Bayesian Factor Analysis

Aaron A. D'Souza

June 8, 2002

1 Introduction

Factor analysis is a well established statistical method that is commonly used to extract lower dimensional manifolds of high dimensional data by exploiting the covariance structure inherent in the data. In a broad sense, factor analysis assumes that the observed high dimensional data is the result of a linear combination of a smaller number of "factors" plus some added noise. In order to apply this technique effectively however, the user is required to supply a significant amount of knowledge *a priori*. This may include an estimate of the underlying dimensionality, as well as the number of mixture components (in the case that mixture models are used for non-linear manifolds). In general, the data obtained is often high dimensional, and an accurate estimate of the non-linearity and intrinsic local dimensionalities is hard to obtain.

Ideally we would like this knowledge to fall out of the inference process itself. It is also desirable that the model be as simple as possible and yet adequately represent the structure in the observed data. The problem is that more complex models (models which assume a larger number of factors, and ones with a larger number of mixture componenets) will always do a better job of fitting data than simpler models, and we must often resort to expensive cross-validation techniques to ensure that overfitting does not take place.

In recent years, researchers have been turning to Bayesian techniques as a viable method of doing data analysis. This framework is particularly appealing since the regularization of model complexity falls naturally out of the Bayesian formalism of integrating over the space of possible models. In order to use this framework however, we must begin by formulating factor analysis as a probabilistic model. We shall start with the simplest formulation of factor analysis, and gradually work up to more complex model structures, as we progress in the complexity and generality of our statistical analysis.

2 The Probabilistic Factor Analysis Model

Mathematically, we can write the generative model for factor analysis as follows:

$$\mathbf{t} = \mathbf{W}\mathbf{x} + \boldsymbol{\epsilon} + \boldsymbol{\mu} \tag{1}$$

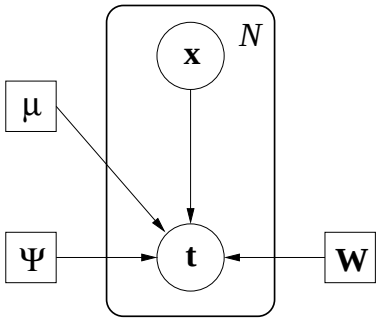

Figure 1: Graphical model for maximum likelihood factor analysis

where we define

$$\begin{aligned}
\mathbf{t} &\rightarrow d \text{ dimensional vector of observed variables} \\
\mathbf{W} &\rightarrow d \times q \text{ matrix of factor loadings} \\
\mathbf{x} &\rightarrow q \text{ dimensional vector of hidden variables with distribution } \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{I}) \text{ [I]} \\
\boldsymbol{\mu} &\rightarrow \text{mean of observed variables} \\
\boldsymbol{\epsilon} &\rightarrow \text{noise with distribution } \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \boldsymbol{\Psi})
\end{aligned}$$

Factor analysis assumes that the covariance matrix $\boldsymbol{\Psi}$ of the noise $\boldsymbol{\epsilon}$ is a diagonal matrix with possibly distinct elements. The consequences of this assumption are twofold. Firstly, by assuming that $\boldsymbol{\Psi}$ is diagonal, we assume that the noise in each dimension is independent, and any correlation between dimensions is accounted for by the factor loading matrix $\mathbf{W}$. Secondly, by allowing the diagonal elements of $\boldsymbol{\Psi}$ to be distinct, we allow the magnitude of the noise in each dimension to be different. It can be proved that if we restrict $\boldsymbol{\Psi}$ to be a multiple of the unit matrix (i.e. $\boldsymbol{\Psi} = \sigma^2\mathbf{I}$) then factor analysis reduces to Principle Component Analysis (PCA).

3 Maximum Likelihood Estimation

We begin by considering a simple version of the factor analysis model in which we do not place distributions over the model parameters $\mathbf{W}$ and $\boldsymbol{\Psi}$. Representing probabilistic systems in terms of graphical models is rapidly becoming a useful tool in Bayesian analysis. The graphical model corresponding to our current forumlation of the factor analysis model is shown in fig.

[I] We use the notation $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the multivariate Normal distribution which is mathematically defined as:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-d/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

where d is the dimensionality of $\mathbf{x}$.

1. Throughout this document we shall adopt the convention that circular nodes in the graph denote variables that have probability distributions over them, while we represent variables that have no distributions with rectangular nodes.

The problem of fitting a factor analysis model to the observed data, can be thought of as equivalent to the problem of determining the values of the model parameters $\mathbf{W}$ and $\mathbf{\Psi}$ which when plugged into a generative factor analyzer are most likely to generate the observed data distribution. In other words we are interested in maximizing the *likelihood* of generating the observed data $p(\mathcal{D}|\mathbf{W}, \mathbf{\Psi})$ given the model parameters $\mathbf{W}$ and $\mathbf{\Psi}$. This approach is called the Maximum Likelihood (ML) framework, and although this method is not truly Bayesian in the sense of using Bayes Rule[II] to infer a posterior distribution over the parameter values, it will lay the probabilistic foundation upon which the more sophisticated factor analysis models are built.

Although the ML approach is a theoretically appealing solution, we often find that the expressions for likelihood are analytically intractable. The Expectation Maximization (EM) algorithm can be used to simplify the math considerably. It is this approach that we shall discuss in some detail in the following sections.

3.1 Estimation of W and Ψ using EM

Given a set of N data points $\mathcal{D} = \{\mathbf{t}_i\}$ we wish to estimate the parameters $\mathbf{W}$ and $\mathbf{\Psi}$. In the EM formalism, instead of maximizing the likelihood of the observed data $p(\mathcal{D}|\mathbf{W}, \mathbf{\Psi})$ (also called the *incomplete* data likelihood), we attempt to maximize the joint likelihood $p(\mathcal{D}, \mathbf{X}|\mathbf{W}, \mathbf{\Psi})$ of the observed data and all unobserved random variables in the model (also known as the *complete* data likelihood). Since this quantity is a function of the random variable $\mathbf{x}$ which we cannot observe, we must work with the *expectation* of this quantity w.r.t. some distribution $Q(\mathbf{X})$. It is easy to show that this expectation is always a lower bound to the incomplete data likelihood for any arbitrary distribution $Q(\mathbf{X})$, and is only equal to the incomplete data likelihood when the expectation is taken w.r.t. the posterior distribution of $\mathbf{X}$ (i.e. when $Q(\mathbf{X}) = p(\mathbf{X}|\mathcal{D}, \mathbf{W}, \mathbf{\Psi})$). The log[III] complete data likelihood can be written as follows:

$$
\begin{aligned}
l_c(\mathbf{W}, \mathbf{\Psi}) &= \log p(\mathcal{D}, \mathbf{X}|\mathbf{W}, \mathbf{\Psi}) \\
&= \log \prod_i^N p(\mathbf{t}_i, \mathbf{x}_i|\mathbf{W}, \mathbf{\Psi}) \\
&= \sum_i^N \log p(\mathbf{t}_i, \mathbf{x}_i|\mathbf{W}, \mathbf{\Psi}) \\
&= \sum_i^N \log p(\mathbf{t}_i|\mathbf{x}_i, \mathbf{W}, \mathbf{\Psi}) + \sum_i^N \log p(\mathbf{x}_i|\mathbf{W}, \mathbf{\Psi})
\end{aligned}
\tag{2}
$$

