View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by UEL Research Repository at University of East London

roar @UEL
research open access repository

University of East London Institutional Repository: http://roar.uel.ac.uk

This paper is made available online in accordance with publisher policies. Please scroll down to view the document itself. Please refer to the repository record for this item and our policy information available from the repository home page for further information.

To see the final version of this paper please visit the publisher's website. Access to the published version may require a subscription.

Author(s): Dybowski, Richard., Weller, Peter.
Article Title: Prediction regions for the visualization of incomplete datasets
Year of publication: 2001
Citation: Dybowski R, Weller P. (2001) "Prediction regions for the visualization of incomplete datasets". Computational Statistics 16 (1), 25-41
Link to published version: (direct link not available: http://www.springer.com/statistics/journal/180)
DOI: (not stated)
ISSN: 0943-4062 (Print) 1613-9658 (Online)

Publisher statement:
http://www.springer.com/open+access?SGWID=0-169302-12-467999-0

Information on how to cite items within roar@uel:
http://www.uel.ac.uk/roar/openaccess.htm#Citing

Prediction Regions for the Visualization of Incomplete Datasets

Richard Dybowski[1] and Peter R. Weller[2]

[1] Intensive Care Group (Division of Medicine), King's College London (St Thomas Hospital Campus), Lambeth Palace Road, London SE1 7EH, U.K.
[2] Centre for Measurement and Information in Medicine, City University, Northampton Square, London, EC1V 0HB, U.K.

Summary

A complication in the visualization of biomedical datasets is that they are often incomplete. A response to this is to multiply impute each missing datum prior to visualization in order to convey the uncertainty of the imputations. In our approach, the initially complete cases in a real-valued dataset are represented as points in a principal components plot and, for each initially incomplete case in the dataset, we use an associated prediction region or interval displayed on the same plot to indicate the probable location of the case. When a case has only one missing datum, a prediction interval is used in place of a region. The prediction region or interval associated with an incomplete case is determined from the dispersion of the multiple imputations of the case mapped onto the plot. We illustrate this approach with two incomplete datasets: the first is based on two multivariate normal distributions; the second on a published, simulated health survey.

Keywords: Visualization, Multiple imputation, Prediction regions, MANET, XGobi

1 Introduction

Many techniques are available for approximating the distribution of multivariate data by a distribution in two-dimensional space. Principal components analysis (PCA) (Hotteling 1933) is the most commonly used visualization technique for real-valued samples, whereas multiple correspondence analysis (e.g., Gower & Hand 1996) is the most popular method when variables are categorical. Other techniques in use include canonical variates analysis (Seal 1964), which attempts to maximize between-class differences relative to within-class differences, and Sammon mapping (Sammon 1969), which attempts to preserve the Euclidean distances between the cases present in feature space. A common requirement of these visualization methods is that the dataset to be visualized must have no missing values, but biomedical databases are prone to a variety of errors (Heitjan 1993, Albert & Horwitz 1995), including absent values.

A response to the problem of missing data is to replace each missing datum with an estimated value (*imputation*). Various imputation methods have been used. These include (a) the use of unconditional means; (b) the use of domain heuristics (cold-deck imputation), such as the assumption of clinical normality (Knaus, Zimmerman, Wagner, Draper & Lawrence 1981); (c) the use of those attribute values contained in the nearest complete case (nearest-neighbour hot-deck imputation) (e.g., Little & Rubin 1987); and (d) the use of conditional expectations based on population parameters estimated by the expectation-maximization (EM) algorithm (Dempster, Laird & Rubin 1977). Once an initially incomplete dataset has been imputed, one can proceed by using a standard visualization method, such as PCA, to view the data. But a problem with replacing a missing datum with a single value is that this process does not indicate how close an imputed value is to the true value. In other words, one is not informed about the associated uncertainty in the estimated value. *Multiple imputation* (Rubin 1987) is a means of conveying this uncertainty. In this approach, a number of possible imputations are assigned to each missing datum instead of just a single value, and the spread of these possible values indicates the extent of the uncertainty.

Multiple imputation is facilitated by several software packages, for example, Shafer's suite of `S-PLUS` functions (Schafer 1998) and `SOLAS` (Statistical Solutions 1998). But how do we visualize a multiply imputed dataset by, say, PCA? If we display all the complete and imputed cases together, the resulting scatterplot could easily be swamped by a mass of multiple imputations. On the other hand, if the set of imputations associated with an incomplete case is replaced by their centroid, we are then faced with the problem of displaying the uncertainty of the imputations on the principal components plot. The answer we propose in this paper is to replace each set of imputations with a prediction region (or interval) centred on the centroid of the set. The prediction region associated with an incomplete case is calculated from the dispersion of the imputations of the case, and it defines an area of the scatterplot within which there is a given probability that the image of the case under the PCA will be present.

The `XGobi` and `MANET` software packages have been developed to provide a graphical examination of incomplete datasets. `XGobi` (Swayne & Buja 1998) uses scatterplots to visualise pairs of selected variables from a multidimensional dataset. Two linked windows are used to present the main data and the missing value information. A linked brushing method is employed to present missing information as a shadow dataset. The main window is used to present the data with missing values replaced by some fixed or imputed values. The shadow window displays four square clusters of binary indicators corresponding to the presence and "missingness" of the data. The pairs of variables of interest are selected from a menu system on the side of the display. The comparison between several schemes can be accomplished by using more windows.

