http://www.psi.toronto.edu

Probabilistic Sparse Matrix Factorization

Delbert Dueck, Brendan J. Frey

September 28, 2004 **PSI TR 2004–023**

Abstract

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. Physically, these prototype vectors may correspond to different hidden variables that play a 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 hard constraints reflecting structural knowledge of the problem and probabilistic relationships between variables that are induced by known uncertainties in the problem. We present soft-decision probabilistic sparse matrix factorization (PSMF) to better account for uncertainties due to varying levels of noise in the data, varying levels of noise in the prototypes used to explain the data, and uncertainty as to which hidden prototypes are selected to explain each expression vector.

Probabilistic Sparse Matrix Factorization

Delbert Dueck, Brendan J. Frey
University of Toronto

1 Introduction

Many information processing problems can be formulated as finding a factorization of a matrix of data, $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$, while taking into account hard constraints reflecting structural knowledge of the problem and probabilistic relationships between variables that are induced by known uncertainties in the problem (e.g., noise in the data). 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.

One approach to analyzing data vectors lying in a low-dimensional linear subspace is to stack them to form a data matrix, $\mathbf{X}$, 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 $\mathbf{X} \approx \mathbf{Y} \cdot \mathbf{Z}$.

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

This type of problem was called "sparse matrix factorization" in [1], and is related to independent component analysis [2]. 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 $\mathbf{Y}$ 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.

On the other hand, our method finds such a factorization while accounting for uncertainties due to (1) different levels of noise in the data, (2) different levels of noise in the factors used to explain the data, and (3) uncertainty as to which hidden prototypes are selected to explain each input vector.

2 Probabilistic sparse matrix factorization (PSMF)

Let $\mathbf{X}$ be the matrix of data such that rows correspond to each of G data points and columns to each of T data dimensions. We denote the collection of unobserved factor profiles as a matrix, $\mathbf{Z}$, with rows corresponding to each of C factors and T columns, as before. Each data point, $\mathbf{x}_g$, can be approximated by a linear combination of a small number (r_g) of these transcription factor profiles, $\mathbf{z}_c$:

$$\mathbf{x}_g \approx \sum_{n=1}^{r_g} y_{g s_{gn}} \mathbf{z}_{s_{gn}} \tag{1}$$

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

The factor profiles contributing to the g^{th} data point 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 isotropic Gaussian sensor noise (with variance ψ_g^2) for each data point, so the likelihood of $\mathbf{x}_g$ is as follows:

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

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 data point is multinomially distributed such that $P(r_g = n) = \nu_n$, where ν is a user-specified N-vector. We make no assumptions about $\mathbf{Y}$ beyond the sparseness constraint, so $P(\mathbf{Y}) \propto 1$.

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

$$\begin{aligned} P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}|\Psi) &= P(\mathbf{X}|\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}, \Psi) \cdot P(\mathbf{Y}) \cdot P(\mathbf{Z}) \cdot P(\mathbf{S}|\mathbf{r}) \cdot P(\mathbf{r}) \\ &\propto \prod_{g=1}^{G} \mathcal{N}\left(\mathbf{x}_g;\ \sum\nolimits_{n=1}^{r_g} y_{gs_{gn}} \mathbf{z}_{s_{gn}},\ \psi_g^2 \mathbf{I}\right) \cdot \prod_{c=1}^{C} \mathcal{N}\left(\mathbf{z}_c;\ \mathbf{0},\ \mathbf{I}\right) \\ &\quad \cdot \prod_{g=1}^{G} \prod_{c=1}^{C} \prod_{n=1}^{N} \left(\tfrac{1}{C}\right)^{\delta(s_{gn}-c)} \cdot \prod_{g=1}^{G} \prod_{n=1}^{N} (\nu_n)^{\delta(r_g - n)} \end{aligned} \tag{3}$$

It is often easier to work with the complete log likelihood, $\ell_{\mathcal{C}}$:

$$\begin{aligned} \ell_{\mathcal{C}} &= \log P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}|\Psi) \\ &= -\frac{1}{2} \sum_{g=1}^{G} \sum_{t=1}^{T} \left[\log 2\pi\psi_g^2 + \left(x_{gt} - \sum\nolimits_{n=1}^{r_g} y_{gs_{gn}} z_{s_{gn}t} \right)^2 / \psi_g^2 \right] \\ &\quad -\frac{1}{2} \sum_{c=1}^{C} \sum_{t=1}^{T} \left(\log 2\pi + z_{ct}^2\right) - \sum_{g=1}^{G} r_g \cdot \log C + \sum_{g=1}^{G} \log \nu_{r_g} \end{aligned} \tag{4}$$