[II] Bayes famous rule: $p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$

[III] Since the log function is monotonic, our analysis is considerably simplified if we maximize the *log*-likelihood

but since the distribution of $\mathbf{x}$ is independent of $\mathbf{W}$ and $\mathbf{\Psi}$

$$l_c(\mathbf{W}, \mathbf{\Psi}) = \sum_i^N \log p(\mathbf{t}_i|\mathbf{x}_i, \mathbf{W}, \mathbf{\Psi}) + \sum_i^N \log p(\mathbf{x}_i) \tag{3}$$

Since the second term in this equation is independent of $\mathbf{W}$ and $\mathbf{\Psi}$ it suffices (for the purposes of maximizing eq. (3) w.r.t. $\mathbf{W}$ and $\mathbf{\Psi}$) to think of the complete log-likelihood as simply:

$$l_c(\mathbf{W}, \mathbf{\Psi}) = \sum_i^N \log p(\mathbf{t}_i|\mathbf{x}_i, \mathbf{W}, \mathbf{\Psi}) \tag{4}$$

Using the definition of our probabilistic factor analysis model — in particular the linear dependence of $\mathbf{t}$ on $\boldsymbol{\epsilon}$ and our assumption of Gaussian noise — we can prove that the distribution of $\mathbf{t}$ given $\mathbf{x}$ is $\mathcal{N}(\mathbf{t}; \mathbf{W}\mathbf{x}, \mathbf{\Psi})$[IV]. Hence we can expand eq. (4) as follows:

$$\begin{aligned} l_c(\mathbf{W}, \mathbf{\Psi}) &= \sum_i^N \log \frac{1}{(2\pi)^{d/2} |\mathbf{\Psi}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \mathbf{\Psi}^{-1}(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)\right\} && (5) \\ &= -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{\Psi}| - \frac{1}{2}\sum_i^N \left(\mathbf{t}_i^T \mathbf{\Psi}^{-1}\mathbf{t}_i - 2\mathbf{t}_i^T \mathbf{\Psi}^{-1}\mathbf{W}\mathbf{x}_i + \mathbf{x}_i^T \mathbf{W}^T \mathbf{\Psi}^{-1}\mathbf{W}\mathbf{x}_i\right) \\ &= k - \frac{N}{2}\log|\mathbf{\Psi}| - \frac{1}{2}\sum_i^N \left(\mathbf{t}_i^T \mathbf{\Psi}^{-1}\mathbf{t}_i - 2\mathbf{t}_i^T \mathbf{\Psi}^{-1}\mathbf{W}\mathbf{x}_i + \mathrm{Tr}\left[\mathbf{W}^T \mathbf{\Psi}^{-1}\mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\right]\right) \text{ [V]} && (6) \end{aligned}$$

3.1.1 The M step

The "M" step in EM takes the expected complete log-likelihood as defined in eq. (7) and maximizes it w.r.t. the parameters that are to be estimated; in this case $\mathbf{W}$ and $\mathbf{\Psi}$.

To estimate $\mathbf{W}$ we start with eq. (6). Taking expectations and differentiating w.r.t $\mathbf{W}$ we get:

$$\begin{aligned} \langle l_c(\mathbf{W}, \mathbf{\Psi})\rangle &= k - \frac{N}{2}\log|\mathbf{\Psi}| - \frac{1}{2}\sum_i^N \left(\mathbf{t}_i^T \mathbf{\Psi}^{-1}\mathbf{t}_i - 2\mathbf{t}_i^T \mathbf{\Psi}^{-1}\mathbf{W}\langle\mathbf{x}_i\rangle \right. \\ &\qquad \left. + \mathrm{Tr}\left[\mathbf{W}^T \mathbf{\Psi}^{-1}\mathbf{W}\left\langle\mathbf{x}_i\mathbf{x}_i^T\right\rangle\right]\right) \\ \frac{\partial\langle l_c(\mathbf{W}, \mathbf{\Psi})\rangle}{\partial \mathbf{W}} &= -\frac{1}{2}\sum_i^N \left(-2\mathbf{\Psi}^{-1}\mathbf{t}_i\langle\mathbf{x}_i\rangle^T + 2\mathbf{\Psi}^{-1}\mathbf{W}\left\langle\mathbf{x}_i\mathbf{x}_i^T\right\rangle\right) \text{ [VI]} \end{aligned} \tag{7}$$

[IV] Since $\langle\mathbf{t}|\mathbf{x}\rangle = \langle(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})|\mathbf{x}\rangle = \mathbf{W}\mathbf{x}$ and $\mathrm{Cov}(\mathbf{t}|\mathbf{x}) = \langle(\mathbf{t} - \mathbf{W}\mathbf{x})(\mathbf{t} - \mathbf{W}\mathbf{x})^T|\mathbf{x}\rangle = \langle\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\mathbf{x}\rangle = \mathbf{\Psi}$

[V] Here we have used the relation $\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathrm{Tr}\left[\mathbf{A}\mathbf{x}\mathbf{x}^T\right]$, where $\mathrm{Tr}[\cdot]$ is the trace operator.

[VI] Where we have used the relations $\frac{\partial}{\partial \mathbf{X}}\mathbf{A}^T\mathbf{X}\mathbf{B} = \mathbf{A}\mathbf{B}^T$, and $\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}\left[\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{B}\right] = \mathbf{A}\mathbf{X}\mathbf{B} + \mathbf{A}^T\mathbf{X}\mathbf{B}^T$.

