BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i144–i151
doi:10.1093/bioinformatics/bti1041

Multi-way clustering of microarray data using probabilistic sparse matrix factorization

Delbert Dueck*, Quaid D. Morris and Brendan J. Frey

Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada M5S 3G4

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Motivation: We address the problem of multi-way clustering of microarray data using a generative model. Our algorithm, probabilistic sparse matrix factorization (PSMF), is a probabilistic extension of a previous hard-decision algorithm for this problem. PSMF allows for varying levels of sensor noise in the data, uncertainty in the hidden prototypes used to explain the data and uncertainty as to the prototypes selected to explain each data vector.

Results: We present experimental results demonstrating that our method can better recover functionally-relevant clusterings in mRNA expression data than standard clustering techniques, including hierarchical agglomerative clustering, and we show that by computing probabilities instead of point estimates, our method avoids converging to poor solutions.

Contact: delbert@psi.toronto.edu

1 INTRODUCTION

Many kinds of data can be viewed as consisting of a set of vectors, each of which is a noisy combination of a small number of noisy prototype vectors. Moreover, these prototype vectors may correspond to different hidden variables that play a meaningful role in determining the measured data. For example, a gene's expression is influenced by the presence of transcription factor proteins, and two genes may be activated by overlapping sets of transcription factors. Consequently, the activity of each gene can be explained by the activities of a small number of transcription factors.

This task can be viewed as the problem of factorizing a data matrix, while taking into account constraints reflecting the structural knowledge of the problem and the probabilistic relationships between variables that are induced by known uncertainties in the problem. A simple example of a technique for finding such a factorization is principal components analysis (PCA). In this paper we study algorithms for finding matrix factorizations, but with a specific focus on sparse factorizations and on properly accounting for uncertainties while computing the factorization.

Our interest in algorithms for 'probabilistic sparse matrix factorization' (PSMF) is motivated by a problem identified during our collaborations with researchers working in the area of molecular biology. By viewing microarray expression profiles as points in a vector space, researchers have been able to use well-known vector-space data analysis techniques to identify new patterns of biological significance and make predictions based on previous studies. In particular, the data matrices of these profiles have been used in the large-scale prediction of gene function for genes with unknown function, based on genes with known function (Hughes *et al*., 2000). Because many biological functions depend upon the coordinated expression of multiple genes, similarity of expression profile often implies similarity of function (Marcotte *et al*., 1999). This relationship has been exploited to predict the function of uncharacterized genes by various researchers (see e.g. Brown *et al*., 2000). These schemes make use of annotation databases like Gene Ontology (Ashburner *et al*., 2000), which assign genes to one or more predefined functional categories.

However, noise in the expression measurements has limited the predictive accuracy of these algorithms, especially in those categories containing a small number of genes. These 'small' categories can be of greater interest because less is typically known about them; furthermore, they make specific (and thus more easily confirmed) predictions about gene function.

Here, we introduce an unsupervised technique that jointly denoises expression profile data and computes a sparse matrix factorization, thereby rendering a multi-way classification of the data. Our technique models the underlying causes of the expression data in terms of a small number of hidden factors. The representation of the expression profile in terms of these hidden factors is less noisy; here we explore whether this representation is more amenable to functional prediction.

Our technique explicitly maximizes a lower bound on the log-likelihood of the data under a probability model. The sparse encoding found by our method can be used for a variety of tasks, including functional prediction and data visualization. We report P-values, which show that our technique predicts functional categories with greater statistical significance than a standard method, hierarchical agglomerative clustering (UPGMA). Also, we show that our algorithm, which

*To whom correspondence should be addressed.

computes probabilities rather than making hard decisions, obtains a higher data log-likelihood than the version of the algorithm that makes hard decisions.

2 METHODS FOR MATRIX FACTORIZATION

One approach to analyzing data vectors lying in a low-dimensional linear subspace is to stack them to form a data matrix, **X**, and then find a low-rank matrix factorization of the data matrix. Given $\mathbf{X} \in \mathcal{R}^{G\times T}$, matrix factorization techniques find a $\mathbf{Y} \in \mathcal{R}^{G\times C}$ and a $\mathbf{Z} \in \mathcal{R}^{C\times T}$ such that $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$.

A variety of techniques have been proposed for finding matrix factorizations, including non-probabilistic techniques such as principal components analysis, independent components analysis (Bell and Sejnowski, 1995) and network component analysis (Liao *et al*., 2003), and also probabilistic techniques that account for noise, such as factor analysis.

Interpreting the rows of **X** as input vectors $\{\mathbf{x}_g\}_{g=1}^{G}$, the rows of **Z** (i.e. $\{\mathbf{z}_c\}_{c=1}^{C}$) can be viewed as vectors that span the C-dimensional linear subspace, in which case the gth row of **Y** contains the coefficients $\{y_{gc}\}_{c=1}^{C}$ that combine these vectors to explain the gth row of **X**.

Gene expression is thought to be regulated by a small (compared to the total number of genes) set of factors which act in combination to maintain the steady-state abundance of specific mRNAs. Some of these factors could represent the binding of one (or more) transcription factors (TFs) to the promoter region(s) of the gene, other factors could include nonsense-mediated mRNA decay induced in varying degrees depending on the abundance of specific splicing factors that generate alternative splicing of the precursor mRNA. We assume that the expression of each gene is influenced by only a small subset of the possible factors and that these factors influence their targets to various degrees.

There is good evidence that the TFs, in particular, have varying effects on the expression of their targets. The TF binding sites for different genes can have different affinities; some genes have multiple binding sites whereas others have only one. It is also well known that factors act combinatorially. Inspired by this, we model the gene expression vector as a weighted combination of a small number of prototype vectors—each prototype representing the influence of a different biological or experimental factor (or factors).

This type of problem is called ‘sparse matrix factorization’ in (Srebro and Jaakkola, 2001), and is related to independent component analysis (ICA). In their model, Srebro and Jaakkola augment the $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$ matrix factorization setup with the sparseness structure constraint that each row of **Y** has at most N non-zero entries[1]. They then describe an iterative algorithm for finding a sparse matrix factorization that makes hard decisions at each step.

However, our method finds such a factorization while accounting for uncertainties due to (1) different levels of noise in each expression profile, (2) different levels of noise in the factors used to explain the data and (3) uncertainty about the hidden prototypes selected to explain each input vector.

2.1 Probabilistic sparse matrix factorization

Let **X** be the matrix of gene expression data such that rows correspond to each of G genes and columns to each of T tissues (i.e. entry x_{gt} represents the amount by which gene g is expressed in cells of tissue type t.) We denote the collection of unobserved factor profiles as a matrix, **Z**, with rows corresponding to each of C factors and T columns corresponding to tissues, as before.

We model each gene expression profile, $\mathbf{x}_g$, as a linear combination of a small number (r_g) of these factor profiles, $\mathbf{z}_c$, plus noise:

$$\mathbf{x}_g = \sum_{n=1}^{r_g} y_{gs_{gn}} \mathbf{z}_{s_{gn}} + \text{noise}. \tag{1}$$

The factor profiles contributing to the gth gene expression profile are indexed by $\{s_{g1}, s_{g2}, \ldots, s_{gr_g}\}$, with corresponding weights $\{y_{gs_{g1}}, y_{gs_{g2}}, \ldots, y_{gs_{gr_g}}\}$. This is identical to the $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$ matrix factorization with $\{\mathbf{S}, \mathbf{r}\}$ representing the sparseness structure constraint. We account for varying levels of noise in the observed data by assuming the presence of gene-specific isotropic Gaussian sensor noise with variance ψ_g^2 so the likelihood of $\mathbf{x}_g$ is as follows:

$$P(\mathbf{x}_g|\mathbf{y}_g, \mathbf{Z}, \mathbf{s}_g, r_g, \psi_g^2) = \mathcal{N}\left(\mathbf{x}_g; \sum_{n=1}^{r_g} y_{gs_{gn}} \mathbf{z}_{s_{gn}}, \psi_g^2 \mathbf{I}\right). \tag{2}$$

We complete the model with prior assumptions that the factor profiles ($\mathbf{z}_c$) are normally distributed and that the factor indices (s_{gn}) are uniformly distributed. The number of causes, r_g, contributing to each gene's profile is multinomially distributed such that $P(r_g = n) = \nu_n$, where ν is a user-specified N-vector. We make no assumptions about **Y** beyond the sparseness constraint, so $P(\mathbf{Y}) \propto 1$.

Multiplying these priors by Equation (2) forms the following joint distribution:

$$\begin{aligned} &P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}|\Psi) \\ &\quad = P(\mathbf{X}|\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}, \Psi) \cdot P(\mathbf{Y}) \cdot P(\mathbf{Z}) \cdot P(\mathbf{S}|\mathbf{r}) \cdot P(\mathbf{r}) \\ &\quad \propto \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{x}_g; \sum_{n=1}^{r_g} y_{gs_{gn}} \mathbf{z}_{s_{gn}}, \psi_g^2 \mathbf{I}\right) \cdot \prod_{c=1}^{C} \mathcal{N}(\mathbf{z}_c; \mathbf{0}, \mathbf{I}) \\ &\qquad \cdot \underbrace{\prod_{g=1}^{G}\prod_{c=1}^{C}\prod_{n=1}^{N} \left(\frac{1}{C}\right)^{[s_{gn}=c]}}_{\text{uniformly distributed}} \cdot \underbrace{\prod_{g=1}^{G}\prod_{n=1}^{N} (\nu_n)^{[r_g=n]}}_{\text{multinomially distributed}}. \end{aligned} \tag{3}$$

[1] When $N = 1$, this scheme degenerates to clustering with arbitrary data vector scaling; $N = C$ yields ordinary low-rank approximation.

Here, the Iverson notation is used where [True] = 1 and [False] = 0.

2.2 Factorized variational inference

Exact inference with Equation (3) is intractable so we utilize a factorized variational method (Jordan *et al.*, 1998, www.ai.mit.edu/research/abstracts/abstracts2001/genomics/Olsrebro.pdf) and approximate the posterior distribution with a mean-field decomposition:

$$P(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}|\mathbf{X},\Psi) \approx \prod_{g=1}^{G} Q(\mathbf{y}_g)\cdot\prod_{c=1}^{C} Q(\mathbf{z}_c) \cdot\prod_{g=1}^{G}\prod_{n=1}^{N} Q(s_{gn})\cdot\prod_{g=1}^{G} Q(r_g). \qquad (4)$$

We introduce variational parameters (λ, ζ, ϕ, σ, ρ) that parameterize the Q-distribution as follows:

$$Q(\mathbf{y}_g) = \overbrace{\prod_{n=1}^{r_g}\delta\left(y_{gs_{gn}} - \lambda_{gs_{gn}}\right)}^{\lambda_{gc}\text{ is a point estimate of }y_{gc}} \cdot \overbrace{\prod_{\substack{c=1\\ c\notin\{s_{g1},s_{g2},\ldots,s_{gr_g}\}}}^{C}\delta\left(y_{gc}\right)}^{\text{but for }\{\mathbf{s}_g,r_g\}}$$

$$Q(z_{ct}) = \mathcal{N}(z_{ct};\zeta_{ct},\ ,\phi_c^2)$$

$$Q(s_{gn}=c) = \sigma_{gnc}$$

$$Q(r_g=n) = \rho_{gn}.$$

Using this approach, inference corresponds to bringing the Q-distribution as close as possible to the P-distribution by setting the variational parameters to minimize the relative entropy, $\mathrm{D}(Q\|P)$:

$$\min_{\{\lambda,\zeta,\phi,\sigma,\rho\}}\int_{\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}} Q(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r})\cdot\log\frac{Q(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r})}{P(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}|\mathbf{X},\Psi)}. \qquad (5)$$

The constraints $\sum_{c=1}^{C}\sigma_{gnc}=1$, $\sum_{n=1}^{N}\rho_{gn}=1$ become Lagrange multipliers in this optimization problem.

There is no closed-form expression for the posterior [denominator in (5)], but we can subtract $\log P(\mathbf{X})$ inside the integral (It is independent of the variational parameters.) to form the readily-minimized free energy, $\mathcal{F}$:

$$\begin{aligned}\mathcal{F} &= \mathrm{D}(Q\|P) - \log P(\mathbf{X})\\ &= \int_{\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}} Q(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r})\cdot\log\frac{Q(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r})}{P(\mathbf{X},\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}|\Psi)}\\ &\ \ \vdots\\ &= \sum_{g=1}^{G}\sum_{n=1}^{N}\rho_{gn}\sum_{n'=1}^{n}\sum_{c=1}^{C}\left(\sigma_{gn'c}\cdot\log\frac{\sigma_{gn'c}}{1/C^n}\right)\\ &\quad+\sum_{g=1}^{G}\sum_{n=1}^{N}\left(\rho_{gn}\cdot\log\frac{\rho_{gn}}{\nu_n}\right)+\frac{T}{2}\sum_{g=1}^{G}\log 2\pi\psi_g^2\\ &\quad-\frac{T}{2}\sum_{c=1}^{C}\left(1+\log\phi_c^2\right)+\frac{1}{2}\sum_{t=1}^{T}\sum_{c=1}^{C}\left(\zeta_{ct}^2+\phi_c^2\right)\\ &\quad+\frac{1}{2}\sum_{g=1}^{G}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{\rho_{gn}}{\psi_g^2}\sum_{c_1=1}^{C}\sigma_{g1c_1}\sum_{c_2=1}^{C}\sigma_{g2c_2}\cdots\\ &\quad\sum_{c_n=1}^{C}\sigma_{gnc_n}\left[\left(x_{gt}-\sum_{n'=1}^{n}\lambda_{gc_{n'}}\zeta_{c_{n'}t}\right)^2+\sum_{n'=1}^{n}\lambda_{gc_{n'}}^2\phi_{c_{n'}}^2\right].\end{aligned}$$

The free energy can be minimized sequentially with respect to each variational parameter (λ, ζ, ϕ, σ, ρ) by analytically finding zeros of the partial derivatives with respect to them. This coordinate descent represents the E-step in variational EM (Jordan *et al.*, 1998) that alternates with a brief M-step, where the global sensor noise is fit by similarly solving $\partial\mathcal{F}/\partial\Psi = 0$. For a more detailed treatment of the algorithm and parameter update equations (Dueck and Frey, 2004, www.psi.toronto.edu).