3 Iterated Conditional Modes (ICM)

Learning globally-optimal settings for variables $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{S}$, $\mathbf{r}$ and the Ψ noise parameter in (4) is intractable, so we resort to approximate techniques. One possibility is to set each variable to its maximum empha posteriori (MAP) value and iterate. For instance, this procedure would modify the $\mathbf{r}$-variable as follows:

$$\begin{aligned} \mathbf{r} &\leftarrow \operatorname*{argmax}_{\mathbf{r}} P(\mathbf{r}|\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \Psi) \\ &= \operatorname*{argmax}_{\mathbf{r}} \overbrace{\{\log P(\mathbf{r}|\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \Psi) + \log P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}|\Psi)\}}^{\ell_{\mathcal{C}}} \end{aligned} \tag{5}$$

The scheme in (5) is known as *iterated conditional modes* [3], and can be implemented by directly maximizing the complete log likelihood, $\ell_{\mathcal{C}}$. This section concludes by outlining the

steps within a single ICM iteration (6)–(10).

$$\forall g \in \{1,\ldots,G\}, \forall n \in \{1,\ldots,r_g\}:$$
$$s_{gn} \leftarrow \underset{s_{gn}\in\{1,\ldots,C\}}{\operatorname{argmin}} \left\{ \sum_{t=1}^{T} \left(x_{gt} - \sum\nolimits_{n=1}^{r_g} y_{gs_{gn}} z_{s_{gn}t} \right)^2 \right\} \tag{6}$$

$$\forall g \in \{1,\ldots,G\}:$$
$$r_g \leftarrow \underset{r_g\in\{1,\ldots,N\}}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{t=1}^{T} \left[\left(x_{gt} - \sum\nolimits_{n=1}^{r_g} y_{gs_{gn}} z_{s_{gn}t} \right)^2 / \psi_g^2 \right] + \log \frac{C^{r_g}}{\nu_{r_g}} \right\} \tag{7}$$

Elements of the $\mathbf{r}$-vector and $\mathbf{S}$-matrix are independent of each other when conditioned on the other variables, so they can be updated element-wise, as in (6)–(7). This is not the case for $\mathbf{Y}$ and $\mathbf{Z}$, whose MAP values can be determined by solving $\partial \ell_{\mathcal{C}} / \partial \mathbf{Y} = \mathbf{0}$ and $\partial \ell_{\mathcal{C}} / \partial \mathbf{Z} = \mathbf{0}$. Rows of $\mathbf{Y}$ are independent of one another (given $\{\mathbf{Z}, \mathbf{S}, \mathbf{r}\}$), as are columns of $\mathbf{Z}$ (given $\{\mathbf{Y}, \mathbf{S}, \mathbf{r}\}$), so the updated values for the g^{th} row of $\mathbf{Y}$ (8) and the t^{th} column of $\mathbf{Z}$ (9) are solutions to the following linear systems:

$$\forall c \in \{s_{g1}, s_{g2}, \ldots, s_{gr_g}\}:$$
$$\sum_{n=1}^{r_g} y_{gs_{gn}} \left\{ \sum_{t=1}^{T} z_{ct} z_{s_{gn}t} \right\} = \sum_{t=1}^{T} x_{gt} z_{ct} \tag{8}$$

$$\forall c \in \{1,\ldots,C\}:$$
$$z_{ct} + \sum_{c'=1}^{C} z_{c't} \left\{ \sum_{g=1}^{G} \frac{y_{gc} y_{gc'}}{\psi_g^2} \right\} = \sum_{g=1}^{G} \frac{x_{gt} y_{gc}}{\psi_g^2} \tag{9}$$

Finally, the parameter Ψ is learned by updating it with its maximum likelihood estimate, obtained by similarly solving $\partial \ell_{\mathcal{C}} / \partial \Psi = \mathbf{0}$.

$$\forall g \in \{1,\ldots,G\}: \quad \psi_g^2 \leftarrow \frac{1}{T} \sum_{t=1}^{T} \left(x_{gt} - \sum\nolimits_{n=1}^{r_g} y_{gs_{gn}} z_{s_{gn}t} \right)^2 \tag{10}$$

4 Factorized Variational Inference

Iterated conditional modes is a simple method of learning latent variables and parameters by directly maximizing the log likelihood, but it is susceptible to getting trapped in local maxima. This occurs because ICM fails to account for uncertainty as it makes hard decisions. Another solution is to utilize a factorized variational method [4] and approximate the posterior distribution (3) with a mean-field decomposition:

$$P(\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r} | \mathbf{X}, \Psi) \approx \prod_{g=1}^{G} Q(\mathbf{y}_g) \cdot \prod_{c=1}^{C} Q(\mathbf{z}_c) \cdot \prod_{g=1}^{G} \prod_{n=1}^{N} Q(s_{gn}) \cdot \prod_{g=1}^{G} Q(r_g) \tag{11}$$

