Variational Bayesian Learning of Cooperative Vector Quantizer Model - The Theory

Xinghua Lu[1,2], Milos Hauskrecht[2] and Roger S. Day[3]

[1]Center for Biomedical Informatics, [2]Department of Computer Science,
[3]Department of Biostatistics
The University of Pittsburgh
xlu@cbmi.upmc.edu, milos@cs.pitt.edu and day@upci.pitt.edu

Abstract

This is the first part of a two-parted report on development of a statistical learning algorithm for a latent variable model referred to as cooperative vector quantizer model. This part presents the theory and mathematical derivations of a variational Bayesian learning algorithm for the model. The model has general applications in the field of machine learning and signal processing. For example it can be used to solve the problem of blind source separation or image separation. Our special interest is in its potential biological application in that we can use the model to simulate signal transduction components regulating gene expression as latent variables. The algorithm is capable of automatically and efficiently determining the number of latent variables of the model, estimating the distribution of the parameters and latent variables. Thus, we can use the model to address following biological questions regarding gene expression regulation: (1) What are the key signal transduction components regulating gene expression in a given kind of cell; (2) How many key components are needed to efficiently encode information for gene expression regulation; (3) What are the states of the key components for a given gene expression data point. Such information will provide insight for understanding the mechanism of information organization of cells, mechanism of diseases and drug effect/toxicity.

1 Introduction

1.1 Biological Motivation

A biological system has a sophisticated signal transduction system. Activation of a signal transduction pathway usually involves change of state of many signal transduction molecules which exert diverse cellular functions. Quite often, the signal is be eventually passed to transcription factors or repressor which, in turn, will activate or depress the transcription of genes. For example, activation of epithelial growth hormone receptor (EGFR) usually activates a cascade of protein kinases, which eventually activate transcription of a set of early response genes. Thus, the correlated expression level of these early response genes simply reflects the state of this signal transduction pathway. However, the biological systems are complicated by the fact that different pathways are inter-weaved. It is not uncommon that expression level of an individual gene is controlled by multiple pathways. A ordinary cell has hundreds to thousands of receptors on it plasma membrane and is constantly bombarded by different signals from surrounding environment. It would be very inefficient if each of these receptors has a distinct pathway controlling expression of individual genes. One can imagine that signals from different receptors will eventually be orchestrated at a certain level such that information is encoded most efficiently and, from this level, information is further disseminated to control the expression of thousands genes. This is analogous to information compression, where large amount of information is compressed, passed through a channel and regenerated at the other end of channel. Let us hypothesize that there exist some signal transduction components (STC) which encode necessary information to control the gene expression for a give cell type. Then, a biologist would tend to ask following questions:

1. What are these STC? Can one identify the STC and map them to a biological entities such as proteins or pathways?

2. How many STC are needed to encode the information in a given kind cell?

3. Can one infer the state of these STC when provided with gene expression data?

Capability of answering these questions will provide insight into a biological system in terms of (1)how information of signal transduction pathways are organized and what are the key components that can efficiently encode information. (2) mechanism of diseases; (3) mechanism of drug effect or toxicity, and

so on. However, these questions also pose serious challenges to computational biologists because such information can not be directly observed from DNA microarray experiments, even though the contemporary DNA microarray technology almost enables one to study the expression of genes almost at whole genome level. Nonetheless, upon given gene expression data, a computational biologist potentially can infer the information based on certain assumption and model. One plausible approach is to model the STCs with latent variables in order to explain how observed data are generated and use statistical learning techniques to infer the parameters of the model. Once equipped with the parameters of model, one can estimate the states of STC when given new microarray data. In this research, we develop a novel learning algorithm for a latent variable generative model based on recent advances in machine learning field to address these questions.

Currently, a variety of techniques have been applied to explore the correlated gene expression patterns to infer the regulation pathways. Among which, clustering algorithms, including the nonparametric hierarchical clustering and model-based mixture of Gaussian models, are most commonly used. These approaches provide useful information about transcription profiles of genes and group genes with similar profile assuming they are regulated by same pathway. However, one key drawback of clustering is that genes are assigned to clusters mutually exclusively, which does not reflect the fact that expression of an individual gene can be regulate by multiple pathways. Other approaches such as principal component analysis (Raychaudhuri et al., 2000), single value decomposition (SVD), independent component analysis (ICA) (Liebermeister, 2002) are also used to analyze gene expression patterns. However, most of the above mentioned approaches can not effectively address the question such as what is optimal number of clusters (or components) to be included in the model. Recently, graphic models like Bayesian network and Boolean network have been used to modeling the genetic regulation pathways (Friedman et al., 2000; Liang et al., 1998). One limitation of such approach is that statistical *dependence* are frequently confounded by existence of latent variable, such as activation state of proteins or pathways which are not explicitly modeled by the approach. Current graphic learning algorithms can not handle latent variables efficiently due to computational complexity.

1.2 Latent Variable Models

Latent variable models have been widely used in statistics, psychology, economics and machine learning researches. They are also frequently referred to as generative models in that generation of observed data are controlled by latent variables.

Popular latent variable models include factor analysis, ICA, independent factor analysis, cooperative vector quantizer (CVQ) and probabilistic principle component analysis (PPCA) (Attias, 1999a; Roweis and Ghahramani, 1999; Ghahramani, 1995; Tipping and Bishop, 1997). As pointed out by Rowies and Ghahramani (Roweis and Ghahramani, 1999), most of these model belong to a unified linear Gaussian model and assume the form of

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \epsilon \tag{1}$$

where $\mathbf{y}$ is a D dimensional vector of observed data, $\mathbf{W}$ is a $D \times K$ "loading matrix", $\mathbf{x}$ is a K dimensional vector of latent "factors/sources" with $K << D$ and ϵ is D dimensional noise which assume Gaussian distribution $\epsilon \sim \mathcal{N}(0, \Lambda)$. The key idea of these model is use latent variables as "*informative lower dimensional projection or explanation of the complicated observations*". When analyzing high dimensional data as in our case, the advantage of using latent variable model is several: (1) Dimension reduction. The dimension D of microarray data range from thousand to ten's thousand, it is very difficult and inefficient to describe the characteristic of a given sample with such high dimension. With reduced dimension, one describe the data more succinctly. Furthermore, reduced dimension means reduced computation complexity and less likely to over-fit data if we further use microarray data to perform classification. (2) Explain the correlation of observed data at latent variables level. As discussed in previous section that correlated expression patterns of genes may simply reflect the fact that they co-regulated by same STC. By putting constrains on the covariance matrix of observed data noise, the latent variable models can capture the correlated expression patterns at latent variable level, thus provide explanation for observed covariance. (3) Inferring state of latent variables. The distribution (or state) of latent variables for a data point can be inferred use statistical techniques. This information is very useful in that we can estimated states of a biological system and such information can be used to perform other tasks, i.e. classification.

However, all above mentioned models are not suitable for answering the questions raised in section 1.1. One of our goal is to determine what biological entities the STC may corresponding to. This requires us to recover the sources uniquely and to determine what genes are controlled by each STC. Then we can use biological knowledge to further infer what these STC might correspond to. This is closely related to a machine learning topic – blind source separation. It is well known that factor analysis and PPCA are inadequate in this respect due to the fact that latent variables in these models are Gaussian distributed and recovered sources are subjected to rotation invariance (Attias, 1999a). Conventional ICA model assumes

non-Gaussian distribution for sources and can perform blind source separation, but the assumptions of the model are too restrictive, i.e. assuming same number of source and observations and noise free system. Recently developed independent factor analysis (Attias, 1999a) try to avoid the rotational invariance of conventional factor analysis by adopting mixture of Gaussian distribution for latent variables. The idea is adopted by ICA community to develop newer versions of EM learning algorithms to perform blind source separation. However, after recovering the distribution of the sources, it will be difficult to interpret the result (a mixture of Gaussian distributions) from biological point of view.

In this research, we adopt the cooperative vector quantizer model and also refer to it as a multiple cause model. We extend the model by performing full Bayesian learning in order to address the questions raised in section 1.1. The EM algorithm for learning parameters of the model was developed by Ghahramani (Ghahramani, 1995) and was demonstrated to be capable of separating sources uniquely. Here, we will briefly introduce the key features of the model and its relevance to biological problem and leave detailed discussion in section 2. In this model, we represent the signal transduction components or sources as a set of latent binary variables which can assume on/off state. The state of a source reflects balanced effect from upstream signal transduction system. The observed DNA microarray data is the result of concerted regulation by these sources. When a source is turned on, it influences gene expression pattern by outputting a weight onto every gene on the microarray, although for most of genes the weight is zero reflecting the fact that the given source has no influence on these genes. On the other hand, if a source outputs a nonzero weight to a subset of genes, it indicates that these genes are co-regulated by the source. Although the fact that this model can be used to identify subset of co-regulated genes sounds similar to clustering, there is a fundamental difference between the approaches because, in our model, expression of an individual gene can be regulated by multiple sources. The main task for our model is to learn/estimate the weight matrix associated with sources. Thus, after learning the weight matrix of the model, we would be able to determine what genes are regulated by a given STC and use biological knowledge to infer what the STC corresponding to. Furthermore, with model parameters learned, we would be able to estimate the states of STC for new data.

1.3 Bayesian Model Selection

The question (2) in the section 1.1 is of great biological interest because it addresses how information is organized inside a biological system as regulation of

gene expression is concerned. This problem was not effectively addressed before either experimentally or computationally due to lack of data to support such study. With advent of DNA microarray technology, one can monitor the gene transcription at whole genome level. With data collected under variety of conditions, e.g. data collected while cells going through cell cycle, a good experiment sample will contain most of the necessary information to address the issue. From biological point view, the merge point of signal transduction does exist and potentially can be at transcription factors/repressors level, because most pathways exert their influence on gene expression through activating or inactivating these proteins. The question is whether it is most efficient to encode all information controlling gene expression at this level and how can one determine the number of STC needed to encode the information. In another word, if we set out to model the DNA microarray data with latent variable model, how many latent variables should we include in the model and whether the number reflects the most efficient information encoding?

This can be address in Bayesian model selection framework (Bishop, 1999; Ghahramani and Beal, 2000a; MacKay, 1995; Kass and Raftery, 1994) which embodies **Occam's Razor**, a principle that states to select the simplest model among those have same description power. That is, if we select model according to Bayesian model selection frame, we automatically recover the model that has minimum number of parameters compared to other models with same description power and recover the number of latent variables that will describe the data most efficiently. Similarly, in information theory, minimal description length (MDL) principle dictates that an encoding system prefers a model that has minimum parameters comparing to models with same power to describe the observed data (Hansen and Yu, 2001). The relation between Bayesian model selection and MDL principle has been point out by several authors (Hansen and Yu, 2001; Attias, 1999b). Critics on Bayesian model selection is that it requires integration of parameters which is intractable for most of practical models. In this research, we will adopt newly developed variational Bayesian approach to overcome such drawback and perform model selection in a efficient way (see details in section 3 and 8.3).

