BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 9 2007, pages 1164–1167
doi:10.1093/bioinformatics/btm069

Gene expression

pcaMethods—a bioconductor package providing PCA methods for incomplete data

Wolfram Stacklies[1], Henning Redestig[2], Matthias Scholz[3], Dirk Walther[2] and Joachim Selbig[2,4,*]

[1]CAS-MPG Partner Institute for Comp. Biology, 320 Yue Yang Road, 200031 Shanghai, China, [2]Max Planck Institute for Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm, [3]Ernst-Moritz-Arndt-University of Greifswald, Competence Center for Functional Genomics, F.L.Jahnstraße 15, 17487, Greifswald and [4]University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany

Received on October 25, 2006; revised on February 13, 2007; accepted on February 21, 2007

Advance Access publication March 7, 2007

Associate Editor: Olga Troyanskaya

ABSTRACT

Summary: *pcaMethods* is a Bioconductor compliant library for computing principal component analysis (PCA) on incomplete data sets. The results can be analyzed directly or used to estimate missing values to enable the use of missing value sensitive statistical methods. The package was mainly developed with microarray and metabolite data sets in mind, but can be applied to any other incomplete data set as well.

Availability: http://www.bioconductor.org

Contact: selbig@mpimp-golm.mpg.de

Supplementary information: Please visit our webpage at http://bioinformatics.mpimp-golm.mpg.de/

1 INTRODUCTION

1.1 Motivation

Microarray data are used in a range of biological research areas although they frequently contain considerable numbers of missing values. Principal component analysis (PCA) is often a first step in the analysis process, however, the standard approach is not tolerant to missing data, because it is based on eigenvalue decomposition of the covariance matrix. Missing value estimation becomes important when subsequent statistical analyses depend on complete data sets, e.g. independent component analysis or various clustering algorithms, such as correlation-based hierarchical clustering. The *pcaMethods* package provides PCA methods that are robust against missing data and that allow for missing value estimation.

2 ALGORITHMS

2.1 Probabilistic PCA (PPCA)

PPCA combines an expectation maximization (EM) approach with a probabilistic model. The EM approach is based on the assumption that the latent variables (scores) as well as the noise come from normal distributions. In standard PCA, data points far from the training set but close to the subspace defined by the principal components fit the model equally well. PPCA, on the other hand, defines a density model such that the likelihood for data points far from the training set is much lower, even if they are close to the principal subspace.

Our implementation of PPCA is based on the Matlab^{TM}ppca script written by *Jakob Verbeek*.[1]

2.2 Bayesian PCA (BPCA)

Similar to PPCA, BPCA uses an EM approach combined with a Bayesian estimation method to calculate the likelihood of an estimated value. BPCA was developed especially for missing value estimation and is based on a variational Bayesian framework (VBF), Bishop (1999), with automatic relevance determination (ARD). In BPCA, ARD leads to a different scaling of the principal components, scores and eigenvalues when compared to standard PCA or PPCA. When the Euclidean norm of a component is small relative to the variance of the noise observed for this component, it will shrink to near zero. This suppresses redundant components, but for medium-sized eigenvalues, the norm of the principal components will be smaller than in PCA. In case of small numbers of observations, the difference between 'real' and predicted eigenvalues may become larger reflecting the lack of information to accurately determine true components from incomplete and noisy data.

Another slight difference from PCA results may arise from the fact that the VBF algorithm does not force orthogonality between principal components. See Oba *et al.* (2003) for a detailed discussion of the BPCA algorithm.

The method provided by the *pcaMethods* package is a port of the bpca Matlab^{TM} script also provided by Oba *et al.*[2]

*To whom correspondence should be addressed.