We parameterize the Q-distribution as follows:

$$Q(\mathbf{y}_g) = \overbrace{\prod_{n=1}^{r_g} \delta\left(y_{gs_{gn}} - \lambda_{gs_{gn}}\right)}^{\lambda_{gc} \text{ is a point estimate of } y_{gc}, \ldots} \cdot \overbrace{\prod_{\substack{c=1 \\ c \notin \{s_{g1}, s_{g2}, \ldots, s_{gr_g}\}}}^{C} \delta\left(y_{gc}\right)}^{\ldots \text{ unless } \{\mathbf{s}_g, r_g\} \text{ force } y_{gc} \text{ to zero.}} \tag{12}$$

$$Q(z_{ct}) = \mathcal{N}(z_{ct};\ \zeta_{ct},\ \phi_c^2); \quad Q(s_{gn} = c) = \sigma_{gnc}; \quad Q(r_g = n) = \rho_{gn} \tag{13}$$

Using this approach, factorization corresponds to bringing the Q-distribution as close as possible to the P-distribution by varying the Q-distribution parameters. For this reason, the parameters in the Q-distribution are called variational parameters [4]. To ensure normalization, the variational parameters are constrained to satisfy $\sum_{c=1}^{C} \sigma_{gnc} = 1$ and $\sum_{n=1}^{N} \rho_{gn} = 1$. Note that ICM can be considered a special case where the Q-distribution consists entirely of point estimates of P-distribution latent variables.

In order to closely approximate the P-distribution, we seek to minimize the relative entropy, $\mathrm{D}(Q\|P)$, between it and the Q-distribution:

$$\{\lambda, \zeta, \phi, \sigma, \rho\} \leftarrow \underset{\{\lambda,\zeta,\phi,\sigma,\rho\}}{\operatorname{argmin}} \int_{\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}} Q(\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}) \cdot \log \frac{Q(\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r})}{P(\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}|\mathbf{X}, \Psi)} \tag{14}$$

There is no closed-form expression for the posterior (denominator in (14)), 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, $\mathcal{F}$:

$$\begin{aligned}
\mathcal{F} = \mathrm{D}(Q\|P) - \log P(\mathbf{X}) &= \int_{\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}} Q(\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}) \cdot \log \frac{Q(\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r})}{P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}|\Psi)} \\
&\vdots \\
&= \sum_{g=1}^{G}\sum_{n=1}^{N} \rho_{gn} \sum_{n'=1}^{n}\sum_{c=1}^{C} \left(\sigma_{gn'c} \cdot \log \frac{\sigma_{gn'c}}{1/C}\right) + \sum_{g=1}^{G}\sum_{n=1}^{N} \left(\rho_{gn} \cdot \log \frac{\rho_{gn}}{\nu_n}\right) \\
&\quad -\frac{T}{2}\sum_{c=1}^{C}\left(1 + \log \phi_c^2\right) + \frac{1}{2}\sum_{t=1}^{T}\sum_{c=1}^{C}\left(\zeta_{ct}^2 + \phi_c^2\right) + \frac{T}{2}\sum_{g=1}^{G}\log 2\pi\psi_g^2 \\
&\quad +\frac{1}{2}\sum_{g=1}^{G}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{\rho_{gn}}{\psi_g^2}\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\cdots\sum_{c_n=1}^{C}\prod_{n'=1}^{n}\sigma_{gn'c_{n'}}\left[\left(x_{gt} - \sum_{n'=1}^{n}\lambda_{gc_{n'}}\zeta_{c_{n'}t}\right)^2 + \sum_{n'=1}^{n}\lambda_{gc_{n'}}^2\phi_{c_{n'}}^2\right]
\end{aligned} \tag{15}$$

Computation of the final term in (15), which sums over all possible configurations of the sparseness enforcement variational parameters ($\{\mathbf{S}, \mathbf{r}\}$), is $\mathcal{O}(GN^2C^NT)$. Considerable computational savings can be realized by taking advantage of complete factorization of $\mathbf{S}$; the $\prod_{n'=1}^{n}\sigma_{gn'c_{n'}}$ probabilities and the $\sum_{c_1=1}^{C}\sum_{c_2=1}^{C}\cdots\sum_{c_n=1}^{C}$ summations can be rearranged to make things $\mathcal{O}(GN^3C^2T)$ as shown in (16).

