

Accounting for Probe-level Noise in Principal Component Analysis of Microarray Data

Guido Sanguinetti^a, Marta Milo^{a,c}, Magnus Rattray^b and Neil D. Lawrence^a

^a Department of Computer Science, Regent Court, 211 Portobello Road, Sheffield, S1 4DP, U.K.,

^b School of Computer Science, University of Manchester, Oxford Road, Manchester, M13 9PL, U.K.,

^c Department of Biomedical Science, Addison Building, Western Bank, Sheffield, S10 2TN, U.K.

ABSTRACT

Motivation: Principal Component Analysis (PCA) is one of the most popular dimensionality reduction techniques for the analysis of high dimensional datasets. However, in its standard form, it does not take into account any error measures associated with the data points beyond a standard spherical noise. This indiscriminate nature provides one of its main weaknesses when applied to biological data with inherently large variability, such as expression levels measured with microarrays. Methods now exist for extracting credibility intervals from the probe-level analysis of cDNA and oligonucleotide microarray experiments. These credibility intervals are gene and experiment specific, and can be propagated through an appropriate probabilistic downstream analysis.

Results: We propose a new model-based approach to PCA that takes into account the variances associated with each gene in each experiment. We develop an efficient EM-algorithm to estimate the parameters of our new model. The model provides significantly better results than standard PCA, while remaining computationally reasonable. We show how the model can be used to ‘denoise’ a microarray data set leading to improved expression profiles and tighter clustering across profiles. The probabilistic nature of the model means that the correct number of principal components is automatically obtained.

Availability: The software used in the paper is available from <http://www.cs.man.ac.uk/~liux/puma>. The microarray data is deposited in the NCBI database.

Contact: neil@dcs.shef.ac.uk

1 INTRODUCTION

The advent of large scale gene expression level analysis through cDNA and oligonucleotide based microarrays has led to an increasing interest in methods of downstream analysis which can succinctly summarise the information content of the data. Different approaches to analysis include hierarchical clustering, the self organising map (SOM) and support vector machines (for references and a review see Baldi and Brunak [2001, Ch. 12]). In practice, expression levels obtained through microarray analysis can be very noisy, particularly for genes which have low expression. A common weakness of the analyses mentioned above is that they cannot take advantage of any credibility intervals provided by the low level analysis that extracts the expression data. Such credibility intervals (or error bars) are becoming more commonly available for both cDNA arrays [Lawrence et al., 2003] and oligonucleotide arrays [Milo et al., 2003, Hein et al., 2005, Liu et al., 2005]. Such low level analyses are often based on probabilistic models and by continuing our downstream analysis in a probabilistic manner we can propagate the uncertainty through

the analysis. In this paper we show how this may be achieved using principal component analysis (PCA).

PCA is one of the most popular techniques for extracting information from high dimensional datasets. It seeks to explain high dimensional data by using low dimensional latent variables. This approach¹ was introduced for the analysis of microarray data by Alter et al. [2000] and has subsequently been used in a number of papers (see references in Girolami and Breitling [2004]). The motivating idea is that variability between gene expression levels should be explained by the (few) physiological processes taking place (*e.g.* response to drug treatment). The principal components (sometimes called ‘eigengenes’) would then represent a physiological process and the components of the eigengene would represent the relative weight of each gene in the process.

PCA implicitly assumes that the uncertainty associated with each gene under each condition is constant. This is often an unreasonable assumption, particularly when we are considering the logarithm of the expression levels. One popular method for dealing with this problem is to non-linearly transform the gene expression levels so that variance across experiments is comparable for each gene [Huber et al., 2002]. A drawback with this approach is that a global transformation does not adequately account for the fact that the same gene may be measured with different precision in different experiments. For example, a gene which has a lower expression level in one condition will typically be measured with relatively less precision in that condition. Another drawback with this approach is that a complex non-linear transformation of the data complicates the interpretation of measurements when compared to a global log transformation which measures fold-changes on a comparable scale for all genes. Instead of transforming the data, we prefer instead to propagate variances through the downstream analysis using probabilistic models.

In this paper we propose a modified approach to PCA which can take account of credibility intervals. This leads to a far more robust downstream analysis. Our approach eliminates the need to heuristically reject those genes which are perceived as unreliable before the downstream analysis is applied and allows us to automatically select the latent subspace dimension in a principled way without recourse to complex Bayesian methods. Our model builds on a latent variable model known as probabilistic PCA [Tipping and Bishop, 1999]. This model has previously been applied to microarray data analysis

¹ PCA is sometimes referred to as singular value decomposition which is an algorithm that can be used to solve the eigenvalue problem that underpins PCA.

and has been extended in order to deal with Bayesian learning [Oba et al., 2003a], missing value estimation [Oba et al., 2003b] and non-Gaussian latent variables [Girolami and Breiting, 2004]. However, in all of these previous applications the noise has been considered spherical. Here we provide the first extension to a model with a non-spherical and non-*i.i.d.* noise distribution. Our approach is quite general and could also be applied to other probabilistic models.