Setting to zero and solving for $\mathbf{W}$ gives us:

$$\mathbf{W} = \left(\sum_i^N \mathbf{t}_i \left\langle \mathbf{x}_i \right\rangle^T \right) \left(\sum_i^N \left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle \right)^{-1} \tag{8}$$

To maximize w.r.t. $\mathbf{\Psi}$ we start with eq. (5). Taking expectations and differentiating w.r.t. $\mathbf{\Psi}^{-1}$ (Note that differentiating w.r.t. $\mathbf{\Psi}^{-1}$ instead of $\mathbf{\Psi}$ makes the analysis simpler) we get:

$$\begin{aligned}
\left\langle l_c(\mathbf{W}, \mathbf{\Psi}) \right\rangle &= \left\langle \sum_i^N \log \frac{1}{(2\pi)^{d/2} \left|\mathbf{\Psi}\right|^{1/2}} \exp\left\{ -\frac{1}{2} (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \mathbf{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i) \right\} \right\rangle \\
&= -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log \left|\mathbf{\Psi}\right| - \frac{1}{2} \sum_i^N \left\langle (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \mathbf{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i) \right\rangle \\
\frac{\partial \left\langle l_c(\mathbf{W}, \mathbf{\Psi}) \right\rangle}{\partial \mathbf{\Psi}^{-1}} &= \frac{N}{2} \mathbf{\Psi} - \frac{1}{2} \sum_i^N \left\langle (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \right\rangle \text{ [VII]} \\
&= \frac{N}{2} \mathbf{\Psi} - \frac{1}{2} \sum_i^N \mathbf{t}_i \mathbf{t}_i^T + \left(\sum_i^N \mathbf{t}_i \left\langle \mathbf{x}_i \right\rangle^T \right) \mathbf{W}^T - \frac{1}{2} \mathbf{W} \left(\sum_i^N \left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle \right) \mathbf{W}^T
\end{aligned}$$

Setting to zero and solving for $\mathbf{\Psi}$ with the help of eq. (8) gives us:

$$\mathbf{\Psi} = \frac{1}{N} \text{diag} \left[\sum_i^N \mathbf{t}_i \mathbf{t}_i^T - \left(\sum_i^N \mathbf{t}_i \left\langle \mathbf{x}_i \right\rangle^T \right) \mathbf{W}^T \right] \tag{9}$$

We have introduced the diag $[\cdot]$ operator in Eq. (9) so that $\mathbf{\Psi}$ is constrained to be a diagonal matrix.

3.1.2 The E step

We are still left with the problem of determining the actual values of $\langle \mathbf{x}_i \rangle$ and $\left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle$. As we mentioned earlier, in order to guarantee that we are indeed maximizing the *incomplete* data likelihood, it is essential that the expected complete log likelihood (which is its lower bound) is maximized by taking the expectation w.r.t. $p(\mathbf{X}|\mathcal{D}, \mathbf{W}, \mathbf{\Psi})$. Hence the expectations $\langle \mathbf{x}_i \rangle$ and $\left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle$ should actually be computed w.r.t. $p(\mathbf{X}|\mathcal{D}, \mathbf{W}, \mathbf{\Psi})$.

In this relatively simplified setting, we can actually obtain an analytical form for the posterior distribution $p(\mathbf{x}|\mathbf{t})$ using Bayes rule as follows:

$$p(\mathbf{x}_i|\mathbf{t}_i) \propto p(\mathbf{t}_i|\mathbf{x}_i) p(\mathbf{x}_i)$$

[VII] Where we have used the relations $\frac{\partial}{\partial \mathbf{X}} \log|\mathbf{X}| = (\mathbf{X}^{-1})^T$, and $\frac{\partial}{\partial \mathbf{X}} \mathbf{A}^T \mathbf{X} \mathbf{B} = \mathbf{A}\mathbf{B}^T$.

Hence

$$\begin{aligned}\log p(\mathbf{x}_i|\mathbf{t}_i) &= \log p(\mathbf{t}_i|\mathbf{x}_i) + \log p(\mathbf{x}_i) + const \\ &= -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Psi}| - \frac{1}{2}(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T\mathbf{\Psi}^{-1}(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i) \\ &\quad - \frac{q}{2}\log 2\pi - \frac{1}{2}\mathbf{x}_i^T\mathbf{x}_i + const \qquad (10) \\ &= -\frac{1}{2}\left(\mathbf{t}_i^T\mathbf{\Psi}^{-1}\mathbf{t}_i - 2\mathbf{x}_i^T\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{t}_i + \mathbf{x}_i^T(\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})\mathbf{x}_i\right) + const \qquad (11)\end{aligned}$$

From the quadratic form we can infer that the posterior distribution of $\mathbf{x}_i$ is Gaussian:

$$p(\mathbf{x}_i|\mathbf{t}_i) = \mathcal{N}\left(\mathbf{x}_i; \mathbf{m}_{\mathbf{x}}^{(i)}, \mathbf{\Sigma}_{\mathbf{x}}\right) \qquad (12)$$

with

$$\begin{aligned}\mathbf{\Sigma}_{\mathbf{x}} &= (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1} \qquad (13) \\ \mathbf{m}_{\mathbf{x}}^{(i)} &= \mathbf{\Sigma}_{\mathbf{x}}\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{t}_i \\ &= (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{t}_i \\ &= \boldsymbol{\beta}\mathbf{t}_i \qquad (14)\end{aligned}$$

where we define $\boldsymbol{\beta} \equiv (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{\Psi}^{-1}$

Given this distribution we can infer the required expectations as follows:

$$\begin{aligned}\langle \mathbf{x}_i \rangle &= \mathbf{m}_{\mathbf{x}}^{(i)} = \boldsymbol{\beta}\mathbf{t}_i \qquad (15) \\ \left\langle \mathbf{x}_i\mathbf{x}_i^T \right\rangle &= \mathbf{\Sigma}_{\mathbf{x}} + \mathbf{m}_{\mathbf{x}}^{(i)}\mathbf{m}_{\mathbf{x}}^{(i)^T} \\ &= (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1} + \boldsymbol{\beta}\mathbf{t}_i\mathbf{t}_i^T\boldsymbol{\beta}^T \\ &= \mathbf{I} - \mathbf{W}^T(\mathbf{\Psi} + \mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W} + \boldsymbol{\beta}\mathbf{t}_i\mathbf{t}_i^T\boldsymbol{\beta}^T \qquad (16)\end{aligned}$$

Using the Sherman-Morrison-Woodbury matrix inversion theorem, we can derive the following result (refer to appendix A.2):

$$\mathbf{W}^T(\mathbf{\Psi} + \mathbf{W}\mathbf{W}^T)^{-1} = (\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{\Psi}^{-1} = \boldsymbol{\beta}$$

Notice that the second form is much easier to evaluate since $(\mathbf{I} + \mathbf{W}^T\mathbf{\Psi}^{-1}\mathbf{W})$ is a smaller matrix than $(\mathbf{\Psi} + \mathbf{W}\mathbf{W}^T)$ and $\mathbf{\Psi}$ is diagonal. Plugging this result into eq. (16) we get:

$$\left\langle \mathbf{x}_i\mathbf{x}_i^T \right\rangle = \mathbf{I} - \boldsymbol{\beta}\mathbf{W} + \boldsymbol{\beta}\mathbf{t}_i\mathbf{t}_i^T\boldsymbol{\beta}^T$$

4 Inferring Underlying Dimensionality — Gaussian Approximation

The preceding section arrives at an extremely elegant solution to the problem of estimating the values of $\mathbf{W}$ and $\mathbf{\Psi}$ in our factor analysis model. However, one must still make an assumption

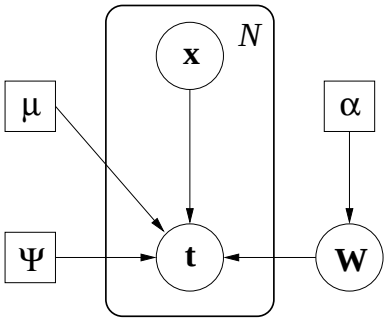

Figure 2: Graphical model for inferring the underlying latent variable dimensionality using a gaussian approximation.

of the dimensionality q of the underlying distribution. In doing so one runs the risk of selecting too high a value of q and overfitting the data by explaining noise, or of selecting too low a value and over generalizing, resulting in not capturing the true data complexity.