`MANET` (Unwin, Hawkins, Hofmann & Siegl 1996) was developed to implement interactive graphics tools for data sets with missing values and to provide a platform for investigating new interactive ideas. The package has a large number of methods for presenting data, each of which has been adapted to include missing data. A missing value chart gives an indication of the proportion of data present for each variable. Missing values can be represented in histograms and barcharts by an additional column, the size of which corresponds to the amount of data missing. Data can be also visualized in scatterplots with missing data represented by the classical method of plotting these points on the appropriate axis. Three additional boxes are included below the plot for data where each or both axis information is missing, again the size of each box gives an indication of the proportion of data missing. The package offers an extensive set of visualization tools with a consistent method of indicating the proportion of missing data. However, as with `XGobi`, there is no feature for imputing estimates for these values within the package or for identifying them in the main plots. In addition, neither `XGobi` nor `MANET` provides prediction regions to graphically indicate the possible imputations for the incomplete cases, and it is this omission that we address in this paper.

In the next section we give a brief description of Markov chain Monte Carlo for multiple imputation. We then consider visualization of multiply imputed datasets by using a principal components plot to display the distribution of the complete cases, and a representation of the uncertainty of the incomplete cases on the same plot by prediction regions and intervals.

2 Multiple Imputation

Let a dataset $\mathbf{X}$ associated with class C consist of observed values $\mathbf{X}_{obs}$ and unobserved (missing) values $\mathbf{X}_{mis}$. Let $\boldsymbol{\theta}$ be the set of parameters defining the population from which $\mathbf{X}$ has been sampled, for example, if the population is multivariate normal then $\boldsymbol{\theta}$ consists of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Since $\boldsymbol{\theta}$ is expected to be different for different classes, the remainder of this section applies to each subset $\mathbf{X}$ resulting from a class-wise partition of a multi-class dataset.

A maximum-likelihood approach to multiple imputation is to first estimate the population parameters $\boldsymbol{\theta}$ from $\mathbf{X}_{obs}$ via the EM algorithm (giving $\boldsymbol{\theta}_{MLE}$), and then obtain M independent random samples $\mathbf{X}_{mis}^{<1>}, \ldots, \mathbf{X}_{mis}^{<M>}$ from the probability distribution $p(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}_{MLE})$. However, these imputations are obtained under the assumption that $\boldsymbol{\theta}_{MLE}$ are the true population parameters.

The Bayesian approach to multiple imputation overcomes the uncertainty in $\boldsymbol{\theta}$ by integrating over all possible $\boldsymbol{\theta}$:

$$\begin{aligned} p(\mathbf{X}_{mis}|\mathbf{X}_{obs}) &= \int p(\mathbf{X}_{mis}, \boldsymbol{\theta}|\mathbf{X}_{obs}) d\boldsymbol{\theta} \\ &= \int p(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{X}_{obs}) d\boldsymbol{\theta} \qquad (1) \end{aligned}$$

Imputations $\mathbf{X}_{mis}^{<1>}, \ldots, \mathbf{X}_{mis}^{<M>}$ are then sampled independently and randomly from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$. From (1), the sampling process is equivalent to randomly selecting $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{X}_{obs})$ and then selecting $\mathbf{X}_{mis}$ from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})$, a new selection from $p(\boldsymbol{\theta}|\mathbf{X}_{obs})$ being made before each single selection from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta})$.

Pseudo-random draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{X}_{obs})$ can be simulated by *Markov chain Monte Carlo* (MCMC). In the context of $p(\boldsymbol{\theta}|\mathbf{X}_{obs})$, MCMC generates a Markov chain $\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(t)}$ such that $\boldsymbol{\theta}^{(t)}$ has the distribution $p(\boldsymbol{\theta}|\mathbf{X}_{obs})$ as $t \to \infty$. A number of MCMC techniques are in existence (Roberts 1996), and the data-augmentation MCMC method for the multiple imputation of dataset $\mathbf{X}$ is as follows:

Algorithm 1. (*Data augmentation* (Tanner & Wong 1987))

```
begin
  $\boldsymbol{\theta}^{(0,1)} \leftarrow \boldsymbol{\theta}_{MLE}|\mathbf{X}_{obs}$;
  for $m = 1$ to $M$ {i.e., M Markov chains} do
      for $t = 0$ to $T - 1$ {i.e., T iterations within each Markov chain} do
          randomly select $\mathbf{X}_{mis}^{(t+1,m)}$ from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}^{(t,m)})$;
          randomly select $\boldsymbol{\theta}^{(t+1,m)}$ from $p(\boldsymbol{\theta}|\mathbf{X}_{obs}, \mathbf{X}_{mis}^{(t+1,m)})$;
      endfor
      randomly select $\mathbf{X}_{mis}^{<m>}$ from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs}, \boldsymbol{\theta}^{(T,m)})$;
  endfor
end
```

By separating successive $\mathbf{X}_{mis}^{<m>}$ in Algorithm 1 with sufficiently long Markov chains (e.g. $T = 50$), imputations $\mathbf{X}_{mis}^{<1>}, \ldots, \mathbf{X}_{mis}^{<M>}$ can be regarded as independent draws from $p(\mathbf{X}_{mis}|\mathbf{X}_{obs})$.

There are two points to note about Algorithm 1. First, it assumes that data are missing at random (as defined by Rubin (1976)). If any missing-data mechanisms are known, a sensitivity analysis of the effect of the mechanisms on the prediction regions should be conducted. Rubin (1987) describes some imputation techniques to use when the mechanism is not ignorable. Second, the following visualization procedure is not restricted to the use of Algorithm 1: other types of multiple imputation could be used. This algorithm is used simply to provide an example of how multiple imputation could be achieved.

3 Visualization

The imputations $\mathbf{X}_{mis}^{<1>}, \ldots, \mathbf{X}_{mis}^{<M>}$ resulting from Algorithm 1 give rise to a collection of completed datasets $\mathbf{X}_{obs} \cup \mathbf{X}_{mis}^{<1>}, \ldots, \mathbf{X}_{obs} \cup \mathbf{X}_{mis}^{<M>}$. In this section we will describe a method for visualizing these datasets when the attributes are real-valued.