In the next section we review probabilistic PCA and describe how the structure of the model may be modified to account for dimension and measurement specific noise in the data. We then discuss how the parameters of our model may be optimised in a practical way using an expectation maximisation algorithm. In Section 3 we investigate a data set derived from oligonucleotide arrays with our method. We show how our approach not only gives a robust version of PCA but also allows us to obtain a ‘denoised’ version of the data set and automatically determine the number of principal components to be retained.

2 METHODS

We will start by briefly reviewing the latent variable model known as probabilistic PCA [Tipping and Bishop, 1999]. In probabilistic PCA it is assumed that each d -dimensional data point \mathbf{y}_n can be reconstructed from a q -dimensional latent point \mathbf{x}_n via a linear transformation W and a corrupting noise vector ϵ_n ,

$$\mathbf{y}_n = W\mathbf{x}_n + \boldsymbol{\mu} + \epsilon_n.$$

The vector $\boldsymbol{\mu}$ is the empirical mean of the data set. In the case of probabilistic PCA the noise vector is assumed to have come from a spherical Gaussian distribution,

$$\epsilon_n \sim N(0, \sigma^2 I),$$

which implies that

$$\mathbf{y}_n | \mathbf{x}_n \sim N(W\mathbf{x}_n + \boldsymbol{\mu}, \sigma^2 I).$$

If the latent variable \mathbf{x}_n is also assumed to be governed by a spherical Gaussian prior, $\mathbf{x}_n \sim N(0, I)$, then the marginal distribution over \mathbf{y}_n is found to be

$$\mathbf{y}_n \sim N(\boldsymbol{\mu}, WW^T + \sigma^2 I). \quad (1)$$

Note that to recover PCA the prior over the latent space need not necessarily be spherical: it may be governed by its own mean and covariance. This leads to a redundant parameterisation of the model which, while mathematically of little importance, can be exploited to improve the algorithmic performance as we outline in the Appendix.

2.1 Relationship to Factor Analysis

Tipping and Bishop [1999] showed that the maximum likelihood solution for W recovers the principal subspace of the data; this model is thereby recognised as a probabilistic interpretation of PCA. The model is strongly related to factor analysis [Bartholomew, 1987]. In factor analysis, however, the distribution governing ϵ_n is a diagonal covariance Gaussian distribution,

$$\epsilon_n \sim N(0, B^{-1}),$$

where B is a diagonal matrix whose i th diagonal element is given by β_i and the variance associated with the i th output dimension is

given by β_i^{-1} . We refer to the inverse variance, β_i , as a precision. This leads to a marginal distribution governing \mathbf{y}_n of the form

$$\mathbf{y}_n \sim N(\boldsymbol{\mu}, WW^T + B^{-1}).$$

This model is slightly more flexible than probabilistic PCA, but this additional flexibility comes with a cost. The precision for each data dimension (which in microarray analysis could either be each experiment or each gene/probe-set) must be separately estimated. This causes the model optimisation to become iterative, in contrast to probabilistic PCA where the likelihood can be maximised through an eigenvalue decomposition.

2.2 Propagating Measurement Uncertainty

Factor analysis allows different variances (or precisions) for each data dimension but here we want to allow the precision to vary for each data point and dimension. In other words, we wish to allow the variance to be different for every gene in each experiment, *i.e.* Nd variances where there are d experiments and N different genes. Such a model is far richer than factor analysis but it involves estimation of Nd precision parameters from a data set of typically only Nd separate values. Fortunately, recent advances in probe level analysis techniques have meant that these precisions can be derived directly from the microarray chip [Lawrence et al., 2003, Milo et al., 2003, Hein et al., 2005]. For example this can be achieved, for Affymetrix arrays, by making use of the information about measurement uncertainty provided by multiple probes in each gene’s probe-set. The probabilistic models that have been developed associate each gene with a level of uncertainty.

We will consider the following modified model. Take \mathbf{y}_n to be a d dimensional vector which represents the true log expression level associated the n th gene under d different conditions. Rather than observing \mathbf{y}_n directly we assume that we observe a corrupted form $\hat{\mathbf{y}}_n$ where

$$\hat{\mathbf{y}}_n = \mathbf{y}_n + \boldsymbol{\nu}_n \quad (2)$$

and $\boldsymbol{\nu}_n$ is noise which is distributed as

$$\boldsymbol{\nu}_n \sim N(0, B_n^{-1}).$$

Here, B_n is a diagonal matrix whose i th diagonal element is given by β_{ni} which is the precision associated with the i th experiment for the n th gene. This precision can be obtained through one of the probabilistic analysis methods mentioned above.