Each column of the $\mathbf{W}$ matrix represents one dimension of the underlying latent variable space. What is required is a way for us to determine how many columns of $\mathbf{W}$ are actually relevant based on the data that is presented to us. In some sense, the number of columns of $\mathbf{W}$ is a measure of our factor analysis model complexity — the larger the value of q, the greater the model complexity since it can explain (or generate) a larger family of data sets.

In order to determine the most appropriate latent variable dimensionality, we start with the maximum possible value of $q = d - 1$, but place a prior distribution over each of the $d - 1$ columns of $\mathbf{W}$ parameterized by a precision parameter $\boldsymbol{\alpha}$ which functions as an inverse spherical covariance for each column. Hence we can write the distribution of $\mathbf{W}$ as follows:

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i}^{d-1} \left(\frac{\alpha_i}{2\pi}\right)^{d/2} \exp\left(-\frac{\alpha_i}{2}\mathbf{w}_i^T\mathbf{w}_i\right) \tag{17}$$

This change in model structure is reflected in our updated graphical model as shown in figure 2. Here we see a new node $\boldsymbol{\alpha}$ being added as a parent to $\mathbf{W}$. Also the $\mathbf{W}$ node has been changed from a rectangular node to a circle, reflecting the fact that we now have a prior distribution over $\mathbf{W}$ and that we can estimate a posterior distribution for this variable using Bayesian analysis rather than merely compute a maximum likelihood estimate of its value. Using Bayes rule we have:

$$p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) \propto p(\mathcal{D}|\mathbf{W})p(\mathbf{W}|\boldsymbol{\alpha}) \tag{18}$$

and hence

$$\log p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) = const + \log p(\mathcal{D}|\mathbf{W}) + \log p(\mathbf{W}|\boldsymbol{\alpha}) \tag{19}$$

Since we have a distribution over $\mathbf{W}$, in practice we would like to find the maximum *a posteriori* value $\mathbf{W}_{MP}$ of this distribution, which means finding the value of $\mathbf{W}$ which maximizes eq. (19).

Since we know that the complete log-likelihood l_c is a lower bound to the true data log-likelihood, we can substitute l_c for $\log p(\mathcal{D}|\mathbf{W})$ and try to maximize this new equation.

$$\log p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) = const + l_c(\mathbf{W}, \boldsymbol{\Psi}) + \log p(\mathbf{W}|\boldsymbol{\alpha}) \tag{20}$$

which using eq. (17) and discarding terms that are independent of $\mathbf{W}$ and $\boldsymbol{\Psi}$ gives us:

$$\log p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) = l_c(\mathbf{W}, \boldsymbol{\Psi}) - \frac{1}{2}\sum_{i}^{d-1} \alpha_i \mathbf{w}_i^T \mathbf{w}_i \tag{21}$$

$$= l_c(\mathbf{W}, \boldsymbol{\Psi}) - \frac{1}{2}\mathrm{Tr}\left[\mathbf{W}\mathbf{A}\mathbf{W}^T\right] \tag{22}$$

Which is simply the complete log-likelihood with a regularization term that penalizes solutions of $\mathbf{W}$ with higher intrinsic dimensionality. EM still applies within this framework and we can differentiate the expectation of this regularized likelihood to derive the update equations for $\mathbf{W}$ and $\boldsymbol{\Psi}$. Substituting from eq. (6) and taking expectations we get:

$$\left\langle \log p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) \right\rangle = -\frac{1}{2}\sum_{i}^{N} \left(\mathbf{t}_i^T \boldsymbol{\Psi}^{-1} \mathbf{t}_i - 2\mathbf{t}_i^T \boldsymbol{\Psi}^{-1} \mathbf{W} \left\langle \mathbf{x}_i \right\rangle + \mathrm{Tr}\left[\mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle \right] \right)$$
$$- \frac{1}{2}\mathrm{Tr}\left[\mathbf{W}\mathbf{A}\mathbf{W}^T\right] - \frac{N}{2}\log|\boldsymbol{\Psi}| + k \tag{23}$$

Maximizing w.r.t. $\mathbf{W}$ we get:

$$\frac{\partial \left\langle \log p(\mathbf{W}|\mathcal{D}, \boldsymbol{\alpha}) \right\rangle}{\partial \mathbf{W}} = -\frac{1}{2}\sum_{i}^{N} \left(-2\boldsymbol{\Psi}^{-1}\mathbf{t}_i \left\langle \mathbf{x}_i \right\rangle^T + 2\boldsymbol{\Psi}^{-1}\mathbf{W}\left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle \right) - \mathbf{W}\mathbf{A} = \mathbf{0}^{\text{VIII}} \tag{24}$$

Hence

$$\mathbf{W}\sum_{i}^{N} \left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle + \boldsymbol{\Psi}\mathbf{W}\mathbf{A} = \sum_{i}^{N} \mathbf{t}_i \left\langle \mathbf{x}_i \right\rangle^T \tag{25}$$

or equivalently

$$\mathbf{W}\mathbf{S} + \boldsymbol{\Psi}\mathbf{W}\mathbf{A} = \mathbf{m} \tag{26}$$

Where we define $\mathbf{S} \equiv \sum_i^N \left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle$ and $\mathbf{m} \equiv \sum_i^N \mathbf{t}_i \left\langle \mathbf{x}_i \right\rangle^T$

Since $\boldsymbol{\Psi}$ is a diagonal matrix, we can obtain a closed form solution for each row of $\mathbf{W}$ individually. For the k^{th} row:

$$\mathbf{w}_k = \mathbf{m}_k(\mathbf{S} + p_k\mathbf{A})^{-1} \tag{27}$$

Where $\mathbf{w}_k$ and $\mathbf{m}_i$ are the k^{th} rows of $\mathbf{W}$ and $\mathbf{m}$ respectively, p_k is the k^{th} diagonal element of $\boldsymbol{\Psi}$, and $1 \leq k \leq d$.