2 Model

In the CVQ model, a set of hidden discrete sources $\mathbf{s} = \{s_1, s_2, \ldots, s_K\}$ controls the generation of a vector of D dimension observed variables $\mathbf{y}$. Each discrete source is an indicator vector of dimension m. Each vector s_k has one nonzero element such that $s_{ki} = 1$ and $s_{kj} = 0, \forall j \neq i$. In this research, we use the sources to model the state of signal transduction components, therefore we restrict the sources to be binary variables, reflecting activation and inactivation states of a component respectively. We can easily extend the current model to accommodate multiple states of component if biological justification exists. When the source $s_k = 1$, it will output a D dimensional weight $\mathbf{w}_k$ to $\mathbf{y}$. We can think the source variable s_k as a switch which, when turned on, allows outflow of weights $\mathbf{w}_k$ to $\mathbf{y}$. More formally

$$\mathbf{y} = \sum_{i=1}^{K} s_k \mathbf{w}_k + \epsilon \tag{2}$$

where s_k is an index function, $\mathbf{w}_k$ is the weight output by source s_k, $\epsilon \sim \mathcal{N}(0, \Lambda)$ is noise of the system.

If the weight $\mathbf{w}_k$ is Gaussian distribution, the linear combination of weight is still a Gaussian distribution (Roweis and Ghahramani, 1999). Thus

$$P(\mathbf{y}|\mathbf{s}) \sim \mathcal{N}\left(\sum_{i=1}^{K} s_k \mathbf{w}_k, \Lambda\right) \tag{3}$$

Parameters (θ) of the model: $\pi = \{\pi_1, \pi_2, \ldots, \pi_K\}$ are the probabilities that $s_k = 1$; $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K\}$ are the weights output by $\mathbf{s}$; $\Lambda = \tau^{-1} I$ is a $D \times D$ diagonal variance matrix, where τ is precision (inversed variance) of observation $\mathbf{y}$.

3 Variational Bayesian Learning

One of our goal is to determine what model structure best describe the data. In this project, one of the main concerns is the number of sources. Let $\boldsymbol{Y} = \{\boldsymbol{y}^n; n = 1, 2, \ldots, N\}$ be observed data; $M = \{M_i; i = 1, 2, \ldots, K\}$ be a set of possible model structures, where M_i is a model with i hidden sources. We can use Bayes' rule to calculate the posterior probability of each models.

$$P(M_i|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|M_i)P(M_i)}{P(\boldsymbol{Y})} \tag{4}$$

Then we can select the model that has the highest posterior probability. The full Bayesian treatment of model selection requires calculating the *evidence* $P(\boldsymbol{Y}|M_i)$ by integrating out all possible setting of parameters $\boldsymbol{\theta}$ for a given model

$$P(\boldsymbol{Y}|M_i) = \int_{\theta} P(\boldsymbol{Y}|\boldsymbol{\theta}, M_i)P(\boldsymbol{\theta}|M_i)d\boldsymbol{\theta} \tag{5}$$

However, the integration is intractable. We can use the variational approximation to achieve the goal, which takes advantage of the fact that log marginal probability of observed data $\mathcal{L}(\boldsymbol{Y})$ [1] results from integrating out hidden variables $(\mathbf{H})$ and parameters $(\boldsymbol{\theta})$ and can be bounded below as following

$$\begin{aligned}
\mathcal{L}(\mathbf{Y}) &= \ln P(\mathbf{Y}) && (6)\\
&= \ln \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} P(\mathbf{Y}, \mathbf{H}|\boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta} && (7)\\
&= \ln \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H}, \boldsymbol{\theta})\frac{P(\mathbf{Y}, \mathbf{H}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{Q(\mathbf{H}, \boldsymbol{\theta})}d\boldsymbol{\theta} && (8)\\
&\geq \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H}, \boldsymbol{\theta}) \ln \frac{P(\mathbf{Y}, \mathbf{H}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{Q(\mathbf{H}, \boldsymbol{\theta})}d\boldsymbol{\theta} \equiv \mathcal{F}(Q) && (9)
\end{aligned}$$

for any distribution $Q(.)$ (Attias, 1999b; Ghahramani and Beal, 2000a) . The inequality is established by Jensen's Inequality. We can demonstrate that the difference between $\mathcal{L}(\mathbf{Y})$ and $\mathcal{F}(Q)$ is the Kullback-Leibler divergence between $Q(\mathbf{H}, \boldsymbol{\theta})$ and true posterior $P(\mathbf{H}, \boldsymbol{\theta}|\mathbf{Y})$.

[1] For the purpose of notation simplicity, we omit conditioning on model M_i. Most of probabilities $P(.)$ mentioned henceforth in the report are conditional probabilities $P(.|M_i)$ implicitly conditioned on a given model.

$$
\begin{aligned}
&\ln P(\mathbf{Y}) - \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H},\boldsymbol{\theta}) \ln \frac{P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{Q(\mathbf{H},\boldsymbol{\theta})} d\boldsymbol{\theta} && (10)\\
&= \ln P(\mathbf{Y}) - \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H},\boldsymbol{\theta}) \ln \frac{P(\mathbf{H},\boldsymbol{\theta}|\mathbf{Y})P(\mathbf{Y})}{Q(\mathbf{H},\boldsymbol{\theta})} d\boldsymbol{\theta} \\
&= \ln P(\mathbf{Y}) - \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H},\boldsymbol{\theta}) \ln \frac{P(\mathbf{H},\boldsymbol{\theta}|\mathbf{Y})}{Q(\mathbf{H},\boldsymbol{\theta})} d\boldsymbol{\theta} - \ln P(\mathbf{Y}) \\
&= \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H},\boldsymbol{\theta}) \ln \frac{Q(\mathbf{H},\boldsymbol{\theta})}{P(\mathbf{H},\boldsymbol{\theta}|\mathbf{Y})} = KL(Q \parallel P) \geq 0 && (11)
\end{aligned}
$$

Thus, maximizing $\mathcal{F}(Q)$ is equivalent to minimizing $KL(Q \parallel P)$. When $Q(\mathbf{H},\boldsymbol{\theta}) = P(\mathbf{H},\boldsymbol{\theta}|\mathbf{Y})$, at which point $KL(Q \parallel P) = 0$, $\mathcal{L}(\mathbf{Y}) = \mathcal{F}(Q)$ and one can use conventional EM algorithm (Dempster et al., 1977) to estimate parameters of the model . In many cases, estimation of the posterior distribution $P(\mathbf{H},\boldsymbol{\theta})$ is infeasible, then one can use an arbitrary distribution $Q(\mathbf{H},\boldsymbol{\theta})$ as approximation of posterior distribution. More specific, if we adopt variational approximation approach [2] and restrict the $Q(\mathbf{H},\boldsymbol{\theta})$ to be factorized as $Q(\mathbf{H},\boldsymbol{\theta}) = Q_H(\mathbf{H})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
& \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q(\mathbf{H},\boldsymbol{\theta}) \ln \frac{P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{Q(\mathbf{H},\boldsymbol{\theta})} d\boldsymbol{\theta} && (12)\\
= & \int_{\boldsymbol{\theta}} \sum_{\mathbf{H}} Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) Q_H(\boldsymbol{H}) \ln \frac{P(\boldsymbol{Y},\boldsymbol{H}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{Q_{\boldsymbol{H}}(\boldsymbol{H})Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} d\boldsymbol{\theta} && (13)\\
= & \int_{\boldsymbol{\theta}} d\boldsymbol{\theta} Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \left[\sum_{\mathbf{H}} Q_H(\mathbf{H}) \ln \frac{P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})}{Q_H(\mathbf{H})} + \ln \frac{P(\boldsymbol{\theta})}{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right] && (14)\\
= & \left\langle \sum_{\mathbf{H}} Q_H(\mathbf{H}) \ln \frac{P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})}{Q_{\mathbf{H}}(\mathbf{H})} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} - \left\langle \ln \frac{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{P(\boldsymbol{\theta})} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} && (15)\\
= & \mathcal{F}_{\boldsymbol{\theta}} - KL(Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \parallel P(\boldsymbol{\theta})) \equiv \mathcal{F}(Q_{\mathbf{H}}(\mathbf{H}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{Y}) && (16)
\end{aligned}
$$

where $\langle . \rangle_{Q(.)}$ is to take expectation with respect to distribution $Q(.)$

To maximize the $\mathcal{F}(Q_{\mathbf{H}}(\mathbf{H}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{Y})$ with respect to $Q_{\mathbf{H}}(\mathbf{H})$ and $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$,

[2] See section 4.

we have [3]

$$Q_H^*(\mathbf{H}) \propto exp\langle \ln P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \quad (17)$$

$$Q_{\boldsymbol{\theta}}^*(\boldsymbol{\theta}) \propto P(\boldsymbol{\theta}) exp\langle \ln P(\mathbf{Y},\mathbf{H}|\boldsymbol{\theta})\rangle_{Q_H(\mathbf{H})} \quad (18)$$

As we can see from (17) and (18) that $Q_{\mathbf{H}}(\mathbf{H})$ and $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ are coupled. We can use EM like iterations to update two distributions and maximize the $\mathcal{F}(Q(\mathbf{H}), Q(\boldsymbol{\theta}), \mathbf{Y})$ function.

- **VBE step**
  Maximize $\mathcal{F}(Q_H(\mathbf{H}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{Y})$ with respect to $Q_H(\mathbf{H})$ using the expected ***natural parameters*** under current $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$.

- **VBM step**
  Maximize $\mathcal{F}(Q_H(\mathbf{H}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{Y})$ with respect to parameter distribution $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. As indicated by equation (18), this amounts to updating the posterior with expected sufficient statistics under $Q_H(\mathbf{H})$.

Iterate through the **VBE step** and **VBM step** until $\mathcal{F}(Q_{\mathbf{H}}(\mathbf{H}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{Y})$ converge. Notice that maximizing $\mathcal{F}(Q_{\mathbf{H}}(\mathbf{H}), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{Y})$ is only maximizing the *lower bound* for marginal log likelihood $\mathcal{L}(\mathbf{Y})$. How tight is the lower bound depends on how well the distribution $Q(\mathbf{H}, \boldsymbol{\theta})$ approximate posterior distribution.

[3] Described in detailed in section 6 and 7.