$$\begin{aligned}
\mathcal{F} = &\sum_{g=1}^{G}\sum_{c=1}^{C}\sum_{n=1}^{N}\left(\sum_{n'=n}^{N}\rho_{gn'}\right)\left(\sigma_{gcn} \cdot \log \frac{\sigma_{gcn}}{1/C}\right) + \sum_{g=1}^{G}\sum_{n=1}^{N}\left(\rho_{gn} \cdot \log \frac{\rho_{gn}}{\nu_n}\right) \\
&-\frac{T}{2}\sum_{c=1}^{C}\left(1 + \log \phi_c^2\right) + \frac{1}{2}\sum_{t=1}^{T}\sum_{c=1}^{C}\left(\zeta_{ct}^2 + \phi_c^2\right) - \frac{1}{2}\sum_{g=1}^{G}\frac{\sum_{n=2}^{N}(n-1)\,\rho_{gn}}{\psi_g^2}\sum_{t=1}^{T}x_{tg}^2 \\
&+\frac{T}{2}\sum_{g=1}^{G}\log 2\pi\psi_g^2 + \frac{1}{2}\sum_{g=1}^{G}\sum_{n=1}^{N}\frac{\sum_{n'=n}^{N}\rho_{gn'}}{\psi_g^2}\sum_{c=1}^{C}\sigma_{gcn}\sum_{t=1}^{T}\left[\left(x_{gt} - \lambda_{gc}\zeta_{ct}\right)^2 + \lambda_{gc}^2\phi_c^2\right] \\
&+\sum_{g=1}^{G}\sum_{n=1}^{N}\sum_{n'=n+1}^{N}\frac{\sum_{n''=n'}^{N}\rho_{gn''}}{\psi_g^2}\sum_{c=1}^{C}\sum_{c'=1}^{C}\sigma_{gcn}\sigma_{gc'n'}\lambda_{gc}\lambda_{gc'}\sum_{t=1}^{T}\zeta_{ct}\zeta_{c't}
\end{aligned} \tag{16}$$

The variational parameters λ, ζ, ϕ, σ, and ρ can be minimized sequentially by analytically finding zeros of partial derivatives similar to the ICM case (8)-(10). First, the update for ρ_{gn} can be obtained by finding the zero of $\partial\left[\mathcal{F}+\mathcal{L}_g\left(1-\sum_{n'=1}^{N}\rho_{gn'}\right)\right]/\partial\rho_{gn}$:

$$\begin{aligned}
&\forall g\in\{1,\ldots,G\},\forall n\in\{1,\ldots,N\}:\\
&\quad\rho_{gn}\propto\nu_n\cdot e^{-\sum_{c=1}^{C}\sum_{n'=1}^{n}\left(\sigma_{gcn'}\cdot\log\frac{\sigma_{gcn'}}{1/C}\right)}\cdot e^{-\frac{1}{2\psi_g^2}\sum_{c=1}^{C}\left(\sum_{n'=1}^{n}\sigma_{gcn'}\right)\sum_{t=1}^{T}\left[-2x_{gt}\lambda_{gc}\zeta_{ct}+\lambda_{gc}^2\left(\zeta_{ct}^2+\phi_c^2\right)\right]}\\
&\qquad\cdot e^{-\frac{1}{\psi_g^2}\sum_{c'=1}^{C}\sum_{c''=1}^{C}\left(\sum_{n'=1}^{n}\sum_{n''=n'+1}^{n}\sigma_{gc'n'}\sigma_{gc''n''}\right)\lambda_{gc'}\lambda_{gc''}\sum_{t=1}^{T}\zeta_{c't}\zeta_{c''t}}
\end{aligned}\tag{17}$$

Part of introducing the $\mathcal{L}_g$ Lagrange multiplier also involves normalizing ρ_{gn} by $\sum_{n'=1}^{N}\rho_{gn'}$ to make it a valid probability distribution (summing to one).

Likewise, finding the zero of $\partial\left[\mathcal{F}+\mathcal{L}_{gn}\left(1-\sum_{c'=1}^{C}\sigma_{gc'n}\right)\right]/\partial\sigma_{gcn}$ yields an update equation for σ_{gcn}:

$$\begin{aligned}
&\forall g\in\{1,\ldots,G\},\forall c\in\{1,\ldots,C\},\forall n\in\{1,\ldots,N\}:\\
&\quad\sigma_{gcn}\propto e^{-\frac{1}{2\psi_g^2}\sum_{t=1}^{T}\left[(x_{gt}-\lambda_{gc}\zeta_{ct})^2+\lambda_{gc}^2\phi_c^2\right]}\cdot e^{-\frac{1}{\psi_g^2}\sum_{c'=1}^{C}\sum_{\substack{n'=1\\n'\neq n}}^{N}\frac{\sum_{n''=\max(n,n')}^{N}\rho_{gn''}}{\sum_{n''=n}^{N}\rho_{gn''}}\sigma_{gc'n'}\lambda_{gc}\lambda_{gc'}\sum_{t=1}^{T}\zeta_{ct}\zeta_{c't}}
\end{aligned}\tag{18}$$

As before, each σ_{gcn} must be normalized by $\sum_{c'=1}^{C}\sigma_{gc'n}$.

