

METHODOLOGY ARTICLE

Open Access

Probabilistic principal component analysis for metabolomic data

Gift Nyamundanda¹, Lorraine Brennan², Isobel Claire Gormley^{1*}

Abstract

Background: Data from metabolomic studies are typically complex and high-dimensional. Principal component analysis (PCA) is currently the most widely used statistical technique for analyzing metabolomic data. However, PCA is limited by the fact that it is not based on a statistical model.

Results: Here, probabilistic principal component analysis (PPCA) which addresses some of the limitations of PCA, is reviewed and extended. A novel extension of PPCA, called probabilistic principal component and covariates analysis (PPCCA), is introduced which provides a flexible approach to jointly model metabolomic data and additional covariate information. The use of a mixture of PPCA models for discovering the number of inherent groups in metabolomic data is demonstrated. The jackknife technique is employed to construct confidence intervals for estimated model parameters throughout. The optimal number of principal components is determined through the use of the Bayesian Information Criterion model selection tool, which is modified to address the high dimensionality of the data.

Conclusions: The methods presented are illustrated through an application to metabolomic data sets. Jointly modeling metabolomic data and covariates was successfully achieved and has the potential to provide deeper insight to the underlying data structure. Examination of confidence intervals for the model parameters, such as loadings, allows for principled and clear interpretation of the underlying data structure. A software package called MetabolAnalyze, freely available through the **R** statistical software, has been developed to facilitate implementation of the presented methods in the metabolomics field.

Background

Metabolomics is the term used to describe the study of small molecules or metabolites present in biological samples. Examples of such metabolites include lipids, amino acids, bile acids, keto-acids. Studies of the concentration levels of these molecules in biological samples aim to enhance understanding of the effect of a particular stimulus or treatment [1-3]. The most commonly applied analytical technologies to metabolomic studies are nuclear magnetic resonance spectroscopy (NMR) [4] and mass spectrometry (MS) [5]. With respect to NMR-based metabolomics the data are usually in the form of spectra which are binned into regions of a specified width. Typically, the data generated by these methods are large and complex. Firstly, the number of observations n in metabolomics

experiments is typically much less than the number of peaks (or variables) p in a spectrum, $n \ll p$. In such a situation, the application of standard parametric statistical methods such as regression is not straight forward as there are insufficient data for parameter estimation. Secondly, many metabolites may not have any relationship with the trait under study and they can induce variation which is not relevant, hampering comprehensive data analysis [6]. In view of these difficulties, when analyzing metabolomic data there is a genuine need for multivariate dimension reducing techniques which can take into account the complexities of the data and expose any underlying relationships. Principal components analysis (PCA) [7] is probably the most widely used technique for analyzing metabolomic data [8]. The popularity of PCA in metabolomics is due to the fact that it is a simple non-parametric method which can project the NMR or MS spectra into lower dimensional space, revealing inherent data structure, and providing a

* Correspondence: Claire.Gormley@ucd.ie

¹School of Mathematical Sciences, University College Dublin, Ireland
Full list of author information is available at the end of the article

reduced dimensional representation of the original data. Despite its widespread use in metabolomics, PCA has several shortcomings. Most significantly, PCA does not have an associated probabilistic model, which makes assessing the fit of PCA to the data difficult and limits the potential to extend the scope of application of PCA. Additionally, PCA can fail to reveal underlying groups of subjects in the data, therefore providing a spurious view of the underlying data structure [9,10]. Other limitations include the inability of PCA to deal with missing data appropriately [11].

Such limitations can be addressed by deriving PCA from a probabilistic framework resulting in probabilistic PCA (PPCA) [12]. The merits of a probabilistic approach are manifold. Firstly, the maximum likelihood solution of the PPCA model corresponds to PCA and hence the familiar characteristics (such as principal scores and loadings) of PCA are retained. With regard to model estimation, closed form solutions for parameter estimates exist. Additionally, the Expectation Maximization (EM) algorithm [13] can be employed to estimate the parameters of the PPCA model. The EM algorithm is computationally efficient and also has the capacity to deal with missing data [14]. The probability density based approach facilitates comparison of different PPCA models to determine the 'best' model for the data using statistically principled approaches. In practice, this allows selection of the number of required principal components in a statistically valid manner. Given the probabilistic footing of the PPCA model, the Bayesian inferential framework [15,16] can be employed for inference, facilitating the inclusion of any prior information the practitioner may have [17].

Perhaps the key advantage of approaching PCA from a probabilistic modeling point of view is the facility to assess the uncertainty associated with the resulting model output. In practice, this translates to the ability to examine the uncertainty in an observations' principal component score, or in the estimated loadings. In this article, the construction of confidence intervals for model parameters such as loadings using the jackknife method is illustrated. Thus a more principled and clear insight to the principal component solution is available under the probabilistic approach than under the traditional covariance matrix decomposition approach.

In general, metabolomics studies generate metabolomic data in addition to other phenotypic data, examples of which include age, gender and bmi (body mass index) in the case of human based studies. Inclusion of these covariates in multivariate models when analyzing metabolomic data is highly desirable in order to allow a comprehensive analysis of the data. In this article a novel extension of the PPCA model called probabilistic

principal components and covariates analysis (PPCCA) is introduced which incorporates covariates into the model and facilitates joint modeling of metabolomic data and covariates. Another crucial benefit of the probability density based approach is that a collection of PPCA models can be combined to form a mixture of PPCA models (MPPCA) for nonlinear modeling purposes [18]. A mixture of PPCA models can be used to cluster subjects whilst facilitating dimensionality reduction of metabolomic data. This model is closely related to the mixture of factor analyzers used to cluster microarray expression data in [19]. The application of MPPCA analysis to metabolomic data is presented and highlights the danger of assuming a single underlying PPCA model in cases where (often unknown) groups of observations are present.

These statistical methods (PPCA, PPCCA and MPPCA) are illustrated through an application to two metabolomic datasets. A software package called *MetabolAnalyze* [20], freely available through the **R** statistical software [21], has been developed to facilitate implementation of the presented methods in the metabolomics community and elsewhere.

Methods

Probabilistic PCA (PPCA) is a probabilistic formulation of PCA based on a Gaussian latent variable model and was first introduced by Tipping and Bishop in 1999 [12]. The PPCA model reduces the dimension of high-dimensional data by relating a p -dimensional observed data point to a corresponding q -dimensional latent variable through a linear transformation function, where $q \ll p$. Given the statistical model underpinning PPCA, extensions of the model are possible, and a wealth of statistical tools can be utilized. Such extensions and tools are detailed in what follows.