So far our approach is rather general and we could take the distribution \mathbf{y}_n to come from any probabilistic model of interest. We will assume a probabilistic PCA model as the marginal distribution for the true expression level \mathbf{y}_n , as given in (1), and obtain,

$$\hat{\mathbf{y}}_n | \mathbf{x}_n \sim N(W\mathbf{x}_n + \boldsymbol{\mu}, \sigma^2 I + B_n^{-1}). \quad (3)$$

We denote collectively $A_n = \sigma^2 I + B_n^{-1}$ and using² $\mathbf{x}_n \sim N(0, I)$ we have the following marginalised likelihood,

$$\hat{\mathbf{y}}_n \sim N(\boldsymbol{\mu}, WW^T + A_n).$$

The corrupted data is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance $C_n = WW^T + A_n$. Traditionally in PCA the data is centred

² Girolami and Breiting [2004] argue that this choice of prior distribution has the problem of implying negative expression levels. This is certainly a problem if one works directly with the expression levels; however when considering log expression levels this is no longer a problem.

and $\boldsymbol{\mu}$ is taken to be zero. This is reasonable as in probabilistic PCA the maximum likelihood solution for $\boldsymbol{\mu}$ is the empirical mean of the data. However, in the modified model we present, as we shall see, the maximum likelihood solution for $\boldsymbol{\mu}$ is no longer the empirical data mean so it does not make sense to work with a centred data set. The parameter $\boldsymbol{\mu}$ must be learnt jointly with the rest of the model.

The log likelihood for the suggested model takes the form

$$\mathcal{L} = -\frac{1}{2} \sum_{n=1}^N (\log(|C_n|) + \text{tr}(C_n^{-1} S_n)) + \text{const} \quad (4)$$

where $S_n = (\hat{\mathbf{y}}_n - \boldsymbol{\mu})(\hat{\mathbf{y}}_n - \boldsymbol{\mu})^T$. There are several important differences from standard probabilistic PCA: firstly the matrix A_n is not proportional to the identity, hence we cannot obtain a closed analytic solution for the maximum likelihood value of the parameters. Secondly, differentiating (4) w.r.t. the mean vector $\boldsymbol{\mu}$ (while keeping the other parameters fixed) yields

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\mu}} \propto \boldsymbol{\mu} \sum_{n=1}^N C_n^{-1} - \sum_{n=1}^N C_n^{-1} \hat{\mathbf{y}}_n$$

which implies a fixed point equation for $\boldsymbol{\mu}$ of the form

$$\boldsymbol{\mu}_{ML} = \left(\sum_{n=1}^N C_n^{-1} \right)^{-1} \left(\sum_{n=1}^N C_n^{-1} \hat{\mathbf{y}}_n \right). \quad (5)$$

Hence the maximum likelihood estimator of the mean vector $\boldsymbol{\mu}$ is no longer the empirical mean of the data, and the solution is interdependent (through C_n) with the orientation of the principal sub-space, W . Note, however, that (5) does indeed return the empirical mean in the special case when all the C_n s are equal (under these conditions standard factor analysis is recovered).

2.3 Efficient Likelihood Optimisation

Given the gradients of the likelihood we can optimise the parameters through a non-linear optimisation such as scaled conjugate gradients. Unfortunately, for our model such an optimisation will have high computational demands because the likelihood and its gradient contains several multiplications of large matrices. In practice, for a typical sized microarray data set, such an approach becomes impractical. A more efficient algorithm can be obtained through an expectation maximisation (EM) approach [Dempster et al., 1977].

Generally EM algorithms lead to a simplified optimisation problem (the M-step) by incorporating an additional step (the E-step). For our corrupted data PCA model this additional step is the computation of the posterior distribution for the latent space. This posterior is obtained through Bayes' theorem

$$\mathbf{x}_n | \hat{\mathbf{y}}_n \sim N \left(M_n W^T A_n^{-1} (\hat{\mathbf{y}}_n - \boldsymbol{\mu}), M_n \right) \quad (6)$$

where we have defined

$$M_n = \left[W^T A_n^{-1} W + I \right]^{-1}.$$

The EM algorithm proceeds by maximising a lower bound on the true likelihood. This bound is dependent on both the posterior distribution in (6) and the parameters of the model. Once the posterior has been updated (the E-step) the expectation of the joint log-likelihood of the data and the latent variables is optimised with

respect to the model parameters (the M-step). These updates are applied iteratively until convergence. While iterative application of these equations seems to be adding to our computation load, the update equations for the parameters in the M-step are quicker than the joint optimisation over all parameters that direct optimisation of (4) would require. If the required number of iterations is not excessive, and the updates in the E-step and M-step may be done efficiently, then the likelihood optimisation can proceed much quicker through the EM algorithm.