3.1 Principal components plots

Given a sample of d-dimensional feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$, *principal components analysis* determines the eigenvalues $\lambda_1, \ldots, \lambda_d$ and the associated eigenvectors $\mathbf{a}_1, \ldots, \mathbf{a}_d$ that satisfy the equation

$$\mathbf{S}\mathbf{a}_i = \lambda_i \mathbf{a}_i,$$

where $\mathbf{S}$ is the sample covariance matrix, and the eigenvectors are of unit length and orthogonal. The *j-th principal component* is the linear combination $\mathbf{a}_j^{\mathrm{T}}\mathbf{x}$ for which λ_j (the sample variance of $\mathbf{a}_j^{\mathrm{T}}\mathbf{x}$) is the j-th largest eigenvalue. In this paper, the scatter plot of the first and second principal components of a dataset will be referred to as the *principal components* (*PC*) *plot* of the dataset.

The M imputations of the incomplete dataset have produced M possible imputations for each of the incomplete cases. The aim is to show, on a PC plot, how the imputations obtained for each incomplete case are distributed about their centroid in feature space. This will be done by performing a single PCA on both the complete cases and on the centroids of the incomplete cases. The imputed cases are then mapped onto the PC plot obtained from the PCA.

Let $\mathbf{X}^{<m>}$ denote the $d \times N$ imputed dataset $\mathbf{X}_{obs} \cup \mathbf{X}_{mis}^{<m>}$ composed of N d-dimensional feature vectors (cases). Let $\mathbf{V}$ denote the columns (feature vectors) of $\mathbf{X}^{<m>}$ that were complete in $\mathbf{X}_{obs}$ and let $\mathbf{W}^{<m>}$ be the columns of $\mathbf{X}^{<m>}$ that were incomplete in $\mathbf{X}_{obs}$. Clearly, $\mathbf{V}$ will be the same for all $\mathbf{X}^{<m>}$. If

$$\overline{\mathbf{W}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{W}^{<m>},$$

the j-th column of $\overline{\mathbf{W}}$ is the centroid of the j-th columns of $\mathbf{W}^{<1>}, \ldots, \mathbf{W}^{<M>}$.

Let $\mathbf{V}_C$ and $\overline{\mathbf{W}}_C$ be those datasets $\mathbf{V}$ and $\overline{\mathbf{W}}$ that are linked with class C. The first two principle components $\mathbf{b}_1^{\mathrm{T}}\mathbf{x}$ and $\mathbf{b}_2^{\mathrm{T}}\mathbf{x}$ resulting from the PCA of data array $[\mathbf{V}_1, \overline{\mathbf{W}}_1, \ldots, \mathbf{V}_K, \overline{\mathbf{W}}_K]$ enable the data to be projected onto the two-dimensional plane defined by eigenvectors $\mathbf{b}_1$ and $\mathbf{b}_2$. The PC plot will consist of points corresponding to the initially complete feature vectors $[\mathbf{V}_1, \ldots, \mathbf{V}_K]$ together with points corresponding to the centroids $[\overline{\mathbf{W}}_1, \ldots, \overline{\mathbf{W}}_K]$ of the imputed feature vectors. If PCA is performed on data array

$$[\mathbf{V}_1, [\mathbf{W}_1^{<1>}, \ldots, \mathbf{W}_1^{<M>}], \ldots, \mathbf{V}_K, [\mathbf{W}_K^{<1>}, \ldots, \mathbf{W}_K^{<M>}]]$$

in place of $[\mathbf{V}_1, \overline{\mathbf{W}}_1, \ldots, \mathbf{V}_K, \overline{\mathbf{W}}_K]$, the distribution of points in the resultant PC plot would be biased by the distribution of the imputed feature vectors

$$[[\mathbf{W}_1^{<1>}, \ldots, \mathbf{W}_1^{<M>}], \ldots, [\mathbf{W}_K^{<1>}, \ldots, \mathbf{W}_K^{<M>}]].$$

3.2 Prediction regions

Given the principal components $\mathbf{b}_1^{\mathrm{T}}\mathbf{x}$ and $\mathbf{b}_2^{\mathrm{T}}\mathbf{x}$ obtained from the above PCA, the imputed feature vectors $[\mathbf{W}^{<1>}, \ldots, \mathbf{W}^{<M>}]$ for each class C can be projected onto the PC plot via the function $(\mathbf{b}_1^{\mathrm{T}}\mathbf{x}, \mathbf{b}_2^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}$. This enables the spread of the different imputations of an initially incomplete feature vector $\mathbf{x}_{obs,j}$ (the j-th columns of $\mathbf{W}^{<1>}, \ldots, \mathbf{W}^{<M>}$) to be examined. From the dispersion of these M points in the PC plot, a prediction region for the true but unobserved feature vector (the one for which $\mathbf{x}_{obs,j}$ is a subset) can be established as follows.