Probabilistic Principal Components Analysis

Let $\underline{x}_i = (x_{i1}, \dots, x_{ip})^T$ be an observed set of variables (eg. an NMR spectrum) for observation i and $\underline{u}_i = (u_{i1}, \dots, u_{iq})^T$ be a latent variable corresponding to observation i in the latent, reduced dimension space. In terms of traditional PCA, \underline{u}_i can be viewed as the principal score of subject i . The PPCA model can be expressed as follows

$$\underline{x}_i = \mathbf{W}\underline{u}_i + \underline{\mu} + \underline{\epsilon}_i$$

where \mathbf{W} is a $p \times q$ loadings matrix, $\underline{\mu}$ is a mean vector and $\underline{\epsilon}_i$ is multivariate Gaussian noise for observation i , i.e. $p(\underline{\epsilon}_i) = MV N_p(\underline{0}, \sigma^2 \mathbf{I})$ where \mathbf{I} denotes the identity matrix. The latent variable \underline{u}_i is also assumed to be multivariate Gaussian distributed, $p(\underline{u}_i) = MV N_q(\underline{0}, \mathbf{I})$. The

conditional distribution of the observed data given the latent variable can then be expressed as

$$p(\underline{x}_i | \underline{u}_i) = MVN_p(\mathbf{W}\underline{u}_i + \underline{\mu}, \sigma^2 \mathbf{I}). \quad (1)$$

The distribution of the observed data, $p(\underline{x}_i)$, also known as the predictive distribution, can be derived from the convolution of $p(\underline{u}_i)$ and $p(\underline{x}_i | \underline{u}_i)$ giving

$$p(\underline{x}_i) = MVN_p(\underline{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2 \mathbf{I}).$$

In contrast to the more conventional view of PCA which is a mapping from the high dimensional data space to a low dimensional latent space, the PPCA framework is based on a mapping from a latent space to the data space. The observed data \underline{x}_i is generated by first drawing a value for the latent variable \underline{u}_i from its unit variance multivariate Gaussian distribution, $p(\underline{u}_i)$. The observed variable \underline{x}_i is then sampled, conditioning on the generated value for \underline{u}_i , from the isotropic distribution defined in (1).

Any observed data point \underline{x}_i can be represented in a latent space by its corresponding q -dimensional latent variable \underline{u}_i . The distribution of the latent variable given the observed data can be derived using Bayes' Theorem to give

$$p(\underline{u}_i | \underline{x}_i) = MVN_q(\mathbf{M}^{-1}\mathbf{W}^T(\underline{x}_i - \underline{\mu}), \sigma^2 \mathbf{M}^{-1}) \quad (2)$$

where \mathbf{M} is a $q \times q$ matrix defined as $\mathbf{M} = \mathbf{W}^T \mathbf{W} + \sigma^2 \mathbf{I}$. A key benefit of the PPCA approach is that, not only is an estimate of the location of each observation in the lower dimensional space available through its expected value $\mathbb{E}(\underline{u}_i) = \mathbf{M}^{-1}\mathbf{W}^T(\underline{x}_i - \underline{\mu})$, an estimate of its associated uncertainty is available through the covariance matrix $\sigma^2 \mathbf{M}^{-1}$ in (2). This is in contrast to conventional PCA where the lower dimensional location (i.e. the score) of an observation is available, but the uncertainty associated with it is not. The parameters (\mathbf{W} , $\underline{\mu}$ and σ^2) of the PPCA model can be estimated using maximum likelihood. Maximizing the (log) likelihood function with respect to model parameters is non-trivial; in [12] it is demonstrated that the estimates do however have closed form solutions. Crucially, the log likelihood of the PPCA model is maximized when the columns of \mathbf{W} span the principal subspace of conventional PCA [12]. Thus the maximum likelihood estimate (MLE) of the loadings matrix $\hat{\mathbf{W}}$ in PPCA corresponds exactly to the loadings matrix in conventional PCA. Hence the model output in PPCA is exactly that obtained in conventional PCA, but with the additional advantages of uncertainty assessment and potential model extensions.

In this article maximum likelihood estimates of the model parameters are derived via the EM algorithm [13] because of its stability and widespread applicability. The EM algorithm is typically used to compute MLEs in probabilistic models when the model depends on unobserved variables or when some data are missing. The algorithm alternates between two steps until convergence: the expectation (E) step and the maximization (M) step. In the E-step, the expected values of the latent variables are estimated given the observed data and the current estimates of the model parameters. In the M-step, the model parameters are re-estimated by maximizing the log likelihood function using the expected values of the latent variables derived in the previous E-step. The two steps are repeated until convergence. Many convergence assessment criteria are available; some criteria are based on log likelihood gain between iterations while others use an estimate of the converged log likelihood value as a basis for stopping. Here Aitken's acceleration procedure [22] is used for convergence assessment. Specific details of the EM algorithm for the PPCA model are given in Additional File 1. An implementation of the algorithm is available in the package MetabolAnalyze through the R statistical software [21].

Probabilistic Principal Components and Covariates Analysis

With its basis in a statistical model, the PPCA model can be extended in several ways. Given the availability and relevance of subject covariates in metabolomic studies, here the PPCA model is extended to facilitate joint modeling of covariates and metabolomic data, giving the probabilistic principal components and covariates analysis (PPCCA) model. This novel model extension is achieved by assuming that the latent variable distribution for observation i follows a multivariate Gaussian distribution centered at $\underline{\delta}_i$ rather than at the origin, i.e. $p(\underline{u}_i) = MVN_q(\underline{\delta}_i, \mathbf{I})$, where

$$\underline{d}_i = \mathbf{a}\underline{c}_i = \begin{bmatrix} \underline{a}_1^T \\ \vdots \\ \vdots \\ \underline{a}_q^T \end{bmatrix} \begin{bmatrix} 1 \\ \vdots \\ c_{iL} \end{bmatrix} \quad (3)$$

Here \mathbf{a} is a $q(L+1)$ matrix of parameters which capture the relationship between the latent variable and the covariates and \underline{c}_i is a $(L+1)$ vector of an intercept term and the L covariates of observation i . The motivation behind this model extension is that a subject's covariates may influence their location in the principal subspace. Conditional on this location, the observed data point is then generated as in (1). Note that through the model

definition (3) the covariates may have different effects on each of the dimensions of the principal subspace through the parameter vectors $\underline{\alpha}_1, \dots, \underline{\alpha}_q$.