The portion of the lower bound that directly depends on the model parameters is given by

$$\begin{aligned} \mathcal{L}_c &= -\frac{1}{2} \sum_{n=1}^N \log |A_n| + \sum_{n=1}^N \text{tr} \left(\langle \mathbf{x}_n \mathbf{x}_n^T \rangle \right) \\ &+ \sum_{n=1}^N (\hat{\mathbf{y}}_n - \boldsymbol{\mu})^T A_n^{-1} (\hat{\mathbf{y}}_n - \boldsymbol{\mu}) \\ &- 2 \sum_{n=1}^N \langle \mathbf{x}_n \rangle^T W^T A_n^{-1} (\hat{\mathbf{y}}_n - \boldsymbol{\mu}) \\ &+ \sum_{n=1}^N \text{tr} \left(W^T A_n^{-1} W \langle \mathbf{x}_n \mathbf{x}_n^T \rangle \right), \end{aligned} \quad (7)$$

where the notation $\langle \cdot \rangle$ denotes an expectation under the posterior distribution over \mathbf{x}_n . The required expectations may be evaluated as

$$\begin{aligned} \langle \mathbf{x}_n \rangle &= M_n W^T A_n^{-1} (\hat{\mathbf{y}}_n - \boldsymbol{\mu}), \\ \langle \mathbf{x}_n \mathbf{x}_n^T \rangle &= M_n + \langle \mathbf{x}_n \rangle \langle \mathbf{x}_n \rangle^T. \end{aligned}$$

Updates of these expectations take place in the E-step. For the M-step we must consider the gradient of (7) with respect to the parameters, W , $\boldsymbol{\mu}$ and σ^2 ,

$$\frac{\partial \mathcal{L}_c}{\partial W} = - \sum_{n=1}^N \left[2 A_n^{-1} (\hat{\mathbf{y}}_n - \boldsymbol{\mu}) \langle \mathbf{x}_n \rangle^T - 2 A_n^{-1} W \langle \mathbf{x}_n \mathbf{x}_n^T \rangle \right],$$

$$\begin{aligned} \frac{\partial \mathcal{L}_c}{\partial \sigma} &= -2 \sum_{n=1}^N \left\{ (\hat{\mathbf{y}}_n - \boldsymbol{\mu})^T A_n^{-2} (\hat{\mathbf{y}}_n - \boldsymbol{\mu}) \right. \\ &- 2 \langle \mathbf{x}_n \rangle^T W^T A_n^{-2} (\hat{\mathbf{y}}_n - \boldsymbol{\mu}) \\ &\left. + W^T A_n^{-2} W \langle \mathbf{x}_n \mathbf{x}_n^T \rangle - \text{tr}(A_n^{-1}) \right\} \sigma, \end{aligned}$$

$$\frac{\partial \mathcal{L}_c}{\partial \boldsymbol{\mu}} = 2 \sum_{n=1}^N A_n^{-1} (\boldsymbol{\mu} - \hat{\mathbf{y}}_n + W \langle \mathbf{x}_n \rangle). \quad (8)$$

These gradients lead to fixed point equations for W and $\boldsymbol{\mu}$,

$$\begin{aligned} \boldsymbol{\mu} &= \left(\sum_{n=1}^N A_n^{-1} \right)^{-1} \sum_{n=1}^N A_n^{-1} (\hat{\mathbf{y}}_n - W \langle \mathbf{x}_n \rangle), \\ W_{jl} &= \sum_{p=1}^q H_{jp} (L_j)_{pl}^{-1}, \end{aligned} \quad (9)$$

where A_{nj} is the j -th diagonal element of A_n and we have introduced the two matrices

$$H = \sum_{n=1}^N A_n^{-1} (\hat{\mathbf{y}}_n - \tilde{\boldsymbol{\mu}}) \langle \mathbf{x}_n \rangle^T, \\ L_j = \sum_{n=1}^N A_{nj}^{-1} \langle \mathbf{x}_n \mathbf{x}_n^T \rangle,$$

in order to give (9) a more readable form.

These update equations can be iterated to find the maximum likelihood solution with respect to W and $\boldsymbol{\mu}$. Unfortunately a fixed point equation for σ^2 (which accounts for any residual variance) is not so straightforward as the gradient with respect to σ^2 is not linear. An efficient update for σ^2 can be obtained by using Newton's method.

When maximising likelihoods through fixed point equations convergence can be slow if there are strong correlations between the parameters. In the modified PCA model, for example, the solution for W is strongly dependent on the solution for $\boldsymbol{\mu}$. When such correlations occur it can be advantageous to introduce apparently redundant parameters to improve the speed of convergence. In the Appendix we describe a redundant parameterisation of this model that allows us to achieve a significant speed up.

2.4 Processing Data

Processing of a microarray data set is undertaken in the following manner. The expression levels are represented in a matrix $\hat{Y} = [\hat{\mathbf{y}}_1 \dots \hat{\mathbf{y}}_N]^T$. The elements of this matrix are the 'corrupted' log expression levels or, in the case of cDNA arrays, log ratios of expression levels. The uncertainty associated with the data is stored in a separate matrix $\hat{B} = [\beta_1 \dots \beta_N]^T$ in the form of precisions. These precisions are the inverse variances which correspond to the log expression levels (or ratios) in \hat{Y} . The algorithm returns the principal subspace, W , the residual variance, σ^2 , and the inferred mean $\boldsymbol{\mu}$ as well as a set of moments under the posterior: $\{\langle \mathbf{x}_n \rangle\}_{n=1}^N$ and $\{\langle \mathbf{x}_n \mathbf{x}_n^T \rangle\}_{n=1}^N$. Note that the dimensionality of the subspace can be automatically determined due to the inclusion of the uncertainty information (see Section 2.5).