Suppose that $\overline{\mathbf{y}}$ is the mean vector of M q-dimensional observations $\mathbf{y}_1, \ldots, \mathbf{y}_M$, and $\mathbf{y}_0$ is unobserved but sampled from the same multivariate normal distribution $N(\boldsymbol{\theta}_y, \boldsymbol{\Sigma}_y)$ as $\mathbf{y}_1, \ldots, \mathbf{y}_M$. If $\mathbf{S}$ is the sample covariance matrix of $\mathbf{y}_1, \ldots, \mathbf{y}_M$ then $(M-1)\mathbf{S}$ has a Wishart distribution with $M-1$ degrees of freedom and parameter $\boldsymbol{\Sigma}_y$, from which it follows that $(\mathbf{y}-\overline{\mathbf{y}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{y}-\overline{\mathbf{y}})$ has $q(M^2-1)/(M(M-q))$ times the F distribution with q and $(M-q)$ degrees of freedom (e.g., Krzanowski 1988). If $F_\alpha(q, M-q)$ is the quantile for this F distribution such that

$$p\left[\frac{M(M-q)}{q(M^2-1)}(\mathbf{y}-\overline{\mathbf{y}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{y}-\overline{\mathbf{y}}) < F_\alpha(q, M-q)\right] = 1-\alpha \qquad (2)$$

then the set of $\mathbf{x}$ satisfying the inequality

$$(\mathbf{y}-\overline{\mathbf{y}})^{\mathrm{T}}\mathbf{S}^{-1}(\mathbf{y}-\overline{\mathbf{y}}) < \frac{q(M^2-1)}{M(M-q)}F_\alpha(q, M-q) \qquad (3)$$

is a $100(1-\alpha)$ percent *prediction region* for $\mathbf{y}_0$.

Note that a prediction region is not the same as a confidence region. Suppose we have a population with parameter $\boldsymbol{\Phi}$ from which a set of observations $\mathcal{D}$ has been drawn randomly. A *confidence region* Δ_c is an estimate of $\boldsymbol{\Phi}$, based on sample $\mathcal{D}$, which has the form $p(\boldsymbol{\Phi} \in \Delta_c|\mathcal{D}) = 1-\alpha$. In contrast, a prediction region Δ_p is an estimate, based on $\mathcal{D}$, of a possible observation $\mathbf{x}$ drawn from the same population as $\mathcal{D}$, the estimate having the form $p(\mathbf{x} \in \Delta_p|\mathcal{D}) = 1-\alpha$.

The predictive distribution $p(\mathbf{y}_0|\overline{\mathbf{y}}, \mathbf{S})$ can be expressed as

$$p(\mathbf{y}_0|\overline{\mathbf{y}}, \mathbf{S}) = \iint p(\mathbf{y}_0|\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y^{-1})p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_y^{-1}|\overline{\mathbf{y}}, \mathbf{S})d\boldsymbol{\mu}, d\boldsymbol{\Sigma}_y^{-1}, \qquad (4)$$

and Geisser (1993) has shown that this Bayesian approach provides an alternative derivation of (2) via the distribution

$$
\begin{aligned}
p(\mathbf{y}_0|\overline{\mathbf{y}}, \mathbf{S}) = {} & \left(\frac{M}{(M+1)\pi}\right)^{q/2} \frac{\Gamma(M/2)\,|(M-1)\mathbf{S}|^{(M-1)/2}}{\Gamma((M-q)/2)} \\
& \times \left|(M-1)\mathbf{S} + \left(\frac{M}{M+1}\right)(\overline{\mathbf{y}} - \mathbf{y}_0)(\overline{\mathbf{y}} - \mathbf{y}_0)^{\mathrm{T}}\right|^{-M/2}.
\end{aligned} \tag{5}
$$

From (3), the boundary of a $100(1-\alpha)$ percent region for $\mathbf{y}_0$ is the set of $\mathbf{y}$ satisfying

$$
(\mathbf{y} - \overline{\mathbf{y}})^{\mathrm{T}} \mathbf{S}^{-1} (\mathbf{y} - \overline{\mathbf{y}}) = \frac{q(M^2-1)}{M(M-q)} F_\alpha(q, M-q). \tag{6}
$$

When $q = 2$, the left-hand side of (6) becomes a non-homogeneous polynomial of degree 2 with respect to both components of $\mathbf{y}$. If the equation is expressed as a quadratic in y_2, with $\mathbf{y} = (y_1, y_2)^{\mathrm{T}}$, the two roots of the quadratic are functions of y_1. If we write these roots as $y_2 = r_1(y_1)$ and $y_2 = r_2(y_1)$, the solution set for (6), which is an ellipse, is the set of real-valued vectors $(y_1, r_1(y_1))^{\mathrm{T}}$ and $(y_1, r_2(y_1))^{\mathrm{T}}$.

A problem with plotting $r_1(y_1)$ and $r_2(y_1)$ over finite increments of y_1 is that gaps will appear at the right- and left-hand ends of the ellipse unless the increments are sufficiently small. This is due to $r_1(y_1)$ and $r_2(y_1)$ taking complex values whenever y_1 is not a component of the vectors of the solution set. Replacing the quadratic equation with a pair of parametric equations for y_1 and y_2 circumvents the problem, and we derived parametric equations for this purpose, as follows.

If $\mathbf{y} = (y_1, y_2)^{\mathrm{T}}$, $\overline{\mathbf{y}} = (\phi_1, \phi_2)^{\mathrm{T}}$ and $\mathbf{S}^{-1} = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ then (6) becomes a 2nd-order non-homogeneous polynomial in y_1 and y_2. Suppose that $\phi_1 = \phi_2 = 0$ and let the rectangular coordinate system for the PC plot be denoted by Oy_1y_2. If Oy_1y_2 is rotated anticlockwise about the origin O by ψ radians, we obtain a new coordinate system $Oy_1'y_2'$ related to Oy_1y_2 by

$$
\left.\begin{aligned}
y_1 &= \cos(\psi)y_1' - \sin(\psi)y_2' \\
y_2 &= \sin(\psi)y_1' + \cos(\psi)y_2'
\end{aligned}\right\}. \tag{7}
$$