Under the PPCCA model, the conditional distribution of \underline{x}_i given \underline{u}_i is the same as that of the PPCA model given in (1). The predictive distribution $p(\underline{x}_i)$ differs from that of the PPCA model and is now defined as

$$p(\underline{x}_i) = MVN_p(\mathbf{W}\underline{\delta}_i + \underline{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})$$

The posterior distribution of the latent variable \underline{u}_i given the observed data \underline{x}_i is also affected by the inclusion of covariates and is defined to be

$$p(\underline{u}_i | \underline{x}_i) = MVN_q(\mathbf{M}^{-1}[\mathbf{W}^T(\underline{x}_i - \underline{\mu}) + \sigma^2\underline{\delta}_i], \sigma^2\mathbf{M}^{-1})$$

The location (or score) of observation i in the latent space is given by $\mathbb{E}(\underline{u}_i) = \mathbf{M}^{-1}[\mathbf{W}^T(\underline{x}_i - \underline{\mu}) + \sigma^2\underline{\delta}_i]$, which depends on both the data point \underline{x}_i and the covariates of observation i through $\underline{\delta}_i$. Thus when representing an observation in a reduced dimensional space the PPCCA model takes account of both the spectra data and the associated covariates. Deeper insight to the true underlying structure of the data is then feasible as possibly influential external factors are explicitly modeled.

The effect of covariates on the q th latent dimension can be explored through examination of the estimated regression parameter vector $\underline{\alpha}_q = (\alpha_{q0}, \dots, \alpha_{qL})^T$; these parameters can be interpreted within the context of the problem to provide insight to the type and strength of influence some covariates may have on the (often interpretable) latent dimensions.

Parameter estimation for the PPCCA model can be achieved via an efficient EM algorithm; Specific details of the EM algorithm for the PPCCA model are given in Additional File 1. An implementation of the algorithm is available in the package MetabolAnalyze through the R statistical software [21].

Mixtures of Probabilistic Principal Components Analysis Models

The models discussed so far assume that the association between the observed data and the latent variable is linear. This assumption can be inadequate in a situation where the observations in the data set have an underlying group structure. In such cases the linearity assumption may not reveal all of the internal structures of the data [23]. Standard PCA also suffers from this phenomenon.

In many high throughput technologies which result in high dimensional data, interest often lies in identifying underlying sub groups within a set of observations. Exploring high dimensional data with underlying

nonlinear structures therefore requires modeling attention. Employing a single PPCA model to model such data is not adequate, since PPCA provides a globally linear projection of the data. A collection of single PPCA models can be combined to obtain a mixture of probabilistic principal components analysis models (MPPCA) [18] which clusters observations into groups and reduces data dimension.

Under a MPPCA model, with probability π_g observation i is modeled as

$$\underline{x}_i = \mathbf{W}_g \underline{u}_{ig} + \underline{\mu}_g + \underline{\epsilon}_{ig}$$

Here \mathbf{W}_g and $\underline{\mu}_g$ are a $p \times q$ loadings matrix and the mean respectively for group g , and $\underline{\epsilon}_{ig}$ is a multivariate Gaussian noise process for observation i , given that i is a member of group g . The latent location for observation i , given that i is a member of group g , is denoted \underline{u}_{ig} . That is, with probability π_g observation i is modeled using a PPCA model with group Specific parameters.

Observation i is assumed to have been drawn from a finite mixture distribution with G components (or groups), i.e.

$$p(\underline{x}_i) = \sum_{g=1}^G \pi_g p(\underline{x}_i | \underline{\mu}_g, \Sigma_g)$$

where $p(\underline{x}_i | \underline{\mu}_g, \Sigma_g)$ is a PPCA model for group g with mean parameter $\underline{\mu}_g$ and covariance matrix $\Sigma_g = \mathbf{W}_g \mathbf{W}_g^T + \sigma^2\mathbf{I}$. The mixing proportion π_g denotes the probability of an observation belonging to group g ($0 \leq \pi_g \leq 1$ and $\sum_{g=1}^G \pi_g = 1$). Note that for reasons of parsimony the error covariance ² has been constrained to be equal for all groups [24].

The MPPCA model can be fitted using a two stage EM algorithm called the Alternating Expectation Conditional Maximization (AECM) algorithm [25]. For clarity, the details of the AECM algorithm for the MPPCA model are deferred to Additional File 1. Under the MPPCA model, in addition to the latent location variable, the unobserved group membership of each observation is also viewed as a latent variable. Specifically, for each observation, a latent binary vector $\underline{z}_i = (z_{i1}, \dots, z_{iG})^T$ is imputed where $z_{ig} = 1$ if observation i belongs to group g and 0 otherwise. At convergence of the AECM algorithm the estimate \hat{z}_{ig} is the posterior probability of observation i belonging to group g . The MPPCA model clusters observations into groups by assigning them to the group to which they have highest posterior probability of membership. Thus clustering of observations and

dimension reduction, through the use of principal components, are achieved simultaneously.

Model Selection

A crucial advantage of working within a probabilistic framework is that a wealth of statistically based model selection tools can be utilized. This allows the determination of the “best” statistical model for the data, i.e. the optimal number of principal components q to retain and, in the case of the MPPCA model, the optimal value of G . Such choices are made on the basis of statistical principles instead of using traditional ad-hoc approaches, such as a scree plot.

The Bayesian Information Criterion (BIC) [26] is a popular model selection tool. The BIC is defined to be

$$\text{BIC} = 2l - K \ln(n) \quad (4)$$

where l is the maximum log likelihood value, K is the number of free parameters in the model and n is the number of observations. The model which yields highest BIC value is deemed the optimal model.

The BIC can be viewed as a criterion that rewards model fit (through the first term in (4)) but penalizes model complexity (through the second term in (4)). The penalization in the BIC is much stronger than that of the widely used Akaike Information Criterion [27] and typically selects more parsimonious models. Within the context of mixture models, the BIC has been widely employed, see [10,28,29].

Despite the tendency for the BIC to select parsimonious models, in high dimensional data settings it often exhibits noisy behaviour—the BIC can be undefined or perform poorly due to the occurrence of singularities for some starting values of the EM algorithm, for some models (i.e. different values of q), or for some numbers of groups (in the case of the MPPCA model). Additionally, diagnosing convergence of the EM algorithm in highly-parameterized models can be difficult, leading to noisy BIC values.