4 Mean Field Approximation

As discussed in previous section, maximization of $\mathcal{F}(Q_H(\mathbf{H}), Q_\theta(\theta), \mathbf{Y})$ corresponds to minimization of KL divergence between the posterior distribution $P(\mathbf{H}, \theta|\mathbf{Y})$ and approximation distribution $Q(\mathbf{H}, \theta)$. For many practical models, including the model in this research, exact calculation of the posterior is intractable. One can use restrict distribution $Q(\mathbf{H}, \theta)$ to approximate posterior. A commonly used approach is the mean field approximation (Jordan et al., 1998). The intuition underlying the mean field approximation is that, when the distribution of a variable v_i is dependent on many other variables $\mathbf{V}_{j\neq i} = \{v_j; j \neq i\}$, change of one individual variable of $\mathbf{V}_{j\neq i}$ may have limited effect on v_i due to influence of other variables. That is, v_i is surrounded by a "mean field". This apparently decouples v_i and $\mathbf{V}_{j\neq i}$ and the joint distribution of the variables can be factored and updated iteratively. Haft et al (Haft et al., 1997) demonstrated that the $KL(Q||P)$ for any two arbitrary distributions $Q(\mathbf{X})$ and $P(\mathbf{X})$, where $\mathbf{X} = \{X_1, X_2, \ldots, X_K\}$, can be minimized model-independently by adopting a mean field approximation. More specifically, for any $P(\mathbf{X})$ and $Q(\mathbf{X})$, if the distribution $Q(\mathbf{X})$ is restricted as following factorized from

$$Q(\mathbf{X}) = \prod_{i=1}^{K} Q_i(X_i) \tag{19}$$

one can minimize $KL(Q||P)$ iteratively with respect to $Q(X_i)$ while fixing others, $Q(X_{j\neq i})$. Define $\mathbf{X}_{j\neq i} = \{X_j, j \neq i\}$ be the set of all X except X_i, then $Q(\mathbf{X}) = Q(X_i)Q(\mathbf{X}_{j\neq i})$. Rewrite $KL(Q||P)$ as

$$\begin{aligned}
KL(Q||P) &= \int Q(\mathbf{X}) \ln \frac{Q(\mathbf{X})}{P(\mathbf{X})} dX \\
&= \int_{X_i} \int_{\mathbf{X}_{j\neq i}} Q(X_i)Q(\mathbf{X}_{j\neq i}) \ln \frac{Q(X_i)Q(\mathbf{X}_{j\neq i})}{P(\mathbf{X})} dX_i d\mathbf{X}_{j\neq i} \\
&= \int_{X_i} \int_{\mathbf{X}_{j\neq i}} Q(X_i)Q(\mathbf{X}_{j\neq i}) \left(\ln Q(X_i) + \ln Q(\mathbf{X}_{j\neq i})\right) dX_i d\mathbf{X}_{j\neq i} \\
&\quad - \int_{X_i} \int_{\mathbf{X}_{j\neq i}} Q(X_i)Q(\mathbf{X}_{j\neq i}) \ln P(\mathbf{X}) dX_i d\mathbf{X}_{j\neq i} \\
&= \int_{X_i} Q(X_i) \ln Q(X_i) dX_i + \int_{\mathbf{X}_{j\neq i}} Q(\mathbf{X}_{j\neq i}) \ln Q(\mathbf{X}_{j\neq i}) d\mathbf{X}_{j\neq i} \\
&\quad - \int_{X_i} \int_{\mathbf{X}_{j\neq i}} Q(X_i)Q(\mathbf{X}_{j\neq i}) \ln P(\mathbf{X}) dX_i d\mathbf{X}_{j\neq i}
\end{aligned} \tag{20}$$

To minimize $KL(Q||P)$ w.r.t $Q(X_i)$, we need to further place constrain over $Q(X_i)$ such that $\int Q(X_i)dX_i = 1$, which can be achieved by Lagrangian optimization. Define Lagrangian function

$$\begin{aligned}\mathcal{L} &= \int_{X_i} Q(X_i)\ln Q(X_i)dX_i + \int_{\mathbf{X}_{j\neq i}} Q(\mathbf{X}_{j\neq i})\ln Q(\mathbf{X}_{j\neq i})d\mathbf{X}_{j\neq i} \\ &\quad -\int_{X_i}\int_{\mathbf{X}_{j\neq i}} Q(X_i)Q(\mathbf{X}_{j\neq i})\ln P(\mathbf{X})dX_i d\mathbf{X}_{j\neq i} \\ &\quad -\lambda\left(\int Q(X_i)dX_i - 1\right) \end{aligned} \tag{21}$$

take derivative and set to zero

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial Q(X_i)} &= \frac{\partial}{\partial Q(X_i)}\int_{X_i} Q(X_i)\ln Q(X_i)dX_i \\ &\quad -\frac{\partial}{\partial Q(X_i)}\int_{X_i}\int_{\mathbf{X}_{j\neq i}} Q(X_i)Q(\mathbf{X}_{j\neq i})\ln P(\mathbf{X})dX_i d\mathbf{X}_{j\neq i} \\ &\quad -\frac{\partial}{\partial Q(X_i)}\left\{\lambda\left(\int Q(X_i)dX_i - 1\right)\right\} \\ &= \ln Q(X_i) + 1 - \langle\ln P(\mathbf{X})\rangle_{Q(\mathbf{X}_{j\neq i})} - \lambda = 0 \end{aligned} \tag{22}$$

solve for $Q(X_i)$

$$Q(X_i) = \frac{1}{exp(1-\lambda)}exp\left\{\langle\ln P(\mathbf{X})\rangle_{Q(\mathbf{X}_{j\neq i})}\right\} \tag{23}$$

To solve for $exp(1-\lambda)$, we time both sides by itself and integrating both sides over space of (X_i), we have

$$exp(1-\lambda) = \int_{X_i} exp\left\{\langle\ln P(\mathbf{X})\rangle_{Q(\mathbf{X}_{j\neq i})}\right\}dX_i \tag{24}$$

then

$$Q(X_i) = \frac{exp\left\{\langle\ln P(\mathbf{X})\rangle_{Q(\mathbf{X}_{j\neq i})}\right\}}{\int_{X_i} exp\left\{\langle\ln P(\mathbf{X})\rangle_{Q(\mathbf{X}_{j\neq i})}\right\}dX_i} \tag{25}$$

Thus, by iterating through $Q(X_i); i = 1, 2, \ldots, K$, we can achieve overall minimization of $KL(Q||P)$.

5 Priors

To perform variational Bayesian learning, we need the priors for the parameters. Our model consists of parameters $\boldsymbol{\pi}$, $\boldsymbol{W}$ and Λ. We choose conjugate priors for these parameters to facilitate calculation of posterior. We follow the strategies used in the research of variational Bayesian PCA and mixture of factor analyzers (Ghahramani and Beal, 2000b; Bishop, 1999) and define the priors as following

1. For each π_k, which is a Bernoulli parameter, its prior is a Beta distribution

$$\pi_k \sim Beta(\alpha, \beta) \tag{26}$$

   Thus, if our model has K sources, we would need K corresponding prior and $P(\boldsymbol{\pi})$ is of form

$$P(\boldsymbol{\pi}) = \prod_{k=1}^{K} P(\pi_k) \tag{27}$$

2. The loading weight $\mathbf{W}$ can be represented as a $D \times K$ matrix where each column $\mathbf{w}_k; k = 1, 2, \ldots, K$ is the weight output by source s_i. The $P(\mathbf{w}_k|\gamma_k)$ assumes Gaussian distribution

$$\mathbf{w}_k \sim \mathcal{N}(0, \gamma_k^{-1}\mathbf{I}) \tag{28}$$

   where $\gamma_k = \frac{1}{\sigma^2}$ is inverse of variance (precision) for column k of matrix, which follows a gamma distribution

$$\gamma_k \sim \mathcal{G}(\gamma_k|a_\gamma, b_\gamma) \tag{29}$$

   $\mathcal{G}(x|a, b)$ is a gamma distribution in form $p(x) = b^a x^{a-1} exp\{-bx\}/\Gamma(a)$. Then

$$P(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \mathcal{G}(\gamma_k|a_\gamma, b_\gamma) \tag{30}$$

   Thus, the prior for the weight matrix $\mathbf{W}$ is governed by a vector of $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_K\}$ and is of following form

$$P(\mathbf{W}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} \left(\frac{\gamma_k}{2\pi}\right)^{D/2} exp\left(-\frac{\gamma_k}{2}||\mathbf{w}_k||^2\right) \tag{31}$$

3. As in variational Bayesian PCA model, we adopt isotropic variance. (for the sake of simplicity or because PPCA use this to constrain W to be eigenvectors?). The system noise σ^2 is governed by a gamma prior distribution. Let $\tau = \sigma^{-2}$, then

$$P(\tau) = \mathcal{G}(\tau | c_\tau, d_\tau) \tag{32}$$

The graphic representation of the model is shown in Figure 1.

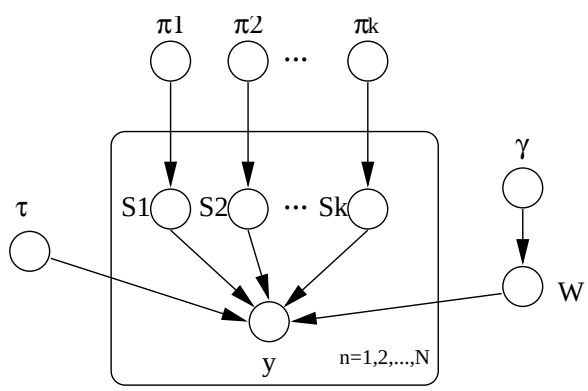

Figure 1: Directed graphic representation of multiple cause model. The square corresponds to an individual data point which contains observed variable $\mathbf{y}$ and latent variables $\mathbf{s}$ which have different instantiation for each data point. $\mathbf{W}$, $\boldsymbol{\gamma}$, τ and α, β are system variables.

6 VBE

As mentioned in section (3), we need to use an EM-like algorithm to iteratively update the distributions $Q_H(s)$ and $Q_\theta(\boldsymbol{\theta})$ to approximate the true posterior distribution. By doing this, we will minimize the KL divergence between the true posterior and approximate distribution and maximize the lower bound for the marginal log likelihood of observed data. In this section, we will discuss how to update the approximate distribution for latent variables.

As we can see from equation (15), to maximize $\mathcal{F}(Q_H(H), Q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y})$ with respect to $Q_H(H)$ is equivalent to maximization of the first term of equation (15) w.r.t $Q_H(H)$

$$\mathcal{F}_\Theta = \left\langle \sum_H Q_H(H) \ln \frac{P(\mathbf{y}, H|\boldsymbol{\theta})}{Q_H(H)} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$$

Inside the $<>$, the term is similar to E step of conventional EM algorithm, which contains the expected complete data likelihood w.r.t $Q_H(H)$ and entropy of $Q_H(H)$. However, we need to further take expectation over the $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. According to the theorem by Ghahramani and Beal (Ghahramani and Beal, 2000a), if the complete likelihood of $P(\mathbf{y}, H|\boldsymbol{\theta})$ belongs to an exponential family, we can rewrite the formula as

$$\mathcal{F}_\Theta = \left\langle \sum_H Q_H(H) \ln \frac{f(\mathbf{y}, H) g(\boldsymbol{\theta}) exp\{\phi(\boldsymbol{\theta})^T u(\mathbf{y}, H)\}}{Q_H(H)} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{33}$$

$$= \sum_H Q_H(H) \ln \frac{f(\mathbf{y}, H) g(\boldsymbol{\theta}) exp\{\overline{\phi}(\boldsymbol{\theta})^T u(\mathbf{y}, H)\}}{Q_H(H)} \tag{34}$$

where $\phi(\boldsymbol{\theta})$ in equation (33) is a vector of natural parameters and $\overline{\phi}(\boldsymbol{\theta})$ in equation (34) is the same vector taken expectation with respect to $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$. Thus, taking expectation of complete likelihood w.r.t $Q_H(H)$ can be done by plugging in $\overline{\phi}(\boldsymbol{\theta})$ and proceeding as conventional E step.