On replacing y_1 and y_2 in the above polynomial with (7), the substitution

$$
\psi = \begin{cases} \frac{1}{2}\tan^{-1}\left(\frac{2b}{a-c}\right) & \text{if } a \neq c \\ \pi/4 & \text{otherwise,} \end{cases}
$$

simplifies the polynomial to

$$
g \cdot (y_1')^2 + h \cdot (y_2')^2 = J \tag{8}
$$

(e.g., Salas & Hille 1982), where

$$g = a\cos(\psi)^2 + 2b\cos(\psi)\sin(\psi) + c\sin(\psi)^2,$$

$$h = a\sin(\psi)^2 - 2b\cos(\psi)\sin(\psi) + c\cos(\psi)^2,$$

and

$$J = \frac{2(M^2-1)}{M(M-2)} F_\alpha(2, M-2).$$

Equation (8) can be parametrized with respect to parameter τ ($\in [0, 2\pi]$) by setting $y_1'(\tau)$ and $y_2'(\tau)$ equal to $\sqrt{\frac{J}{g}}\sin(\tau)$ and $\sqrt{\frac{J}{h}}\cos(\tau)$, respectively, whereupon, from (7), we obtain the required parametric equations:

$$\left.\begin{array}{l} y_1(\tau) = \phi_1 + \cos(\psi)\sqrt{\frac{J}{g}}\sin(\tau) - \sin(\psi)\sqrt{\frac{J}{h}}\cos(\tau) \\ y_2(\tau) = \phi_2 + \sin(\psi)\sqrt{\frac{J}{g}}\sin(\tau) + \cos(\psi)\sqrt{\frac{J}{h}}\cos(\tau) \end{array}\right\}, \tag{9}$$

where ϕ_1 and ϕ_2 provide the necessary translation of the ellipse.

3.3 Prediction intervals

It is possible for a case in a dataset to have only one missing datum. Multiple imputation of this case will result in a set of vectors lying along a line in feature space, parallel to one of the coordinate axes. Consequently, the mapping of these points onto the PC plot will result in a line of points on the PC plot. In this situation, a *prediction interval* is used in place of a prediction region.

Let $\mathbf{y}_1, \ldots, \mathbf{y}_M$ be a linear arrangement of points in a PC plot arising from a multiple imputation of an incomplete case, and let $y_1^{(i)}$ be the first element of $\mathbf{y}_i (i = 1, \ldots, M)$. If $y_1^{(1)}, \ldots, y_1^{(M)}$ are assumed to be sampled from a normal distribution, a $100(1-\alpha)$ percent prediction interval for $\mathbf{y}_0$ is the line between the vectors

$$\left(\begin{array}{c} \overline{y} \pm t_\alpha(M-1)s\sqrt{\frac{M-1}{M}} \\ k_1\left[\overline{y} \pm t_\alpha(M-1)s\sqrt{\frac{M-1}{M}}\right] + k_2 \end{array} \right), \tag{10}$$

where $\overline{y}$ and s are the sample mean and standard deviation of $\{y_1^{(1)}, \ldots, y_1^{(M)}\}$, $t_\alpha(M-1)$ is the $100(1-\alpha)$ percentile of Student's t-distribution with $M-1$ degrees of freedom, k_1 is the slope of $\{\mathbf{y}_1, \ldots, \mathbf{y}_M\}$ in the PC plot relative to the first principal component, and k_2 is the attendant intercept. Geisser (1993) provides a derivation of $\overline{y} \pm t_\alpha(M-1)s\sqrt{(M-1)/M}$.

4 Examples

Two examples are given to illustrate the foregoing proposal of prediction regions and intervals. The first is based on a pair of multivariate normal distributions, and the second is taken from a previously published simulation of a health survey.

4.1 Example 1

The first example is based on the artificial dataset shown in Table 1. This consists of 40 vectors drawn randomly from two 4-variate normal distributions. The population mean vectors for the two classes are $(2,2,2,2)^{\mathrm{T}}$ and $(5,5,5,5)^{\mathrm{T}}$, respectively. Both populations have a covariance matrix equal to the 4×4 identity matrix. Each class has five incomplete cases.

The dataset was imputed using the `NORM` multiple imputation package provided by Schafer (1998) as freeware for use within the `S-PLUS` statistical environment (Venables & Ripley 1997). This package implements the data-augmentation procedure shown in Algorithm 1, and it assumes that data originate from a multivariate normal distribution. The number of imputations by data augmentation was set to 10, with 50 iterations within each Markov chain.

Figure 1 shows a visualization resulting from this dataset using equations (9) for the regions and vectors (10) for the intervals. Each class has three cases with only one missing value; therefore, three prediction intervals per class are displayed. In addition, there are two prediction regions for each class. This figure is based on 75% prediction regions and intervals; however, a better approach is to depict the prediction distributions as contour diagrams consisting of nested prediction regions, as this avoids an arbitrary choice of $100(1-\alpha)$. Figure 2 shows only two such contour diagrams for clarity.

The multiple imputations for an incomplete case will lie on the line or (hyper)plane Ξ defined by the variables associated with the missing values. The distribution of the multiple imputations on Ξ is determined by the uncertainty of the imputed values. It is the combination of the orientation of Ξ in feature space, the distribution of the points on Ξ, and the orientation of the PC plane in feature space that defines the direction and shape of the line or ellipsoid resulting from the projected points.

4.2 Example 2

The second example is a previously published dataset (Table 6.14. in Schafer (1997)), namely, a simulation based on a health survey provided by the Na-

Table 1: Artificial incomplete dataset from two 4-variate normal distributions.

Class 1				*Class 2*			
U_1	U_2	U_3	U_4	U_1	U_2	U_3	U_4
2.88	2.73	2.85	1.70	3.21	4.32	4.66	5.80
?	?	3.53	3.34	4.83	5.72	4.73	4.35
1.77	3.27	1.98	-0.03	6.35	?	3.28	5.22
0.60	3.03	2.80	1.20	5.02	5.63	7.00	4.55
?	4.43	0.89	1.51	5.03	4.38	5.66	4.07
2.77	1.54	2.22	3.88	4.41	4.57	?	4.40
3.16	?	?	?	5.31	4.90	4.85	6.30
1.86	2.52	1.03	2.25	6.19	3.94	4.74	4.19
0.93	0.61	2.87	2.58	?	5.48	5.55	4.80
2.64	2.35	2.56	4.44	4.81	6.36	5.07	6.78
3.05	2.16	3.59	0.74	7.62	3.57	5.02	4.04
0.80	0.73	?	1.46	5.16	4.97	5.84	6.46
-1.08	1.59	3.34	1.87	4.75	?	?	?
3.67	1.42	2.64	2.31	4.98	5.86	4.64	4.00
1.97	2.68	1.56	1.93	5.55	4.52	5.79	6.20
3.01	2.83	2.74	0.97	6.03	5.24	6.30	4.95
1.91	2.45	3.27	1.69	3.75	4.90	6.05	6.01
2.83	2.39	1.18	0.56	?	?	7.02	4.97
1.70	?	2.36	0.22	5.90	6.20	5.85	5.64
3.47	1.54	3.12	2.26	3.42	3.66	5.17	4.44

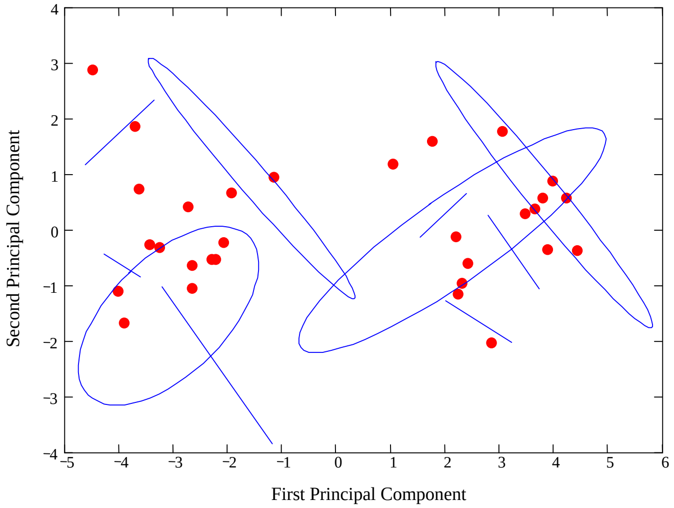

Figure 1: Visualization of the dataset given in Table 1. The PC plot is based on both the complete cases present in the dataset (*dots*) and on the centroids (*not shown*) of the imputed cases (*not shown*). 75% prediction regions (*ellipses*) and intervals (*lines*) for the incomplete cases are displayed.

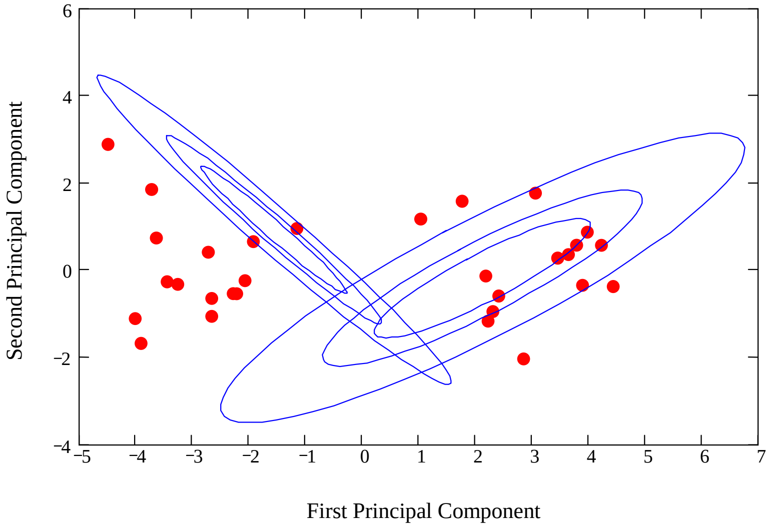

Figure 2: The 50%, 75% and 95% prediction regions (*inner ellipses*, *middle ellipses and outer ellipses*, *respectively*) for two of the incomplete cases present in Table 1 (Compare with Figure 1).

tional Center for Health Statistics (1994). The dataset consists of 25 cases with four attributes: (i) *age_group* (ordinal; 100% complete); (ii) *hypertension* (binary; 68% complete); (iii) *body_mass_index* (continuous; 64% complete); (iii) *cholesterol_level* (continuous; 60% complete). Twelve of the cases are incomplete.

Given the presence of a binary attribute (*hypertension*), the joint probability distribution for the dataset is clearly not multivariate normal; nevertheless, Schafer (1997) found that he could successfully apply data augmentation for the normal model to it. Therefore, we first imputed the entire dataset by this method 10 times using the `NORM` package (Schafer 1998). We then rounded the continuous imputes for *hypertension* to the nearest category, and each missing *hypertension* value was imputed with the mode of the 10 rounded estimates obtained for it. This resulted in the dataset having 20 cases with *hypertension* = 0 and five cases with *hypertension* = 1.

With these imputed values of *hypertension* held fixed and regarded as true, the dataset was again imputed 10 times so that 10 new imputations could be obtained for the continuous-valued attributes *body_mass_index* and *cholesterol_level*. The resulting prediction regions and intervals arising from this second set of 10 imputations are shown in Figure 3. The three complete cases, the prediction region and the prediction interval associated with *hypertension* = 1 are seen to be clustered in the top, right quadrant of the PC plot, indicating that probabilistic classification between the two levels of *hypertension* may be possible.

5 Problems and tentative solutions

The conceptual simplicity of the centroid-based approach was our motivation for investigating it, but there are drawbacks to it.