To eradicate this issue, here a regularized version of the BIC [30,31] is employed as a model selection tool. This modified version of the BIC evaluates the likelihood at the maximum a posteriori (MAP) estimator instead of the MLE. The MAP estimator is derived within the EM algorithm framework where a conjugate prior is included, and the convolution of the likelihood and prior are maximized at the M step. Here, a conjugate inverse gamma prior on σ^2 is employed throughout—the reported BIC values are based on the MAP estimate for σ^2 rather than on the MLE. Further details are provided in Additional File 1. This approach avoids singularities, and performs similarly to the BIC when such issues are absent. It also has the effect of

smoothing noisy behavior of the BIC, which is often observed when parameter estimation is unstable.

Jackknife resampling

The loadings of any probabilistic principal components based model can be used to identify variables responsible for the structure in the data. Rather than examining the (typically large number of) point estimates of the loadings alone, a gauge of the uncertainty associated with the loadings can be obtained through estimation of their standard errors.

Here the jackknife resampling method [32,33] is implemented. Standard errors are estimated by recomputing the loadings (i.e. re fitting the relevant PPCA model) with the i th observation removed from the dataset giving the loadings matrix \mathbf{W}^{-i} , for $i = 1, \dots, n$. The jackknife standard error for the j th loading on the k th principal component is then estimated as

$$SE(\hat{w}_{jk}) = \sqrt{\frac{n-1}{n} \sum_{i=1}^n [w_{jk}^{-i} - \bar{w}_{jk}]^2}$$

$$\text{Where } \bar{\mathbf{W}} = \frac{1}{n} \sum_{i=1}^n \mathbf{W}^{-i}.$$

The standard errors can then be used to compute 95% confidence intervals (CIs) for the individual loadings. Such CIs can be used to identify loadings which differ significantly from zero on a selected principal component in the optimal model. Those variables whose loadings are significantly different from zero relate to the variables responsible for the structure within the data. This approach therefore provides a sparse list of relevant variables.

Computation time is often an issue when using the EM algorithm to fit statistical models, and employing the jackknife technique to obtain standard errors would clearly exacerbate this problem. In practice computation times are considerably reduced by choosing good starting values for the algorithm for each of the n runs when using the jackknife. Here the maximum likelihood estimates of the model parameters when fitted to the entire data set are employed as starting values for each of the jackknife runs, considerably reducing computational costs.

Metabolomic datasets

The datasets used here were derived from a study previously reported [34]. Briefly, animals were randomly assigned to two treatment groups and treated with pentylenetetrazole (PTZ, treated group) or saline (control group) for a period of 4 weeks. A third treatment group consisted of animals who received one injection only and these data are not used within this paper. Throughout

the treatment period urine was collected from the animals in collection tubes containing 1% sodium azide surrounded by ice. The animals had no access to food during this time but had free access to water. At the end of the treatment period brain regions were isolated and metabolites extracted as previously described [34]. NMR spectra were acquired and the spectra were integrated into bin regions of 0.04 ppm using AMIX (Bruker) excluding the water regions (4.0-6.0 ppm).

The urine dataset used herein was constructed from NMR spectra acquired from urine collected on day ten of the study; it consists of 18 spectral profiles (from 9 treated and 9 control animals) over 189 spectral bin regions.

The brain dataset comes from animals in the control group only with spectra acquired from tissues from four brain regions: the pre-frontal cortex, hippocampus, cerebellum and brainstem. In total, there are 33 spectral profiles over 164 spectral bin regions.

Results

Application of PPCA to metabolomic data

To explore the effect of treatment with PTZ, a PPCA model was fitted to the urinary metabolomic data. Parameter estimation was achieved via the EM algorithm. A number of PPCA models with varying numbers of principal components was fitted; a modified Bayesian Information Criterion (BIC) was used to aid selection of the optimal model (that is, the required number of principal components q) where a higher value of the criterion indicates a preferable model.

The fitted PPCA model is illustrated in Figure 1. Figure 1A shows that the modified BIC is maximized by a model with five principal components (PCs); such a model explains 84% of the variation within the urine spectra data. Should the principal scores be required as a reduced dimensional input to further statistical modeling of the data, the modified BIC clearly indicates that a five dimensional representation is optimal. Overall, it represents an unambiguous means of selecting the optimal number of principal components. For clarity, the scores and loadings on the first two principal dimensions are illustrated. The scores plot (Figure 1B) reveals that grouping of animals with respect to their treatment status is evident on the first principal component. The 95% posterior sets are small, indicating little uncertainty in the scores. The associated loadings (which in turn correspond to metabolites) are presented in Additional file 2.

The 95% confidence intervals (CIs) of the individual loadings are estimated using the jackknife technique – these CIs are used to identify loadings which are significantly different from zero. Of the 189 spectral bins in the urine spectra dataset, 86 have loadings on PC 1 significantly different from zero.

In order to further identify metabolites that strongly influence the separation of the treatment and control groups and which will serve as markers for treatment response, significant loadings greater (in absolute value) than 0.8 were selected. The cutoff value of 0.8 was chosen by examining a frequency plot of the (absolute) loading values of the significant spectral bins. A region in the plot in which the number of selected significant spectral bins drops steeply while the loading values remain relatively constant may be used as an indication of a cutoff point. In the current analysis, the plot (in Additional file 2) drops at the value 0.8.

Seventeen spectral bins had (absolute) loading values greater than the cutoff point of 0.8 and are illustrated in Figure 1C, along with their 95% CIs. Further analysis was performed to identify which of these bin regions differ significantly between the two treatment groups (using a t-test, correcting for multiple testing). Of the seventeen spectral bins, ten had signal intensities which were significantly different between the two groups. Included in these changes were bin regions due to the drug administered (1.74 ppm, 1.86 ppm, 1.9 ppm, 2.22 ppm, and 3.1 ppm). Taurine levels (3.30 ppm and 3.46 ppm), dimethylamine (2.74 ppm) and one unassigned peak (6.06 ppm) were significantly lower in the treated group, while isocitrate levels were significantly higher (2.98 ppm) in the treated group.