ϕ_c is updated as follows:

$$\phi_c^2\leftarrow\left(1+\sum_{g=1}^{G}\frac{\lambda_{gc}^2}{\psi_g^2}\sum_{n=1}^{N}\sigma_{gcn}\sum_{n'=1}^{N}\rho_{gn'}\right)^{-1}\tag{19}$$

Analogous to the ICM case, rows of λ are independent of one another as are columns of ζ, given all other variational parameters. Updated values for the g^{th} row of λ and the t^{th} column of ζ are solutions to (20), (21).

$$\begin{aligned}
&\forall c\in\{1,\ldots,C\}:\\
&\quad\sum_{c'=1}^{C}\lambda_{gc'}\left\{\left(\sum_{n=1}^{N}\sum_{n'=n+1}^{N}\left(\sigma_{gcn}\sigma_{gc'n'}+\sigma_{gcn'}\sigma_{gc'n}\right)\sum_{n''=n'}^{N}\rho_{gn''}\right)\left(\sum_{t=1}^{T}\zeta_{ct}\zeta_{c't}\right)\right\}\\
&\qquad+\lambda_{gc}\left\{\left(\sum_{n=1}^{N}\sigma_{gcn}\sum_{n'=n}^{N}\rho_{gn'}\right)\left(\sum_{t=1}^{T}\left(\zeta_{ct}^2+\phi_c^2\right)\right)\right\}=\left(\sum_{n=1}^{N}\sigma_{gcn}\sum_{n'=n}^{N}\rho_{gn'}\right)\left(\sum_{t=1}^{T}x_{gt}\zeta_{ct}\right)
\end{aligned}\tag{20}$$

$$\begin{aligned}
&\frac{\zeta_{ct}}{\phi_c}+\sum_{c'=1}^{C}\zeta_{c't}\left\{\sum_{g=1}^{G}\frac{\lambda_{gc'}\lambda_{gc}}{\psi_g^2}\sum_{n=1}^{N}\sum_{n'=n+1}^{N}\left(\sigma_{gcn}\sigma_{gc'n'}+\sigma_{gcn'}\sigma_{gc'n}\right)\sum_{n''=n'}^{N}\rho_{gn''}\right\}\\
&\quad=\sum_{g=1}^{G}\frac{x_{gt}\lambda_{gc}}{\psi_g^2}\sum_{n=1}^{N}\sigma_{gcn}\sum_{n'=n}^{N}\rho_{gn'}
\end{aligned}\tag{21}$$

Finally, parameter Ψ is again learned by solving $\partial\mathcal{F}/\partial\Psi = \mathbf{0}$.

$$\forall g \in \{1,\ldots,G\}:$$
$$\psi_g^2 \leftarrow \frac{1}{T}\sum_{c=1}^{C}\left(\sum_{n=1}^{N}\sigma_{gcn}\sum_{n'=n}^{N}\rho_{gn'}\right)\sum_{t=1}^{T}\left[(x_{gt}-\lambda_{gc}\zeta_{ct})^2+\lambda_{gc}^2\phi_c^2\right] - \frac{1}{T}\left(\sum_{n=2}^{N}(n-1)\,\rho_{gn}\right)\sum_{t=1}^{T}x_{tg}^2$$
$$+\frac{2}{T}\sum_{c=1}^{C}\sum_{c'=1}^{C}\left(\sum_{n=1}^{N}\sum_{n'=n+1}^{N}\sigma_{gcn}\sigma_{gc'n'}\sum_{n''=n'}^{N}\rho_{gn''}\right)\lambda_{gc}\lambda_{gc'}\sum_{t=1}^{T}\zeta_{ct}\zeta_{c't} \tag{22}$$