In conventional EM algorithm, if we set $Q_H(H) = P(H|\mathbf{y}, \boldsymbol{\theta})$, to maximize (34) is to maximize log likelihood $\mathcal{L}(\boldsymbol{\theta})$. In other latent variable models, such as the factor analysis or probabilistic PCA, the posterior distribution of hidden variables can be solved analytically. However, in our model, the estimation of true posterior $P(H|\mathbf{y}, \boldsymbol{\theta})$ is intractable. One more time, we resort to variational approach to approximate the true posterior. More specifically, we decouple the

hidden variables and factor the $Q_H(H)$ as following

$$Q_H(H) = \prod_{k=1}^{K} Q_H(H_k) \tag{35}$$

The hidden variables in our model are the source $\mathbf{s} = \{s_1, s_2, \ldots, s_K\}$. Thus, $Q_H(\mathbf{s}) = \prod_{k=1}^{K} Q_H(s_k)$. The parameters of our model are: $\pi_k, \mathbf{w}_k$ and Λ, where π_k is probability $s_k = 1$, $\mathbf{w}_k$ is th weight associated with s_k, $\Lambda = \tau^{-1} I$ is a diagonal variance matrix of noise. Thus, we can write the complete log likelihood of an individual data point of our model as following

$$\begin{aligned}
\ln P(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta}) &= \ln P(\mathbf{s}|\boldsymbol{\theta}) P(\mathbf{y}|\mathbf{s}, \boldsymbol{\theta}) && (36)\\
&= \sum_{k=1}^{K} \left(s_k \ln \pi_k + (1 - s_k) \ln(1 - \pi_k)\right) - \frac{1}{2} \ln |\Lambda| \\
&\quad - \frac{1}{2} \left(\mathbf{y} - \sum_k s_k \mathbf{w}_k\right)^T \Lambda^{-1} \left(\mathbf{y} - \sum_k s_k \mathbf{w}_k\right) + c \\
&= \sum_{k=1}^{K} \left(\ln \frac{\pi_k}{(1 - \pi_k)} s_k + \ln(1 - \pi_k)\right) - \frac{1}{2} \ln |\Lambda| \\
&\quad - \frac{1}{2} \left(\mathbf{y}^T \Lambda^{-1} \mathbf{y} - 2\mathbf{y}^T \Lambda^{-1} \sum_k s_k \mathbf{w}_k + \sum_k \sum_j s_k s_j \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_j\right) \\
&\quad + c && (37)
\end{aligned}$$

where c is a constant. We can see that $\ln \frac{\pi_k}{(1-\pi_k)}$, $\ln |\Lambda|$, Λ^{-1}, $\Lambda^{-1}\mathbf{W}$ and $\mathbf{W}^T \Lambda^{-1} \mathbf{W}$ are the natural parameters

Taking expectation w.r.t $Q_H(\mathbf{s})$ and $Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ respectively as indicated in equa-

tion (15) ‘

$$
\begin{aligned}
&\left\langle \sum_{k=1}^{K} \left(\ln \frac{\pi_k}{(1-\pi_k)} s_k + \ln(1-\pi_k) \right) - \frac{1}{2} \ln |\Lambda| \right\rangle_{Q_H(\mathbf{s}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\
&- \left\langle \frac{1}{2} \left(\mathbf{y}^T \Lambda^{-1} \mathbf{y} - 2\mathbf{y}^T \Lambda^{-1} \sum_k s_k \mathbf{w}_k + \sum_k \sum_j s_k s_j \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_j \right) \right\rangle_{Q_H(\mathbf{s}) Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \qquad (38) \\
= \; & \sum_{k=1}^{K} \left(\left\langle \ln \frac{\pi_k}{(1-\pi_k)} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \langle s_k \rangle_{Q_H(\mathbf{s})} + \langle \ln(1-\pi_k) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) - \frac{1}{2} \langle \ln |\Lambda| \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\
&- \frac{1}{2} \left(\mathbf{y}^T \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \mathbf{y} - 2\mathbf{y}^T \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \sum_k \langle s_k \rangle_{Q_H(\mathbf{s})} \langle \mathbf{w}_k \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) \\
&- \frac{1}{2} \left(\sum_k \sum_j \langle s_k s_j \rangle_{Q_H(\mathbf{s})} \left\langle \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_j \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) + c \qquad (39)
\end{aligned}
$$

Thus, in VBE step, we need to optimize the $\langle s_k \rangle_{Q_H(\mathbf{s})}$ and $\langle s_k s_j \rangle_{Q_H(\mathbf{s})}$. We define a mean field parameter $\langle s_k \rangle_{Q_H(\mathbf{s})} = \lambda_k$ and $\langle s_k s_j \rangle = \lambda_k \lambda_j + \delta_{kj}(\lambda_k - \lambda_k^2)$. We can rewrite the $\mathcal{F}_{\boldsymbol{\theta}}$ and maximize w.r.t λ_k

$$
\begin{aligned}
\mathcal{F}_{\Theta} \; = \; & \sum_{k=1}^{K} \left(\left\langle \ln \frac{\pi_k}{(1-\pi_k)} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \lambda_k + \langle \ln(1-\pi_k) \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) - \frac{1}{2} \langle \ln |\Lambda| \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\
&- \frac{1}{2} \left(\mathbf{y}^T \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \mathbf{y} - 2\mathbf{y}^T \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \sum_k \lambda_k \langle \mathbf{w}_k \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) \\
&- \frac{1}{2} \left(\sum_k \sum_j \left(\lambda_k \lambda_j \left\langle \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_j \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) + \sum_k \delta_{kj}(\lambda_k - \lambda_k^2) \left\langle \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_k \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) \\
&- \sum_{k=1}^{K} \left(\lambda_k \ln \lambda_k + (1-\lambda_k) \ln(1-\lambda_k) \right) + c \qquad (40)
\end{aligned}
$$

Take derivative w.r.t λ_k

$$\begin{aligned}\frac{\partial \mathcal{F}_\Theta}{\partial \lambda_k} &= \left\langle \ln \frac{\pi_k}{(1-\pi_k)} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} - \ln \frac{\lambda_k}{(1-\lambda_k)} \\ &\quad + \mathbf{y}^T \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \left\langle \mathbf{w}_k \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} - \sum_{j \neq k} \lambda_j \left\langle \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_j \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\ &\quad - \frac{1}{2} \left\langle \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_k \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \end{aligned} \tag{41}$$

Set to zero, we have [4]

$$\begin{aligned}\ln \frac{\lambda_k}{(1-\lambda_k)} &= \left\langle \ln \frac{\pi_k}{(1-\pi_k)} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \\ &\quad + \mathbf{y}^T \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \left\langle \mathbf{w}_k \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} - \sum_{j \neq k} \lambda_j tr \left(\left\langle \mathbf{w}_j \mathbf{w}_k^T \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) \\ &\quad - \frac{1}{2} tr \left(\left\langle \mathbf{w}_k \mathbf{w}_k^T \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right) \end{aligned} \tag{42}$$

where λ_k can be solved using a logistic function. Thus, we can optimize $Q_H(s_k)$ analytically and we can update $Q_H(\mathbf{s})$ by iteratively optimizing λ_k until $\mathcal{F}_\Theta$ converge.

To update λ_k using the equation, we need to calculate the *expected natural parameters*. Since the natural parameters are from decoupled distributions [5] respectively, we can take expectation independently.

As mentioned before, $\pi_k, k = 1, 2, \ldots, K$ is a Bernoulli parameter and $Q_{\boldsymbol{\theta}}(\pi_k)$ is a beta distribution $Q_{\boldsymbol{\theta}}(\pi_k) \sim Beta(\tilde{\alpha}_k, \tilde{\beta}_k)$. Therefore,

$$\begin{aligned}\left\langle \ln \frac{\pi_k}{(1-\pi_k)} \right\rangle &= \int_0^1 \frac{\Gamma(\tilde{\alpha}_k + \tilde{\beta}_k)}{\Gamma(\tilde{\alpha}_k)\Gamma(\tilde{\beta}_k)} \times \\ &\quad \ln \left(\frac{\pi_k}{(1-\pi_k)} \right) \pi_i^{\tilde{\alpha}_k - 1} (1-\pi_k)^{\tilde{\beta}_k - 1} d\pi_k \\ &= \Psi(\tilde{\alpha}_k) - \Psi(\tilde{\beta}_k) \end{aligned} \tag{43}$$

[4] We rearranged $\left\langle \mathbf{w}_k^T \Lambda^{-1} \mathbf{w}_k \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$ to $tr \left(\left\langle \mathbf{w}_k \mathbf{w}_k^T \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \left\langle \Lambda^{-1} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \right)$ so that we can take expectation over $\mathbf{W}$ and Λ^{-1} separately.