3 EXPERIMENTAL RESULTS

To experimentally analyze the performance of our algorithm, we use the recently-published expression dataset from Zhang *et al.* (2004, http://hugheslab.med.utoronto.ca/Zhang). This dataset is one of the most comprehensive mammalian gene expression datasets now available, containing profiles of over 40 000 known and predicted genes (of which they determined that 22 709 contained a clear expression) across a set of 55 mouse tissues, organs and cell types. In addition to measuring the profiles, Zhang *et al*. associated Gene Ontology Biological Process (GO-BP) annotations with each gene and showed that a supervised learning algorithm can be trained to predict these annotations with high precision. These results were obtained using profiles in which the expression measurement of a gene in each tissue is represented by the approximate log ratio of the normalized intensity of the gene in the given tissue to the gene's median normalized intensity across the 55 tissues. Observing that the majority of genes expressed in any tissue were expressed in less than half of the tissues, Zhang *et al.* set ratios less than one equal to one, reasoning that these ratios represent noise rather than downregulation. We use the same representation of the data as Zhang *et al*., so the data is entirely non-negative. The factor loading and factor profile matrices (**Y** and **Z**), however, are free to contain positive or negative entries, consistent with the notion that factors can act to increase or reduce expression levels.

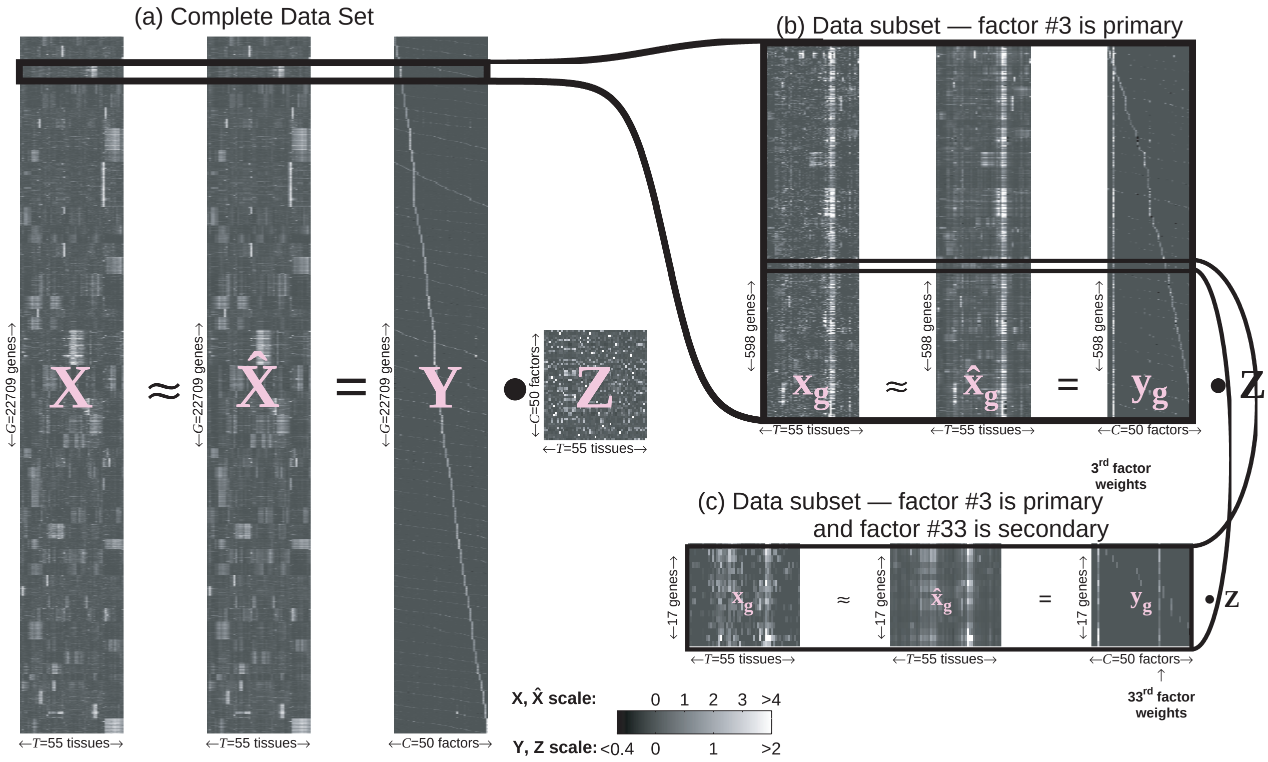

Fig. 1. Data matrix **X** approximated by $\hat{\mathbf{X}}$, the product of sparse **Y** and low-rank **Z**. Gene expression profiles appear as row vectors in **X** and $\hat{\mathbf{X}}$ sorted by primary class (s_{g1}), secondary class (s_{g2}), etc. (**a**) shows the complete dataset. (**b**) shows only those genes where factor 3 is primary (i.e. $\{\forall g \mid s_{g1} = 3\}$—note the vertical line in $\mathbf{y_g}$). (**c**) shows those genes where factor 3 is primary and factor 33 is secondary (i.e. $\{\forall g \mid s_{g1} = 3 \cap s_{g2} = 33\}$—note the vertical lines in $\mathbf{y_g}$).

The functional category labels for the genes with known biological function were taken from (Zhang *et al*., 2004). An example of a category label is ‘visual perception’, which indicates the gene expresses a protein that is involved in ‘the series of events required for an organism to receive a visual stimulus, convert it to a molecular signal, and recognize and characterize the signal’. These labels are derived from GO-BP category labels (Ashburner *et al*., 2000) assigned to genes by the European Bioinformatics Institute and Mouse Genome Informatics.

Among the 22 709 clearly expressed genes in the Zhang *et al*. database, 9499 have annotations in one or more of 992 different GO categories. The category sizes range from 3 to 456 genes, with more than half the categories having <20 genes.

We present results for the 22 709 genes × 55 tissues datasets shown in Figure 1. The data matrix, **X**, is shown alongside the model’s approximation, $\hat{\mathbf{X}}$, also expressed as $\hat{\mathbf{X}} = \mathbf{Y} \cdot \mathbf{Z}$. A total of $C = 50$ factors were used, and the user-specified prior on the number of factors (r_g) explaining each expression profile was set to $\nu = \begin{bmatrix} .55 & .27 & .18 \end{bmatrix}$, making $N = 3$. A uniform prior ν (reflecting no knowledge about the distribution of **r**) would give equal preference to all values of a particular r_g. For any given $r_g < N$, a factor can almost always be found that, if present with infinitesimal weight (y_{gc}), will imperceptibly improve the cost function ($\mathcal{F}$), with the end result that almost all r_g would then be maximized (equal to N). Weighting the prior towards lower values ensures that factors will only be included if they make a non-negligible difference (we only choose $\nu_n \propto 1/n$, $\forall n \leq N$ for simplicity).

Gene expression profiles (row vectors) in Figure 1 are first sorted by ‘primary’ factor (s_{g1}); next, within each s_{g1} grouping, genes are sorted by ‘secondary’ factor (s_{g2}), and so on. This organization is easily visualized in the hierarchical diagonal structure of **Y**.

Figure 1b zooms in on the 598 genes whose primary factor is 3 (i.e. $s_{g1} = 3$). Genes with this primary factor show high expression in colon, small intestine and large intestine (the prominent vertical expression band in the $\mathbf{x_g}$ panel in Figure 1b). The most significantly enriched GO-BP category in this category are lipid metabolism [GO:0006629] (P-value $<10^{-10}$). Figure 1c zooms in further on to a subset of 17 genes whose secondary factor is 33 (i.e. $s_{g2} = 33$). Genes associated with factor 33 show high expression in tissues associated with the immune system (like thymus and bone marrow) and are most significantly enriched for the GO-BP category response to pest/pathogen/parasite [GO:0009613] (P-value $< 10^{-13}$). Indeed, the set of genes shown in Figure 1c contains six

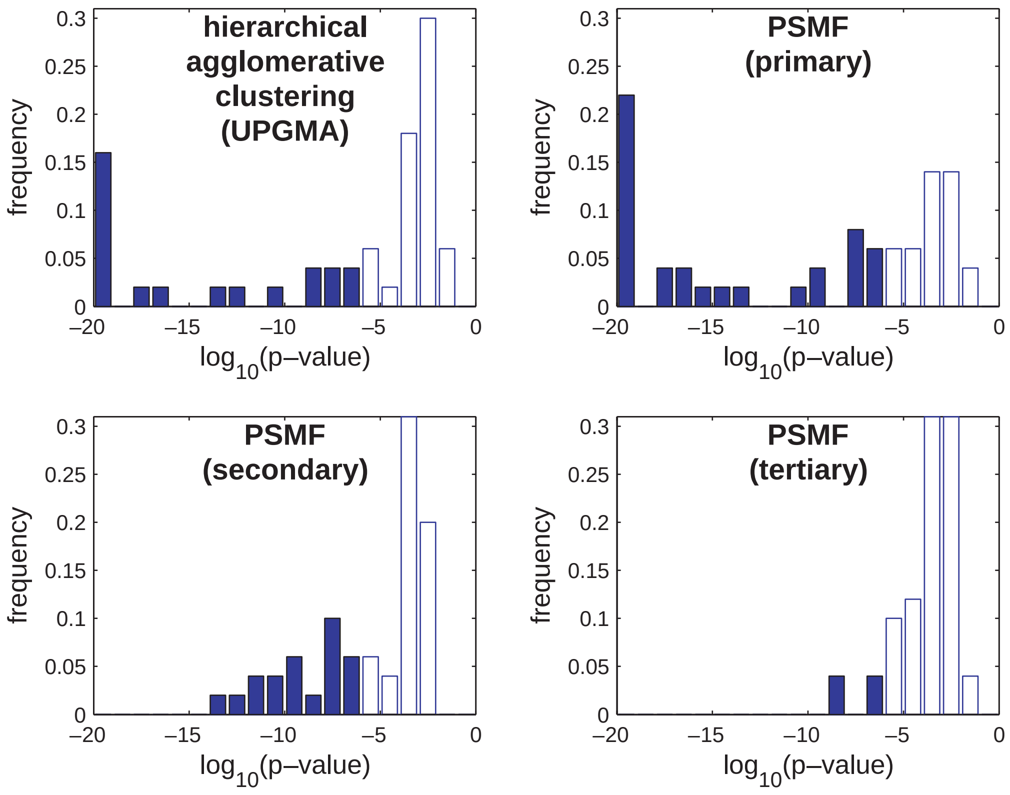

Fig. 2. P-values for UPGMA and probabilistic sparse matrix factorization ($C=50$, $N=3$). Significantly enriched factors/clusters (at $\alpha = 0.05$ after Bonferroni correction) are shown as solid bars.

(of twelve annotated) genes involved in the immune response including two genes involved in the major histocompatibility complex (H2-Q1, H2-T23), a portion of a T-cell receptor, and a subunit of the proteasome (Psme1). Here, PSMF appears to have identified a subset of the genes expressed in the lower digestive system that are involved in immune system activity occurring there.

3.1 Unsupervised characterization of mRNA data

The overall objective for this research is to develop a model that captures the functionally relevant hidden factors that explain gene expression data. As such we can gauge success by measuring the similarity of the hidden factor assignments between genes with similar functions.

We cluster genes on the basis of their factor assignments and use P-values calculated from the hypergeometric distribution to measure the enrichment of these clusters for genes with similar functions (Tavazoie *et al*., 1999). For a given gene cluster of size M of which K elements had a given functional annotation, the hypergeometric P-value is the probability that a random set of genes of size M (selected without replacement from all genes in the dataset) would have K or more elements with given functional annotation. For $N=3$, we generate three different clustering of the genes: by 'primary' factor, 'secondary' factor (for those genes where $r_g \geq 2$), and 'tertiary' factor assignments (where $r_g \geq 3$). So, for example, genes g and g' are in the same 'primary' factor cluster if $s_{g1} = s_{g'1}$. We label each cluster with the GO-BP category having the lowest (i.e. best) hypergeometric P-value for the genes in that cluster, making this the P-value associated with the entire cluster. Histograms of these P-values, as well as those for hierarchical agglomerative clustering, are shown in Figure 2.

For reference, we also compute the functional enrichment of gene clusters generated randomly and those generated using UPGMA, sparse matrix factorization (SMF), PCA and ICA. To generate the UPGMA clusters, we used average linkage with Pearson's correlation coefficient as the distance measure on the expression profiles. We generated SMF clusters in the same way as we generated the PSMF clusters. Note that unlike UPGMA, PSMF and SMF allow us to generate 'primary', 'secondary' and 'tertiary' clusters.

For PCA, we cluster genes by their dominant principal component measured by the absolute value. To generate the ICA clusters, we used the FastICA package (Hyvärinen, 1999, www.cis.hut.fi/projects/ica/fastica) to recover 55 source vectors (i.e. factors) for X, each of length 22 709. We ignore the mixing matrix recovered by FastICA, and cluster the genes by the index of the largest absolute value among the corresponding 55-dimensional slice in the matrix of source vectors.

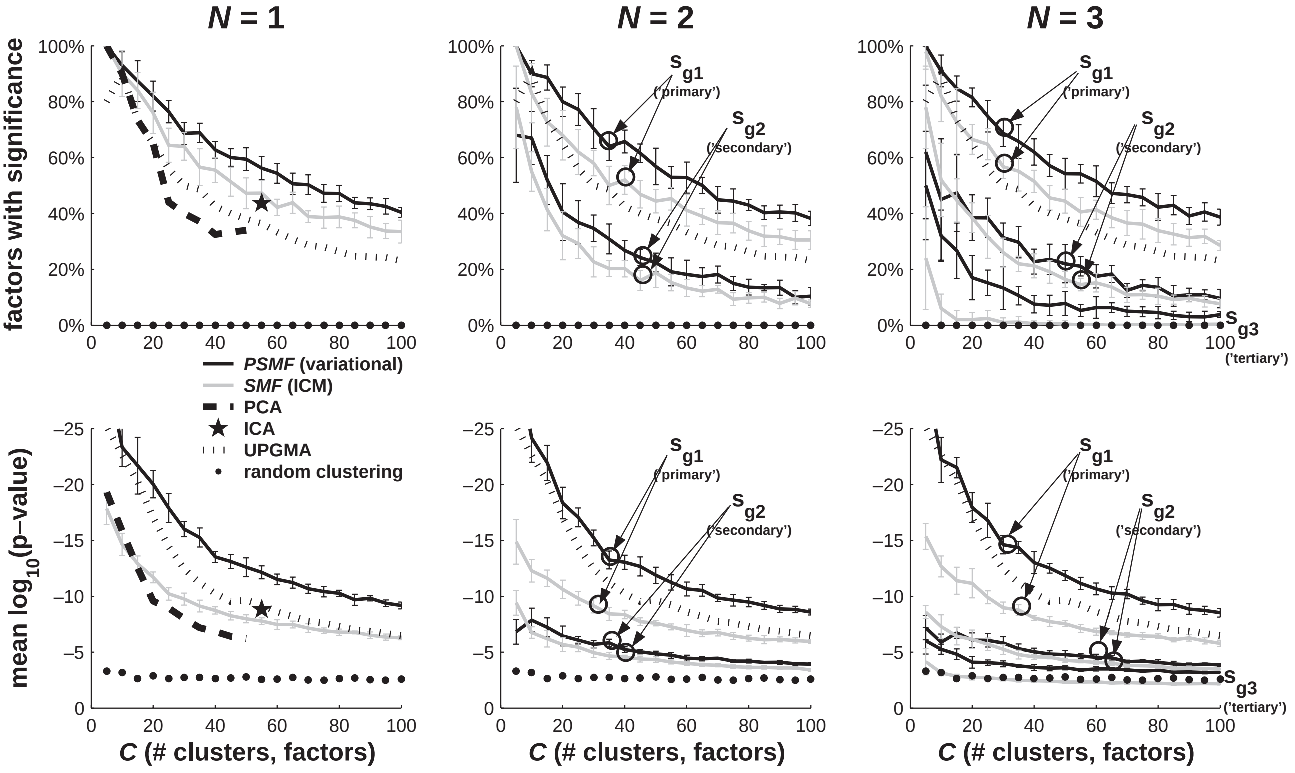

Fig. 3. Fraction of clusters/factors with significance (top row) and mean $\log_{10}$ P-value (bottom row). PSMF (Dueck and Frey, 2004) and SMF (Srebro and Jaakkola, 2001) are shown for $N = 1$ (left), $N = 2$ (center) and $N = 3$ (right) along with UPGMA and random clustering over a range of C-values. PCA is shown only for $C < T = 55$ after which it would overfit the data; ICA is shown for $C = 55$ only. Error bars represent the sample standard deviation over 10 trials. Results for UPGMA and random clustering are shown on all plots for comparison.

Table 1. Fraction of clusters that have significant functional enrichment (for $C = 50$ clusters or factors)

	Primary	Secondary	Tertiary
PSMF (N = 1)	0.59		
PSMF (N = 2)	0.57	0.22	
PSMF (N = 3)	0.54	0.22	0.078
SMF (N = 1)	0.47		
SMF (N = 2)	0.44	0.19	
SMF (N = 3)	0.44	0.16	0.004
PCA	0.34		
ICA ($C = 55$)	0.44		
UPGMA	0.38		
Random	0		

This method (as opposed to transposing matrices or not taking absolute values) was selected because it yielded results with the greatest significance.

We summarize these histograms in two ways: the proportion of clusters that are significantly enriched for at least one functional category at $\alpha = 0.05$ (after a Bonferroni correction) and the mean $\log_{10}$ (P-value). Table 1 shows the proportion of clusters with significance for $C = 50$ factors and Figure 3 shows both these quantities for $N = \{1, 2, 3\}$ over $0 < C \leq 80$.

The plot shows that PSMF with $N = 1$ outperforms UPGMA and clustering by dominant sources with ICA. For $N > 1$, the functional enrichment of 'primary' clusters for PSMF and SMF remains constant but the 'secondary' and 'tertiary' clusters are also more functionally enriched than random. Note that for all values of N, the PSMF clusters are more enriched than the corresponding SMF clusters.

3.2 Avoiding local minima by accounting for uncertainty

The soft-decision factorized variational method (PSMF) finds better hidden factors than the hard-decision iterated conditional modes (SMF) (Srebro and Jaakkola, 2001; Besag, 1986). This may be the case because SMF is too inflexible to properly allow factor groupings to evolve, as shown in the likelihood plots of Figure 4. Note that these plots compare complete log-likelihoods [i.e. log of Equation (3)] for both PSMF and SMF—not the more favorable free energy for PSMF—so the number of parameters is the same.

As an example, consider the case where there is uncertainty about gene g having as its primary factor (s_{g1}) profile c_1 or

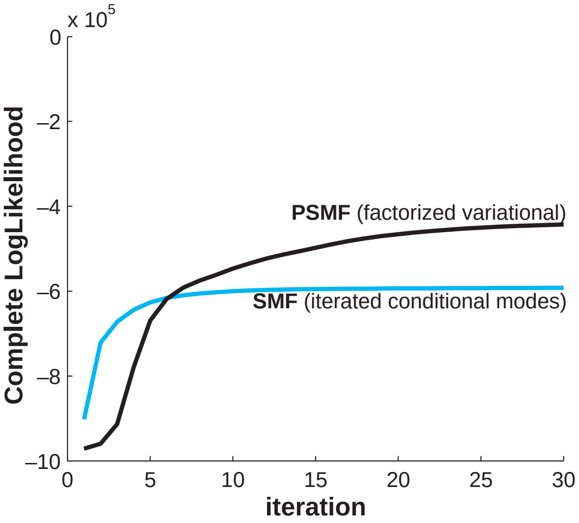

Fig. 4. Direct complete log-likelihood maximization of SMF and indirect complete log-likelihood maximization (via free energy minimization—see Section 2.2) of PSMF. The complete log-likelihood is calculated after each iteration of both ICM and variational EM using maximum-likelihood latent variable estimates.

c_2 (with probabilities 0.49 and 0.51, respectively). When the factor profiles ($\mathbf{Z}$) are updated to reflect their membership, SMF makes the maximum-likelihood hard decision that $s_{g1} = c_2$ and gene g updates factor profile $\mathbf{z}_{c_2}$ only. On the other hand, soft-decision PSMF accounts for the uncertainty by updating profiles $\mathbf{z}_{c_1}$ and $\mathbf{z}_{c_2}$ in proportion to their probabilities (roughly equally, in this case).

In our simulations, we observe that SMF appears to get trapped in a local log-likelihood maximum immediately after the first several iterations—a consequence of making hard decisions early on in the optimization. An additional advantage of using probabilistic inference is that the free energy provides a tighter bound on the marginal log-likelihood of the data (not shown), which is greater than the complete log-likelihood.

4 SUMMARY

Many kinds of data vectors can most naturally be explained as a linear combination of a selection of prototype vectors, which can be viewed as a computational problem of finding a sparse matrix factorization. While most work on biological gene expression arrays has focused on clustering techniques and methods for dimensionality reduction, there is recent interest in performing these tasks jointly, which corresponds to sparse matrix factorization. Our algorithm computes a sparse matrix factorization, but instead of making point estimates (hard decisions) for factor selections (Srebro and Jaakkola, 2001), our algorithm computes probability distributions. We find that this enables the algorithm to avoid local minima found by iterated conditional modes.

There are clearly >50 hidden factors that contribute to the gene expression in mouse. Indeed, a recent study identified 779 likely TFs alone (Zhang *et al*., 2004). We have, however, limited ability to identify factors by the available data. We cannot, for example, distinguish factors that do not have an observably different effect on mRNA levels in the profiled tissues. Also, the data we modeled were drawn primarily from adult tissue whereas many TFs are most active during embryogenesis. With additional data, we would be able to model more factors. However, the fact that, with just 50 hidden factors, we were able to extract functionally-relevant representations is a strong validation of our approach.

Compared to a standard technique used in the bioinformatics community for clustering gene expression profiles, UPGMA, our technique finds clusters that have higher statistical significance (lower P-values) in terms of enrichment of gene function categories. An additional advantage of our method over standard clustering techniques is that the secondary and higher-order labels found for each expression vector can be used for more refined visualization and functional prediction. Currently we are using this method to analyze a new genome-wide, exon resolution, mouse genome dataset containing over 12 million measurements. We are currently using this method to analyze a new genome-wide exon-tiling microarray dataset, in conjunction with our GenRate algorithm (Frey *et al*., 2005).

ACKNOWLEDGEMENTS

The authors thank Tim Hughes from the Banting and Best Department of Medical Research at the University of Toronto for making the data available, and Sam Roweis, Tommi Jaakkola and Nati Srebro for helpful discussions. We also gratefully acknowledge the financial support of Natural Sciences and Engineering Research Council of Canada and a cash gift from Microsoft Corporation. Q.D.M. was supported by a Canadian Institute for Health Research grant. B.J.F. was supported by a Premier's Research Excellence award.

REFERENCES

Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bell,A.J. and Sejnowski,T.J. (1995) An information maximization approach to blind separation and blind deconvolution. *Neural Comput.*, **7**, 1129–1159.

Besag,J. (1986) On the statistical analysis of dirty pictures. *J. R. Stat. Soc. B*, **48**, 259–302.

Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.,Jr. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Nat. Acad. Sci. USA*, **97**, 262–267.

Dueck,D. and Frey,B. (2004) Probabilistic sparse matrix factorization. Technical report PSI-2004-23, University of Toronto, available at www.psi.toronto.edu

Frey,B.J., Morris,Q.D., Mohammad,N., Robinson,M., Zhang,W., and Hughes,T.R. (2005) Finding novel transcripts in high-resolution genome-wide microarray data using the GenRate model. To appear in *Proc. Ninth International Conference on Research in Computational Molecular Biology (RECOMB)*, Cambridge, MA.

Hyvärinen,A. (1999) Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Netw.*, **10**, 626–634.

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.

Jordan,M.I., Ghahramani,Z., Jaakkola,T.S. and Saul,L.K. (1998) An introduction to variational methods for graphical models. In Jordan,M.I. (ed.), *Learning in Graphical Models*. Kluwer Academic Publishers, Norwell, MA.

Liao,J.C., Boscolo,R., Yang,Y.L., Tran,L.M., Sabatti,C. and Roychowdhury,V.P. (2003) Network component analysis: reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.

Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.

Srebro,N. and Jaakkola,T. (2001) Sparse Matrix Factorization of Gene Expression Data. Unpublished note, MIT Artificial Intelligence Laboratory.

Tavazoie,S., Hughes,J.D., Campbell,M.J., Cho,R.J., and Church,G.M. (1999) Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, 213–215.

Zhang,W., Morris,Q.D., Chang,R., Shai,O., Bakowski,M.A., Mitsakakis,N., Mohammad,N., Robinson,M.D., Zirngibl,R., Somogyi,E. *et al*. (2004) The functional landscape of mouse gene expression. *J. Biol.*, **3**, 21.