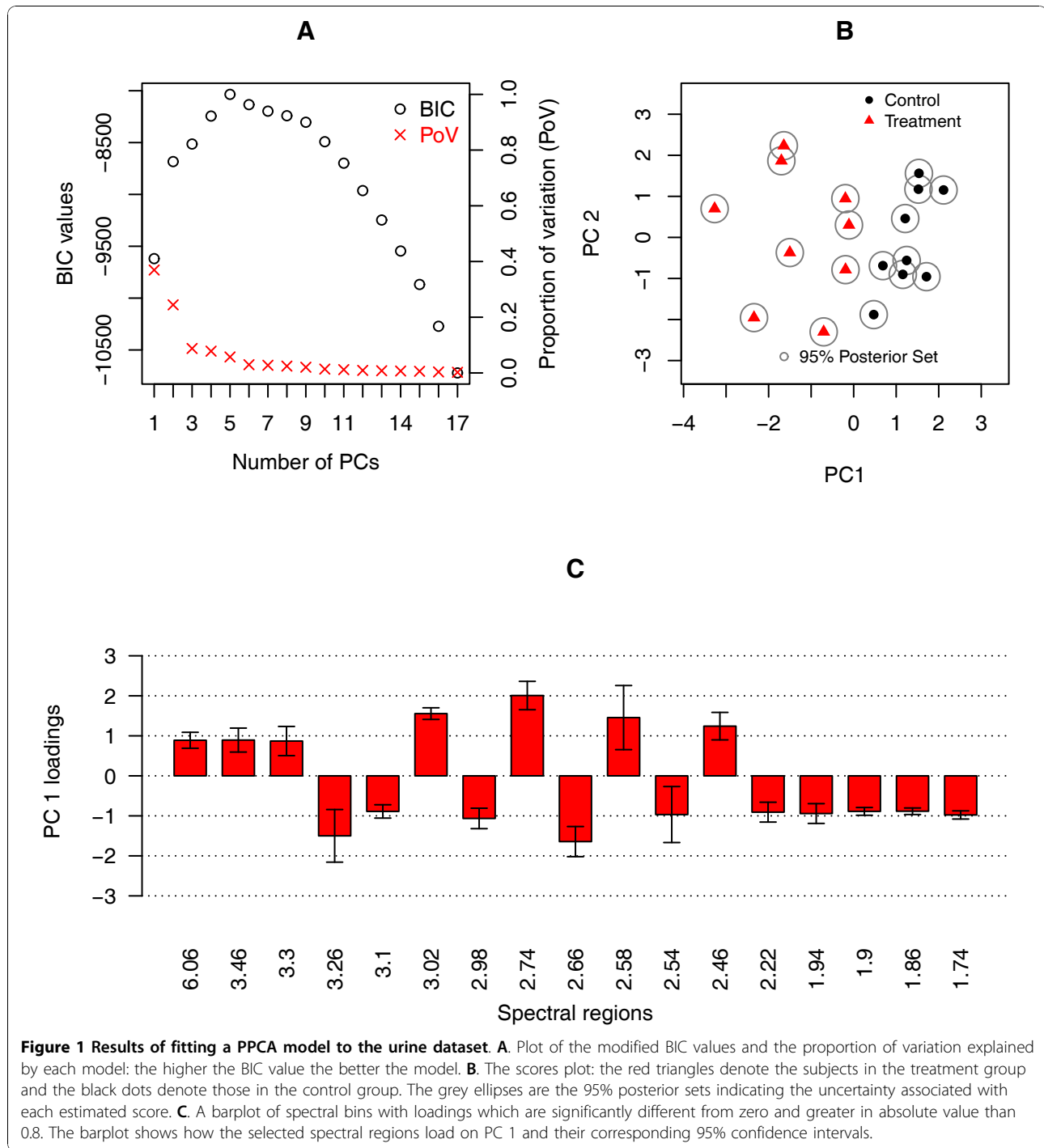
Application of PPCCA to metabolomic data

In addition to the urine metabolomic data the weight of each of the eighteen animals was recorded. Inclusion of this covariate in the analysis was achieved using the PPCCA model. Specifically, the covariate is incorporated to the PPCA model by allowing it to influence the score of each animal in the principal subspace. From (3), the expected value of the score of animal i , δ_i , is modeled as a linear function of its covariate (i.e. weight):

$$\delta_i = \alpha C_i = \begin{bmatrix} \alpha_{10} & \alpha_{11} \\ \vdots & \vdots \\ \alpha_{q0} & \alpha_{q1} \end{bmatrix} \begin{bmatrix} 1 \\ c_{i1} \end{bmatrix}$$