[5] See Section (7) for detailed discussion

wher $\Psi(x) = \frac{\partial \ln \Gamma(x)}{\partial x}$ is digamma function. Similarly,

$$\begin{aligned} \langle \ln(1-\pi_k) \rangle &= \int_0^1 \frac{\Gamma(\tilde{\alpha}_k + \tilde{\beta}_k)}{\Gamma(\tilde{\alpha}_k)\Gamma(\tilde{\beta}_k)} \ln(1-\pi_k) \pi_i^{\tilde{\alpha}_k - 1} (1-\pi_k)^{\tilde{\beta}_k - 1} d\pi_k \\ &= \Psi(\tilde{\beta}_k) - \Psi(\tilde{\alpha}_k + \tilde{\beta}_k) \end{aligned} \tag{44}$$

As for $\langle \mathbf{w}_k \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$, we directly plug in the mean $\tilde{m}_{\mathbf{w}}^{(k)}$ of current $Q_{\boldsymbol{\theta}}(\mathbf{w}_k)$, which is a Gaussian distribution. For $\left\langle \mathbf{w}_k \mathbf{w}_k^T \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$

$$\left\langle \mathbf{w}_k \mathbf{w}_k^T \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} = \tilde{m}_{\mathbf{w}}^{(k)} (\tilde{m}_{\mathbf{w}}^{(k)})^T + \boldsymbol{\Sigma}_{\mathbf{w}}^{(k)} \tag{45}$$

where $\tilde{m}_{\mathbf{w}}^{(k)}$ is mean of column k of current $Q(\mathbf{W})$ and $\boldsymbol{\Sigma}_{\mathbf{w}}^{(k)}$ is the covariance matrix for the column. Similarly $\left\langle \mathbf{w}_j \mathbf{w}_k^T \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$ can be calculated as

$$\left\langle \mathbf{w}_j \mathbf{w}_k^T \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} = \langle \mathbf{w}_j \rangle \left\langle \mathbf{w}_k^T \right\rangle = \tilde{m}_{\mathbf{w}}^{(j)} \left(\tilde{m}_{\mathbf{w}}^{(k)} \right)^T + \boldsymbol{\Sigma}_{\boldsymbol{w}}^{(jk)} \tag{46}$$

For $\langle \Lambda^{-1} \rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}$, we can plug in the mean of $Q_{\boldsymbol{\theta}}(\tau)$ (see equation 59) as

$$\left\langle \Lambda^{-1} \right\rangle = \langle \tau \rangle I \tag{47}$$

7 VBM

7.1 Variational Approximation of Parameter Distributions

As discussed in section (3), the goal of variational Bayesian learning is to lower bound the marginal log likelihood of data $\mathcal{L}(\mathbf{Y})$. We recover $\mathcal{L}(\mathbf{Y})$ by minimizing the KL divergence of true posterior distribution for parameters $P(\boldsymbol{\theta}|\mathbf{Y})$ and approximation distribution $Q(\boldsymbol{\theta})$. However, to calculate the KL divergence requires evaluation of true posterior distribution of parameters which is infeasible. Therefore, we need to rewrite the KL divergence as

$$\begin{aligned} KL(Q(\boldsymbol{\theta})||P(\boldsymbol{\theta}|\mathbf{Y})) &= \int Q(\boldsymbol{\theta}) \ln \frac{Q(\boldsymbol{\theta})}{P(\boldsymbol{\theta}|\mathbf{Y})} \\ &= \int Q(\boldsymbol{\theta}) \ln \frac{Q(\boldsymbol{\theta})P(\mathbf{Y})}{P(\boldsymbol{\theta},\mathbf{Y})} \\ &= \int Q(\boldsymbol{\theta}) \ln \frac{Q(\boldsymbol{\theta})}{P(\boldsymbol{\theta},\mathbf{Y})} + \ln P(\mathbf{Y}) \geq 0 \end{aligned} \tag{48}$$

Since $\ln P(\mathbf{Y})$ is a constant, thus minimizing $KL(Q(\boldsymbol{\theta})||P(\boldsymbol{\theta}|\mathbf{Y}))$ divergence is equivalent to minimizing the divergence between distribution $Q(\boldsymbol{\theta})$ and joint distribution $P(\boldsymbol{\theta},\mathbf{Y})$ until the difference equals $-\ln P(\mathbf{Y})$, at which point $KL(Q||P) = 0$. As demonstrated in section (4), we can minimize the KL divergence between the two distribution by adopting mean field approximation. More specifically, we restrict $Q(\boldsymbol{\theta})$ to be following form

$$Q(\boldsymbol{\theta}) = Q(\boldsymbol{\pi}, \boldsymbol{W}, \gamma, \boldsymbol{\Lambda}) \tag{49}$$

$$= Q(\boldsymbol{\pi})Q(\boldsymbol{W})Q(\gamma)Q(\boldsymbol{\Lambda}) \tag{50}$$

Then we can achieve overall minimization of KL by iteratively minimizing KL divergence with respect to factorial distributions $Q(\theta_i)$. The optimal $Q(\theta_i)$ is of form

$$Q(\theta_i) = \frac{exp\left\{\langle \ln P(\boldsymbol{\theta},\mathbf{Y})\rangle_{Q(\theta_{j\neq i})Q(\mathbf{s})}\right\}}{\int exp\left\{\langle \ln P(\boldsymbol{\theta},\mathbf{Y})\rangle_{Q(\theta_{j\neq i})Q(\mathbf{s})}\right\} d\theta_i} \tag{51}$$

where the complete likelihood can be analytically solved. In our case, we need further include latent variables into consideration. After VBE step, we already updated $Q(\mathbf{s})$ and obtained the expectation of instantiation of latent variables $\mathbf{S} =$

$\{\langle \mathbf{s}^n \rangle ; n = 1, 2, \ldots, N\}$ for the data points w.r.t distribution $Q(\mathbf{s})$. This enables us to plug in these values into the complete likelihood in equation (51) to derive following results:

$$\begin{aligned} P(\mathbf{Y}, \mathbf{S}, \boldsymbol{\theta}) &= P(\mathbf{Y}|\mathbf{S}, \boldsymbol{\theta}) P(\mathbf{S}|\theta) P(\boldsymbol{\theta}) \\ &= P(\mathbf{Y}|\mathbf{S}, \mathbf{W}, \boldsymbol{\Lambda}) P(\mathbf{S}|\pi) P(\boldsymbol{\pi}) P(\mathbf{W}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}) P(\boldsymbol{\Lambda}) \end{aligned} \tag{52}$$

take expectation of log complete likelihood w.r.t $Q(\boldsymbol{\theta})$

$$\langle \ln \{P(\mathbf{Y}|\mathbf{S}, \mathbf{W}, \boldsymbol{\Lambda}) P(\mathbf{S}|\pi) P(\boldsymbol{\pi}) P(\mathbf{W}|\boldsymbol{\gamma}) P(\boldsymbol{\gamma}) P(\boldsymbol{\Lambda})\} \rangle_{Q_\theta(\theta)} \tag{53}$$

$$\begin{aligned} = & \int Q(\boldsymbol{\pi}) Q(\mathbf{W}) Q(\gamma) Q(\boldsymbol{\Lambda}) \ln \left[\prod_{n=1}^{N} P(\mathbf{y}^n | \langle s^n \rangle, \mathbf{W}, \boldsymbol{\Lambda}) \right] d\boldsymbol{\pi} d\mathbf{W} d\gamma d\boldsymbol{\Lambda} \\ & + \int Q(\mathbf{s}) \ln P(\mathbf{S}|\pi) d\mathbf{S} \\ & + \int Q(\boldsymbol{\pi}) \ln P(\boldsymbol{\pi}) d\boldsymbol{\pi} \\ & + \int Q(\mathbf{W}) Q(\gamma) \ln P(\mathbf{W}|\boldsymbol{\gamma}) d\mathbf{W} \\ & + \int Q(\gamma) \ln P(\gamma) d\gamma \\ & + \int Q(\boldsymbol{\Lambda}) \ln P(\boldsymbol{\Lambda}) d\boldsymbol{\Lambda} \end{aligned} \tag{54}$$

where n is indicator for data point.

With this form, it become obviouse that, when minimizing $KL(Q(\boldsymbol{\theta}||P(\mathbf{Y}, \boldsymbol{\theta}))$ using equation (51) with respect to individual parameter distribution $Q(\theta_i)$, many terms unrelated to $Q(\theta_i)$ become irrelevant during normalization process and can be dropped. Eventually, the optimal distribution will look like

$$Q(\theta_i) \propto exp \left\{ \left\langle \ln \left[\prod_{n=1}^{N} P(y^n | \langle s^n \rangle, \mathbf{W}, \boldsymbol{\Lambda}) \right] \right\rangle_{Q(\theta_{j \neq i})} + \ln P(\theta_i) \right\} \tag{55}$$

which is an approximate posterior distribution $P(\theta_i|\mathbf{Y})$.

The factorial distributions is of following form

$$Q(\boldsymbol{\pi}) = \prod_{k=1}^{K} Beta(\pi_k|\widetilde{\alpha}_k, \widetilde{\beta}_k) \tag{56}$$

$$Q(\mathbf{W}) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{w}_d|\widetilde{\mathbf{m}}_{\mathbf{w}}^{(d)}, \widetilde{\boldsymbol{\Sigma}}_{\mathbf{w}}^{(d)}) \tag{57}$$

$$Q(\boldsymbol{\gamma}) = \prod_{k=1}^{K} \mathcal{G}(\gamma_k|\widetilde{a}_{\gamma}^{(k)}, \widetilde{b}_{\gamma}^{(k)}) \tag{58}$$

$$Q(\tau) = \mathcal{G}(\tau|\widetilde{c}_{\boldsymbol{\tau}}, \widetilde{d}_{\boldsymbol{\tau}}) \tag{59}$$

Since all of the factorized distributions $Q(\theta_i)$ and their prior distributions belong to exponential family, we can take advantages of characteristics of exponential family distributions during updating the approximate posterior distributions. If the prior distributions for $Q_\theta(\theta)$ for our model belong to exponential family and can be written as

$$P(\theta|\eta, \nu) = h(\boldsymbol{\eta}, \boldsymbol{\nu})g(\boldsymbol{\theta})^{\eta} exp\left\{\phi(\boldsymbol{\theta})^T \boldsymbol{\nu}\right\} \tag{60}$$

where η and ν are hyper-parameters. Furthermore, the complete data log likelihood is of the form

$$P(\boldsymbol{y}, \boldsymbol{s}|\theta) = f(\boldsymbol{y}, \boldsymbol{s})g(\theta)exp\left\{\phi(\boldsymbol{\theta})u(\boldsymbol{s}, \boldsymbol{y})\right\} \tag{61}$$

where f, g and u are functions defined in exponential family and ϕ are a vector of *natural parameters*. According to Ghahramani and Beal (Ghahramani and Beal, 2000a), at the maxima of $\mathcal{F}(Q_H(H), Q_\theta(\theta), \boldsymbol{y})$, $Q_\theta(\theta)$ assume following form

$$Q_\theta(\boldsymbol{\theta}) = h(\widetilde{\eta}, \widetilde{\boldsymbol{\nu}})g(\boldsymbol{\theta})^{\widetilde{\eta}} exp\left\{\phi(\boldsymbol{\theta})^T \widetilde{\boldsymbol{\nu}}\right\} \tag{62}$$

where $\widetilde{\eta} = \eta + n$, $\widetilde{\boldsymbol{\nu}} = \boldsymbol{\nu} + \sum_{i=1}^{n} \overline{u}(\boldsymbol{y}_i)$ and $\overline{u}(\boldsymbol{y}_i) = E_{Q_H(\boldsymbol{s})}(u(s_i, \boldsymbol{y}_i))$. In another word, we use expected sufficient statistics under the $Q_H(\boldsymbol{s})$ to update the approximated posterior distribution $Q_\theta(\boldsymbol{\theta})$.