[VIII] Where we have used the relation $\frac{\partial}{\partial \mathbf{X}}\mathrm{Tr}\left[\mathbf{X}\mathbf{A}\mathbf{X}^T\right] = \mathbf{X}\left(\mathbf{A} + \mathbf{A}^T\right)$

4.1 Estimation of $\boldsymbol{\alpha}$

From the graphical model in figure 2 we see that in order to compute the likelihood of the data given the hyperparameters $\boldsymbol{\alpha}$ we must integrate over the distribution of $\mathbf{W}$

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\alpha}) &= \int p(\mathcal{D}|\mathbf{W}, \boldsymbol{\alpha}) p(\mathbf{W}|\boldsymbol{\alpha}) d\mathbf{W} \\ &= \int p(\mathcal{D}|\mathbf{W}) p(\mathbf{W}|\boldsymbol{\alpha}) d\mathbf{W} \end{aligned} \tag{28}$$

Now since we assume that our observed data is Independently Identically Distributed (IID) we can write:

$$p(\mathcal{D}|\mathbf{W}) = \prod_i^N p(\mathbf{t}|\mathbf{W}) \tag{29}$$

$$= \left[\frac{1}{(2\pi)^{d/2} |\mathbf{C}|^{1/2}}\right]^N \exp\left\{-\frac{1}{2}\sum_i^N \mathbf{t}_i^T \mathbf{C}^{-1} \mathbf{t}_i\right\} \tag{30}$$

$$= \left[\frac{1}{(2\pi)^{d/2}}\right]^N \exp\left\{-\frac{1}{2}\sum_i^N \left(\mathbf{t}_i^T \mathbf{C}^{-1} \mathbf{t}_i + \log|\mathbf{C}|\right)\right\} \tag{31}$$

Where $\mathbf{C} = \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi}$ is the covariance matrix of the observed data[IX]. Hence using eq. (17) for $p(\mathcal{D}|\mathbf{W})$ along with eq. (31) in eq. (28) we get:

$$p(\mathcal{D}|\boldsymbol{\alpha}) \propto \left[\prod_i^{d-1} \left(\frac{\alpha_i}{2\pi}\right)^{d/2}\right] \int \exp\left\{-\frac{1}{2}\sum_i^N \left(\mathbf{t}_i^T \mathbf{C}^{-1} \mathbf{t}_i + \log|\mathbf{C}|\right) - \frac{1}{2}\sum_i^{d-1} \alpha_i \mathbf{w}_i^T \mathbf{w}_i\right\} d\mathbf{W} \tag{32}$$

$$= \left[\prod_i^{d-1} \left(\frac{\alpha_i}{2\pi}\right)^{d/2}\right] \int \exp\left\{-S(\mathbf{W})\right\} d\mathbf{W} \tag{33}$$

Where we define

$$S(\mathbf{W}) \equiv \frac{1}{2}\sum_i^N \left(\mathbf{t}_i^T \mathbf{C}^{-1} \mathbf{t}_i + \log|\mathbf{C}|\right) + \frac{1}{2}\sum_i^{d-1} \alpha_i \mathbf{w}_i^T \mathbf{w}_i \tag{34}$$

[IX] This is trivially shown:

$$\begin{aligned} \langle \mathbf{t} \rangle &= \langle \mathbf{W}\mathbf{x} + \boldsymbol{\epsilon} \rangle = \mathbf{W} \langle \mathbf{x} \rangle + \langle \boldsymbol{\epsilon} \rangle = \mathbf{0} \\ \mathrm{Cov}(\mathbf{t}) = \left\langle \mathbf{t}\mathbf{t}^T \right\rangle &= \left\langle (\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})(\mathbf{W}\mathbf{x} + \boldsymbol{\epsilon})^T \right\rangle \\ &= \mathbf{W} \left\langle \mathbf{x}\mathbf{x}^T \right\rangle \mathbf{W}^T + \mathbf{W} \left\langle \mathbf{x}\boldsymbol{\epsilon}^T \right\rangle + \left\langle \boldsymbol{\epsilon}\mathbf{x}^T \right\rangle \mathbf{W}^T + \left\langle \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T \right\rangle \\ &= \mathbf{W}\mathbf{W}^T + \mathbf{0} + \mathbf{0} + \boldsymbol{\Psi} \\ &= \mathbf{W}\mathbf{W}^T + \boldsymbol{\Psi} \equiv \mathbf{C} \end{aligned}$$

Approximate $S(\mathbf{W})$ with a second-order Taylor series expansion around the extremal point $\mathbf{W}_{MP}$. Since the first order derivative at an extremal is zero, our expansion does not contain a linear term.

$$S(\mathbf{W}) \approx S(\mathbf{W}_{MP}) + \frac{1}{2}(\mathbf{W} - \mathbf{W}_{MP})^T \mathbf{H} (\mathbf{W} - \mathbf{W}_{MP}) \tag{35}$$

Where $\mathbf{H}$ is the $d(d-1) \times d(d-1)$ Hessian matrix of $S(\cdot)$ evaluated at $\mathbf{W}_{MP}$. Using this approximation to $S(\mathbf{W})$ we can now evaluate the integral

$$\begin{aligned}
\int \exp\{-S(\mathbf{W})\}\, d\mathbf{W} &\approx \int \exp\left\{-S(\mathbf{W}_{MP}) - \frac{1}{2}(\mathbf{W} - \mathbf{W}_{MP})^T \mathbf{H} (\mathbf{W} - \mathbf{W}_{MP})\right\} d\mathbf{W} \\
&= \exp\{-S(\mathbf{W}_{MP})\} \int \exp\left\{-\frac{1}{2}(\mathbf{W} - \mathbf{W}_{MP})^T \mathbf{H} (\mathbf{W} - \mathbf{W}_{MP})\right\} d\mathbf{W} \\
&= \exp\{-S(\mathbf{W}_{MP})\} (2\pi)^{d(d-1)/2} \left|\mathbf{H}^{-1}\right|^{1/2}
\end{aligned} \tag{36}$$

Where the value of the integral is now simply the normalizing constant for a gaussian distribution in $\mathbf{W}$ with covariance $\mathbf{H}^{-1}$. Substituting this result back into eq. (33) we get:

$$p(\mathcal{D}|\boldsymbol{\alpha}) \propto \left[\prod_i^{d-1} \left(\frac{\alpha_i}{2\pi}\right)^{d/2}\right] \exp\{-S(\mathbf{W}_{MP})\} (2\pi)^{d(d-1)/2} \left|\mathbf{H}^{-1}\right|^{1/2} \tag{37}$$

or equivalently

$$\log p(\mathcal{D}|\boldsymbol{\alpha}) = const + \frac{d}{2}\sum_i^{d-1} \log \alpha_i - S(\mathbf{W}_{MP}) - \frac{1}{2}\log|\mathbf{H}| \tag{38}$$

Differentiating w.r.t. α_k we get:

$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\alpha})}{\partial \alpha_k} = \frac{\partial}{\partial \alpha_k}\left[\sum_i^{d-1} \frac{d}{2}\log \alpha_i\right] - \frac{\partial}{\partial \alpha_k} S(\mathbf{W}_{MP}) - \frac{1}{2}\frac{\partial}{\partial \alpha_k}\log|\mathbf{H}| \tag{39}$$

Now using eq. (34) we have:

$$\frac{\partial}{\partial \alpha_k} S(\mathbf{W}_{MP}) = \frac{1}{2}\|\mathbf{w}_k^{MP}\|^2 \tag{40}$$

In order to compute the partial derivative of $\mathbf{H}$ w.r.t. α_k let us express $S(\mathbf{W})$ as follows:

$$S(\mathbf{W}) = E_{\mathbf{W}} + E_{\boldsymbol{\alpha}} \tag{41}$$

where we define $E_{\mathbf{W}} \equiv \frac{1}{2}\sum_i^N \left(\mathbf{t}_i^T \mathbf{C}^{-1} \mathbf{t}_i + \log|\mathbf{C}|\right)$ and $E_{\boldsymbol{\alpha}} \equiv \frac{1}{2}\sum_i^{d-1} \alpha_i \mathbf{w}_i^T \mathbf{w}_i$. Hence we can write the Hessian of $S(\mathbf{W})$ as:

$$\mathbf{H} = \nabla\nabla E_{\mathbf{W}} + \nabla\nabla E_{\boldsymbol{\alpha}} \tag{42}$$

Now if we assume that $\mathbf{W}$ is structured such that each column is lined up to form a large vector of dimensionality $d(d-1)$, then we can write:

$$\nabla\nabla E_{\boldsymbol{\alpha}} = \begin{bmatrix} \alpha_1\mathbf{I}_d & \mathbf{0} & \ldots\ldots\ldots \\ \mathbf{0} & \alpha_2\mathbf{I}_d & \mathbf{0} & \ldots\ldots \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \ldots\ldots\ldots & \mathbf{0} & \alpha_{d-1}\mathbf{I}_d \end{bmatrix} \tag{43}$$

Hence we can view $\nabla\nabla E_{\boldsymbol{\alpha}}$ as a block diagonal matrix, where each $d \times d$ block along the diagonal is a unit matrix scaled by a corresponding α_i. Let λ_{ij} be the j^{th} eigenvalue of the i^{th} diagonal submatrix of $\nabla\nabla E_{\mathbf{W}}$. Since the determinant of a matrix is equal to the product of it's eigenvalues[X], we can write:

$$\begin{aligned} \frac{\partial \log|\mathbf{H}|}{\partial\alpha_k} &= \frac{\partial}{\partial\alpha_k}\log\prod_i^{d-1}\prod_j^{d}(\lambda_{ij}+\alpha_i) \\ &= \frac{\partial}{\partial\alpha_k}\sum_i^{d-1}\sum_j^{d}\log(\lambda_{ij}+\alpha_i) \\ &= \sum_j^{d}\frac{1}{\lambda_{kj}+\alpha_k} \\ &= \mathrm{Tr}_k\left[\mathbf{H}^{-1}\right] \end{aligned} \tag{44}$$

Hence substituting from eqs. (40) and (44) back into eq. (39) gives us:

$$\frac{\partial \log p(\mathcal{D}|\boldsymbol{\alpha})}{\partial\alpha_k} = \frac{d}{2\alpha_k} - \frac{1}{2}\|\mathbf{w}_k^{MP}\|^2 - \frac{1}{2}\mathrm{Tr}_k\left[\mathbf{H}^{-1}\right] = 0 \tag{45}$$

Solving for α_k we arrive at the update equation:

$$\alpha_k = \frac{d}{\|\mathbf{w}_k^{MP}\|^2 + \mathrm{Tr}_k\left[\mathbf{H}^{-1}\right]} \tag{46}$$

If we make the assumption that the parameter $\mathbf{W}$ is well determined, then its posterior distribution will be sharply peaked, which means $\mathbf{H}$ will have large eigenvalues. This also implies that $\mathrm{Tr}_k\left[\mathbf{H}^{-1}\right]$ will be very small. Under this assumption, eq. (46) for α_k reduces to:

$$\alpha_k = \frac{d}{\|\mathbf{w}_k^{MP}\|^2} \tag{47}$$

This avoids costly computation and manipulation of the $d(d-1) \times d(d-1)$ Hessian matrix.

[X]This can be proved trivially; we can decompose a matrix $\mathbf{A}$ into the product $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^T$, where $\mathbf{V}$ is a matrix of eigenvectors, and $\mathbf{D}$ is a diagonal matrix of the corresponding eigenvalues λ_i. Since $\mathbf{V}$ is orthonormal (implying $|\mathbf{V}| = 1$), we have $|\mathbf{A}| = |\mathbf{D}| = \prod_i \lambda_i$.

When doing the Taylor series expansion of $S(\mathbf{W})$ in eq. (35) we assumed that our current estimate of $\mathbf{W}$ is a (possibly local) extremum. This assumption was important since it allowed us to eliminate the linear term in the expansion and retain only the quadratic term, making it a Gaussian approximation to the posterior distribution. Algorithmically this means that we should perform our EM iterations to update $\mathbf{W}$ (and $\mathbf{\Psi}$) given a fixed current estimate of the value of $\boldsymbol{\alpha}$. When these iterations converge then we know that we have reached an extremum, and we can use the current value of $\mathbf{W}$ to re-estimate $\boldsymbol{\alpha}$. Thus our algorithm performs the EM updates with an outer loop that periodically re-estimates the value of $\boldsymbol{\alpha}$.

In general since EM operates in a maximum likelihood framework, it will favour higher dimensionalities of $\mathbf{W}$ since this will always result in an increase in the likelihood. However, by using the $\boldsymbol{\alpha}$ as a precision parameter on each of the columns of $\mathbf{W}$ we create a penalized likelihood which seeks to limit the model complexity (dimensionality). By formulating the problem this way EM results in a compromise between maximizing the dimensionality to increase likelihood, and reducing the penalizing term (which increases with the dimensionality).

5 Inferring Underlying Dimensionality — Variational Approximation

If anything, the previous section should give us a hint that as the model complexity increases, integrating over the model parameter distributions becomes increasingly more complex, and indeed eventually analytically intractable. In the previous section itself, we fit a Gaussian distribution to the posterior distribution of $\mathbf{W}$ in order to be able to integrate over it. We could also adopt a sampling approach and use Monte Carlo methods to give us an approximation to the true distribution. In general however, sampling approaches tend to be expensive; both in terms of computation and in terms of storage since the probability distributions are effectively represented by a collection of samples.

In this section we will explore Variational Methods as another method of approximating the posterior distributions of model parameters. Variational methods have long been used in statistical physics, and have recently been getting significant attention from the statistical learning community. In essence, variational methods allow us to create a bound on the function of interest (in our case the log-evidence for the observed data). Subsequent analysis then works towards minimizing the difference between the bound and the true function.

Let us begin by augmenting our factor analysis model to place a probability distribution over all parameters whose cardinality scales with model complexity. As shown in fig. 3, we have now placed a probability distribution over the $\boldsymbol{\alpha}$ precision parameters as well. Since these parameters cannot be negative (being inverse covariances for each column vector of $\mathbf{W}$), we cannot place a Gaussian distribution over them. Instead we place a Gamma prior over the $\boldsymbol{\alpha}$ variables:

$$p(\boldsymbol{\alpha}) = \prod_{i}^{q} \mathcal{G}\left(\alpha_i; a_\alpha, b_\alpha\right) \text{ XI} \tag{48}$$

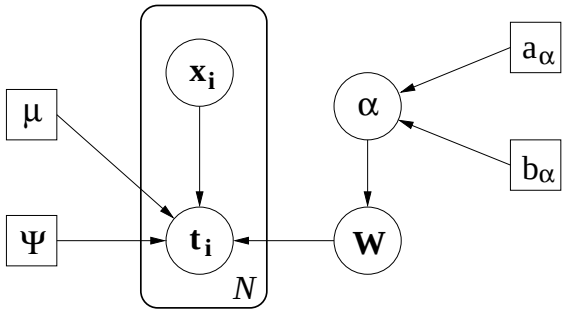

Figure 3: Graphical model for learning a factor analyzer with automatic dimensionality estimation

Let us now look at the log probability of the observed data $\mathcal{D}$ (also known as the *evidence*). This can be obtained by marginalizing over all the model parameters and hidden variables as follows:

$$\log p(\mathcal{D}) = \log \int p(\mathcal{D}, \mathbf{X}, \mathbf{W}, \boldsymbol{\alpha}) d\mathbf{X} d\mathbf{W} d\boldsymbol{\alpha} \tag{49}$$

Using Jensen's inequality, we can lower-bound this quantity as follows:

$$\log p(\mathcal{D}) \geq \int Q(\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha}) \log \frac{p(\mathcal{D}, \mathbf{X}, \mathbf{W}, \boldsymbol{\alpha})}{Q(\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha})} d\mathbf{X} d\mathbf{W} d\boldsymbol{\alpha} = \mathcal{F}(Q) \tag{50}$$

for any arbitrary distribution $Q(\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha})$.

Maximizing the functional $\mathcal{F}(Q)$ is equivalent to minimizing the Kullback-Liebler divergence between Q and the true posterior distribution $p(\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha}|\mathcal{D})$[XII]. There are two ways of assuming a functional form for Q. One is to assume a parameterized version of the distribution which simplifies its analytical form at the expense of introducing extra "variational" parameters that must be optimized. Another approach is to assume a factorized form for the distribution. This is the approach we shall take here. We assume the factorization:

$$Q(\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha}) = Q(\mathbf{X}) Q(\mathbf{W}) Q(\boldsymbol{\alpha}) \tag{51}$$

[XI] We use the notation $\mathcal{G}(x; a, b)$ to denote the Gamma distribution which is mathematically defined as:

$$\mathcal{G}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx)$$

[XII] This can be easily proved as follows:

$$\begin{aligned} \log p(\mathcal{D}) &= \int Q(\boldsymbol{\theta}) \log p(\mathcal{D}) d\boldsymbol{\theta} = \int Q(\boldsymbol{\theta}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} = \int Q(\boldsymbol{\theta}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} + \int Q(\boldsymbol{\theta}) \log \frac{Q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathcal{D})} d\boldsymbol{\theta} \\ &= \int Q(\boldsymbol{\theta}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} + KL\left\{Q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}|\mathcal{D})\right\} \end{aligned}$$

Using the calculus of variations we can prove (see appendix B) that the solution for each of the individual Q distributions that maximizes the functional $\mathcal{F}(Q)$ is of the form:

$$Q_i(\theta_i) = \frac{\exp \langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{k \neq i}}}{\int \exp \langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{k \neq i}} d\theta_i} \tag{52}$$

or equivalently

$$\log Q_i(\theta_i) = \langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{k \neq i}} + const \tag{53}$$

Where $\boldsymbol{\theta} = \{\mathbf{X}, \mathbf{W}, \boldsymbol{\alpha}\}$ and $\langle \cdot \rangle_{Q_{k \neq i}}$ denotes expectation taken with respect to every distribution other than $Q_i(\theta_i)$.

Let us first determine the expression for $p(\mathcal{D}, \boldsymbol{\theta})$. The graphical model makes this easy to express:

$$p(\mathcal{D}, \boldsymbol{\theta}) = \left[\prod_i^N p(\mathbf{t}_i | \mathbf{x}_i, \mathbf{W}) p(\mathbf{x}_i) \right] p(\mathbf{W}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) \tag{54}$$

and hence

$$\log p(\mathcal{D}, \boldsymbol{\theta}) = \sum_i^N \log p(\mathbf{t}_i | \mathbf{x}_i, \mathbf{W}) + \sum_i^N \log p(\mathbf{x}_i) + \log p(\mathbf{W}|\boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha}) \tag{55}$$

$$\begin{aligned} &= -\frac{N}{2} \log |\boldsymbol{\Psi}| - \frac{1}{2} \sum_i^N (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i) \\ &\quad - \frac{1}{2} \sum_i^N \mathbf{x}_i^T \mathbf{x}_i \\ &\quad + \frac{d}{2} \sum_i^q \log \alpha_i - \frac{1}{2} \sum_i^q \alpha_i \mathbf{w}_i^T \mathbf{w}_i \\ &\quad + \sum_i^q (a_\alpha - 1) \log \alpha_i - \sum_i^q b_\alpha \alpha_i + const \end{aligned} \tag{56}$$

5.1 Estimation of $Q(\boldsymbol{\alpha})$

To obtain an expression for $Q(\boldsymbol{\alpha})$ using eq. (53), we take expectations of eq. (56) w.r.t. the distribution $Q(\mathbf{W})Q(\mathbf{X})$. All terms not involving $\boldsymbol{\alpha}$ can conveniently be clubbed into a *const* term at the end of the equation, and contribute to the normalizing constants in the distribution

of $\boldsymbol{\alpha}$.

$$\begin{aligned}\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q(\mathbf{W})Q(\mathbf{X})} &= \frac{d}{2} \sum_i^q \log \alpha_i - \frac{1}{2} \sum_i^q \alpha_i \left\langle \|\mathbf{w}_i\|^2 \right\rangle \\ &\quad + \sum_i^q (a_\alpha - 1) \log \alpha_i - \sum_i^q b_\alpha \alpha_i + const \\ &= \sum_i^q (a_\alpha + \frac{d}{2} - 1) \log \alpha_i - \sum_i^q \left(b_\alpha + \frac{\left\langle \|\mathbf{w}_i\|^2 \right\rangle}{2} \right) \alpha_i + const \end{aligned} \tag{57}$$