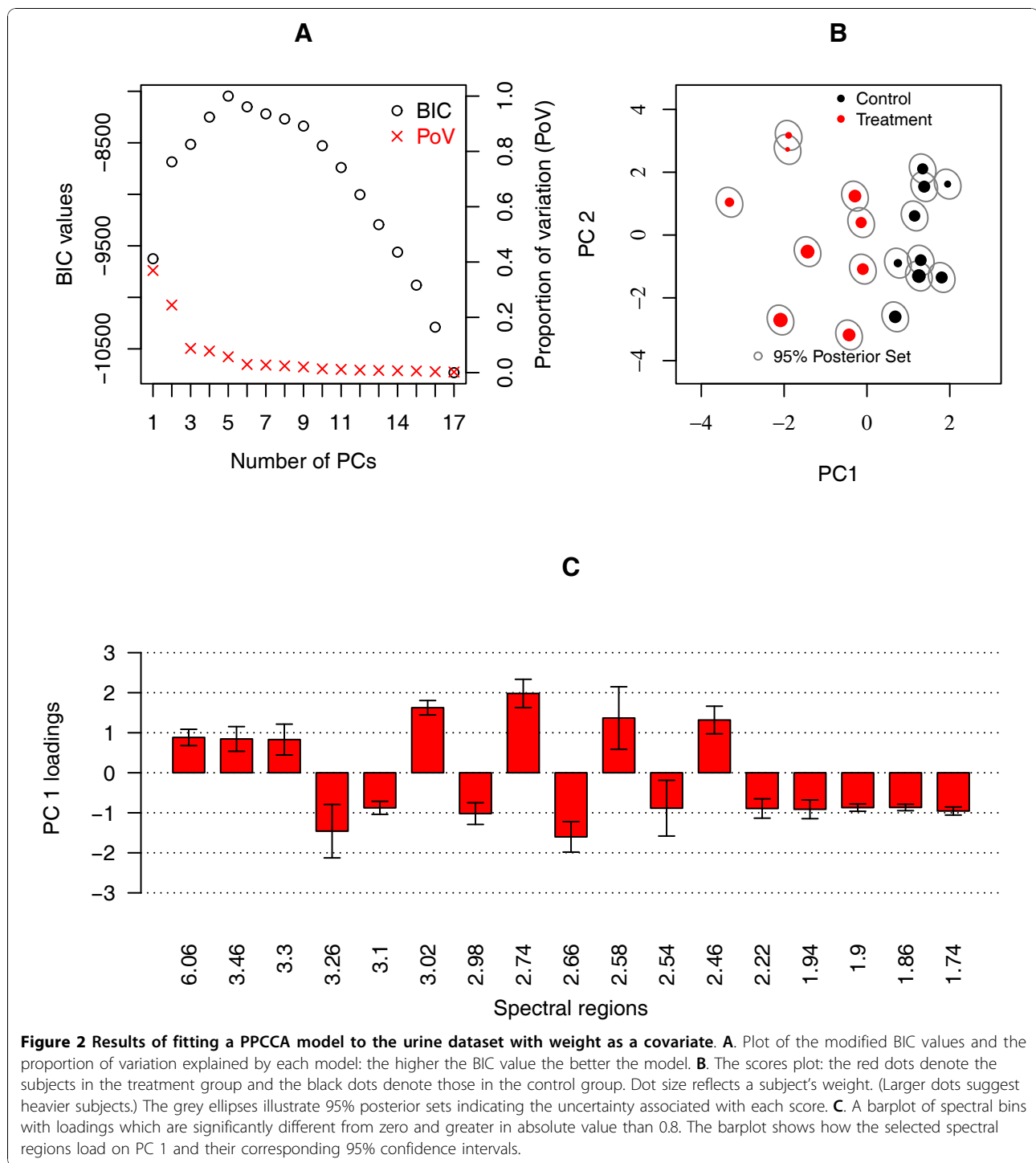
where $\alpha_{10}, \dots, \alpha_{q0}$ are intercept parameters for each dimension of the principal subspace and $\alpha_{11}, \dots, \alpha_{q1}$ are slope parameters for the weight covariate (denoted c_{i1} for animal i) for each dimension of the principal subspace. Hence the influence of an animal's weight on their score in the principal subspace is explicitly modeled and can be interpreted through the parameter matrix α .

The EM algorithm was employed to fit the PPCCA model to the urine spectra and the weight covariate. Figure 2A shows that the modified BIC is maximized by



a model with five principal components (PCs). Examination of the (two dimensional for clarity) scores plot (Figure 2B) and the loadings plot (in Additional file 2) indicates that while the parameter estimates on the first principal component dimension remain relatively unchanged from the fitted PPCA model, the estimates on the second principal component dimension differ slightly. The general structure of the scores and loadings

remains relatively unchanged suggesting that the animal's weights are not influencing the separation between treated and control animals on PC1. Additionally, it is apparent (Figure 2B) that the uncertainty associated with the estimated scores increases under the fitted PPCA model. Selection of inferential bin regions identified the same seventeen regions as those obtained using the PPCA model (Figure 2C).



The influence of the weight covariate can be quantified by examining the associated regression parameter matrix, detailed in Table 1. Standard errors of the PCCA regression parameters were estimated using jackknife resampling. The parameter estimates and the associated 95% CIs show that an animal's weight has a significant negative effect on the second principal

component only ($\alpha_{20} = 4.08$ and $\alpha_{21} = -6.28$) and is not contributing to the treatment effect observed on PC1.

Application of MPPCA to metabolomic data

MPPCA was applied to the brain metabolomic data in order to determine the number of inherent groups (if any) in the data. This application also illustrates the

Table 1 Regression parameter estimates from the fitted PCCA model

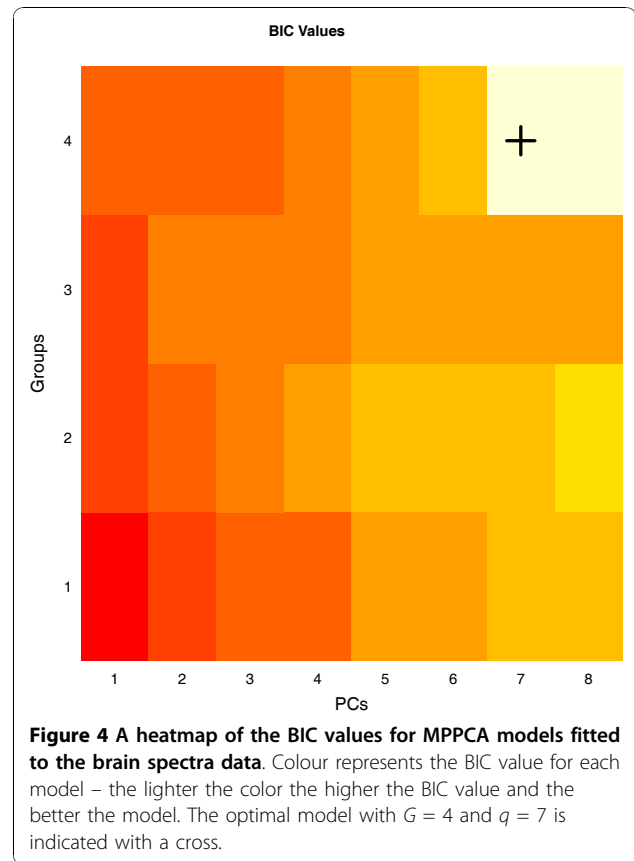
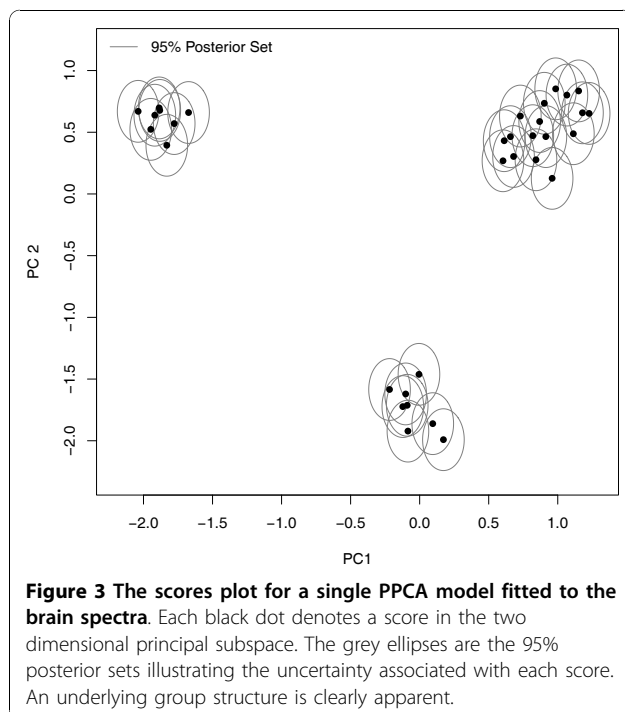
	Intercept	Slope
PC 1 (α_1)	-0.87 (-3.34, 1.61)	1.34 (-2.47, 5.15)
PC 2 (α_2)	4.08 (2.44, 5.72)	-6.28 (-8.57, -3.99)
PC 3 (α_3)	0.86 (-0.07, 1.78)	-1.32 (-2.77, 0.14)
PC 4 (α_4)	0.18 (-1.03, 1.38)	-0.28 (-2.13, 1.58)
PC 5 (α_5)	0.03 (-2.20, 2.26)	-0.04 (-3.32, 3.23)