7.2 Update $Q(\pi)$

Once the we obtain expected instantiation of latent variables, it is straight forward to update $Q(\pi)$ in that $\mathbf{s}$ d-sparate π from other variables. To update the each of $Q_\theta(\pi_k)$ which is a beta distribution, we can use expected sufficient statistics $\langle s_k \rangle_{Q_H(\mathbf{s})}$ as calculated in section 6 combining with prior $Beta_k(\alpha_k, \beta_k)$ to update posterior

$$Q_\theta(\pi_k) \sim Beta\left(\alpha_k + \sum_{n=1}^{N} \langle s_k^n \rangle_{Q_H(\mathbf{s})}, \beta_k + N - \sum_{n=1}^{N} \langle s_k^n \rangle_{Q_H(\mathbf{s})}\right) \quad (63)$$

$$\sim Beta\left(\alpha_k + \sum_{n=1}^{N} \lambda_k^n, \beta + N - \sum_{n=1}^{N} \lambda_k^n\right) \quad (64)$$

where $\langle s_k^n \rangle_{Q_H(\mathbf{s})} = \lambda_k^n$ is expectation of source $s_k = 1$ for data point n.

7.3 Update $Q(\mathbf{W})$ and $Q(\gamma)$

To update the $Q(\mathbf{W})$, we have

$$Q(\mathbf{W}) \propto exp\left\{\langle \ln P(\mathbf{Y}|\mathbf{S},\mathbf{W},\Lambda)\rangle_{Q(\theta_{j\neq\mathbf{W}})} + < \ln P(\mathbf{W}|\gamma) >_{Q(\gamma)}\right\} \tag{65}$$

where $Q(\theta_{j\neq\mathbf{W}})$ is the joint distribution of all parameters except $\mathbf{W}$; the $\mathbf{W}$ is a $D \times K$ weight matrix with each column $\mathbf{w_k}$ being a D dimensional weight output by source s_k; the $\mathbf{s}$ is K dimensional column vector of sources; the $\Lambda = \tau^{-1}I$ is a diagonal covariance matrix for $\mathbf{y}$ and τ inverse of variance (precision) of observation.

7.3.1

Since the Λ is diagonal, each component of $\mathbf{y}$ is independent of others. Thus, we can write $\ln P(\mathbf{Y}|\mathbf{S},\mathbf{W},\Lambda)$ as

$$\begin{aligned}
\ln P(\mathbf{Y}|\mathbf{S},\mathbf{W},\Lambda) &= \ln\left[\prod_{n=1}^{N}\prod_{d=1}^{D} P(y_d^n|\mathbf{s}^n,\mathbf{W},\tau)\right] && (66)\\
&= \sum_{d=1}^{D}\sum_{n=1}^{N} \ln P(y_d^n|\mathbf{s}^n,\mathbf{W},\tau) \\
&\propto \frac{1}{2}ND\ln\tau + \sum_{d=1}^{D}\sum_{n=1}^{N}\left[-\frac{1}{2}\tau\left(y_d^n - \sum_{k=1}^{K} s_k^n w_{dk}\right)^2\right] && (67)
\end{aligned}$$

Define $\mathbf{w}_d$ as a column vector corresponding to the dth row of weight matrix $\mathbf{W}$ and $\mathbf{s}^n$ be the vector of sources for the data point n. Then, we can rewrite $\sum_{k=1}^{K} s_k^n w_{dk} = \mathbf{s}^{nT}\mathbf{w}_d$. Then equation (67) is as

$$\frac{1}{2}ND\ln\tau + \sum_{d=1}^{D}\sum_{n=1}^{N}\left[-\frac{1}{2}\tau\left(y_d^n - \mathbf{s}^{nT}\mathbf{w}_d\right)^2\right] \tag{68}$$

7.3.2

As discussed in section (5), the prior $P(\mathbf{W}|\gamma)$ is of the following form:

$$P(\mathbf{W}|\boldsymbol{\gamma}) = \prod_{k=1}^{K} P(\mathbf{w}_k|\gamma_k) \tag{69}$$

where $\mathbf{w}_k$ is column of the weight matrix.

Our model further defines that covariance matrix for each column of $\mathbf{W}$ is of $\Sigma_w^k = \gamma_k^{-1} I$, as indicated by equation (28). Then, each component of $\mathbf{w}_k$ is independent of others and the prior $P(\mathbf{W})$ can be fully factorized. If we define $\mathbf{w}_d$ as a K dimensional column vector corresponding to dth row of the $\mathbf{W}$, it's covariance matrix will be of the form $\boldsymbol{\Sigma}_w^{(d)} = \boldsymbol{\gamma}^{-1}\mathbf{I}$, of which the diagonal components are $\gamma_1^{-1}, \gamma_2^{-1}, \ldots, \gamma_K^{-1}$. Here the γ_k is the inverse of variance of kth column of $\mathbf{W}$. Then, the log of prior $\ln P(\mathbf{W})$ can be rewritten as

$$\begin{aligned}
\ln P(\mathbf{W}) &= \ln \left[\prod_{d=1}^{D} P(\mathbf{w}_d|\boldsymbol{\Sigma}_w^{(d)}) \right] && (70) \\
&= \sum_{d=1}^{D} \ln P(\mathbf{w}_d|\boldsymbol{\Sigma}_w^{(d)}) \\
&\propto \sum_{d=1}^{D} \left[-\frac{1}{2} \ln |\boldsymbol{\Sigma}_w^{(d)}| - \frac{1}{2} \mathbf{w}_d^T \boldsymbol{\Sigma}_w^{(d)-1} \mathbf{w}_d \right] && (71)
\end{aligned}$$

Combining equations (67) and (71), we can see that both parameter prior and data likelihood are Gaussian distributions. It naturally follows that we recover the $Q(\mathbf{W})$ as Gaussian posterior distribution. Since likelihood function (67) is a linear regression function, updating $Q(\mathbf{W})$ turns out to be well defined problem of Bayesian learning for parameters of linear regression model (Box and Tiao, 1973; Tanner, 1996). Rewrite equation (65) as

$$\ln Q(\mathbf{W}) \propto \sum_{d=1}^{D} \left[\sum_{n=1}^{N} \left\{ -\frac{1}{2}\tau \left(y_d^n - \mathbf{s}^{nT} \mathbf{w}_d \right)^2 \right\} - \frac{1}{2} \mathbf{w}_d^T \boldsymbol{\Sigma}_w^{(d)-1} \mathbf{w}_d \right] \tag{72}$$

Now it is obvious that we can update $Q(\mathbf{W})$ row-wisely, because each component of observed data $\mathbf{y}_d$ is linear regression with independent variables $\mathbf{s}$ and

parameters $\mathbf{w}_d$. We have

$$
\begin{aligned}
& -\sum_{n=1}^{N} \frac{1}{2}\tau \left(y_d^n - \mathbf{s}^{nT}\mathbf{w}_d\right)\left(y_d^n - \mathbf{s}^{nT}\mathbf{w}_d\right) - \frac{1}{2}\mathbf{w}_d^T \boldsymbol{\Sigma}_w^{(d)-1}\mathbf{w}_d \\
= & -\sum_{n=1}^{N} \frac{1}{2}\tau \left\{(y_d^n - \mathbf{s}^{nT}\widehat{\mathbf{w}}_d) + \mathbf{s}^{nT}(\widehat{\mathbf{w}}_d - \mathbf{w}_d)\right\}\left\{(y_d^n - \mathbf{s}^{nT}\widehat{\mathbf{w}}_d) + \mathbf{s}^{nT}(\widehat{\mathbf{w}}_d - \mathbf{w}_d)\right\} \\
& -\frac{1}{2}\mathbf{w}_d^T \boldsymbol{\Sigma}_w^{(d)-1}\mathbf{w}_d \qquad (73) \\
= & -\sum_{n=1}^{N} \frac{1}{2}\tau \left(y_d^n - \mathbf{s}^{nT}\widehat{\mathbf{w}}_d\right)^2 \\
& -\frac{1}{2}\left(\mathbf{w}_d - \widehat{\mathbf{w}}_d\right)^T \tau \sum_{n=1}^{N} \mathbf{s}^n\mathbf{s}^{nT}\left(\mathbf{w}_d - \widehat{\mathbf{w}}_d\right) - \frac{1}{2}\mathbf{w}_d^T \boldsymbol{\Sigma}_w^{(d)-1}\mathbf{w}_d \qquad (74)
\end{aligned}
$$

where $\widehat{\mathbf{w}}_d = \left(\sum_{n=1}^{N} \mathbf{s}^n\mathbf{s}^{nT}\right)^{-1} \sum_{n=1}^{N} \mathbf{s}^n y_d^n$ is the least-squares estimation of $\mathbf{w}_d$. Note that $(y_d^n - \mathbf{s}^{nT}\widehat{\mathbf{w}}_d)\mathbf{s}^{nT}(\widehat{\mathbf{w}}_d - \mathbf{w}_d) = 0 = \mathbf{s}^{nT}(\widehat{\mathbf{w}}_d - \mathbf{w}_d)(y_d^n - \mathbf{s}^{nT}\widehat{\mathbf{w}}_d)$. To derive the posterior distribution, we only need to include the terms containing $\mathbf{w}_d$, then

$$
\begin{aligned}
& \left\langle -\frac{1}{2}\left(\left(\mathbf{w}_d - \widehat{\mathbf{w}}_d\right)^T \tau \sum_{n=1}^{N} \mathbf{s}^n\mathbf{s}^{nT}\left(\mathbf{w}_d - \widehat{\mathbf{w}}_d\right) + \mathbf{w}_d^T \boldsymbol{\Sigma}_w^{(d)-1}\mathbf{w}_d\right)\right\rangle_{Q(\theta_{j\neq w})Q(\mathbf{s})} \\
= & -\frac{1}{2}\left(\mathbf{w}_d - \tilde{\mathbf{m}}_w^{(d)}\right)^T \widetilde{\Sigma}_w^{(d)-1}\left(\mathbf{w}_d - \tilde{\mathbf{m}}_w^{(d)}\right) \qquad (75)
\end{aligned}
$$

where

$$
\begin{aligned}
\widetilde{\boldsymbol{\Sigma}}_w^{(d)} &= \left(\boldsymbol{\Sigma}_w^{(d)-1} + \langle\tau\rangle \sum_{n=1}^{N} \left\langle \mathbf{s}^n\mathbf{s}^{nT}\right\rangle\right)^{-1} \\
&= \left(diag(\langle\boldsymbol{\gamma}\rangle) + \langle\tau\rangle \sum_{n=1}^{N} \left\langle \mathbf{s}^n\mathbf{s}^{nT}\right\rangle\right)^{-1} \qquad (76) \\
\tilde{\mathbf{m}}_w^{(d)} &= \widetilde{\boldsymbol{\Sigma}}_w^{(d)} \langle\tau\rangle \sum_{n=1}^{N} \langle\mathbf{s}^n\rangle y_d^n \qquad (77)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
Q(\mathbf{W}) &= \prod_{d=1}^{D} Q\left(\mathbf{w}_d | \widetilde{\mathbf{\Sigma}}_w^{(d)}\right) && (78)\\
&= \prod_{d=1}^{D} \mathcal{N}\left(\tilde{\mathbf{m}}_w^{(d)}, \widetilde{\mathbf{\Sigma}}_w^{(d)}\right) && (79)
\end{aligned}
$$