We can now recover a posterior estimate for the *uncorrupted* log expression levels, $Y = [\mathbf{y}_1 \dots \mathbf{y}_N]^T$. The mean of this estimate for the n th gene is given by

$$\bar{\mathbf{y}}_n = W \langle \mathbf{x}_n \rangle + \boldsymbol{\mu}$$

and the covariance of each estimate is given by

$$\Sigma_n = W \left[\langle \mathbf{x}_n \mathbf{x}_n^T \rangle - \langle \mathbf{x}_n \rangle \langle \mathbf{x}_n^T \rangle \right] W^T + \sigma^2 I.$$

The new uncertainty associated with the n th gene in the i th experiment is then given by the i th diagonal element of Σ_n . As we shall show in Section 3 these cleaned up expression levels can be used in a further analysis stage, such as hierarchical clustering, where they lead to more consistent clusters and expression profiles with reduced levels of uncertainty.

2.5 Number of Principal Components

The usual approach when implementing PCA for microarray data is to retain a reduced number of principal directions, q , and project the log expression levels along these directions before further processing. Typically the directions retained are those associated with

the largest q eigenvalues of the data covariance. This approach has a natural interpretation: the directions associated with the smaller eigenvalues are assumed to arise from noise.

In general, the number of principal components retained is pre-determined according to the specific problem under consideration. In our model, however, the probabilistic treatment of the probe-level uncertainty allows us to obtain the maximum number of principal components that can be retained. Intuitively, a direction will be discarded if the variation in the data is not statistically significant given the measurement noise; the model will then return an eigenvalue 0 associated with that direction.

3 RESULTS

To demonstrate the efficacy of our approach we considered a data set that consisted of a temporal assay of Affymetrix GeneChip arrays that measured the gene expression profiles of a conditionally immortal cell line, UB/OC-1, from mouse cochlear epithelial cells at embryonic day 13.5 (E13.5), across 14 days of differentiation. The data set aims to discover gene expression patterns associated with early differentiation of mammalian auditory hair cells [Rivolta et al., 2002]. The experiments consisted of 12 samples obtained at 12 different time points during 14 days of differentiation. Up to 'day 0' the cells were cultured under proliferating conditions at 33°C. Differentiation was induced by shifting the temperature to 39°C. The dataset is then a temporal profile of 12 data points with no replicates. Of particular interest in this study is the identification of targets regulated by the transcription factor *gata-3*, which is essential for normal inner ear development. Also important in inner ear development is the protein kinase inhibitor *p27^{kip1}*. In vivo the expression values of *gata-3* and *p27^{kip1}* are low before day 4 when they start to rise. They peak at day 8-9 and after a couple of days the expression level decreases again to then stabilize around a constant value.

In a probabilistic setting, where gene expression is extracted with a different level of sensitivity [Milo et al., 2003, Liu et al., 2005] it is possible to analyse robustly low expressed genes, like transcription factors, that are crucial in development and regulation of gene networks. For this reason we used the *gata-3* expression profile to test our model. Our results show that it is possible to improve in a principled setting the detection of the activity of genes that work at low level but have a crucial role in gene networking. This improvement allows them to be included in any downstream analysis for the identification of related targets.

The raw data was processed using a modified version of the gMOS algorithm [Milo et al., 2003] in which the scales of the gMOS gamma distributions models were shared across probe pairs in different experiments, rather than across the probe-set. The means and variances of the log expression levels were then derived from the resulting gamma distribution.

3.1 Profile Reconstruction

Our first aim was to assess the sensitivity of the profile reconstruction (as described in Section 2.4) to changes in the associated variances. We therefore considered two sub-sampled datasets containing only 500 of the 13,178 genes. The first dataset contained *gata-3* and the second *p27^{kip1}*. In each case we modelled the data three times with our modified PCA. To show the effect of reduced uncertainty in the data we divided the variances by 1, 4 and 9. The corrected expression profiles are shown in Figures 1 and 2. Note that

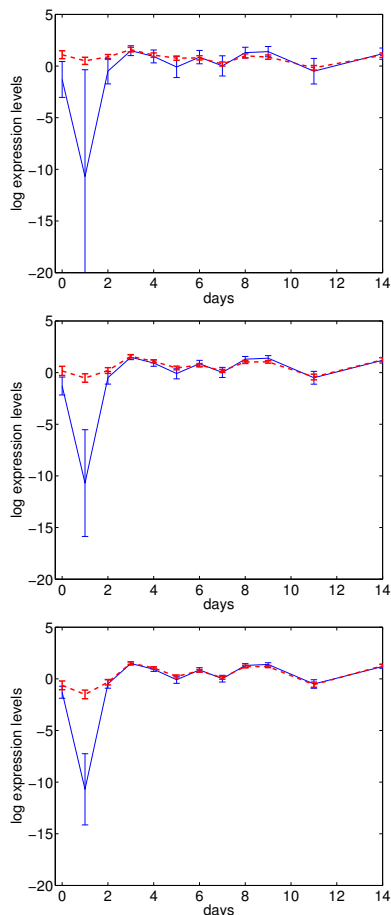


Figure 1. Corrected profile (thick dashed line) and original profile (thin solid line) for the *gata3* gene (a transcription factor) *top*: corrected profile based using the original uncertainties; *middle*: corrected profile with the uncertainty halved and *bottom*: corrected profile with a third of the original uncertainty.