95% CIs are given in parentheses. Those estimates significantly different from zero are highlighted in bold.

pitfalls of assuming a single PCCA model when exploring heterogeneous data.

The scores plot resulting from fitting a single PCCA model to the brain spectral data is illustrated in Figure 3-it is immediately clear that there is a grouping (or clustering) structure within the set of 33 observations. With such a strong clustering structure it would seem extremely unlikely that the same set of principal axes would be relevant to each group. Fitting a MPPCA model simultaneously clusters the data into groups and reduces the dimension of the data within each group.

Thirty-two different MPPCA models were fitted to the 33 spectra by varying the number G of PCCA models in the mixture (each representing a group) from 1 to 4 and the number of principal components q from 1 to 8. Figure 4 is a heat map illustrating the modified BIC value for each fitted MPPCA model; the BIC suggests that the optimal model is the MPPCA model with four groups and seven principal components. This model can be



used to cluster the observations into four groups and to visualize the data in each group within its principal subspace, hence exploring the structure relevant to each group. This method provides an objective means of identifying groups within the data.

In this illustrative example of the clustering and dimension reducing ability of the MPPCA model, the origin of each of the spectra was known. Thus, treating the brain region of origin as an observations' 'true' group, the clustering performance of the method can be assessed, where each observation is assigned to the group for which they have largest posterior probability of membership, under the optimal MPPCA model. In the current example, the model correctly clusters all observations into their brain group of origin (Table 2).

Table 2 Cross tabulation of the group membership of subjects based on the estimated MPPCA model and the brain region of origin

	Cerebellum	Brain stem	Pre-frontal cortex	Hippocampus
Group 1	8	0	0	0
Group 2	0	8	0	0
Group 3	0	0	9	0
Group 4	0	0	0	8

Furthermore, the model correctly separates the prefrontal cortex and hippocampus samples which overlap in the scores plot under the PPCA model (Figure 3).

The usefulness of this approach in the metabolomics field lies in its application to studies where the number of underlying groups and the group membership of each subject is unknown-the MPPCA model can be used to identify G and the members of each group within a study. Examples of such studies include the identification of disease phenotypes or treatment responsive phenotypes.

Conclusions

Principal components analysis is the dominant statistical method currently employed within the field of metabolomics. Principal components analysis has many merits and is particularly well used and understood by metabolomic researchers. However, the scope of principal components analysis is limited and extensions (to make use of additional data sources, for example) are not possible, due to the lack of an underlying statistical model. As metabolomic research becomes more prevalent and data intensive, the development of methods which retain the familiar characteristics of principal components analysis while having additional analytical properties is of immediate importance.

This article demonstrates how probability density based methods can be used in the analysis of data resulting from metabolomics studies. Probabilistic principal components analysis (PPCA), and its equivalence with traditional PCA, is introduced in [12]. Thus PPCA retains the familiar and useful properties of PCA, but is based on a flexible statistical model. Standard statistical tools are then available for use – in this article a model selection criterion is employed as a principled approach to selecting the number of principal components to retain. Additionally, uncertainty in the model estimates is assessed and standard errors are derived through the use of the jackknife technique. This provision of standard errors further aids model interpretation as inference on important model parameters such as loadings can be performed. Here, standard errors are employed to construct confidence intervals which are then used to indicate which loadings (and in turn metabolites) underlie the data structure.

In this article a novel model extension for PPCA, principal components and covariates analysis (PPCCA), is proposed. The PPCCA model offers a flexible way of including informative additional information in the PPCA model. In the context of metabolomics, this of particular interest, as covariates can be hugely influential on the metabolomic profile. Jointly modeling such data in conjunction with metabolomic data is essential to

facilitate comprehensive data analysis and understand the true metabolic changes occurring as a result of a particular stimulus. Overall, incorporating covariates in the PPCA model directly models any variation due to the covariates, thus ensuring that the principal components provide a clear picture of the structure underlying the data.

The use of a mixture of PPCA models as a simultaneous clustering and dimension reduction technique for metabolomic data was demonstrated successfully. This application represents a robust approach to identifying the number of groups within a dataset. It has great potential use within the metabolomics field for identifying metabolotypes which are responsive to certain treatments. Additionally, a mixture of probabilistic principal components and covariates analyzers is an intuitive model extension which would provide clustering, dimension reduction and covariate modeling capabilities.

Overall, the present study details novel methods for analysis of metabolomics data which are freely available through a software package called MetabolAnalyze [20]; the package provides the facility to fit a PPCA model, a PPCCA model or a MPPCA model to metabolomic data, or indeed any other suitable data set.

Additional material

Additional file 1: Statistical details of model fitting.

Additional file 2: Loadings plots and plots to aid selection of the number of influential spectral bins.

Acknowledgements

This research was supported by the Irish Research Council for Science Engineering and Technology (IRCSET) funded PhD Programme in Bioinformatics and Systems Biology in University College Dublin, Ireland (<http://bioinformatics.ucd.ie/PhD/>) and by HRB Ireland (RP/2006/117).

Author details

¹School of Mathematical Sciences, University College Dublin, Ireland. ²School of Agriculture, Food Science and Veterinary Medicine, Conway Institute, University College Dublin, Ireland.

Authors' contributions

LB was involved in the study hypothesis, data interpretation and manuscript writing. CG was involved in the study hypothesis, data interpretation and manuscript writing. NG was involved in implementation and manuscript writing. All authors read and approved the final manuscript.

Authors information

Nyamundanda Gift is PhD candidate in the PhD in Bioinformatics and Systems Biology programme in University College Dublin. Dr. Lorraine Brennan is a lecturer in Nutritional Biochemistry in the School of Agriculture, Food Science and Veterinary Medicine, Conway Institute, University College Dublin. Dr. Isobel Claire Gormley is a lecturer in Statistics in the School of Mathematical Sciences, University College Dublin.

Received: 27 August 2010 Accepted: 23 November 2010

Published: 23 November 2010

References

1. Brennan L: **Session 2: Personalised nutrition. Metabolomic applications in nutritional research.** *Proceedings of the Nutrition Society* 2008, **67**(4):404-408.
2. Keun HC: **Metabonomic modeling of drug toxicity.** *Pharmacology and Therapeutics* 2006, **109**(12):92-106.
3. Gibney MJ, Walsh M, Brennan L, Roche HM, German B, van Ommen B: **Metabolomics in human nutrition: opportunities and challenges.** *American Journal of Clinical Nutrition* 2005, **82**(3):497-503.
4. Reo NV: **Metabonomics based on NMR spectroscopy.** *Drug and Chemical Toxicology* 2002, **25**(4):375-382.
5. Dettmer K, Aronov PA, Hammock BD: **Mass spectrometry-based metabolomics.** *Mass Spectrometry Reviews* 2007, **26**:51-78.
6. van den Berg RA, Hoefsloot HJ, Westerhuis JA, Smilde AK, van der Werf MJ: **Centering, scaling, and transformations: improving the biological information content of metabolomics data.** *BMC Genomics* 2006, **7**:142.
7. Jolliffe IT: *Principal Component Analysis*. 2 edition. New York: Springer; 2002.
8. Walsh MC, Brennan L, Malthouse JPG, Roche HM, Gibney MJ: **Effect of acute dietary standardization on the urinary plasma and salivary metabolomic profiles of healthy humans.** *American Journal of Clinical Nutrition* 2006, **84**(3):531-539.
9. Chang D: **On using principal components before separating a mixture of two multivariate normal distributions.** *Applied Statistics* 1983, **32**:267-275.
10. McLachlan GJ, Peel D: *Finite Mixture Models* New York: Wiley; 2000.
11. Roweis S: **EM Algorithms for PCA and SPCA.** *Neural Information Processing Systems* 1998, **10**:626-632.
12. Tipping ME, Bishop CM: **Probabilistic principal component analysis.** *Journal of the Royal Statistical Society, Series B* 1999, **61**(3):611-622.
13. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm.** *Journal of the Royal Statistical Society, Series B (Methodological)* 1977, **39**:1-38.
14. Bishop CM: *Pattern Recognition and Machine Learning* New York: Springer; 2006.
15. Lee PM: *Bayesian Statistics. An Introduction* London: Arnold; 2004.
16. O'Hagan A, Forster JJ: *Bayesian Inference. Volume 2B of "Kendall's Advanced Theory of Statistics"*. 2 edition. London: Arnold; 2004.
17. Bishop CM: **Bayesian PCA.** In *Advances in Neural Information Processing Systems* Edited by: Kearns MS, Solla SA, Cohn DA 1999, **11**:382-388.
18. Tipping ME, Bishop CM: **Mixtures of Probabilistic Principal Component Analysers.** *Neural Computation* 1999, **11**(2):443-482.
19. McLachlan GJ, Bean RW, Peel D: **A mixture model based approach to the clustering of microarray expression data.** *Bioinformatics* 2002, **18**(3):413-422.
20. Nyamundanda G, Gormley IC, Brennan L: *MetabolAnalyze. Probabilistic latent variable models for metabolomic data* 2010, [R package version 1.0].
21. R Development Core Team: *R. A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria; 2009 [http://www.R-project.org].
22. McLachlan GJ, Krishnan T: *The EM algorithm and Extensions* New York: Wiley; 1997.
23. Bishop CM, Tipping ME: **A Hierarchical Latent Variable Model for Data Visualization.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998, **20**(3):281-293.
24. McNicholas PD, Murphy TB: **Parsimonious Gaussian mixture models.** *Statistics and Computing* 2008, **18**(3):285-296.
25. Meng XL, van Dyk D: **The EM algorithm-an old folk song sung to the fast tune (with discussion).** *Journal of the Royal Statistical Society, Series B* 1997, **4**(59):511-567.
26. Schwarz GE: **Estimating the dimension of a model.** *Annals of Statistics* 1978, **6**(2):461-464.
27. Akaike H: **A new look at the statistical model identification.** *IEEE Transactions on Automatic Control* 1974, **19**(6):716-723.
28. Fraley C, Raftery AE: **Model-Based Clustering, Discriminant Analysis, and Density Estimation.** *Journal of the American Statistical Association* 2002, **97**:611-631.
29. Gormley IC, Murphy TB: **Exploring Voting Blocs Within the Irish Electorate: A Mixture Modeling Approach.** *Journal of the American Statistical Association* 2008, **103**(483):1014-1027.
30. Fraley C, Raftery AE: **Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering.** *Journal of Classification* 2007, **24**:155-181.
31. Costa IG, Roepcke S, Hafemeister C, Schliep A: **Inferring differentiation pathways from gene expression.** *Bioinformatics* 2008, **24**(3):156-164.
32. Efron B: **Bootstrap Methods: Another Look at the Jackknife.** *The Annals of Statistics* 1979, **7**:1-26.
33. Bharat BB, Paul AT, U JL: **Use of Jackknife Resampling Techniques to Estimate the Confidence Intervals of fMRI Parameters.** *Journal of Computer Assisted Tomography* 2001, **25**:113-120.
34. Carmody S, Brennan L: **Effects of pentylentetrazole-induced seizures on metabolomic profiles of rat brain.** *Neurochemistry International* 2010, **56**(2):340-344.

doi:10.1186/1471-2105-11-571

Cite this article as: Nyamundanda et al.: Probabilistic principal component analysis for metabolomic data. *BMC Bioinformatics* 2010 **11**:571.

Submit your next manuscript to BioMed Central
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