One problem with the method is that, in using centroids for the sample covariance matrix $\boldsymbol{S}_x$, some of true variation of $\mathbf{X}^{<1>}, ..., \mathbf{X}^{<M>}$ is missed. This can be overcome by basing the PCA on the mean sample covariance matrix $\overline{\mathbf{S}}_x$ in place of $\boldsymbol{S}_x$. The mean sample covariance matrix for multiple imputations, which was proposed by Rubin (1987, pp. 75-76), is defined by

$$\overline{\mathbf{S}}_x = \frac{1}{M} \sum_{m=1}^{M} \mathbf{S}_x^{<m>},$$

where $\mathbf{S}_x^{<m>}$ is the sample covariance matrix obtained from $\mathbf{X}^{<m>}$.

Another problem with our approach is that the graphical representation depicts only the marginal distributions of the possible imputations. In other words, the displayed prediction regions refer to the possible imputation of

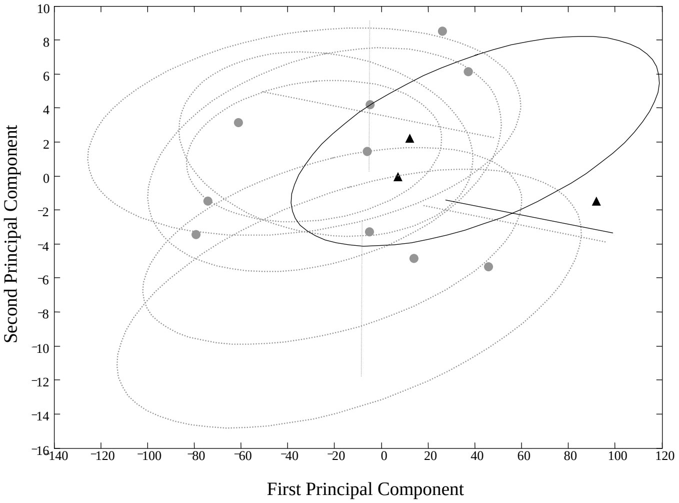

Figure 3: PC plot of the simulated health survey data showing the 75% prediction regions and intervals associated with *hypertension* $= 0$ (*light grey ellipses and lines*) and *hypertension* $= 1$ (*black ellipse and line*). The points from the complete cases with *hypertension* $= 0$ (*light grey dots*) and *hypertension* $= 1$ (*black triangles*) are also shown.

one case at a time, with joint behaviour between cases being disregarded. However, it could be that whenever point $\mathbf{y}_i^{<m>}$ projected from $\mathbf{X}^{<m>}$ is in the upper half of its ellipsoid, another point $\mathbf{y}_j^{<m>}$ from $\mathbf{X}^{<m>}$ tends to be in the lower half of its ellipsoid. A solution to depicting this type of independence is to use a brushing technique. For example, if $\mathbf{Y}$ is a set of projected multiple-imputation points, this could be done by displaying the elements of $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ and, if a user touches an element of, say, $\mathbf{Y}_1$, the points in $\mathbf{Y}_2, \ldots, \mathbf{Y}_m$ originating from the same imputation as that in $\mathbf{Y}_1$ are highlighted.

The use of (2), and thus (6), for constructing prediction regions is based on the assumption that imputations $\mathbf{x}^{<1>}, \ldots, \mathbf{x}^{<M>}$ for an incomplete row $\mathbf{x}_{obs}$ are randomly drawn from a multivariate normal distribution. This assumption has enabled us to explore the idea of using prediction regions and intervals for incomplete datasets; however, a counterexample to the normality assumption can be provided as follows. Consider a hypothetical dataset sampled from a multivariate normal distribution. The dataset consists of complete cases $\mathbf{x}_1, \ldots, \mathbf{x}_M$ and an incomplete case $\mathbf{x}_0$ in which all values are missing. Note that, here, M refers to the number of complete cases, not the number of imputations. If $\overline{\mathbf{x}}$ and $\mathbf{S}$ are the mean vector and covariance matrix for $\mathbf{x}_1, \ldots, \mathbf{x}_M$, comparison of (4) with (1) implies that imputation of the incomplete case by data augmentation with respect to the complete cases will have distribution (5). But this distribution is not a normal distribution, and its tails are broader than those of the multivariate t-distribution with parameters $\overline{\mathbf{x}}$, $\mathbf{S}$ and M degrees of freedom; therefore, if the cases of a dataset are drawn randomly from a multivariate normal distribution, the prediction regions constructed by (9) underestimate the size of the regions. Thus, an extension of this work is to base prediction regions on broad-tailed distributions, such as those within the family of elliptical distributions (Kelker 1970, Muirhead 1982).

An alternative to assuming a particular statistical distribution for a set $\mathbf{Y}$ of projected multiple-imputation points is to use a model-free method. One such alternative is to define a ‘prediction region’ by the *convex hull* for $\mathbf{Y}$. If the elements of $\mathbf{Y}$ are regarded as pins partially stuck into the PCA plane, the convex hull is the boundary defined by placing an elastic band tightly around all the pins. The convex hull for $\mathbf{Y}$ can be determined in $O(|\mathbf{Y}| \ln |\mathbf{Y}|)$ time using Graham’s (1972) scan algorithm.

6 Discussion

In spite of the deficiencies stated in Section 5, our interim approach has enabled us to demonstrate the efficacy of using prediction regions and intervals to visualize incomplete datasets. In the above examples, our method successfully conveyed the presence of clusters consisting of points, ellipses and lines.