as the uncertainty in the original profile is decreased the corrected profile tends to stay closer to its original course. As can be seen from the plots, any point with large associated uncertainty (such as the day 1 point for the *gata-3* profile) can be significantly changed and this can lead to a large decrease in the associated uncertainty. However, the only data point that is reconstructed at a significantly different level due to the reduced uncertainties is the expression level at day 1 for the *gata-3* profile, showing a certain robustness to changes in the estimates of the uncertainties.

3.2 Clustering

Clustering is a widely used technique for summarising expression levels obtained from gene array data. Because of the large number of genes measured and the complexity of the associated gene networks, identifying groups of genes that behave similarly in the dataset can be a useful exploratory technique for finding functional analogues.

One suggested use of PCA in microarray analysis is as a pre-processing step before cluster analysis. The use of PCA before clustering can be justified by the fact that the larger principal-components are expected to capture the structure in the data set.

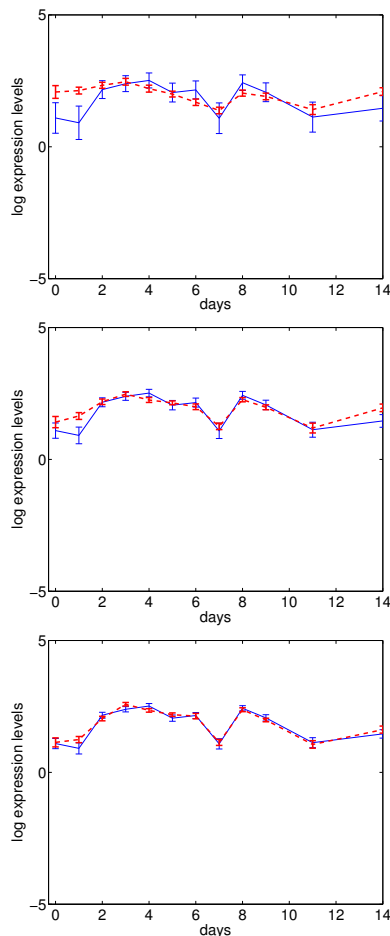


Figure 2. Corrected profile (thick dashed line) and original profile (thin solid line) for the *p27^{kip1}* gene (a cycline-dependent kinase inhibitor) *top*: corrected profile based using the original uncertainties; *middle*: corrected profile with the uncertainty halved and *bottom*: corrected profile with a third of the original uncertainty.

In practice, when using standard PCA as a pre-processing step, it is necessary to manually determine the dimension of the latent space, q . Furthermore the use of a standard PCA does not always improve the clustering but often degrades it [Yeung et al., 2001]. This is due to the fact that the dominant components, which contain most of the variation in the data, are highly influenced by the very noisy data points. Therefore they do not necessarily capture the data's structure. This is often avoided by introducing arbitrary thresholds in order to reject the low expressed genes; the advantage of our method is to give a principled automatic way to select the relevant genes. By accounting for the variance in the log expression levels we can ensure that the components we extract accurately reflect the structure of the data. In Figures 3, 4 and 5 we perform hierarchical clustering on the first fifty genes. Figure 3 uses the reconstructed profiles obtained as described in the previous section, Figure 4 uses the non-reconstructed profiles while Figure 5 uses the (non-reconstructed) profiles of the genes obtained by standard PCA. In both cases these components were derived by considering the entire data set of 13,178 probes. As a result, standard PCA was severely compromised by the low expressed genes. However,

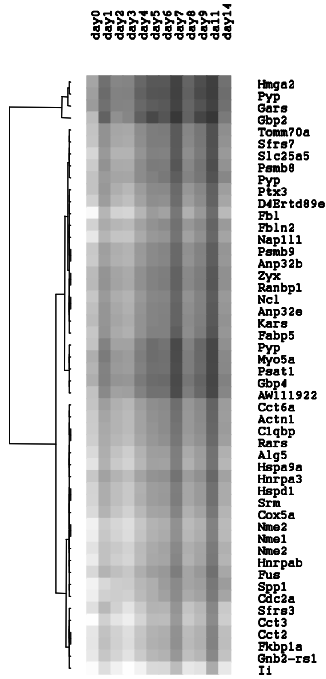


Figure 3. Hierarchical clustering of the corrected profiles. The fifty genes selected are those associated with the second component of our modified probabilistic PCA model. Clustering was carried out using the posterior mean of the corrected profiles, \bar{y}_n . The processed profiles appear to be much more tightly clustered than the corrupted profiles in Figure 4.