[1]http://lear.inrialpes.fr/~verbeek/
[2]http://hawaii.aist-nara.ac.jp/%7Eshige-o/tools/

2.3 Inverse non-linear PCA

Non-linear PCA (NLPCA) (Scholz *et al.*, 2005) is especially suitable for data from experiments where the studied response is non-linear. Examples of such experiments are ubiquitous in biology-enzyme kinetics are inherently non-linear as are gene expression responses influenced by the cell cycle or diurnal oscillations. The inverse version of the NLPCA algorithm works by training an auto-associative neural network composed of a component layer which serves as the 'bottle-neck', a hidden non-linear layer, and an output layer corresponding to the reconstructed data. Missing values in the training data are simply ignored when calculating the error during backpropagation. The input to this network can intuitively be seen as the scores of the resulting principal components which have their corresponding loadings hidden in the neural network.

2.4 Nipals PCA

Nipals (Non-linear estimation by iterative partial least squares), Wold *et al.* (1966) is an algorithm at the root of PLS regression which can execute PCA with missing values by simply leaving them out from the appropriate inner products. It is tolerant to small amounts (generally not more than 5%) of missing data.

2.5 SVDimpute

An implementation of the SVDimpute algorithm as proposed by Troyanskaya *et al.* (2001). The idea behind the algorithm is to estimate the missing values as a linear combination of the k most significant eigengenes, Alter *et al.* (2000), where the most significant eigengene is the one with the greatest eigenvalue.

2.6 LLSimpute

Although the scope of the package is to provide a collection of PCA-based methods, a non-PCA missing value estimation method was included to allow users to better rate and compare the results.

The LLSimpute algorithm for missing value estimation as proposed in Kim *et al.* (2005) works similar to KNNimpute, Troyanskaya *et al.* (2001). For each incomplete variable (gene) G, k similar variables are selected by the Pearson correlation coefficient. Then missing values are imputed by a linear combination of the k selected variables.

Let be G_{comp} the set and g the number of complete observations in G. Then the linear combination is determined as the solution of the least squares problem formulated as

$$\min_{z} \| A^T z - G_{\text{comp}} \|_{\text{Pearson.}} \quad (1)$$

Where, A is a $k \times g$ matrix formed by the corresponding g observations of the k neighbors. On the data used for testing in the original publication, LLSimpute could outperform KNNimpute and compared favorably well with BPCA.

2.7 Performance

BPCA, SVDimpute, NLPCA and Nipals all contain iterative steps which make them less time efficient. During each iteration, BPCA and SVDimpute make use of singular value decomposition (SVD) whose complexity grows cubically with the number of dimensions. This may lead to performance problems when data sets are of high dimensionality.

PPCA is the fastest method and is thus recommended for large data sets. The runtime is linear in the number of data points, data dimensions and components to estimate, thus providing satisfactory performance even on large data sets.

3 PREPROCESSING

Normalization is a critical step for making measurements from different variables or experiments comparable. Generally, there is no straightforward approach, because adequate normalization largely depends on the data of interest, see Huber *et al.* (2005) for a detailed discussion. Here, we will only consider the two standard procedures mean centring and variance scaling.

PCA requires mean centring, because it is based on the calculation of the covariance matrix. Thus, the mean must be subtracted before estimating missing values and added again afterwards. This is done automatically by the methods presented here.

Variance normalization will strongly affect PCA results. Scaling to unit variance may be useful when variables of different units or intensity ranges are compared. For example, if one is interested in the correlation structure between transcripts and metabolites simultaneously. For microarray data, however, one often has expression estimates for genes that are not expressed at all and these must be removed before any scaling is done or they will add unnecessary noise to the PCA model.

4 MISSING VALUE ESTIMATION

If not all principal components are used for projection, which is usually the case, PCA can be seen as a data reduction process. When only the first k components are used for data reduction, the projection can be written as

$$X = 1 \times \bar{x}^T + TP^T + V, \quad (2)$$

where, the term $1 \times \bar{x}^T$ represents the original variable averages, X denotes the observations, $T = t_1, t_2, \ldots, t_k$ the scores, $P = p_1, p_2, \ldots, p_k$ the components and V the residual matrix. An estimate of the complete data set ($\hat{X}$) is obtained by projecting the scores back into the original data space; $\hat{X} = 1 \times \bar{x}^T + TP^T$. This will only produce reasonable results if the residuals V are sufficiently small, implying that most of the important information is captured by the first k components. This is generally the case if the data show only few large eigenvalues. Figure 1A shows the error of prediction for two gene expression data sets with different eigenvalue structure, 5% of the data were removed for testing. The first data are marker genes of the human cell cycle, the second a complete subset of breast cancer related genes; both published in Whitfield *et al.* (2002).