We have concentrated on visualization through the use of PC plots but, as mentioned in the introduction, other visualization methods are available. In the case of canonical variates analysis (CVA), imputed cases can be mapped to the scatter plot of the first and second canonical variates obtained from the complete cases and centroids. This is done via the function $(\mathbf{c}_1^{\mathrm{T}}\mathbf{x},\mathbf{c}_2^{\mathrm{T}}\mathbf{x})^{\mathrm{T}}$, where $\mathbf{c}_1$ and $\mathbf{c}_2$ are the CVA eigenvectors associated with the two largest eigenvalues. An analogous approach can be used with orthogonal CVA (Krzanowski 1995).

As regards Sammon mappings, a Sammon plot can be obtained first with respect to the complete cases and centroids. Whilst the final positions of these on the Sammon plot are fixed, the positions of the imputed cases on the same plot can be determined iteratively with respect to each other and to the locations of the complete cases and centroids.

The categorical and continuous attributes of Example 2 were imputed under the normal model, but a more appropriate scheme is to perform multiple imputation under the Olkin-Tate location model for mixed data (Olkin & Tate 1961). This approach has been discussed by Schafer (1997) and applied by Raghunathan & Grizzle (1995) and Raghunathan & Siscovick (1996), but there remains the challenge of deriving and displaying the associated prediction regions.

Acknowledgements

We thank the two anonymous referees for their constructive comments.

References

Albert, R.H. & W. Horwitz (1995), ‘Incomplete datasets: Coping with inadequate databases’, *Journal of the AOAC International* **78**, 1513–1515.

Dempster, A.P., N.M. Laird & D.B. Rubin (1977), ‘Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)’, *Journal of the Royal Statistical Society* **B39**, 1–38.

Geisser, S. (1993), *Predictive Inference: An Introduction*, Chapman and Hall, New York.

Gower, J.C. & D.J. Hand (1996), *Biplots*, Chapman and Hall, London, pp. 53–61.

Graham, R.L. (1972), 'An efficient algorithm for determining the convex hull of a finite planar set', *Information Processing Letters* **1**, 132–133.

Heitjan, D.F. (1993), 'Ignorability and coarse data: Some biomedical examples', *Biometrics* **49**, 1099–1109.

Hotteling, H. (1933), 'Analysis of a complex of statistical variables into principal components', *Journal of Educational Psychology* **24**, 417–441,498–520.

Kelker, D. (1970), 'Distribution theory of spherical distributions and a location-scale parameter generalization', *Sankhya A* **32**, 419–430.

Knaus, W.A., J.E. Zimmerman, P.P. Wagner, E.A. Draper & D.E. Lawrence (1981), 'APACHE - acute physiology and chronic health evaluation: A physiologically based classification system', *Critical Care Medicine* **9**(8), 591–597.

Krzanowski, W.J. (1988), *Principles of Multivariate Analysis: A User's Perspective*, Clarendon Press, Oxford.

Krzanowski, W.J. (1995), 'Orthogonal canonical variates for discrimination and classification', *Journal of Chemometrics* **9**(6), 509–520.

Little, R.J.A. & D.B. Rubin (1987), *Statistical Analysis with Missing Data*, Wiley, New York.

Muirhead, R.J. (1982), *Aspects of Multivariate Statistical Theory*, John Wiley, New York, pp. 32–40.

National Center for Health Statistics (1994), Plan and operation of the Third National Health and Nutrition Examination Survey. Vital and Health Statistics Series 1, No. 32, NCHS.

Olkin, I. & R.F. Tate (1961), 'Multivariate correlation models with mixed discrete and continuous variables', *Annals of Mathematical Statistics* **32**, 448–465.

Raghunathan, T.E. & D.S. Siscovick (1996), 'A multiple-imputation analysis of a case-control study of the risk of primary cardiac arrest among pharmcologically treated hypertensives', *Applied Statistics* **45**(3), 335–352.

Raghunathan, T.E. & J.E. Grizzle (1995), 'A split questionnaire survey design', *Journal of the American Statistical Society* **90**(429), 54–63.

Roberts, G.O. (1996), Markov chain concepts related to sampling algorithms, *in* W.Gilks, S.Richardson & D.Spiegelhalter, eds, 'Markov Chain Monte Carlo in Practice', Chapman and Hall, London, pp. 45–57.

Rubin, D.B. (1976), ‘Inference and missing data’, *Biometrika* **63**, 581–592.

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, John Wiley, New York.

Salas, S.L. & E. Hille (1982), *Calculus: One and Several Variables, with Analytical Geometry*, John Wiley, New York, pp. 400–404.

Sammon, J.W. (1969), ‘A nonlinear mapping for data structure analysis’, *IEEE Transactions in Computing* **C-18**, 401–409.

Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, Chapman & Hall, London.

Schafer, J.L. (1998), *Software for Multiple Imputation* [online]. Available from: http://www.stat.psu.edu/∼jls/misoftwa.html [Accessed 18 June 1998].

Seal, H.L. (1964), *Multivariate Statistical Analysis for Biologists*, Methuen, London.

Statistical Solutions (1998), *The Solution for Missing Values in your Data* [online]. Available from: http://www.statsol.ie/solas.html [Accessed 1 Dec 1998].

Swayne, D.F. & A. Buja (1998), ‘Missing data in interactive high–dimensional data visualization’, *Computational Statistics* **13**(1), 15–26.

Tanner, M.A. & W.H. Wong (1987), ‘The calculation of posterior distributions by data augmentation (with discussion)’, *Journal of the American Statistical Association* **82**, 528–550.

Unwin, A.R., G. Hawkins, H. Hofmann & B. Siegl (1996), ‘Interactive graphics for data sets with missing values - MANET’, *Journal of Computational and Graphical Statistics* **5**, 113–122.

Venables, W.N. & B.D. Ripley (1997), *Modern Applied Statistics with S-PLUS*, 2nd edn, Springer, New York.