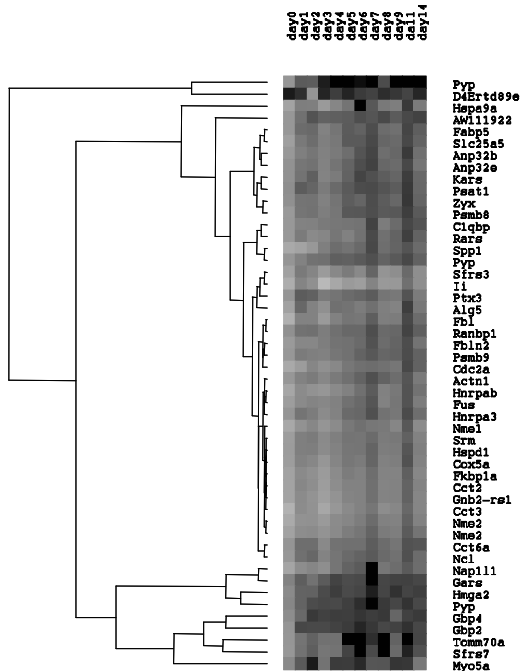


Figure 4. Hierarchical clustering of the uncorrected profiles. As in Figure 3 the fifty genes selected for clustering are those associated with the second component of our modified probabilistic PCA model. However, here clustering was carried out using the ‘corrupted profile’, \hat{y}_n . The profiles are less tightly clustered than those in Figure 3.

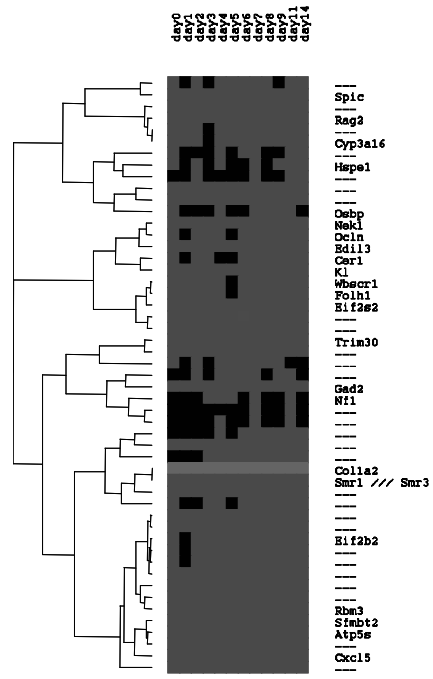


Figure 5. Hierarchical clustering of the data associated with the second principal component for the PCA analysis. The highly negative expression levels are no longer associated with high uncertainty and as a result they dominate. The genes selected for clustering are of little or no relevance to the underlying biological system..

by taking the variances into account, our modified model is able to down-weight the influence of these points thereby reducing their influence. Furthermore, the model was able to automatically determine the requisite number of retained components for this data set ($q = 3$).

Clustering was performed on the selected genes using the Gene Cluster software from the Eisen Lab (available from <http://rana.lbl.gov/EisenSoftware.htm>). As expected, there was little information in the genes selected for clustering by standard PCA (Figure 5). However for the genes selected by the modified model (Figures 3 and 4) there are high functional correlations within the clusters. In particular the clustering of the genes with corrected profiles produces three very distinct functionally related groups. For example the two larger groups are related to cell proliferation and cell cycle regulation. The heat shock proteins (Hsp) cluster together with the chaperonin proteins (Cct2 and Cct3) and are involved in cell growth and survival. Very important is the role of the mortalin mitochondrial heat shock protein that is highly involved in cell cycle regulation and cellular senescence specification. The second largest group contains more cell migration and cell proliferation related genes, like PTX3 and NAP-1 which seem to be more related to the delamination of these sensory epithelial cells. The groups are highlighting the main primary biological processes in this development, which are cell proliferation and migration that will be then followed by a more specific differentiation stage, where the final fate of the cells is determined.

4 DISCUSSION

We have shown how the noise which is inherent in microarray data may be accounted for, in a principled manner, through a probabilistic model that considers the variances associated with each gene's expression level in each condition. We presented a model that performs a modified form of PCA on the data, automatically determining the required number of components for describing the data structure. We have shown, for an example using Affymetrix GeneChips, how the model can recover an improved estimate of the gene expression profiles through denoising.

We were also able to show that the improved profiles can lead to tighter and more coherent clusters by applying hierarchical clustering to both the original profiles and the corrected profiles. We expect similar results to be achievable for cDNA arrays where the variances can be extracted through the image processing techniques suggested in Lawrence et al. [2004].

One of the features of the proposed method is that the importance of the genes with large associated variance is reduced in the downstream analysis. This can obviously be a problem if the estimates of the uncertainties are systematically too large (leading to too many genes being smoothed away) or too small (leading irrelevant genes to dominate the results). It is therefore important that the probe-level analysis required to extract confidence intervals is carried out as accurately as possible.

One aspect of microarray studies is often to provide a list of significant target genes in a given experimental system. In order to provide this methods like scoring and cut-off thresholding are normally used. One of the benefits of the proposed method is that it automatically implements a cut-off by downweighting genes with high associated variance. Current work includes how to use the variance information to produce a list of significant targets and assess differential gene expressions under different experimental conditions.