Once we update $Q(\mathbf{W})$, it is straight forward to update $Q(\gamma)$ because $\mathbf{W}$ d-separate γ from other variables. Recall that $\gamma = [\gamma_1, \gamma_2, \ldots, \gamma_K]^T$, then the posterior for individual element γ_k is as following

$$
\begin{aligned}
Q(\gamma_k) &\propto exp\left\{ \langle \ln P(\mathbf{w}_k|\gamma_k)\rangle + \ln P(\gamma_k) \right\} && (80)\\
&= \gamma_k^{(a_\gamma - 1)} exp\{-b_\gamma \gamma_k\} \gamma_k^{D/2} exp\left\{ -\frac{\gamma_k}{2} \mathbf{w}_k^T \mathbf{w}_k \right\}\\
&= \gamma_k^{\left(a_\gamma + \frac{D}{2} - 1\right)} exp\left\{ -\left(b_\gamma + \frac{||\mathbf{w}_k||^2}{2} \right) \gamma_k \right\} && (81)\\
&= \mathcal{G}(\gamma_k | \tilde{a}_{\gamma k}, \tilde{b}_{\gamma k})
\end{aligned}
$$

then

$$
\begin{aligned}
Q(\boldsymbol{\gamma}) &= \prod_{k=1}^{K} \mathcal{G}(\gamma_k | \tilde{a}_{\gamma k}, \tilde{b}_{\gamma k}) && (82)\\
\tilde{a}_{\gamma k} &= a_{\gamma k} + \frac{D}{2} && (83)\\
\tilde{b}_{\gamma k} &= b_{\gamma k} + \frac{\langle ||\mathbf{w}_k||^2 \rangle}{2} && (84)
\end{aligned}
$$

7.4 Update $Q(\tau)$

Once we have estimated $Q(\mathbf{W})$, we can update $Q(\tau)$.

$$Q(\tau) \propto exp\left\{\ln P(\tau) + \left\langle \ln \prod_{n=1}^{N} P(\mathbf{y}^n|\mathbf{s}^n, \mathbf{W}, \tau)\right\rangle_{Q(\theta j \neq \tau)Q(\mathbf{s})}\right\} \tag{85}$$

inside the curly bracket and omit expectation temporarily

$$\begin{aligned}
& (c_\tau - 1)\ln\tau - d_\tau\tau + \sum_{n=1}^{N}\left(\frac{D}{2}\ln\tau - \frac{\tau}{2}(\mathbf{y}^n - \mathbf{W}\mathbf{s}^n)^T(\mathbf{y}^n - \mathbf{W}\mathbf{s}^n)\right) \\
= & \left[\left(c_\tau + \frac{ND}{2} - 1\right)\ln\tau\right] \\
& - \left[\tau\left\{d_\tau + \frac{1}{2}\sum_{n=1}^{N}(\mathbf{y}^n - \mathbf{W}\mathbf{s}^n)^T(\mathbf{y}^n - \mathbf{W}\mathbf{s}^n)\right\}\right] \qquad (86) \\
= & \left[\left(c_\tau + \frac{ND}{2} - 1\right)\ln\tau\right] \\
& - \left[\tau\left\{d_\tau + \frac{1}{2}\sum_{n=1}^{N}(||\mathbf{y}^n||^2 - 2\mathbf{y}^T\mathbf{W}\mathbf{s}^n + tr\left(\mathbf{W}^T\mathbf{W}\mathbf{s}^n\mathbf{s}^{nT}\right)\right\}\right] \qquad (87)
\end{aligned}$$

We can see the kernel of a gamma distribution in equation (87). The first bracket is shape parameter and second bracket contains shape parameter. After taking expectation w.r.t $Q(\theta j \neq \tau)Q(\mathbf{s})$, we have

$$Q(\tau) = \mathcal{G}(\tau|\tilde{c}_\tau, \tilde{d}_\tau) \tag{88}$$

where

$$\tilde{c}_\tau = c_\tau + \frac{ND}{2} \tag{89}$$

$$\tilde{d}_\tau = d_\tau + \frac{1}{2}\sum_{n=1}^{N}\left\{||\mathbf{y}^n||^2 - 2\mathbf{y}^{nT}\left\langle\mathbf{W}\right\rangle\left\langle\mathbf{s}^n\right\rangle + tr\left(\left\langle\mathbf{W}^T\mathbf{W}\right\rangle\left\langle\mathbf{s}^n\mathbf{s}^{nT}\right\rangle\right)\right\} \tag{90}$$

8 Ensemble Learning

In previous two sections we have described methods to update different distributions. In this section, we will address sequence of updating distributions, monitoring convergence and dimension reduction during variational Bayesian learning process.

8.1 Sequential Update distributions

Apparently, the distributions of latent variables and parameters are coupled and we need to update them in a right sequences. We should use following sequence:

1. $Q(\mathbf{s})$
2. $Q(\boldsymbol{\pi})$
3. $Q(\boldsymbol{W})$
4. $Q(\boldsymbol{\gamma})$
5. $Q(\tau)$

8.2 Monitoring Convergence

In variational Bayesian learning, we maximize $\mathcal{F}(Q(\mathbf{H}), Q(\boldsymbol{\theta}), \mathbf{Y})$, which is the lower bound of log marginal likelihood, at each iteration with respect to different distributions. Therefore, the $\mathcal{F}(Q(\mathbf{H}), Q(\boldsymbol{\theta}), \mathbf{Y})$ should monotonically increase until converge to a local maxima. One advantage of VB learning is that we can monitor the process of converging. Here we can rewrite the equation (15) as

$$\mathcal{F}(Q(\mathbf{H}), Q(\boldsymbol{\theta}), \mathbf{Y}) = \left\langle \sum_{\mathbf{H}} Q_{\mathbf{H}}(\mathbf{H}) \ln \frac{P(\mathbf{Y}, \mathbf{H}|\boldsymbol{\theta})}{Q_{\mathbf{H}}(\mathbf{H})} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} - \left\langle \ln \frac{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{P(\boldsymbol{\theta})} \right\rangle_{Q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} \tag{91}$$

$$\begin{aligned} &= \langle \ln(\mathbf{Y}|\mathbf{S}, \mathbf{W}, \Lambda) \rangle + \langle \ln P(\mathbf{S}|\boldsymbol{\pi}) \rangle - \langle \ln(Q(\mathbf{S}) \rangle \\ &\quad + \langle \ln P(\boldsymbol{\pi}) \rangle - \langle \ln Q(\boldsymbol{\pi}) \rangle \\ &\quad + \langle \ln P(\mathbf{W}) \rangle - \langle \ln Q(\mathbf{W}) \rangle \\ &\quad + \langle \ln P(\boldsymbol{\gamma}) \rangle - \langle \ln Q(\boldsymbol{\gamma}) \rangle \\ &\quad + \langle \ln P(\tau) \rangle - \langle \ln Q(\tau) \rangle \end{aligned} \tag{92}$$

where

$$\begin{aligned}\langle \ln(\mathbf{Y}|\mathbf{S},\mathbf{W},\Lambda)\rangle &= -\frac{ND}{2}\ln(2\pi)+\frac{ND}{2}\langle\ln\tau\rangle-\frac{\tau}{2}\sum_{n=1}^{N}\Big\{\mathbf{y}^{nT}\mathbf{y}^{n}-2\mathbf{y}^{nT}\langle\mathbf{W}\rangle\langle\mathbf{s}^{n}\rangle \\ &\quad +tr\left(\left\langle\mathbf{W}^{T}\mathbf{W}\right\rangle\left\langle\mathbf{s}^{n}\mathbf{s}^{nT}\right\rangle\right)\Big\} \qquad (93)\end{aligned}$$

$$\langle \ln P(\mathbf{S}|\boldsymbol{\pi})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\left\{\left\langle\ln\frac{\pi_k}{1-\pi_k}\right\rangle\langle s_k^n\rangle+\langle\ln(1-\pi_k)\rangle\right\} \qquad (94)$$

$$\langle \ln Q(\mathbf{S})\rangle = \sum_{n=1}^{N}\sum_{k=1}^{K}\left(\lambda_k\ln\lambda_k+(1-\lambda_k)\ln(1-\lambda_k)\right) \qquad (95)$$

$$\begin{aligned}\langle \ln P(\boldsymbol{\pi})\rangle &= \sum_{k=1}^{K}\left\langle\ln\left\{\frac{\Gamma(\alpha_k+\beta_k)}{\Gamma(\alpha_k)\Gamma(\beta_k)}\pi_k^{\alpha_k-1}(1-\pi_k)^{\beta_k-1}\right\}\right\rangle \\ &= \sum_{k=1}^{K}\Big\{\ln\Gamma(\alpha_k+\beta_k)-\ln\Gamma(\alpha_k)-\ln\Gamma(\beta_k) \\ &\quad +(\alpha_k-1)\langle\ln\pi_k\rangle+(\beta_k-1)\langle\ln(1-\pi_k)\rangle\Big\} \qquad (96)\end{aligned}$$

$$\begin{aligned}\langle \ln Q(\boldsymbol{\pi})\rangle &= \sum_{k=1}^{K}\left\langle\ln\left\{\frac{\Gamma(\tilde{\alpha_k}+\tilde{\beta}_k)}{\Gamma(\tilde{\alpha_k})\Gamma(\tilde{\beta}_k)}\pi_k^{\tilde{\alpha_k}-1}(1-\pi_k)^{\tilde{\beta}_k-1}\right\}\right\rangle \\ &= \sum_{k=1}^{K}\Big\{\ln\Gamma(\tilde{\alpha_k}+\tilde{\beta}_k)-\ln\Gamma(\tilde{\alpha_k})-\ln\Gamma(\tilde{\beta}_k) \\ &\quad +(\tilde{\alpha_k}-1)\langle\ln\pi_k\rangle+(\tilde{\beta}_k-1)\langle\ln(1-\pi_k)\rangle\Big\} \qquad (97)\end{aligned}$$

$$\begin{aligned}\langle \ln P(\mathbf{W})\rangle &= \sum_{k=1}^{K}\left\langle-\frac{D}{2}\ln(2\pi)+\frac{D}{2}\ln\gamma_k-\frac{\gamma_k}{2}\mathbf{w}_k^T\mathbf{w}_k\right\rangle \\ &= \sum_{k=1}^{K}\left\{-\frac{D}{2}\ln(2\pi)+\frac{D}{2}\langle\ln\gamma_k\rangle-\frac{\langle\gamma_k\rangle}{2}\left\langle\mathbf{w}_k^T\mathbf{w}_k\right\rangle\right\} \qquad (98)\end{aligned}$$