5 Structured Variational Inference

The mean field decomposition in (11) assumes a significant degree of independence among $\{\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}\}$. More accurate Q-distributions approximating (3) are possible if more structure is introduced. For instance, if the number of factors associated with each data point (r_g) and the indices of these factors ($s_{g1},\ldots,s_{gr_g}$) are coupled, the Q-distribution assumes the following structure:

$$P(\mathbf{Y},\mathbf{Z},\mathbf{S},\mathbf{r}|\mathbf{X},\Psi) \approx \prod_{g=1}^{G}Q(\mathbf{y}_g)\cdot\prod_{c=1}^{C}Q(\mathbf{z}_c)\cdot\prod_{g=1}^{G}Q(\mathbf{s}_g,r_g) \tag{23}$$

An obvious parameterization is as follows:

$$Q(\mathbf{y}_g) = \overbrace{\prod_{n=1}^{r_g}\delta\left(y_{gs_{gn}}-\lambda_{gs_{gn}}\right)}^{\lambda_{gc}\text{ is a point estimate of }y_{gc},\ \ldots} \cdot \overbrace{\prod_{\substack{c=1\\ c\notin\{s_{g1},s_{g2},\ldots,s_{gr_g}\}}}^{C}\delta(y_{gc})}^{\ldots\text{ unless }\{\mathbf{s}_g,r_g\}\text{ force }y_{gc}\text{ to zero.}} \tag{24}$$

$$Q(z_{ct}) = \mathcal{N}\left(z_{ct};\,\zeta_{ct},\,\phi_c^2\right);\quad Q\left(\mathbf{s}_g=\left[\begin{array}{cccc}c_1 & c_2 & \cdots & c_n\end{array}\right],\,r_g=n\right) = \sigma_{gc_1c_2\cdots c_n} \tag{25}$$

Calculating $\mathrm{D}(Q\|P)-\log P(\mathbf{X})$ as before (15) yields the new free energy:

$$\mathcal{F} = \sum_{g=1}^{G}\sum_{n=1}^{N}\sum_{c_1=1}^{C}\cdots\sum_{c_{n-1}=1+c_{n-2}}^{C}\sum_{c_n=1+c_{n-1}}^{C}\sigma_{gc_1\cdots c_{n-1}c_n}\cdot\log\frac{\sigma_{gc_1\cdots c_{n-1}c_n}}{\nu_n\Big/\sum_{n'=1}^{N}\nu_{n'}\binom{C}{n'}}$$
$$-\frac{T}{2}\sum_{c=1}^{C}\left(1+\log\phi_c^2\right)+\frac{1}{2}\sum_{t=1}^{T}\sum_{c=1}^{C}\left(\zeta_{ct}^2+\phi_c^2\right)+\frac{T}{2}\sum_{g=1}^{G}\log 2\pi\psi_g^2$$
$$+\frac{1}{2}\sum_{g=1}^{G}\frac{1}{\psi_g^2}\sum_{n=1}^{N}\sum_{c_1=1}^{C}\cdots\sum_{c_n=1+c_{n-1}}^{C}\sigma_{gc_1\cdots c_n}\sum_{t=1}^{T}\left[\left(x_{gt}-\sum_{n'=1}^{n}\lambda_{gc_{n'}}\zeta_{c_{n'}t}\right)^2+\sum_{n'=1}^{n}\lambda_{gc_{n'}}^2\phi_{c_{n'}}^2\right] \tag{26}$$

Because the factor probabilities are not fully factorized as before, no meaningful simplification of this $\mathcal{O}(GN^2C^NT)$ expression is possible.

It should be noted that since the factors for each data point are no longer independent, elements of each s_g must be sorted in ascending order for the sake of uniqueness. This is built into the indexing of the $\sum_{c_1=1}^{C}\cdots\sum_{c_{n-1}=1+c_{n-2}}^{C}\sum_{c_n=1+c_{n-1}}^{C}\sigma_{gc_1\cdots c_n}$ summations. The other implication is that while factor choices are still uniformly distributed, the σ variational parameters are no longer uniform because the $P(r_g=n)=\nu_n$ prior is mixed in, as shown in the first line of (26).

Finally, an even less-factored approximating Q-distribution discards the assumption of independence between $\mathbf{Y}$ and $\{\mathbf{S}, \mathbf{r}\}$. Its Q-distribution is:

$$P(\mathbf{Y}, \mathbf{Z}, \mathbf{S}, \mathbf{r}|\mathbf{X}, \Psi) \approx \prod_{c=1}^{C} Q(\mathbf{z}_c) \cdot \prod_{g=1}^{G} Q(\mathbf{y}_g, \mathbf{s}_g, r_g) \tag{27}$$

with new parameterization:

$$Q\left(\mathbf{y}_g,\ \mathbf{s}_g = \left[\begin{array}{cccc} c_1 & c_2 & \cdots & c_n \end{array}\right],\ r_g = n\right) = \sigma_{gc_1c_2\cdots c_n} \cdot \prod_{n'=1}^{n} \delta\left(y_{gc_{n'}} - \lambda_{gn'c_1c_2\cdots c_n}\right) \cdot \prod_{\substack{c=1 \\ c\notin\{c_1,c_2,\ldots,c_n\}}}^{C} \delta(y_{gc}) \tag{28}$$