ACKNOWLEDGEMENTS

GS, MR and NL gratefully acknowledge support from a BBSRC award "Improved processing of microarray data with probabilistic models". MM is supported by an Advanced Training Fellowship from the Wellcome Trust. We thank the anonymous reviewers for useful suggestions.

REFERENCES

- O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*, 97(18):10101–10106, 2000.
- P. Baldi and S. Brunak. *Bioinformatics: The Machine Learning Approach*. MIT Press, Cambridge, MA, 2001.
- D. J. Bartholomew. *Latent Variable Models and Factor Analysis*. Charles Griffin & Co. Ltd, London, 1987.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.
- M. Girolami and R. Breitling. Biologically valid linear factor models of gene expression. *Bioinformatics*, 20(17):3021–3033, 2004.
- A.-M. K. Hein, S. Richardson, H. C. Causton, G. K. Ambler, and P. J. Green. BGX: a fully Bayesian gene expression index for Affymetrix GeneChip data. *Biostatistics*, 6:349–373, 2005.

- W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18:S96–S104, 2002.
- N. D. Lawrence, M. Milo, M. Niranjan, P. Rashbass, and S. Soullier. Bayesian processing of microarray images. In C. Molina, T. Adali, J. Larsen, M. V. Hulle, S. Douglas, and J. Rouat, editors, *Neural Networks for Signal Processing XIII*, pages 71–80. IEEE, 2003.
- N. D. Lawrence, M. Milo, M. Niranjan, P. Rashbass, and S. Soullier. Reducing the variability in cDNA microarray image processing by Bayesian inference. *Bioinformatics*, 20(4):518–526, 2004.
- X. Liu, M. Milo, N. D. Lawrence, and M. Rattray. A tractable probabilistic model for Affymetrix probe-level analysis across multiple chips. To appear in *Bioinformatics*, 2005.
- M. Milo, A. Fazeli, M. Niranjan, and N. D. Lawrence. A probabilistic model for the extraction of expression levels from oligonucleotide arrays. *Biochemical Transactions*, 31(6):1510–1512, 2003.
- S. Oba, M. Sato, and S. Ishii. Prior hyperparameters in Bayesian PCA. In *ICANN/ICONIP 2003*, pages 271–279, 2003a.
- S. Oba, M. Sato, I. Takemasa, M. Monden, K. ichi Matsubara, and S. Ishii. A Bayesian missing value estimation method for gene expression data. *Bioinformatics*, 19(16):2088–2096, 2003b.
- M. N. Rivolta, A. Halsall, C. Johnson, M. Tones, and M. C. Holley. Genetic profiling of functionally related groups of genes during conditional differentiation of a mammalian cochlear hair cell line. *Genome Research*, 12(7):1091–1099, 2002.
- M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society B*, 6(3):611–622, 1999.
- K. Y. Yeung, C. Fraley, A. Murua, A. E. Raftery, and W. L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17(10):977–87, 2001.

APPENDIX — FAST IMPLEMENTATION OF THE EM ALGORITHM

The speed of convergence of the EM algorithm can be improved by introducing a redundant parameterisation as follows. We introduce non-spherical priors in the latent space so that $\mathbf{x}_n \sim N(\mathbf{m}, C)$. The E-step in the EM algorithm will now be modified giving

$$\mathbf{x}_n | \hat{\mathbf{y}}_n \sim N(\bar{\mathbf{x}}_n, \Sigma_x),$$

as the posterior for \mathbf{x}_n , where

$$\Sigma_x = (W A_n^{-1} W + C^{-1})^{-1},$$

$$\bar{\mathbf{x}}_n = \Sigma_x (W A_n^{-1} (\hat{\mathbf{y}}_n - \mu) + C^{-1} \mathbf{m}).$$

The M-step update with respect to \mathbf{m} and C are then given by

$$\mathbf{m} = \frac{\sum_{n=1}^N \langle \mathbf{x}_n \rangle}{N},$$

$$C = \sum_{n=1}^N \frac{\langle \mathbf{x}_n \mathbf{x}_n^T \rangle}{N} - 2 \sum_n \frac{\langle \mathbf{x}_n \rangle \mathbf{m}^T}{N} + \mathbf{m} \mathbf{m}^T.$$

The redundancy in the new parameter representation is straightforward to see if we consider the marginalised likelihood for this new

model

$$\hat{\mathbf{y}}_n \sim N\left(\boldsymbol{\mu} + W\mathbf{m}, WCW^T + I\sigma^2 + B_n\right).$$

It is directly equivalent to a marginalised likelihood for a model of the form

$$\hat{\mathbf{y}}_n \sim N\left(\boldsymbol{\mu}', W'W'^T + I\sigma^2 + B_n\right),$$

where $\boldsymbol{\mu}' = \boldsymbol{\mu} + W^T\mathbf{m}$ and $W' = U\Lambda^{\frac{1}{2}}R^T$. Here R is an arbitrary rotation matrix, U are the eigenvectors of the matrix WCW^T and Λ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues.