



Biologically valid linear factor models of gene expression

Mark Girolami^{1,*} and Rainer Breitling^{1,2}

¹Bioinformatics Research Centre, Department of Computing Science and ²Molecular Plant Sciences Group, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

Received on April 23, 2004; revised on May 19, 2004; accepted on May 31, 2004

Advance Access publication June 16, 2004

ABSTRACT

Motivation: The identification of physiological processes underlying and generating the expression pattern observed in microarray experiments is a major challenge. Principal component analysis (PCA) is a linear multivariate statistical method that is regularly employed for that purpose as it provides a reduced-dimensional representation for subsequent study of possible biological processes responding to the particular experimental conditions. Making explicit the data assumptions underlying PCA highlights their lack of biological validity thus making biological interpretation of the principal components problematic. A microarray data representation which enables clear biological interpretation is a desirable analysis tool.

Results: We address this issue by employing the probabilistic interpretation of PCA and proposing alternative linear factor models which are based on refined biological assumptions. A practical study on two well-understood microarray datasets highlights the weakness of PCA and the greater biological interpretability of the linear models we have developed.

Availability: The model estimation routines are currently implemented as Matlab routines and these, as well as data and results reported, are available from the following URL: <http://www.dcs.gla.ac.uk/~girolami/lfm/index.html>

Contact: girolami@dcs.gla.ac.uk

1 INTRODUCTION

One of the main biological challenges when analyzing large-scale expression experiments is the identification of the active cellular processes that combine to generate the measured differences in gene expression. One of the classical multivariate data analysis approaches to this problem is the straightforward application of principal component analysis (PCA), which identifies components, factors or processes that contribute to the observed expression pattern (Raychaudhuri *et al.*, 2000; Alter *et al.*, 2000), for more recent examples (see also Bleharski *et al.*, 2003; Mao *et al.*, 2003; Segal

et al., 2003; Tarte *et al.*, 2003; Tsunoda *et al.*, 2003). Here, we exploit a probabilistic interpretation of PCA (Tipping and Bishop, 1997) that allows us to identify several inherent weaknesses of this particular approach and to devise simple modifications that lead to dramatically improved results with higher biological interpretability and validity while maintaining computational efficiency and statistical rigor.

The requirement to devise a means of analysis of gene expression experiments which yields valid insight into the active cellular processes (and associated interactions at the regulatory level) provides the motivation for this paper (Alter *et al.*, 2000; Gasch *et al.*, 2000; Spellman *et al.*, 1998). It is posited here that groups of genes are active to varying degrees in a number of cellular processes. The most common approach in identifying such gene groups is to cluster the gene expression across experimental conditions (Eisen *et al.*, 1998). Although clustering can effectively identify group structure of genes across arrays, it has been argued and widely accepted that the mutually exclusive class membership of experiments under the clustering model is highly restrictive in that samples may indeed be influenced by varying combinations of different basic cellular processes (Segal *et al.*, 2003). In the clustering approach, the expression pattern in each experiment would be described by the prototypical average expression pattern of the cluster to which it belongs. It is obvious that this is a very rigid description. A far more flexible approach can be achieved when each experimental expression pattern is considered as a combination of several prototypical expression patterns (i.e. 'physiological processes'), and this approach will be the focus of the rest of this paper.

2 LINEAR METHODS FOR GENE EXPRESSION ANALYSIS

In the approaches described in this paper, the expression levels will be represented using a standard statistical linear factor model. This allows the definition of a number of gene groups or processes that have specific responses under certain experimental conditions. It is then given that each gene

*To whom correspondence should be addressed.

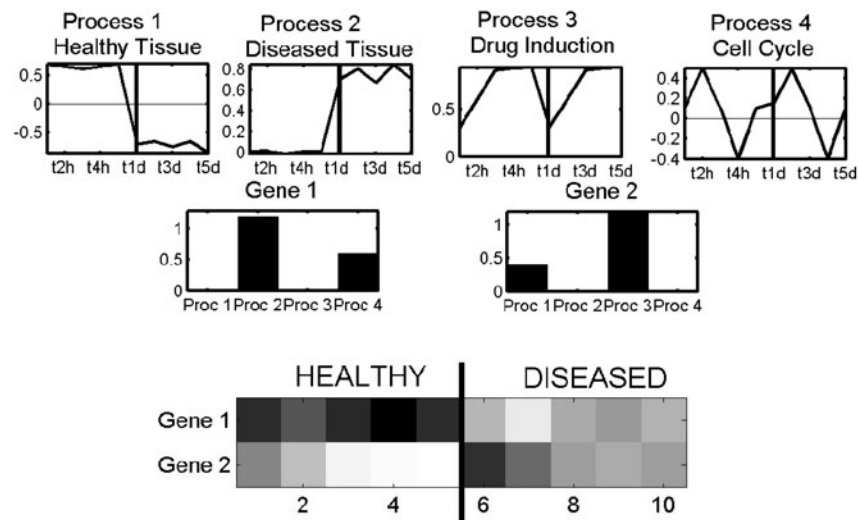


Fig. 1. In this fictitious experiment, differential gene expression was measured at five time points during drug exposure in healthy (t1h–t5h) and diseased tissue (t1d–t5d). Reasonable processes (A_p) in this case would correspond, e.g. to differences between healthy and diseased samples (processes 1 and 2), to induction in response to the drug (process 3) or to various cell cycle stages (assuming that the sample was synchronized initially; process 4). The gene process representation θ_g is shown for two representative genes (middle): Gene 1 is strongly overexpressed in diseased tissue, and also shows cyclical expression during the cell cycle. Gene 2 is slightly repressed in diseased tissue, but is strongly induced in response to drug treatment. The bottom row shows the reconstructed expression pattern for these two genes (e_g) obtained from the linear combination of processes (top row) weighted by gene process representation (middle row).

may be differentially expressed to a certain degree in each of these defined processes. It is hoped that, given biologically valid assumptions, these model processes will correspond to physiological processes. The measured expression levels for a particular array and gene will be some function of a combination of all the defined process responses and the gene participation level in each of these. The simplest functional form to represent the process responses and gene process participation is a linear combination¹. Each gene will be described by its differential expression in each of the possible processes, and each experimental sample will be described by how active the various processes are—the characteristic expression of the gene in the sample is then just a simple weighted sum of its activity in the active processes (Figs 1 and 2). Figure 1 shows an illustrative example of the linear factor model applied to a fictitious experiment, to illustrate the biological interpretation of the obtained linear factor (process) representation.

PCA has been employed as a means of summarizing and analyzing results of microarray experiments (Raychaudhuri *et al.*, 2000; Bleharski *et al.*, 2003; Mao *et al.*, 2003; Segal *et al.*, 2003; Tarte *et al.*, 2003; Tsunoda *et al.*, 2003) as has the equivalent linear matrix singular value decomposition (SVD)

¹In the absence of additional information regarding the nonlinear effects, which may govern gene expression there is little motivation to consider explicit nonlinear models and given that the gene expression data themselves are log-transformed relative measurements, we will only consider linear factor style modeling in the remainder of this paper.

(Alter *et al.*, 2000). A generalization of SVD has also been employed in the assignment of genes to overlapping transcription modules (Ihmels *et al.*, 2002). The main weakness of straightforward application of such decompositions, in the absence of an explicit probabilistic data model, is that there is no straightforward way of objectively assessing model performance (see Tipping and Bishop (1997) for an extended discussion of this point) or indeed assessing whether the distributional assumptions underlying the models are appropriate for the data being considered.

The probabilistic basis of linear PCA was established in Tipping and Bishop (1997) and independent component analysis (ICA), a linear transformation viewed as a generalization of PCA, has a well-defined probabilistic foundation (Roberts and Everson, 2003; Hyvarinen *et al.*, 2001; Girolami, 2000). ICA has recently been applied directly to the analysis of microarray experiments (Martoglio *et al.*, 2002; Lee and Batzoglou, 2003).

In this paper rather than taking a standard linear multivariate analysis tool such as PCA or ICA and applying it directly to the analysis of microarray experiments, we begin with a number of specific biologically valid assumptions concerning the generative processes underlying the expression profiles measured. These biological assumptions are then encoded, explicitly employing the probabilistic formalism, in straightforward linear factor models and the emerging representations are then applied to the analysis of gene expression profiles with very promising results in terms of predictive and biologically valid representational power.

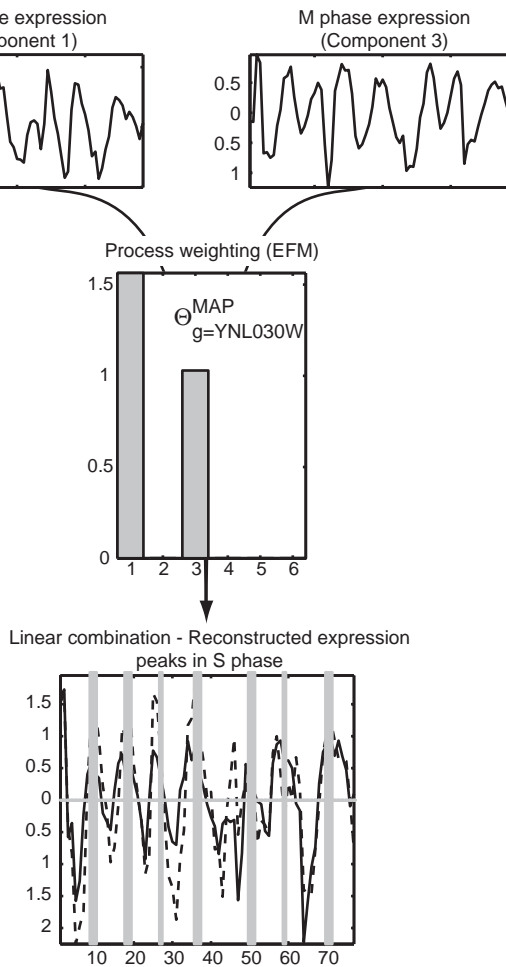


Fig. 2. A linear combination of G_1 phase and M phase processes yields the expression profile of an S phase-specific gene, in this case the $H4$ histone gene $HHF2$ (YNL030W). The top row shows the two non-zero weighted components which a linear model employing an exponential prior defines for this gene. The horizontal axis in these plots are the indices of the experimental arrays sorted by synchronization method and then by time within each of these. The vertical axis in each of the these two plots is the contribution of the particular model component to the overall measured differential expression level. The bar chart in the middle row shows the level to which the gene is active in each of the components. We can see that only components 1 and 3 are active for this particular gene. Finally, the figure in the bottom row shows the actual measured differential expression levels (dashed line) and the reconstructed levels from the model (solid line), the axes correspond to the array ordering as defined for the components in the top row.

The following section introduces the notation employed in the linear factor models which we had developed.

3 LINEAR FACTOR MODELS

In a microarray there are \mathcal{G} probed genes and \mathcal{A} experimental conditions or time points. The differential expression levels

of the microarray are represented by a matrix \mathbf{E} of dimension $\mathcal{G} \times \mathcal{A}$ where each element e_{ga} corresponds to the differential expression measured for gene g in condition (or time point) a . We assume that the differential expression levels of each microarray across genes have zero-mean value after a previous normalization. At the most general level, the expression level e_{ga} of each gene can be reconstructed by the following standard linear expansion

$$e_{ga} = \sum_{p=1}^{\mathcal{P}} A_{pa} \theta_{gp} + n_{ga}. \quad (1)$$

This means that the observed expression of each gene g in array a is the sum of its activities in each of \mathcal{P} hypothesized processes p , denoted by θ_{gp} , weighted by the activity of this process in condition (or time point) a , denoted by A_{pa} , plus some condition-specific noise n_{ga} .

The $\mathcal{P} \times \mathcal{A}$ dimensional matrix \mathbf{A} has elements A_{pa} and has \mathcal{P} rows, \mathbf{A}_p , which determine the response of the process p to the experimental conditions at each of the indices a . Likewise the $\mathcal{G} \times \mathcal{P}$ dimensional matrix Θ which has elements θ_{gp} and has \mathcal{G} rows, Θ_g , which define the levels of activity of each gene g within each of the \mathcal{P} processes. The matrix \mathbf{N} defines the reconstruction error or the linearly additive noise, possibly biological and experimental, which is added to the combined process responses for the genes and arrays under consideration. We can represent this in compact matrix format as

$$\mathbf{E} = \Theta \mathbf{A} + \mathbf{N}. \quad (2)$$

Thus, there are only three sets of factors (Θ , \mathbf{A} , \mathbf{N}) in the model that need to be defined to obtain a description that hopefully will be both powerful in explaining the observed data and biologically meaningful.

Now that the standard linear form of the level of gene representation within the combination of process responses to experimental conditions has been defined, modeling assumptions regarding the form of gene representation within processes and the process responses themselves have to be made explicit. Various biologically motivated approaches to this are described in the following section. The Appendix gives full details of the linear models which are discussed in the remainder of this paper.

4 GENE PROCESS REPRESENTATION DISTRIBUTIONAL ASSUMPTIONS

The assumptions made here will have significant impact on the level of biological validity and interpretability of the emerging model. We now consider a number of assumptions in order to increasing biological validity.

4.1 Assumption 1: Independence of gene representation in processes

We assume that the representation of an individual gene in one process has no effect on its level of representation in

any other available process. From a biological perspective this is a reasonable assumption, as expression of a gene can in principle be quite independently regulated from various independent regulatory sites (enhancers/repressors), and thus the same gene could be recruited to any combination of processes as required. Thus expression in one process does not restrict the possibility of participating in a second, quite unrelated process. Of course, this aspect of regulatory independence does not restrict the possibility that several interacting promoter sites contribute to the expression within each process.

4.2 Assumption 2: Gene process representation levels follow a Gaussian distribution

Let us consider the case where we assume that the representation of each gene in each process (i.e. its characteristic expression change) is distributed as a standardized Gaussian, $\theta_{gp} \sim \mathcal{N}(0, 1)$. Such a model would be equivalent to probabilistic PCA (Tipping and Bishop, 1997). From Assumption 1, if the representation of a gene in one process is independent of its representation in any other, then

$$p(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \prod_{p=1}^{\mathcal{P}} \mathcal{N}_{\theta_p}(0, 1), \quad (3)$$

where $\boldsymbol{\theta}$ denotes a $1 \times \mathcal{P}$ dimensional vector corresponding to a row of the matrix $\boldsymbol{\Theta}$, and $\boldsymbol{\alpha}$ denotes the vector of distribution parameter values (in this case zero-mean and unit variance for all \mathcal{P} dimensions). As such $\boldsymbol{\Theta}_g \sim \prod_{p=1}^{\mathcal{P}} \mathcal{N}(0, 1) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Making this distributional assumption makes explicit Assumption 1 that the representation of an individual gene in one process has no effect on its level of representation in any other available process (under the model).

Although the assumption of independence is biologically well motivated, it is difficult to fully justify the assumption of zero-mean unit variance Gaussianity (which yields PCA), given that this will lead to gene representation levels which are both positive and negative. The assumed symmetry about the origin immediately brings difficulties in terms of interpretation of negative gene representation, as it would mean that expression changes have to be conceptualized with reference to an imaginary sample of mean expression, so that up- and down-regulation are equally likely. Biologically, however, it may often be more suitable to interpret expression in the various processes as relative to a minimal basic expression, e.g. a set of genes that are active in all samples, while the remaining genes are then recruited (or not) specifically for selected processes (i.e. under the various experimental conditions). In such a situation, representing expression changes as occurring in just one direction may significantly facilitate the interpretation of the results. Another example of this type of asymmetry occurs when one process is, e.g. dominated by induction of one specific transcriptional repressor. Again,

expression changes would be most suitably represented as a uni-directional rather than symmetric phenomenon. This then leads us to our first biologically motivated refinement of Assumption 2.

4.3 Assumption 3: Gene process representation levels follow a Bernoulli distribution

The main weakness of the assumption made in the previous section is that a consistent biological interpretation of negative gene process representation is somewhat difficult to achieve. We made the first most biologically appropriate assumption that a gene will either be represented ('switched on') in a process or it will not. For \mathcal{P} processes, this assumption of active gene presence or absence can be encoded by independent draws (Assumption 1) from \mathcal{P} Bernoulli distributions. Each process, p , will have a Bernoulli expectation of gene representation denoted by α_p and of non-representation of $1 - \alpha_p$.

$$p(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \prod_{p=1}^{\mathcal{P}} (1 - \alpha_p) \left(\frac{\alpha_p}{1 - \alpha_p} \right)^{\theta_p}, \quad (4)$$

where $\boldsymbol{\theta} \in \{0, 1\}^{\mathcal{P}}$, i.e. it takes on binary values rather than continuous positive or negative ones as in the previous case. In addition $p(\boldsymbol{\theta} | \boldsymbol{\alpha})$ now denotes a discrete probability distribution. This form of prior has been employed previously in a particular instantiation of the probabilistic relational modeling (PRM) paradigm for decomposing gene expression into overlapping cellular processes (Segal *et al.*, 2003). It is interesting to note that a linear factor model employing such a Bernoulli prior will be identical in form to the specific form of PRM developed in Segal *et al.* (2003) (see Appendix for details).

4.4 Assumption 4: Gene process representation levels follow a uniform distribution

Moving from the assumption of isotropic Gaussian distributed gene representation random variables to discrete presence or absence Bernoulli random variables will improve the biological interpretation of gene representations under the linear factor model. Indeed a PRM based on such a prior has been shown to provide a more informative decomposition of gene expression than hierarchical clustering (e.g. Segal *et al.*, 2003).

However, it should be further noted that a gene will in fact be represented either strongly or weakly (or not at all) in a physiological process and it can then be assumed that this level of representation can be numerically defined by values in the range from zero, indicating no differential expression of the gene, to some fixed value, say α , indicating strong participation in the process function. This then provides a further biologically motivated refinement to the adoption of the previous Bernoulli prior in the model. We now assume that

all levels from inactive to fully active are all equally probable a priori.

Making explicit the assumption of independent and uniformly probable distributions across processes we can then employ the following prior:

$$p(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \prod_{p=1}^{\mathcal{P}} \frac{1}{\alpha_p} \quad (5)$$

in place of the discrete Bernoulli prior and the isotropic Gaussian prior.

4.5 Assumption 5: Gene process representation levels follow an exponential distribution

A further refinement on the prior distributional assumptions of representational activity of genes within processes can be made by noting that of all the possible genes in a process a large number will be inactive, in which case the prior probability of gene representation being zero will be higher than a finite level of activity. This can be encoded by skewing the prior distribution such that the majority of probability mass will be allocated to zero and decaying in an exponential manner for example. In this case, the distribution encoding the above model assumptions is simply

$$p(\boldsymbol{\theta} | \boldsymbol{\alpha}) = \prod_{p=1}^{\mathcal{P}} \frac{1}{\alpha_p} \exp\left(-\frac{1}{\alpha_p} \theta_p\right). \quad (6)$$

Given this range of possible priors on the levels of gene representation, it now remains to discuss how the specific forms of the model can be estimated.

5 MODEL PARAMETER ESTIMATION

The main difference between the possible variants of linear factor models discussed lies in the representation of the prototypical gene expression in each process. If the distribution of the expression levels is assumed to be Gaussian, i.e. each gene will be up-regulated in some processes, down-regulated in others (with equal probability) and is most likely to be unchanged then the linear model that these assumptions define is probabilistic PCA (refer to the Appendix for details).

In the binary model, each gene is either fully active or fully inactive in each of the processes, but no intermediate expression will be allowed. In the uniform model, more realistic intermediate expression levels are allowed, all with equal probability. In the exponential model, it is assumed that it is more likely that a gene is unchanged than that it is changed (processes being characterized by highly specific changes), so that log-relative expression levels close to zero are most probable. Having presented a number of assumed prior distributions for gene representation in processes, the issue of defining the final model has to be considered.

For the isotropic Gaussian prior (PCA), Assumptions 1 and 2, the model parameters can be defined by employing

a straightforward eigenvalue decomposition of the expression data. It is shown in the Appendix that a series of simple iterations of unconstrained (least-squares) and constrained quadratic optimizations are required for the linear models adopting the gene-process representation priors of Assumptions 4 and 5. Assumption 3 requires least squares and an unconstrained quadratic 0–1 optimization over a discrete domain (see Appendix). These iterative routines were employed in the experiments which are now reported.

6 BIOLOGICAL VALIDATION EXPERIMENTS

For the experimental evaluation, we employ the Yeast Stress and Cell-Cycle microarrays of Gasch *et al.* (2000) and Spellman *et al.* (1998). These large experimental datasets are particularly suited for validation purposes as they comprise a very diverse set of well-characterized physiological processes. Of the available genes in each data set the $\mathcal{G} = 1000$ genes with the most variable differential expression levels across experiments were chosen as the working set for the experiments conducted.

Figure 2 shows an illustrative example of how the linear factor models employing biologically motivated prior distributions can provide a biologically valid and interpretable representation. A six factor linear model employing an exponential prior (EFM) was estimated using the Cell-Cycle data. The process representation values for a particular histone gene (YNL030W) are shown as a bar chart in the center. This chart indicates that this gene participates in only two processes (all others are zero). For these two processes the process factors are shown, and it can be seen that they peak in G_1 and M phase of the cell-cycle, respectively. These processes are then weighted by their activity for the YNL030W histone gene and linearly combined to reproduce faithfully the overall level of differential expression as measured experimentally, as can be seen in the lower part of the figure. This combined reconstructed pattern now peaks at S phase, i.e. between G_1 and M phase, as is appropriate for this S phase-specific gene. Note that such a reconstruction would not be possible in a simple clustering approach, where the gene would need to be assigned exclusively to one specific process.

Owing to the probabilistic nature of the developed models the question regarding the appropriate number of processes to be included in the model is straightforwardly answered by objectively assessing the predictive power of each model with a given number of processes \mathcal{P} . The following section experimentally illustrates the estimation of the number of processes to be included in the various models.

6.1 Cross-validation of predictive likelihood

As a first objective assessment of model performance the predictive performance of each model was determined. In other words, given a new gene, which the model has not previously

been presented with, how well can the model reconstruct the levels of differential expression observed. This is assessed by measuring the probability (likelihood) of expression profiles of unseen or ‘held-out’ genes under each model. A 10-fold cross-validation is employed to provide an estimate of the predictive likelihood of each of the models. Superior models of the data generation process will yield higher levels of predictive likelihood than models which poorly reflect the generative mechanisms of the observed data. This provides an objective assessment of the predictive quality of the linear models and in addition allows an objective assessment of the most appropriate number of processes (components factors) to be employed in the model. As the number of model processes increases, the corresponding explanatory power will increase until the model starts to overfit the data. By measuring the cross-validated predictive likelihood for different values of \mathcal{P} the ‘optimal’ value can be identified. We then employ this value in defining processes² for further biological analysis.

In Figure 3, we show the cross-validated predictive likelihood on the Cell-Cycle data for a series of linear factor models with an isotropic Gaussian prior (PPCA), the linear factor model with the Bernoulli prior (BFM), the uniform prior (UFM) and the exponential prior (EFM) over a range of model-orders (number of processes \mathcal{P}). For the models with the non-Gaussian priors the optimal number of processes is found to be around six, while for PCA the turning point is not reached until 25 processes are added.

If we consider the maximum value achieved under each model it can be seen that the performance under BFM, UFM and EFM are similar, and PCA achieves a slightly higher predictive performance on this dataset, albeit for a much larger number of retained components. The reverse is the case on the Yeast Stress data (Fig. 4) where BFM, UFM and EFM significantly outperform PCA to a similar extent. As with the Cell-Cycle dataset, PCA requires a very large number of components (~ 35) to achieve its best performance and we argue that the simpler models with smaller number of processes provide a more parsimonious representation. In the next section, we will show that the simpler models also provide a biologically more valid and interpretable representation of the data.

6.2 Biological analysis

To obtain an objective assessment of the biological meaning of the various process representations, we employed iterative group analysis (iGA) (Breitling *et al.*, 2004). iGA is a statistical technique which enables an automated annotation for the processes which define the linear models. Given that each process will be defined by a group of \mathcal{G} genes which will

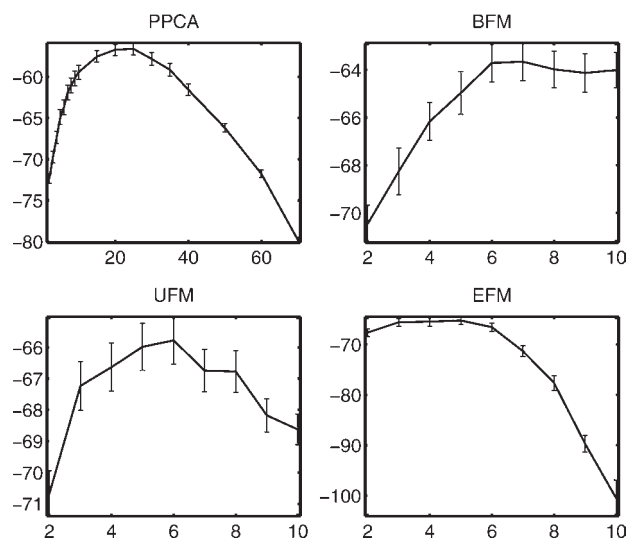


Fig. 3. The cross-validated predictive likelihood for PCA, BFM, UFM and EFM linear factor models of the Cell-Cycle data. The horizontal axis is the number of processes \mathcal{P} in each model and the vertical axis is the cross-validated log-likelihood. Significantly higher values indicate superior predictive models.

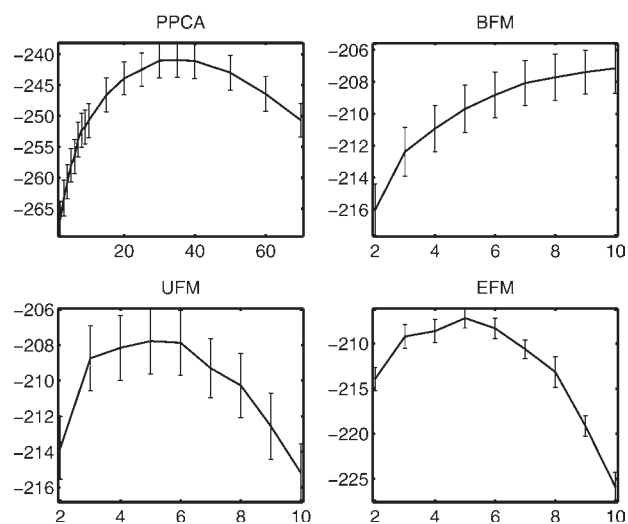


Fig. 4. The cross-validated predictive likelihood for PCA, BFM, UFM and EFM linear factor models of the Yeast Stress data. The horizontal axis is the number of processes \mathcal{P} in each model and the vertical axis is the cross-validated likelihood. BFM, UFM, and EFM achieve their best performance already for about five process, with BFM asymptotically converging after six processes. PCA reaches its maximum only after more than 35 processes are added. Significantly higher values indicate superior predictive models.

be represented (differentially expressed) to a greater or lesser degree a ranking of the representation levels of the genes can be obtained (in the case of PCA, UFM and EFM—see below for the special treatment required for BFM). The iGA method

²When referring to ‘processes’ in the remainder of the paper, it is meant that the ‘processes’ refer to the components or factors in the induced linear models. We employ the term ‘processes’ to emphasize the idea that the model factors can correspond to actual physiological processes.

is given below and the reader should refer to the original publication (Breitling *et al.*, 2004) for a less formal presentation of this statistical technique.

Let us assume that we have an already annotated functional class \mathcal{F} (from Gene Ontology for example) and \mathcal{M} of the \mathcal{G} genes from the process belong to this class. As a statistical test our null hypothesis is that there is no relationship between the ranking of genes within the process and the alternate hypothesis is that a relationship may exist. We then wish to obtain the probability of observing at least x examples from \mathcal{F} given m selections from the top of the ordered list of genes. We use the definition $x = \sum_{i=1}^m \mathcal{I}_i(\mathcal{F})$ where $\mathcal{I}_i(\mathcal{F})$ is an indicator function such that

$$\mathcal{I}_i(\mathcal{F}) = \begin{cases} 1 & g_p^i \in \mathcal{F}, \\ 0 & g_p^i \notin \mathcal{F}, \end{cases}$$

where g_p^i is the identity of the i th ranked gene in process p . Let the random variable X denote the number of examples from \mathcal{F} observed up to rank m then the cumulative distribution of the Hypergeometric distribution can be obtained iteratively using the following expression.

$$P(X \geq x | m, \mathcal{G}, \mathcal{M}) = P(X \geq x-1 | m, \mathcal{G}, \mathcal{M}) - \frac{\binom{m}{x} \binom{\mathcal{G}-m}{\mathcal{M}-x}}{\binom{\mathcal{G}}{\mathcal{M}}} \quad (7)$$

The number of genes, m , ranked by level of representation within the process, which yields a minimum P -value, $P_{\mathcal{F}}$, under the null hypothesis is then obtained by

$$P_{\mathcal{F}} = \min_m P \left(\sum_{i=1}^m \mathcal{I}_i(\mathcal{F}) | m, \mathcal{G}, \mathcal{M} \right). \quad (8)$$

The approach of using the Hypergeometric distribution to determine enriched functional classes \mathcal{F} in each process can also be used on the results from BFM, which does not provide a ranked list of genes in each process but a binary assignment of membership. However, because of this limitation the iterative minimization process of iGA is not applicable, so that the determined significance values tend to be much weaker.

6.3 Yeast cell cycle

This dataset contains the results of four experiments that trace gene expression in yeast through the cell cycle (Spellman *et al.*, 1998). The four experiments differ in the method that was used for the initial synchronization of the cell populations (alpha factor arrest, cdc15 and cdc28 mutants, elutriation). It was expected that a good, i.e. biologically meaningful, process decomposition would on the one hand distinguish between different synchronization methods, and on the other hand might recover the cyclical expression corresponding to the various cell cycle phases. The latter would be very difficult to achieve with simpler clustering techniques, as it can be expected that most samples will be intermediate between

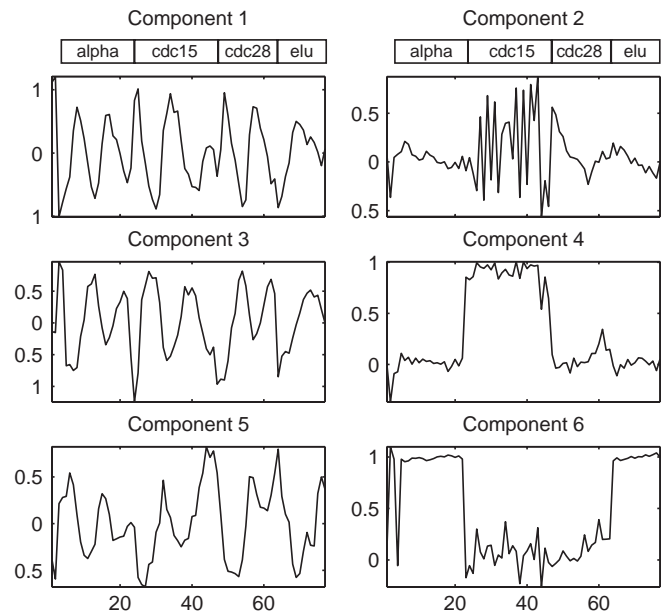


Fig. 5. Process expression profiles obtained from EFM, the description of the axes in each plot can be found in the caption to Figure 2. Three processes show a distinct cyclical activity (see Figure 6 for details). One process (Component 6) clearly distinguishes cdc15/cdc28 samples from the other two synchronization methods, one (Component 4) discriminates cdc15 samples from all others, and the last process (Component 2) is highly represented in every other cdc15 sample, which might indicate a technical artifact that was previously undetected.

phases. In this case, a linear model that allows combinations of processes contributing to each sample will be far more realistic.

Indeed, we find that the physiological processes detected by the factor model correspond to three cycling components (G_1 phase, M phase and G_1/M phase = exit from mitosis) (Figs 5 and 6), and to two treatment-specific components (distinguishing between even/odd samples in cdc15, between cdc15 versus other synchronization methods, and between cdc15/28 versus synchronization by alpha-factor/elutriation). The sixth process peaks in every other cdc15-synchronized sample and may be due to a previously undetected technical artifact. The linear factor processes can thus be used to classify samples, which is very useful, but go way beyond that in their ability to distinguish processes that are present in variable combinations in the biological samples. The iGA interpretation of the processes confirms that they are indeed biologically meaningful and contain the genes that are expected for the corresponding cell-cycle phases (Table 1).

It is particularly interesting to note that processes 1 and 3 both express (the same) histone genes (S phase genes) and a linear combination of these two processes indeed peaks at S phase (Figs 2 and 7).

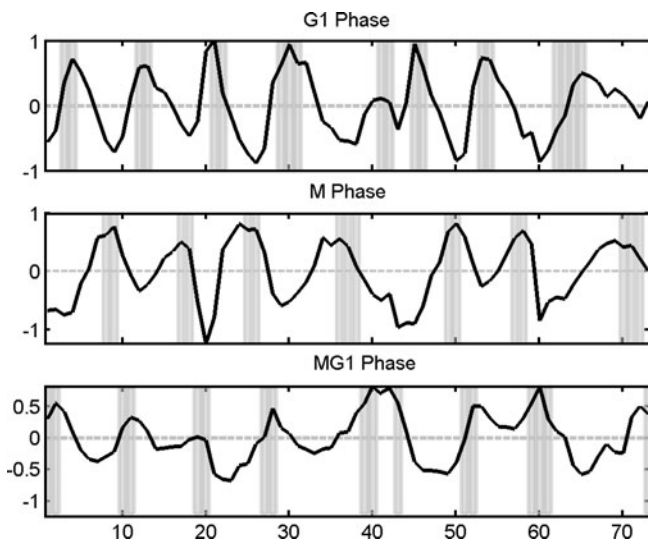


Fig. 6. Expression profile of the three cyclic processes obtained by EFM, the description of the axes in each plot can be found in the caption to Figure 2. Each of these processes corresponds to one distinct phase of the cell cycle (shaded) as determined by Spellman *et al.* based on morphological and molecular criteria (their Figure 1). Together these three processes and their linear combinations cover the whole cell cycle (Fig. 2).

Table 1. Correspondence between factor model processes and cell cycle phases

Process	G ₁	S	G ₂	M	M/G ₁	NA	Total
1(G₁)	22 DR DS CC	7 H	—	—	—	6 CY	35
2	—	—	—	—	—	6 SP	6
3 M	—	8 H	1	7 CC CY	—	0	16
4	—	—	—	—	—	0	0
5(M/G₁)	2 MF CY	—	—	3 MF	10 CW CY MF	19 HS	34
6	—	—	—	—	—	12 RS	12
Total	119	37	34	61	41	—	292

We performed iGA with Gene Ontology annotations on the gene process representations from EFM applied to the yeast cell cycle. For each process, we determined to which cell cycle phase the detected genes had been assigned by Spellman *et al.* (their Figure 7). It can be seen clearly that each of the three cycling processes (boldface) is significantly enriched in the appropriate genes. It can also be seen that both Process 1 (G₁ phase) and Process 3 (M phase) contain histone expression (S phase-specific), and their linear combination results in an S phase expression profile (Fig. 7). The most common biological functions within each group are indicated. The total number in the last row corresponds to the gene numbers in Figure 7 of Spellman *et al.* Abbreviations: DR, DNA repair; DS, DNA synthesis; CC, cell cycle control; H, Histones; MF, mating factor signalling; CW, cell wall; CY, cytokinesis; HS, heat shock chaperones; RS, Ribosomes; SP, sporulation; NA, not assigned.

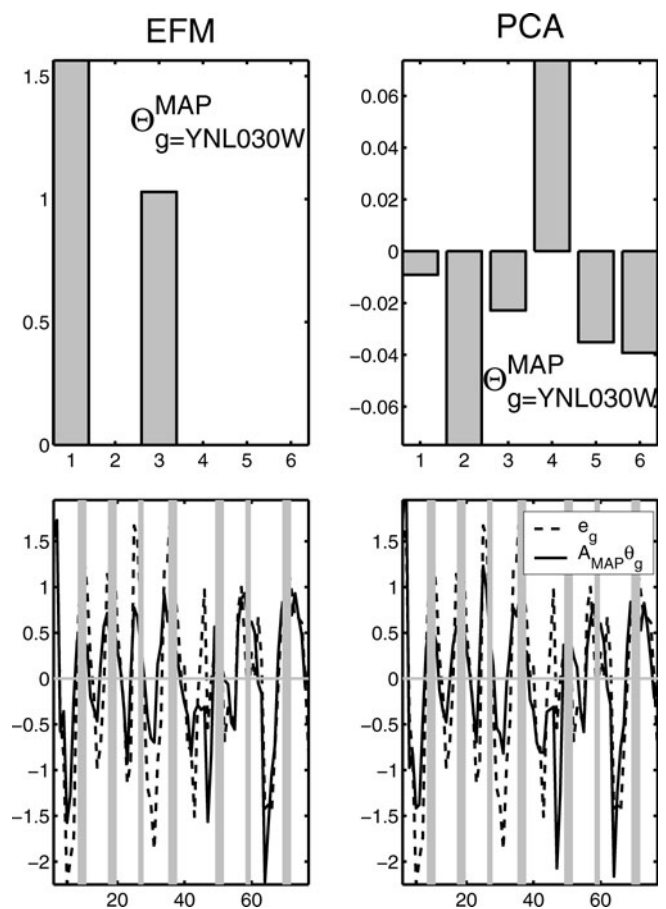


Fig. 7. Process representation of an S phase-specific histone gene in EFM, and PCA. In EFM this gene is represented by about equal amounts of Process 1 (G₁ phase) and Process 3 (M phase) (Top left-hand chart), and their linear combination indeed results in a reconstructed expression pattern that peaks at S phase, i.e. exactly between G₁ and M phase (Bottom left-hand chart); In PCA, the reconstructed expression fits equally well (Bottom right-hand chart); however, the process representation (Top right-hand chart) is much less distinct and impossible to interpret biologically.

All non-Gaussian factor models (BFM, UFM, EFM) yield similar results in this respect. The main difference between them is the weakness of BFM for finding significant changes by iGA: even when using a strict Bonferroni correction for UFM/EFM, and a relaxed one for BFM, the former find more significant annotations and are able to annotate more of the processes (2 annotated processes, 8 functional classes for BFM; 5 processes, 29 functions for EFM). As expected, the same trend is observed in the stress experiments discussed below. The reason is the strict (binary) classification of genes in the process representations, which does not allow the sensitive iterative significance estimations of iGA.

In contrast, the results clearly show that PCA fails miserably to separate the biologically relevant information (cf. Figs 5 and 8). Only one of the first six principal components is clearly

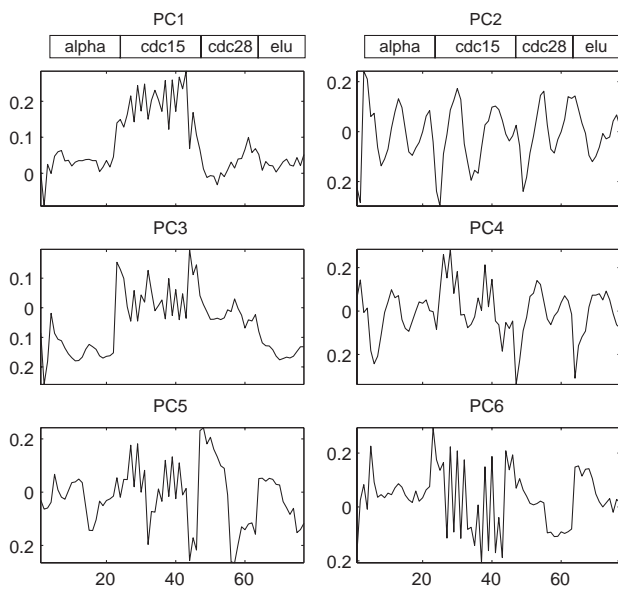


Fig. 8. Process expression profiles obtained by PCA, the description of the axes in each plot can be found in the caption to Figure 2. Only one of the processes (PC2) is clearly cyclical, and one (PC1) shows a relatively clear distinction between *cdc15* samples and the others. The remaining four processes are hardly interpretable biologically.

cyclical, and one distinguishes between *cdc15* and the other synchronization methods, albeit not as clearly as the factor models. But the majority of the principal components do not show a distinct pattern or relate to any explicable biological processes.

6.4 Yeast stress

In this dataset, yeast cells were exposed to various environmental stresses (e.g. heat and cold shock, various oxidative stresses, nitrogen and amino acid starvation, diauxic shift, stationary phase) (Gasch *et al.*, 2000). Most of the stresses were examined along a time series and some were also varied in intensity. As many stresses are known to lead to similar molecular effects (e.g. starvation, protein unfolding) they are expected to activate overlapping physiological responses (e.g. efficient nutrient management, protection against misfolded proteins). For this dataset, PCA is not that obviously inferior to the factor models, although its predictive performance was clearly worse than that of the factor models (Fig. 4). The reason for that is the physiological complexity of the processes underlying this experiment. But even here PCA fails to detect some expected physiological processes, e.g. the prominent and characteristic up-regulation of genes upon entry into stationary phase.

The following five major processes (Fig. 9) were identified in the stress data by the linear factor models:

- (1) Progressive down-regulation during nitrogen depletion and up-regulation in oxidative stress.

- (2) Progressive up-regulation during nitrogen depletion and amino acid starvation.
- (3) Up-regulation in response to heat shock (with a peak at 20–30 min) with more intense response when the shock ‘gradient’ is steeper. This process is also activated in early diamide treatment, i.e. in response to the destabilization of proteins caused by this sulfhydryl oxidizing agent, which has a similar effect as the heat treatment).
- (4) Up-regulation in stationary phase (including late diauxic shift and late nitrogen starvation experiments).
- (5) A relatively unspecific response, dominated by down-regulation in stationary phase and heat shock, but also reacting to many other treatments.

As the experimental treatment in this dataset is far more complex than in the cell-cycle study, the iGA interpretation of the five processes is not as straightforward. However, it can be seen from iGA that Process 1, which peaks during oxidative stress and is minimal during nitrogen starvation, is dominated by up-regulation of oxidative stress responses (glutathione peroxidases, thioredoxin peroxidases) and down-regulation of allantoin catabolism (which in nitrogen-starved cells breaks down nitrogen-rich substrates to recover NH_3 for re-use). Process 2, which peaks in nitrogen and amino acid starvation, also up-regulates oxidative stress responses, but also heat-shock proteins, and down-regulates amino acid permeases (which would lead to a loss of valuable nitrogen-containing substances in nitrogen-starved conditions) and translational elongation (due to lack of substrate during nitrogen starvation). The heat-shock process (Process 3) expectedly up-regulates a variety of heat-shock responses. It also down-regulates various nucleolar components (as does Process 1) it seems that this pattern, as well as the apparent induction of oxidative stress in both Process 1 and 2, may be for a similar reason as the histone effect in the cell cycle study, i.e. a linear combination of the two results in an oxidative stress-specific pattern. Process 4 is rather non-descript in the iGA analysis. It shows up-regulation of some osmotic stress genes, but not the dramatic expression changes that are known to occur during entry into stationary phase. Process 5, on the other hand, is characterized by an up-regulation of ribosomal biogenesis (which is known to be drastically down-regulated in stationary phase) and down-regulation of carbohydrate metabolism (which may be related to the fact that stationary phase as well as the diauxic shift is characterized by a depletion of carbon sources). Three of the factor model processes map quite closely to the first three principal components of PCA (Component 2 and PC3; Component 3 and PC2; Component 5 and PC1), the other two do not. In contrast to the cell cycle results the difference in biological content is not as striking, but it can still be seen that the linear models identify processes that are more distinct than several of the principal components.

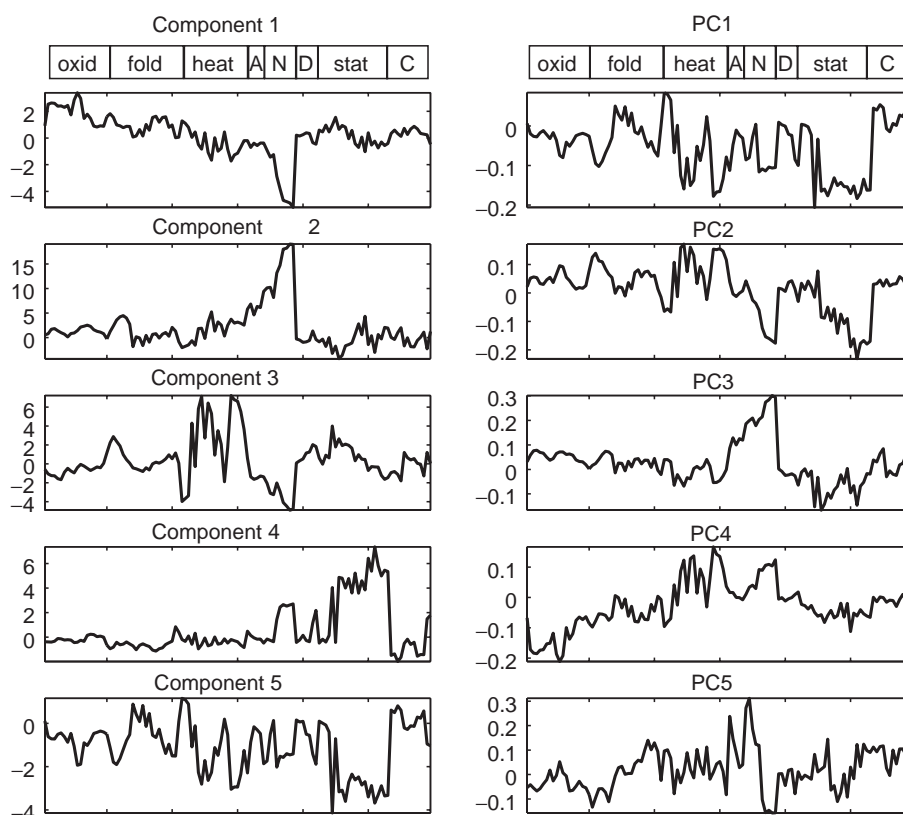


Fig. 9. Process expression profiles obtained by employing the factor models (left-hand column) and process expression profiles obtained by PCA (right-hand column). The horizontal axis in each figure is ordered by the treatments employed by Gasch *et al.* (2000) and within each treatment ordering is either by intensity or time. The vertical axis in each of these two plots is the contribution of the particular model component to the overall measured differential expression level. The labels above each correspond to the following: Oxid, oxidative stress (H_2O_2 and menadione); fold, protein unfolding (dtf and diamide); heat, heat shock; A, amino acid starvation; N, nitrogen starvation; D, diauxic shift; Stat, stationary phase; and C, carbon source variation.

7 CONCLUSION

We conclude by noting that using linear factor models employing appropriate, biologically valid prior distributions for gene process representation instead of standard PCA leads to improved biological validity of the expression data decomposition. Among the tested linear factor models, the more realistic UFM and EFM have the important advantage that their gene process representation can be analyzed by regular iGA to identify biologically important functional annotations. In the test cases presented here, BFM, UFM and EFM all have very similar cross-validated predictive likelihood as would naturally be expected. The level of biological interpretation facilitated by the factor models is superior to that of straight application of PCA. This indicates that the linear factor models incorporating the more realistic biological assumptions are indeed highly recommendable alternatives to PCA for the analysis and interpretation of microarray results. As each biologically meaningful linear component is expected to be dominated by one or a few transcriptional regulatory modules,

such an analysis can be considered a first step towards the reconstruction of regulatory networks underlying the observed expression pattern.

ACKNOWLEDGEMENTS

This work was supported in part by a grant from the BBSRC (17/GG17989).

REFERENCES

- Alter, O., Brown, P.O. and Botstein, D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci., USA*, **97**, 10101–10106.
- Blehariski, J.R., Li, H., Meinken, C., Graeber, T.G., Ochoa, M.T., Yamamura, M., Burdick, A., Sarno, E.N., Wagner, M., Rollinghoff, M., *et al.* (2003) Use of genetic profiling in leprosy to discriminate clinical forms of the disease. *Science*, **301**, 1527–1530.
- Breitling, R., Amtmann, A. and Herzyk, P. (2004) Iterative group analysis (iGA): a simple tool to enhance sensitivity and facilitate

- interpretation of microarray experiments. *BMC Bioinformatics*, **5**, 34.
- Chardaire, P. and Sutter, A. (1995) A decomposition method for quadratic 0-1 programming. *Manage. Sci.*, **41**, 704–712.
- Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.
- Gasch, A.P., Spellman, P.T., Kao, C.M., Carmel-Harel, O., Eisen, M.B., Storz, G., Botstein, D. and Brown, P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.
- Girolami, M. (ed.) (2000) *Advances in Independent Component Analysis*. Springer-Verlag.
- Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D. *et al.* (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.
- Hyvarinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*. Wiley.
- Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y. and Barkai, N. (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.
- Lawson, C.L. and Hanson, B.J. (1974) *Solving Least Squares Problems*. Prentice-Hall.
- Lee, S.I. and Batzoglou, S. (2003) Application of independent component analysis to microarrays. *Genome Biol.*, **4**, R76.
- Mao, R., Zielke, C.L., Zielke, H.R. and Pevsner, J. (2003) Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain. *Genomics*, **81**, 457–467.
- Martoglio, A.M., Miskin, J.W. and MacKay, D.J.C. (2002) A decomposition model to track gene expression signatures: preview on observer-independent classification of ovarian cancer. *Bioinformatics*, **18**, 1617–1624.
- Palubeckis, G. (1992) Heuristics with a worst-case bound for unconstrained quadratic 0-1 programming. *Informatica*, **3**, 225–240.
- Raychaudhuri, S., Stuart, J.M. and Altman, R.B. (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Pacific Symposium on Biocomput.* pp. 455–466.
- Roberts, S. and Everson, R. (eds) (2001) *Independent Component Analysis Principles and Practice*. Cambridge University Press, Cambridge.
- Segal, E., Battle, A. and Koller, D. (2003) Decomposing gene expression into cellular components. In *Pacific Symposium on Biocomputing*. pp. 89–100.
- Segal, N.H., Pavlidis, P., Noble, W.S., Antonescu, C.R., Viale, A., Wesley, U.V., Busam, K., Gallardo, H., DeSantis, D., Brennan, M.F. *et al.* (2003) Classification of clear-cell sarcoma as a subtype of melanoma by genomic profiling. *J. Clin. Oncol.*, **21**, 1775–1781.
- Spellman, P.T., Sherlock, G., Zhang, M.Q., Vishwanath, R.I., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.
- Tarte, K., Zhan, F., De Vos, J., Klein, B. and Shaughnessy, J., Jr (2003) Gene expression profiling of plasma cells and plasmablasts: toward a better understanding of the late stages of B-cell differentiation. *Blood*, **102**, 592–600.
- Tipping, M.E. and Bishop, C.M. (1997) Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B*, **61**, 611–622.
- Tsunoda, T., Koh, Y., Koizumi, F., Tsukiyama, S., Ueda, H., Taguchi, F., Yamaue, H., Saijo, N. and Nishio, K. (2003) Differential gene expression profiles and identification of the genes relevant to clinicopathologic factors in colorectal cancer selected by cDNA array method in combination with principal component analysis. *Int. J. Oncol.*, **23**, 49–59.

8 APPENDIX

8.1 Linear model with Gaussian prior

Here, we make explicit the distributional assumptions of the linear model. First, the noise or reconstruction error can be attributed to a number of experimental artifacts (noise) as well as possible biological fluctuations in expression level that are unrelated to the experimental treatment of interest (see e.g. Hughes *et al.*, 2000).³ which will have an additive effect. This additive effect can reasonably be modeled using a Gaussian distribution.

Second, the possible values which the process responses A_{pa} may take can be represented by noting that each row of \mathbf{A} , denoting the response of a process in all array conditions, can have a mean response of zero such that $A_{pa} \sim \mathcal{N}(0, 1/\beta_p)$ where $\mathcal{N}(b, c)$ denotes a Gaussian distribution with mean b and variance c . By assuming each element of the process, responses is independent and making the further simplifying assumption that the precision β_p is the same for all processes then

$$p(\mathbf{A}|\beta) = \left(\frac{\beta}{2\pi}\right)^{PA/2} \exp\left\{-\frac{\beta}{2}\|\mathbf{A}\|^2\right\}, \quad (9)$$

where $\|\cdot\|$ denotes the Frobenius norm.

The distribution of the noise term can be represented as an isotropic zero-mean Gaussian with a variance σ^2 , in which case, $p(\mathbf{E}_g|\Theta_g, \mathbf{A}, \sigma^2) = \mathcal{N}_{\mathbf{E}_g}(\Theta_g \mathbf{A}, \mathbf{I}\sigma^2)$, where \mathbf{I} denotes an identity matrix of appropriate dimension, in this case $\mathcal{A} \times \mathcal{A}$. For the case where the degenerate distribution over \mathbf{A} is taken, i.e. $\beta \rightarrow 0$ and the isotropic Gaussian on gene process representation is adopted then this linear model of gene array differential expression levels is a probabilistic PCA (PPCA) (Tipping and Bishop, 1997). By this we can then explicitly state the model assumptions implicit in the analysis undertaken (Alter *et al.*, 2000; Raychaudhuri *et al.*, 2000).

The probability of the differential expression given the value of the process activity matrix \mathbf{A} and noise variance σ^2 can

³As the expression data is typically the \log_2 transformed differential expression with respect to a reference sample, it can be argued that the effects of noise are neither additive nor Gaussian, however to make progress in providing a parsimonious representation of the data a linear effect with additive noise is a reasonable modeling assumption to adopt.

be obtained by marginalizing the gene process representation variable

$$\begin{aligned} p(\mathbf{E}_g | \mathbf{A}, \sigma^2) &= \int p(\mathbf{E}_g | \boldsymbol{\theta}, \mathbf{A}, \sigma^2) p(\boldsymbol{\theta} | \boldsymbol{\alpha}) d\boldsymbol{\theta}, \\ &= \int \mathcal{N}_{\mathbf{E}_g}(\boldsymbol{\theta} \mathbf{A}, \mathbf{I} \sigma^2) \mathcal{N}_{\boldsymbol{\theta}}(\boldsymbol{\theta}, \mathbf{I}) d\boldsymbol{\theta}, \\ &= \mathcal{N}_{\mathbf{E}_g}(\mathbf{0}, \mathbf{A}^\top \mathbf{A} + \mathbf{I} \sigma^2), \end{aligned}$$

where $\boldsymbol{\theta}$ is a random variable representing values of a row from $\boldsymbol{\Theta}$. Employing Gaussian integrals it is straightforward to obtain the posterior distribution of the gene process representation variable given the differential expression levels and values of \mathbf{A} and σ^2 .

$$p(\boldsymbol{\theta} | \mathbf{E}_g, \mathbf{A}, \sigma^2) = \mathcal{N}_{\boldsymbol{\theta}} \left(\left[(\mathbf{A} \mathbf{A}^\top + \sigma^2 \mathbf{I})^{-1} \mathbf{A} \mathbf{E}_g^\top \right]^\top, (\mathbf{A} \mathbf{A}^\top + \sigma^2 \mathbf{I})^{-1} \sigma^2 \right), \quad (10)$$

where $\mathcal{N}_a(b, c)$ denotes the probability of a value a under a normal distribution with mean b and variance/covariance c .

The maximum-likelihood solution \mathbf{A}_{ML} can be obtained in closed form employing an eigenvalue decomposition of the sample covariance matrix (Tipping and Bishop, 1997) of the differential array expressions such that $\frac{1}{G} \mathbf{E}^\top \mathbf{E} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^\top$

$$\mathbf{A}_{ML} = \mathbf{U}_{\mathcal{P}} (\boldsymbol{\Lambda}_{\mathcal{P}} - \mathbf{I} \sigma^2)^{1/2} \mathbf{R}, \quad (11)$$

where $\mathbf{U}_{\mathcal{P}}$ is the \mathcal{P} dominant columns of the matrix of eigenvectors \mathbf{U} , $\boldsymbol{\Lambda}_{\mathcal{P}}$ is the diagonal matrix of the \mathcal{P} principal eigenvalues and \mathbf{R} is an arbitrary rotation matrix (Tipping and Bishop, 1997).

We can define the posterior mean, which is the maximum a posteriori (MAP) value of $\boldsymbol{\theta}$, as

$$\begin{aligned} \boldsymbol{\Theta}_g^{\text{MAP}} &= E_{p(\boldsymbol{\theta} | \mathbf{E}_g)} \{\boldsymbol{\theta} | \mathbf{E}_g\} \\ &= \left[(\mathbf{A}_{ML} \mathbf{A}_{ML}^\top + \sigma^2 \mathbf{I})^{-1} \mathbf{A}_{ML} \mathbf{E}_g^\top \right]^\top. \end{aligned} \quad (12)$$

The maximum-likelihood solution of PPCA for the process responses \mathbf{A}_{ML} should provide a biologically interpretable profile for each process across the measured experimental conditions and the MAP value of the gene process representation $\boldsymbol{\Theta}_g^{\text{MAP}}$ should provide information to make inferences regarding the nature of the process groupings identified by the model. Each $\boldsymbol{\Theta}_g^{\text{MAP}}$ could be regarded as a characteristic expression pattern of a certain process, corresponding to the hypothetical expression that would be measured if this one process could be observed in isolation in an experiment. Thus, biological analysis of this expression pattern using methods developed for the interpretation of microarray results should be able to identify the biological nature of each process.

8.2 Linear models with biologically motivated priors

Although a full Bayesian treatment of the linear models under consideration is straightforward, in this study we do not consider the parameter β as a random variable but make assumptions on possible values it may take and consider this fixed for each model. We then have the following sets of unknowns to estimate \mathbf{A} , $\boldsymbol{\Theta}$, σ^2 , and $\boldsymbol{\alpha}$ given the measured experimental results \mathbf{E} and fixed value of β . The posterior probability of the process activity matrix can be obtained from

$$\begin{aligned} p(\mathbf{A} | \boldsymbol{\Theta}, \boldsymbol{\alpha}, \sigma^2, \beta, \mathbf{E}) &= \frac{p(\mathbf{A}, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \sigma^2, \beta, \mathbf{E})}{p(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \sigma^2, \beta, \mathbf{E})}, \\ &= \frac{p(\mathbf{E} | \mathbf{A}, \boldsymbol{\Theta}, \sigma^2) p(\mathbf{A} | \beta) p(\boldsymbol{\Theta} | \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\sigma^2) p(\beta)}{p(\mathbf{E}, \boldsymbol{\Theta} | \boldsymbol{\alpha}, \beta, \sigma^2) p(\boldsymbol{\alpha}) p(\sigma^2) p(\beta)}, \\ &= p(\mathbf{E} | \mathbf{A}, \boldsymbol{\Theta}, \sigma^2) p(\mathbf{A} | \beta) \frac{p(\boldsymbol{\theta} | \boldsymbol{\alpha})}{p(\mathbf{E}, \boldsymbol{\Theta} | \boldsymbol{\alpha}, \beta, \sigma^2)}, \end{aligned} \quad (13)$$

where it is assumed that $\mathbf{A} | \beta$ and $\boldsymbol{\Theta} | \boldsymbol{\alpha}$ are independent, and the joint probability of $\boldsymbol{\alpha}$, β and σ^2 is factorable. Similarly, the posterior probability for the gene process representations can be given as the following expression.

$$\begin{aligned} p(\boldsymbol{\Theta} | \mathbf{A}, \boldsymbol{\alpha}, \sigma^2, \beta, \mathbf{E}) &= \frac{p(\boldsymbol{\Theta}, \mathbf{A}, \boldsymbol{\alpha}, \sigma^2, \beta, \mathbf{E})}{p(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \sigma^2, \beta, \mathbf{E})} \\ &= \frac{p(\mathbf{E} | \mathbf{A}, \boldsymbol{\Theta}, \sigma^2) p(\mathbf{A} | \beta) p(\boldsymbol{\Theta} | \boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\sigma^2) p(\beta)}{p(\mathbf{E}, \mathbf{A} | \boldsymbol{\alpha}, \beta, \sigma^2) p(\boldsymbol{\alpha}) p(\sigma^2) p(\beta)} \\ &= p(\mathbf{E} | \mathbf{A}, \boldsymbol{\Theta}, \sigma^2) p(\boldsymbol{\theta} | \boldsymbol{\alpha}) \frac{p(\mathbf{A} | \beta)}{p(\mathbf{E}, \mathbf{A} | \boldsymbol{\alpha}, \beta, \sigma^2)} \end{aligned} \quad (14)$$

and the posterior for σ^2 follows as

$$\begin{aligned} p(\sigma^2 | \mathbf{A}, \boldsymbol{\Theta}, \boldsymbol{\alpha}, \beta, \mathbf{E}) &= p(\mathbf{E} | \mathbf{A}, \boldsymbol{\Theta}, \sigma^2) p(\sigma^2) \frac{p(\mathbf{A} | \beta) p(\boldsymbol{\theta} | \boldsymbol{\alpha})}{p(\mathbf{E}, \mathbf{A} | \boldsymbol{\alpha}, \beta)}. \end{aligned} \quad (15)$$