The free energy is identical to (26) except that $\lambda_{gc_{n'}}$ is replaced with $\lambda_{gn'c_1c_2\cdots c_n}$. This factorization improves accuracy, but requires N times the memory as the first structured variational technique. In the former, the bottleneck is $\sigma_{gc_1c_2\cdots c_n}$, requiring storage for $G \cdot \sum_{n=1}^{N} \binom{C}{n}$ floating point values. In the latter, the bottleneck is $\lambda_{gn'c_1c_2\cdots c_n}$, with $G \cdot N \cdot \sum_{n=1}^{N} \binom{C}{n}$ values.

6 Simulation Results

We present the results of applying the optimization algorithms presented in the previous sections to sample data. This data was generated with the following MATLAB code:

```
% Generate Sample Data in MATLAB
Z = randn(C,T);
Y = zeros(G,C);
for g = 1:G,
    n = ceil(N*rand);
    Y(g,ceil(C*rand(1,n))) = -5 + 10*rand(1,n);
end;
X = Y * Z + randn(G,T);
```

Since the data is artificially generated directly from the generative model (-), this is not a valid test of model applicability to real-world situations. This synthetic data can, however, be used to assess the relative performance of each Q-distribution factorization depth.

Each of the algorithms (ICM, (11), (23), (27)) was run for 100 iterations on a data set with $G = 1000$, $T = 15$, $C = 25$, and $N = 3$. The variational methods used $\nu = \left[\begin{array}{ccc} .55 & .27 & .18 \end{array}\right]$ (i.e. $\nu_n \propto 1/n$), though other priors weighted more towards low n-values[2] would produce similar values. The plots showing the free-energy minimization (and the related log likelihood (4) maximization) are shown in Figure 1.

Clearly, iterated conditional modes is outperformed by mean-field factorization, and the structured factorizations in turn outperform mean-field. Interestingly, the $Q(\mathbf{Y}) \cdot Q(\mathbf{S}, \mathbf{r})$ and $Q(\mathbf{Y}, \mathbf{S}, \mathbf{r})$ structured factorizations end up with similar log likelihoods. They differ greatly, however, in computational complexity, which is shown by plotting free energy and log likelihood as a function of running time (as opposed to iteration) in Figure 2.

For this particular data set size ($G = 1000$, $T = 15$, $C = 25$, $N = 3$), the tradeoff between computational complexity and Q-distribution accuracy is such that each step up in algorithm complexity involves roughly an order of magnitude increase in running time. The

[2] A uniform prior ν (reflecting no knowledge about the distribution of $\mathbf{r}$) 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 equal N. Weighting the prior towards lower values ensures that factors will only be included if they make a noteworthy difference.

