 5 dimensional hyper plane is aligned with the coordinate axis. We arrange the eigenvectors in a M by M matrix in $\vec{\vec{W}}$ in order of increasing eigen value. In the resulting rotated system, the first M-5 components of a real track have values that, after a simple translation, are essentially zero.* The five remaining non-zero components fully characterize the track. The rotation and translation which accomplish the reduction in the number of non-zero components can be pre-computed using a Monte Carlo model of the apparatus or using data collected in special runs for which the operating conditions are such as to assure only one track per event.

But if the spark chambers are at all inefficient, there will be tracks for which one or more of the original coordinates are missing and it is then not possible to perform the rotation. It is possible to reconsider the whole problem with one layer missing i.e. M-1 dimension and repeat the whole analysis M times, once for each of the layers missing and introduce M additional rotation matrices of degree M-1. And to handle the possibility of two coordinates being missing simultaneously one would introduce $M(M-1)/2$ new rotation matrices each of degree M-2, and so forth until one has covered all possibilities of missing data. For a spectrometer with 16 layers and allowing up to

* This translation, to the center of gravity, we will denote by $\vec{Y}_c$

4 missing coordinates among the 16, one must precompute and store hundreds of thousands coefficients. This seems to me an unrealistic approach to the problem.

Instead we notice that we can use the inherent redundancy in the data to construct a best estimate of the value of the missing coordinate (or coordinates) and then apply the M-dimension rotation matrix to the full set of M-"measurements". (In quotes because some are real and some are constructed). We motivate this construction as follows:

Define δ to be the distance from a specific point $\vec{X}$ to the hyper plane fitted through all realistic track points

Let

$$\vec{Y} = \vec{\vec{W}}\,\vec{X} - \vec{Y}_c$$

so that $$\delta^2 = \sum_{i=1}^{M-5} (Y_i)^2$$

For real tracks $\delta^2 \approx 0$. For tracks with missing components we adjust the value of the unknown component(s) to minimize δ^2

$$0 = \frac{\partial}{\partial X_j}(\delta^2) = 2\sum_{i=1}^{M-5}(Y_i)\frac{\partial}{\partial X_j}Y_i = 2\sum_{i=1}^{M-5} Y_i\ W_{ij}$$

$$0 = \sum_{i=1}^{M-5} W_{ij}\sum_{k=1}^{M} W_{ik}X_k = \sum_{k=1}^{M} X_k \sum_{i=1}^{M-5} W_{ij}\ W_{ik}$$

Define $C_{jk} = \sum_{i=1}^{M-5} W_{ij}\ W_{ik}$ and the equation becomes

$$0 = \sum_{k=1}^{M} X_k\ C_{jk}$$

There is one such linear equation for each missing component. The M x M matrix C_{jk} is constructed from the matrix $\vec{\vec{W}}$ once per experiment and because

it is symmetric, it requires a trivial amount of storage. The coefficients in any set of linear equations corresponding to any combination of missing data are obtained by indexing into this small matrix. For K missing components one must solve an K-fold set of simultaneous linear equations using standard techniques.

Since K, the number of missing components, is supposedly a small number, we thus invert a K by K matrix for each event with K missing components. The elements of the matrix to be inverted are $C_{j_1 j_2}$ where j_1 and j_2 label the missing components.

One can, in principal, anticipate all possible combinations of missing data and invert the corresponding matrix apriori. However, because inverting this small matrix is a relatively small computation, it can be done once for each event in which there is missing data.

If the errors in the individual components are Gaussian distributed with unit variance, the variable δ^2 will be χ^2 distributed. With no missing data there are M-5 degrees of freedom for this χ^2 distribution. With K constructed components there are M-5-K degress of freedom. In fact, the errors in individual components are probably not Gaussian distributed so the above statement is not rigorously true for real data. On the other hand it is probably not so wrong as to be misleading for real present day spectrometers.

I wish to thank Dr. J. Friedman of the Stanford Linear Accelerator Center Computation Group and Dr. H. Wind of CERN for helpful and illuminating discussions.

1) H. Wind, The Use of Tabulated Trajectories, CERN Report NP 72-8.

2) H. Wind, Function Parametrization, 1972 CERN Computing and Data Processing School (CERN 72-21).