$$
\begin{aligned}
\langle \ln Q(\mathbf{W}) \rangle &= \sum_{d=1}^{D} \left\langle -\frac{K}{2}\ln(2\pi) - \frac{1}{2}\ln|\tilde{\mathbf{\Sigma}}_w^{(d)}| - \frac{1}{2}(\mathbf{w}_d - \tilde{\mathbf{m}}_w^{(d)})^T \tilde{\mathbf{\Sigma}}_w^{(d)-1}(\mathbf{w}_d - \tilde{\mathbf{m}}_w^{(d)}) \right\rangle \\
&= \sum_{d=1}^{D} \Bigg\{ -\frac{K}{2}\ln(2\pi) - \frac{1}{2}\ln|\tilde{\mathbf{\Sigma}}_w^{(d)}| - \frac{1}{2} tr\left(\left\langle \mathbf{w}_d \mathbf{w}_d^T \right\rangle \tilde{\mathbf{\Sigma}}_w^{(d)-1} \right) \\
&\quad + tr\left(\tilde{\mathbf{m}}_w^{(d)} \left\langle \mathbf{w}_d^T \right\rangle \tilde{\mathbf{\Sigma}}_w^{(d)-1} \right) - \frac{1}{2} tr\left(\tilde{\mathbf{m}}_w^{(d)} \tilde{\mathbf{m}}_w^{(d)T} \tilde{\mathbf{\Sigma}}_w^{(d)-1} \right) \Bigg\} \\
&= -\frac{DK}{2}\ln(2\pi) - \frac{D}{2}\ln|\tilde{\mathbf{\Sigma}}_w^{(d)}| \qquad (99)
\end{aligned}
$$

$$
\begin{aligned}
\langle \ln P(\boldsymbol{\gamma}) \rangle &= \sum_{k=1}^{K} \left\langle a_\gamma \ln b_\gamma + (a_\gamma - 1)\ln\gamma_k - b_\gamma \gamma_k - \ln\Gamma(a_\gamma) \right\rangle \\
&= \sum_{k=1}^{K} \Bigg\{ a_\gamma \ln b_\gamma + (a_\gamma - 1)\langle \ln\gamma_k \rangle \\
&\quad - b_\gamma \langle \gamma_k \rangle - \ln\Gamma(a_\gamma) \Bigg\} \qquad (100)
\end{aligned}
$$

$$
\begin{aligned}
\langle \ln Q(\boldsymbol{\gamma}) \rangle &= \sum_{k=1}^{K} \left\langle \tilde{a}_\gamma^{(k)} \ln \tilde{b}_\gamma^{(k)} + (\tilde{a}_\gamma^{(k)} - 1)\ln\gamma_k - \tilde{a}_\gamma^{(k)}\gamma_k - \ln\Gamma(\tilde{a}_\gamma) \right\rangle \\
&= \sum_{k=1}^{K} \Bigg\{ \tilde{a}_\gamma^{(k)} \ln \tilde{b}_\gamma^{(k)} + (\tilde{a}_\gamma^{(k)} - 1)\langle \ln\gamma_k \rangle \\
&\quad - \tilde{b}_\gamma^{(k)} \langle \gamma_k \rangle - \ln\Gamma(\tilde{a}_\gamma^{(k)}) \Bigg\} \qquad (101)
\end{aligned}
$$

$$
\begin{aligned}
\langle \ln P(\tau) \rangle &= \left\langle c_\tau \ln d_\tau + (c_\tau - 1)\ln\tau - d_\tau \tau - \ln\Gamma(c_\tau) \right\rangle \\
&= c_\tau \ln d_\tau + (c_\tau - 1)\langle \ln\tau \rangle - d_\tau \langle \tau \rangle - \ln\Gamma(c_\tau) \qquad (102)
\end{aligned}
$$

$$
\begin{aligned}
\langle \ln Q(\tau) \rangle &= \left\langle \tilde{c}_\tau \ln \tilde{d}_\tau + (\tilde{c}_\tau - 1) \ln \tau - \tilde{d}_\tau \tau - \ln \Gamma(\tilde{c}_\tau) \right\rangle \\
&= \tilde{c}_\tau \ln \tilde{d}_\tau + (\tilde{c}_\tau - 1) \langle \ln \tau \rangle - \tilde{d}_\tau \langle \tau \rangle - \ln \Gamma(\tilde{c}_\tau) \qquad (103)
\end{aligned}
$$

All expectations are taken w.r.t. correpsonding variational posterior distributions. Note that $\langle \ln x \rangle = \Psi(a) - \ln(b)$, where $x \sim \mathcal{G}(x|a, b)$.

8.3 Model Selection and Dimension Reduction

We have discussed in the section (3) that we should choose a model M_i that has the highest poterior probability among the candidate models. Bayesian model selection automatically embodies *Occam's Razor* in that complex models is penalized by assigning lower posterior probability (MacKay, 1995). When selecting models according to Bayesian approach, in our case is the number of latent variables, we will automatically recover the model with minimum number of latent variables that explains data well. One way is to select models according to equation (4) by discretely testing models with different number of latent variables K up to a maximum number, e.g. the number of dimension $D - 1$. However, such discrete search is quite computational expensive. To avoid such search, MacKay and Neal introduced the concept of automatic relevance determination (ARD) that can be used for Bayesian model selection. The technique has been successfully applied to determine the number of latent variables for different models in several recent researches (Lawrence and Bishop, 2000; Bishop, 1999; Ghahramani and Beal, 2000b).

We will explain the basic idea of ARD applied to our model using a hypothesized example. Suppose we start with a large number of latent variables, say $D - 1$, and begin to maximize the $\mathcal{F}(Q(\mathbf{H}), Q(\boldsymbol{\theta}), \mathbf{Y})$. If not all latent variables are needed to explain data, we will find the the column of the posterior distribution of weight matrix corresponding to the unused source will shrink toward the mean of prior of column which is zero. Smaller weight in the column of loading matrix will leads to peaking of corresponding posterior distribution of hyper-parameter γ_i at larger value, indicating $\mathbf{w}_i$ is almost invariant. The latter will cause the weight to further shrink toward prior mean during next iteration of updating loading matrix. Thus the weight of the whole column will quickly deminish toward 0 which effectively shut down the source.

Overall, ARD approach provide another advantage for the model selection - avoid discrete search for the models. In some case, the saving may be tremendous. For example, for a mixture model with K potential components, discretely

selecting model from the space of all possible combination of components will be prohibitively expensive. Variational Bayesian approach enables us to use ARD technique to achieve the goal of dimension reduction and model selection.

9 Summary

In this research, we have developed a variational Bayesian learning algorithm for a latent variable model - cooperative vector quantizer model. Although the main motivation in this research is using the model to simulate the biological pathways, the model itself has wide application in machine learning and signal processing, e.g. image separation and blind source separation. As demonstrated by previous researches (Ghahramani, 1995; Dunmur and Titterington, 1997; Miskin, 2000), the model is capable of identify the weight matrix uniquely. Thus, we can use the model to estimate what genes are regulated by a given signal transduction component by studying the weight matrix of the model. Combining such information with biological knowledge, we can potentially map the latent variables to biological entities. Especially, the taxonomic knowledge regarding proteins from the Gene Ontology database will be very useful in the task of mapping latent variables to biological pathways. Furthermore, once having learned the parameters of the model, we can estimate the expected states of latent variables for each DNA microarray data point. Thus, we can potentially describe a system with the states of pathways which will provide insight to many interest biological questions, e.g. the mechanisms of diseases etc. We can also use this information to perform classification or diagnosis. With reduced dimension, a classifier is less like to over-fit the data and, therefore, generalizes well.

We have further extended the model to automatically determine the number of latent variables needed to efficiently explain the generation of data. This address the issue of how information is organized in the cells - a question that has not been well studied before due to the limitation of data and method. We believe the our model can potentially address the question efficiently and provide insight into cellular system from system biology point of view. In next part of the report, we will present the results of using the model to study the DNA microarray data and discuss the strengths and limitations of the model.

References

Attias, H. (1999a). Independent Factor Analysis. *Neural Computation*, 11(4):803–851.

Attias, H. (1999b). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Uncertainty in AI Conference*, pages 21–30.

Bishop, C. M. (1999). Variational principal components. In *Proceedings of Ninth International Conference on Artificial Neural Networks*, volume 1, pages 509–514. ICANN.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood estimation from incomplete data via EM algorithm (with discussion). *Journal of Royal Statistics Society*, B 39:1 – 38.

Dunmur, A. P. and Titterington, D. M. (1997). On a modification to the mean field EM algorithm in factorial learning. In Mozer, M. C., Jordan, M. I., and Petsche, T., editors, *Advances in Neural Information Processing Systems*, volume 9, page 431. The MIT Press.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–20.

Ghahramani, Z. (1995). Factorial learning and EM algorithm. In *Advances in Neural Information Processing Systems 7*. Morgan Kaufmann Publishers.

Ghahramani, Z. and Beal, M. J. (2000a). Propagation algorithms for variational bayesian learning. In *Advances in Neural Information Processing Systems 12*, pages 507–513. MIT Press.

Ghahramani, Z. and Beal, M. J. (2000b). Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems 12*, Cambridge, MA. MIT Press.

Haft, M., Hofmann, R., and Tresp, V. (1997). Modelindependent mean field theory as a local method for approximate propagation of information.

Hansen, M. H. and Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96(454):746.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. (1998). An introduction to variational methods for graphical models. Technical Report CSD-98-980, MIT.

Kass, R. E. and Raftery, A. E. (1994). Bayes Factors. Technical Report Technical Report No 254, Dept. of Statistics and Techical Report No 571, Dept. of Statistics, Univ. of Washington and Carnegie Mellon Univ.

Lawrence, N. D. and Bishop, C. M. (2000). Variational Bayesian independent component analysis. Technical report, Computer Laboratory, University of Cambridge.

Liang, S., Fuhrman, S., and Somogyi, R. (1998). REVEAL, A general reverse engineering algorithm for inference of genetic network architectures. In *Proceeding of Pacific Symposium of Biocomputing*, volume 3, pages 18–29.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18:51–60.

MacKay, D. (1995). Probable networds and plausible predictions - a review of practical Baysian methods for supervised nerual networkds. *Network: Computation in Neural Systems*, 6(3):469–505.

Miskin, J. W. (2000). *Ensemble Learning for Independent Component Analysis*. PhD thesis, Selwyn College, University of Cambridge.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. In *Proceeding of Pacific Symposium on Biocomputing*, pages 455–66.

Roweis, S. and Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Computation*, 11:305–345.

Tanner, M. A. (1996). *Tools for Statistical Inference*. Springer-Verlag, New York.

Tipping, M. and Bishop, C. (1997). Probabilistic principal component analysis.