We can adopt a MAP approach here by noting that the following standard iterations will yield a monotonic increase in the likelihood of the data given the model

$$\begin{aligned} \mathbf{A}_{\text{MAP}} &= \underset{\mathbf{A}}{\operatorname{argmax}} \log p(\mathbf{A} | \boldsymbol{\Theta}_{\text{MAP}}, \boldsymbol{\alpha}, \sigma_{\text{MAP}}^2, \beta, \mathbf{E}) \\ &= \underset{\mathbf{A}}{\operatorname{argmax}} \log \left\{ p(\mathbf{E} | \mathbf{A}, \boldsymbol{\Theta}_{\text{MAP}}, \sigma_{\text{MAP}}^2) p(\mathbf{A} | \beta) \right\} \\ &= \underset{\mathbf{A}}{\operatorname{argmax}} \sum_{g=1}^G \log \left\{ \mathcal{N}_{\mathbf{E}_g}(\boldsymbol{\Theta}_g^{\text{MAP}} \mathbf{A}, \mathbf{I} \sigma_{\text{MAP}}^2) \mathcal{N}_{\mathbf{A}}(\mathbf{0}, \beta \mathbf{I}) \right\}. \end{aligned} \quad (16)$$

$$\begin{aligned}
 \Theta_g^{\text{MAP}} &= \operatorname{argmax}_{\theta} \log p(\Theta | \mathbf{A}_{\text{MAP}}, \alpha, \sigma_{\text{MAP}}^2, \beta, \mathbf{E}) \\
 &= \operatorname{argmax}_{\theta} \log \left\{ p(\mathbf{E} | \mathbf{A}_{\text{MAP}}, \Theta, \sigma_{\text{MAP}}^2) p(\theta | \alpha) \right\} \\
 &= \operatorname{argmax}_{\theta} \log \left\{ \mathcal{N}_{\mathbf{E}_g}(\theta \mathbf{A}_{\text{MAP}}, \mathbf{I} \sigma_{\text{MAP}}^2) p(\theta | \alpha) \right\}.
 \end{aligned} \tag{17}$$

$$\sigma_{\text{MAP}}^2 = \operatorname{argmax}_{\sigma^2} \sum_{g=1}^G \log \left\{ \mathcal{N}_{\mathbf{E}_g}(\Theta_g^{\text{MAP}} \mathbf{A}_{\text{MAP}}, \mathbf{I} \sigma^2) p(\sigma^2) \right\}. \tag{18}$$

These simplify to the following optimization steps where the form of the prior $p(\theta | \alpha)$ differentiates each eventual model. This interleaving of iterations yields a particularly simple form, where $\mathbf{H} = \mathbf{A}_{\text{MAP}} \mathbf{A}_{\text{MAP}}^{\text{T}}$, such that

$$\mathbf{A}_{\text{MAP}} = (\Theta_{\text{MAP}}^{\text{T}} \Theta_{\text{MAP}} + \mathbf{I} \beta \sigma_{\text{MAP}}^2)^{-1} \Theta_{\text{MAP}}^{\text{T}} \mathbf{E} \tag{19}$$

$$\begin{aligned}
 \Theta_g^{\text{MAP}} &= \operatorname{argmin}_{\theta} \frac{1}{2} \theta \mathbf{H} \theta^{\text{T}} \\
 &\quad - \theta \mathbf{A}_{\text{MAP}} \mathbf{E}_g^{\text{T}} - \sigma_{\text{MAP}}^2 \log p(\theta | \alpha)
 \end{aligned} \tag{20}$$

$$\sigma_{\text{MAP}}^2 = \frac{1}{\mathcal{AG}} \sum_{g=1}^G \|\mathbf{E}_g - \Theta_g^{\text{MAP}} \mathbf{A}_{\text{MAP}}\|^2 \tag{21}$$

subject to various appropriate constraints on each θ . We have assumed a uniform prior for σ^2 in which case the MAP estimate is simply the maximum-likelihood solution, the values for α are obtained by maximum-likelihood.

8.2.1 Bernoulli prior Substituting the product of \mathcal{P} Bernoulli distributions in the above $\log p(\theta | \alpha)$ then we obtain a classical unconstrained 0-1 quadratic optimization problem for each gene (Chardaire and Sutter, 1995).

$$\Theta_g^{\text{MAP}} = \operatorname{argmin}_{\theta} \frac{1}{2} \theta \mathbf{H} \theta^{\text{T}} - \theta (\mathbf{A}_{\text{MAP}} \mathbf{E}_g^{\text{T}} + \mathbf{b}^{\text{T}}), \tag{22}$$

where each of the \mathcal{P} elements of the row vector \mathbf{b} is $\sigma_{\text{MAP}}^2 \log \alpha_p / (1 - \alpha_p)$. This is an NP-hard problem and can be solved trivially by exhaustively enumerating all possible $\theta \in \{0, 1\}^{\mathcal{P}}$ as is adopted by (Segal *et al.*, 2003) or by employing algorithms such as those developed by Chardaire and Sutter (1995), however this very soon becomes intractable for reasonable sized \mathcal{P} and heuristic algorithms are required, (see, e.g. Palubeckis, 1992). The estimated α_p terms

are obtained by the standard maximum-likelihood solution i.e. $\alpha_p = \frac{1}{G} \sum_{g=1}^G \theta_{gp}^{\text{MAP}} \forall p$.

8.2.2 Uniform prior Let us assume a priori that each θ_{pg} lies in the pre-set range $[0, \alpha_p]$ then we require to solve

$$\Theta_g^{\text{MAP}} = \operatorname{argmin}_{\theta} \frac{1}{2} \theta \mathbf{H} \theta^{\text{T}} - \theta \mathbf{A}_{\text{MAP}} \mathbf{E}_g^{\text{T}} \tag{23}$$

subject to $0 \leq \theta_p \leq \alpha_p \forall p$; interestingly, this turns out to be a straightforward quadratic program having positivity constraints and an upper-bound on the feasible solution.

8.2.3 Exponential prior Assuming that the mean values of the exponentials is $\alpha_p \forall p$ then the generic nonlinear optimization reduces to the following quadratic form

$$\Theta_g^{\text{MAP}} = \operatorname{argmin}_{\theta} \frac{1}{2} \theta \mathbf{H} \theta^{\text{T}} - \theta \left(\mathbf{A}_{\text{MAP}} \mathbf{E}_g^{\text{T}} - \frac{\sigma_{\text{MAP}}^2}{\alpha} \right) \tag{24}$$

subject to $\theta_p \geq 0 \forall p$, where α denotes a \mathcal{P} -dimensional row vector whose values are each α_p and the fractional term $\sigma_{\text{MAP}}^2 / \alpha$ denotes elementwise division. This boils down to a straightforward non-negative least-squares problem and standard algorithms can be employed in finding a solution satisfying the constraints (Lawson and Hanson, 1974). The maximum-likelihood values of each α_p are obtained in the usual manner, i.e. $\alpha_p = \frac{1}{G} \sum_{g=1}^G \theta_{gp}^{\text{MAP}} \forall p$.

Although in the experiments reported the value of α is pre-set to unit value, it should be noted that the smaller this value is set (smaller mean value) the sparser will be the solution for Θ and hence this provides a means of controlling the level of solution sparsity.

The likelihood under the BFM model can be computed by complete marginalization of the possible gene-process representation values θ , while numerical sampling from the priors is employed to obtain likelihood estimates for UFM and EFM.

8.3 Computational scaling

The required matrix inversion to obtain \mathbf{A}_{MAP} will yield an $\mathcal{O}(\mathcal{P}^3)$ scaling per iteration and the constrained quadratic programs for the Uniform and Exponential priors will yield, in the worst case, $\mathcal{O}(\mathcal{G}\mathcal{P}^3)$ cubic scaling while the combinatorial minimization for the Bernoulli prior employing complete enumeration will yield $\mathcal{O}(\mathcal{G}2^{\mathcal{P}})$ exponential scaling if no appropriate heuristics are employed.