Different approaches for missing value estimation have been proposed, e.g. Kim *et al.* (2005) Sehgal *et al.* (2005) and others. The cited papers also contain comparisons between several methods. From the literature, we conclude that among the

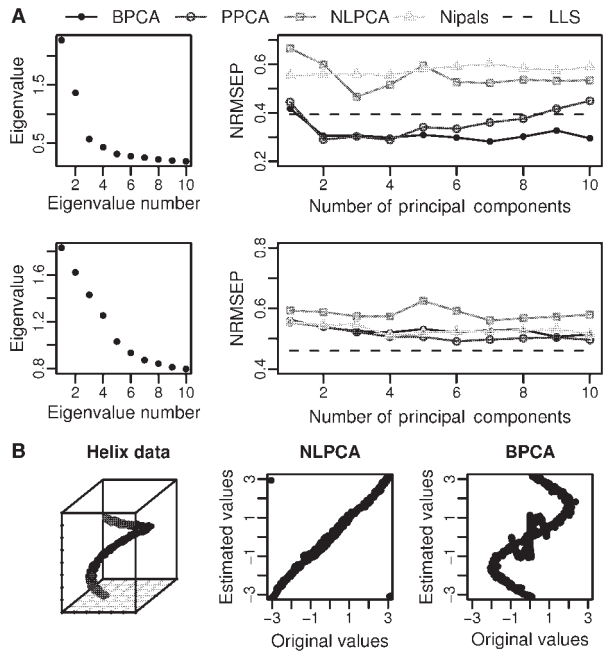

Fig. 1. (**A**) Eigenvalues and the NRMSEP obtained with different numbers of principal components. The lowest NRMSEP as obtained with LLSimpute is plotted in as a dashed line. The first data set shows only one dominant eigenvalue and thus the lower error of prediction. Here, BPCA performs best whereas in the second example with no dominant eigenvalue LLSimpute produces better results. (**B**) A comparison of NLPCA and BPCA applied to artificial 3D helix data from which 30% of all data points have been removed. In contrast to BPCA, NLPCA is able to capture the non-linear structure of the data and thus to produce an accurate estimate.

algorithms provided in *pcaMethods*, BPCA has, on average, the best missing value estimation accuracy. When the data contain strong non-linear dependencies, NLPCA may be superior. Figure 1B shows a comparison of NLPCA and BPCA applied to an artificial 3D helix data set. In contrast to BPCA, NLPCA is able to exactly capture the non-linear structure.

When, on the other hand, an exact PCA solution is needed, both methods are less adequate. Here, PPCA provides good results, also overcoming the performance problems of the other methods. SVDimpute and Nipals both are widely used standard approaches and were included for comparison.

The reader should keep in mind that the estimation accuracy of a certain method depends on the structure of the data it is used on. A detailed comparison on different data sets is beyond the scope of this article.

4.1 Parameter estimation

A common problem is the choice of the optimal number of principal components (or neighbors for LLSimpute). One wants to include the relevant information, but choosing too many components will also include artifacts or noise. Cross validation can be used to determine this parameter. The package provides two such methods.

The first is for internal cross validation that allows to estimate the level of structure in the data and to optimize the choice of the number of components. The level of structure, Q^2 is defined as:

$$Q^2 = 1 - \frac{\sum(x_{ij} - \hat{x}_{ij})^2}{\sum(x_{ij})^2} \tag{3}$$

The maximum value for Q^2 is 1 which means that all variance is represented in the predictions; $X = \hat{X}$.

The second method estimates the optimal number of components in terms of the error of prediction. Q^2 or the normalized root mean square error (NRMSEP) proposed by Feten *et al.* (2005) is used as an error measure:

$$\mathrm{NRMSEP}_k = \sqrt{\frac{1}{g}\sum_{j\in G}\frac{\sum_{i\in O_j}(x_{ij} - \hat{x}_{ij})^2}{o_j s^2_{x_j}}} \tag{4}$$

where G is the set and g the number of incomplete variables, O_j is the set and o_j the number of missing observations in variable j and $s^2_{x_j} = \sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2/(n-1)$ is the variance associated with a given variable.

This error function normalizes the error of prediction by the variance observed for the predicted variable. Hence, the NRMSEP will be lower if the internal variance is larger. This assumption is rational for noisy data. However, if the number of samples is small, the variance may be an unstable criterion and Q^2 should be used instead, also if variance normalization was applied. An advantage of the NRMSEP when used together with cross validation is, that for imputation by averages, it roughly equals $\sqrt{nObs/(nObs-1)}$, where *nObs* is the number of observations. Thus, if the NRMSEP exceeds 1, then missing value imputation by mean substitution is preferable.

5 IMPLEMENTATION

5.1 Availability

The *pcaMethods* package is written in the R language developed within the R Project for Statistical Computing, R Development Core Team (2004). It is part of the Bioconductor suite of packages related to life science applications, Gentleman *et al.* (2004). The version presented here is part of the 2.0 (development) release.

5.2 Data formats

Input data must be provided as an `exprSet` object or as a numerical matrix or data frame already read into R. All methods return a common object called `pcaRes` providing maximum interoperability.

5.3 Visualization

The package also offers methods for visualization of the results, e.g. for plotting an arbitrary number of scores/loadings side by side.

6 WEB INTERFACE

The Bioinformatics group at the Max Planck Institute for Molecular Plant Physiology provides MetaGeneAlyse (MGA), Daub *et al*. (2003). MGA is a web-based tool for the visualization and analysis of large-scale transcript and metabolite profile data sets, available under http://metagenealyse.mpimp-golm.mpg.de/. We included *pcaMethods* in MGA, the user may now perform PCA on incomplete data and estimate missing values via an easy-to-use web interface. Hereby, the user can choose between PPCA, BPCA, NipalsPCA and SVDimpute.

ACKNOWLEDGEMENTS

The authors thank Shigeyuki Oba and Jakob Verbeek for providing the Matlab scripts BPCA and PPCA are based on. Funding to pay the Open Access charges was provided by the Max Planck Society.

Conflict of Interest. none declared.

REFERENCES

Alter,O. *et al*. (2000) Singular value decomposition for genome wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, **97**, 10101–10106.

Bishop,C.M. (1999) Variational principal components. In *IEE Conference Publication on Artificial Neural Networks*, pp. 509–514.

Daub,C.O. *et al*. (2003) MetaGeneAlyse: analysis of integrated transcriptional and metabolite data. *Bioinformatics*, **19**, 2332–2333.

Feten,G. *et al*. (2005) Prediction of missing values in microarray and use of mixed models to evaluate the predictors. *Stat. Appl. Genet. Mol. Biol.*, **4**, Article 10.

Gentleman,R. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Huber,W. *et al*. (2005) An introduction to low-level analysis methods of DNA microarray data. *Bioconductor ProjectWorking Papers*, Working Paper 9.

Kim,H. *et al*. (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **21**, 187–198.

Oba,S. *et al*. (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

R Development Core Team (2004) *R: A Language and Environment for Statistical Computing*. Manual of the R Foundation for Statistical Computing, Vienna, Austria.

Scholz,M. *et al*. (2005) Non-linear pca: a missing data approach. *Bioinformatics*, **21**, 3887–3895.

Sehgal,M.S.B. *et al*. (2005) Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, **21**, 2417–2423.

Troyanskaya,O. *et al*. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Whitfield,M.L. *et al*. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

Wold,H. (1966) Estimation of principal components and related models by iterative least squares. In Krishnaiah,P.R. (ed.), *Multivariate Analysis*, Academic Press, NY, pp. 391–420.