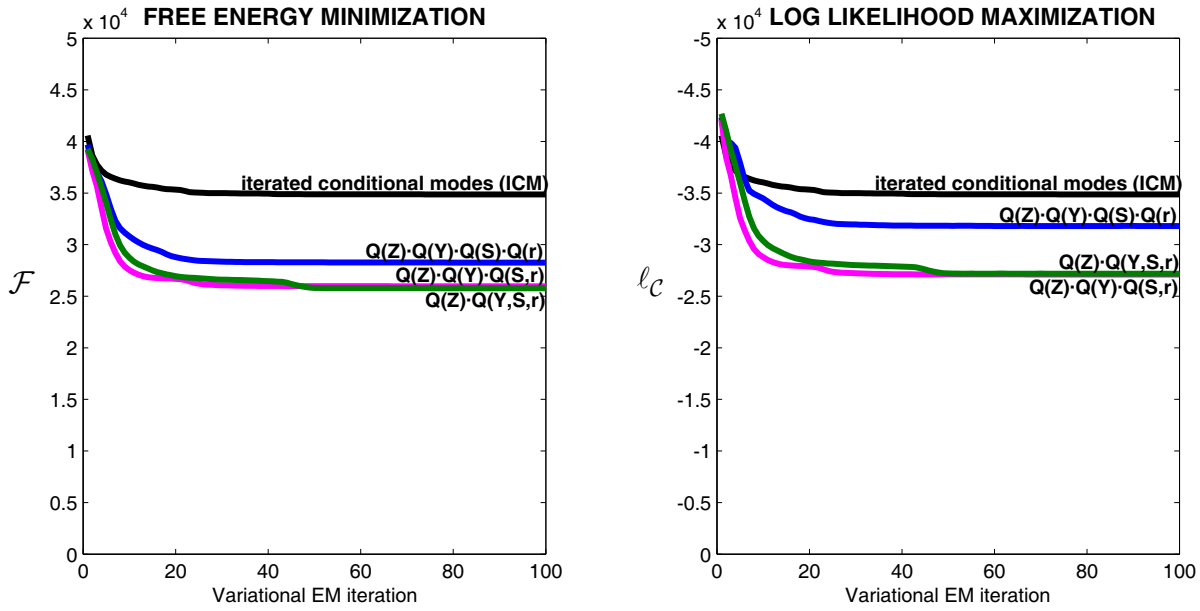

Figure 1: Free energy ($\mathcal{F}$) minimization and model log likelihood ($\ell_{\mathcal{C}}$) maximization as a function of EM iteration. Iterated conditional modes and factorized variational methods with three different factorization levels are shown.

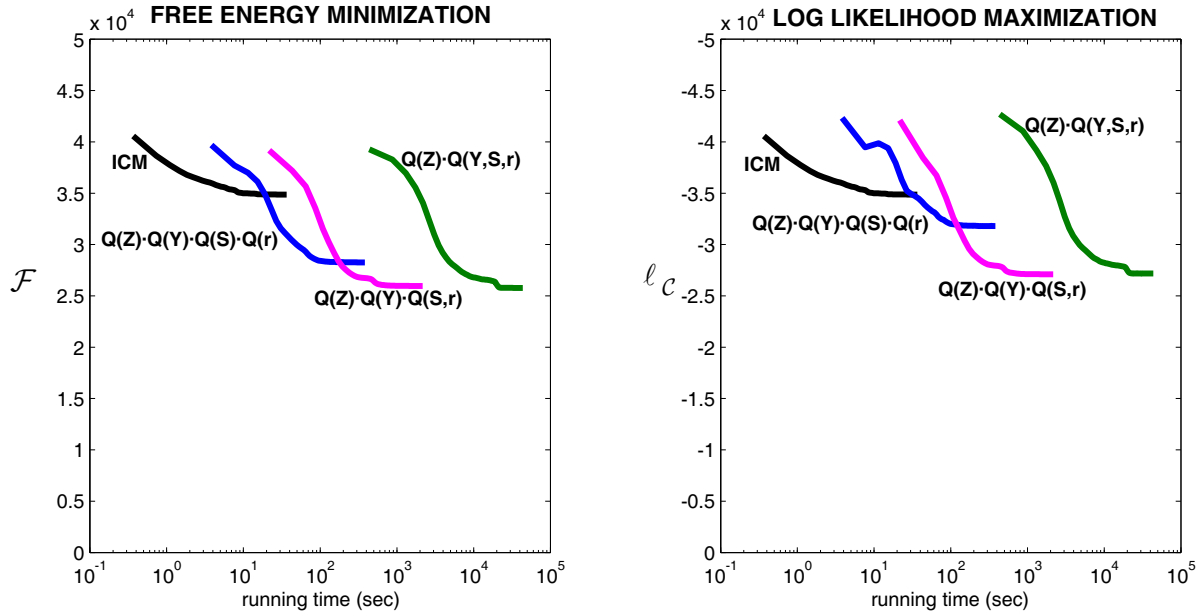

Figure 2: Free energy ($\mathcal{F}$) minimization and model log likelihood ($\ell_{\mathcal{C}}$) maximization as a function of running time. Iterated conditional modes and factorized variational methods with three different factorization levels are shown.

key running-time quantity to keep in mind is C^2 (for ICM and mean-field) versus C^N (for more-structured Q).

7 Summary

Many kinds of data vectors can most naturally be explained as an additive combination of a selection of prototype vectors, which can be viewed as computational problem of finding a sparse matrix factorization. While most work 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. Like [1], our algorithm computes a sparse matrix factorization, but instead of making point estimates (hard decisions) for factor selections, our algorithm computes probability distributions. We find that this enables the algorithm to avoid local minima found by iterated conditional modes.

References

[1] Srebro, N. and Jaakkola, T. (2001) Sparse Matrix Factorization of Gene Expression Data. Unpublished note, MIT Artificial Intelligence Laboratory. Available at www.ai.mit.edu/-research/abstracts/abstracts2001/genomics/01srebro.pdf

[2] Bell, A. J. and Sejnowski, T. J. (1995) An information maximization approach to blind separation and blind deconvolution. *Neural Computation* **7**: 1129-1159.

[3] Besag, J. (1986) On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society B* **48**: 259–302.

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