Since we have

$$\log Q(\boldsymbol{\alpha}) = \langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q(\mathbf{W})Q(\mathbf{X})} + const \tag{58}$$

we can infer that $Q(\boldsymbol{\alpha})$ is of the form:

$$Q(\boldsymbol{\alpha}) = \prod_i^q Q(\alpha_i) \tag{59}$$

$$= \prod_i^q \mathcal{G}\left(\alpha_i; \hat{a}_\alpha, \hat{b}_\alpha^{(i)}\right) \tag{60}$$

where

$$\hat{a}_\alpha = a_\alpha + \frac{d}{2} \tag{61}$$

$$\hat{b}_\alpha^{(i)} = b_\alpha + \frac{\left\langle \|\mathbf{w}_i\|^2 \right\rangle}{2} \tag{62}$$

5.2 Estimation of $Q(\mathbf{W})$

In order to estimate $Q(\mathbf{W})$ we start with eq. (56) and retain only the terms that contain $\mathbf{W}$

$$\log p(\mathcal{D}, \boldsymbol{\theta}) = -\frac{1}{2} \sum_i^N (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i) - \frac{1}{2} \sum_i^q \alpha_i \mathbf{w}_i^T \mathbf{w}_i + const \tag{63}$$

This equation can be rewritten as follows:

$$\begin{aligned}\log p(\mathcal{D}, \boldsymbol{\theta}) &= -\frac{1}{2} \sum_i^N \sum_k^d p_k \left(t_{ik} - \mathbf{w}_k^T \mathbf{x}_i\right)^2 - \frac{1}{2} \sum_k^d \mathbf{w}_k^T \mathbf{A} \mathbf{w}_k + const \\ &= -\frac{1}{2} \sum_i^N \sum_k^d p_k \left(t_{ik}^2 - 2 t_{ik} \mathbf{w}_k^T \mathbf{x}_i + \mathbf{w}_k^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{w}_k\right) - \frac{1}{2} \sum_k^d \mathbf{w}_k^T \mathbf{A} \mathbf{w}_k + const \\ &= -\frac{1}{2} \sum_k^d p_k \left[\sum_i^N t_{ik}^2 - 2 \mathbf{w}_k^T \left(\sum_i^N t_{ik} \mathbf{x}_i \right) + \mathbf{w}_k^T \left(\sum_i^N \mathbf{x}_i \mathbf{x}_i^T + \frac{1}{p_k} \mathbf{A} \right) \mathbf{w}_k \right] + const \end{aligned} \tag{64}$$

where $\mathbf{w}_k$ is a column vector corresponding to the k^{th} row of $\mathbf{W}$, $\mathbf{A} = \text{diag}\,[\boldsymbol{\alpha}]$, and p_k is the k^{th} diagonal element of $\boldsymbol{\Psi}^{-1}$. Taking expectations according to $Q(\mathbf{X})Q(\boldsymbol{\alpha})$ we get:

$$\begin{aligned}\left\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \right\rangle_{Q(\mathbf{X})Q(\boldsymbol{\alpha})} = -\frac{1}{2} \sum_k^d p_k \left[\sum_i^N t_{ik}^2 - 2\mathbf{w}_k^T \left(\sum_i^N t_{ik} \left\langle \mathbf{x}_i \right\rangle \right) \right. \\ \left. + \mathbf{w}_k^T \left(\sum_i^N \left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle + \frac{1}{p_k} \left\langle \mathbf{A} \right\rangle \right) \mathbf{w}_k \right] + const \end{aligned} \tag{65}$$

Since this is a quadratic in $\mathbf{w}_k$ and since we know that

$$\log Q(\mathbf{W}) = \left\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \right\rangle_{Q(\mathbf{X})Q(\boldsymbol{\alpha})} + const \tag{66}$$

we can infer that $Q(\mathbf{W})$ is of the form:

$$Q(\mathbf{W}) = \prod_k^d Q(\mathbf{w}_k) \tag{67}$$

$$= \prod_k^d \mathcal{N}\left(\mathbf{w}_k; \mathbf{m}_{\mathbf{w}}^{(k)}, \boldsymbol{\Sigma}_{\mathbf{w}}^{(k)} \right) \tag{68}$$

where

$$\boldsymbol{\Sigma}_{\mathbf{w}}^{(k)} = \left(p_k \sum_i^N \left\langle \mathbf{x}_i \mathbf{x}_i^T \right\rangle + \left\langle \mathbf{A} \right\rangle \right)^{-1} \tag{69}$$

$$\mathbf{m}_{\mathbf{w}}^{(k)} = p_k \boldsymbol{\Sigma}_{\mathbf{w}}^{(k)} \left(\sum_i^N t_{ik} \left\langle \mathbf{x}_i \right\rangle \right) \tag{70}$$

5.3 Estimation of $Q(\mathbf{X})$

Take expectations of eq. (56) w.r.t. $Q(\mathbf{W})Q(\boldsymbol{\alpha})$ and (for simplicity) retain only the terms that contain $\mathbf{x}_i$.

$$\begin{aligned}\left\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \right\rangle_{Q(\mathbf{W})Q(\boldsymbol{\alpha})} &= -\frac{1}{2} \sum_i^N \left(\mathbf{t}_i^T \boldsymbol{\Psi}^{-1} \mathbf{t}_i - 2\mathbf{x}_i^T \left\langle \mathbf{W}^T \right\rangle \boldsymbol{\Psi}^{-1} \mathbf{t}_i + \mathbf{x}_i^T \left\langle \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \right\rangle \mathbf{x}_i \right) \\ &\quad - \frac{1}{2} \sum_i^N \mathbf{x}_i^T \mathbf{x}_i \end{aligned} \tag{71}$$

$$= -\frac{1}{2} \sum_i^N \left(\mathbf{t}_i^T \boldsymbol{\Psi}^{-1} \mathbf{t}_i - 2\mathbf{x}_i^T \left\langle \mathbf{W}^T \right\rangle \boldsymbol{\Psi}^{-1} \mathbf{t}_i + \mathbf{x}_i^T \left(\mathbf{I} + \left\langle \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \right\rangle \right) \mathbf{x}_i \right) \tag{72}$$

Since we have

$$\log Q(\mathbf{X}) = \langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q(\mathbf{W})Q(\boldsymbol{\alpha})} + const \tag{73}$$

we can infer that $Q(\mathbf{X})$ is of the form:

$$Q(\mathbf{X}) = \prod_i^N Q(\mathbf{x}_i) \tag{74}$$

$$= \prod_i^N \mathcal{N}\left(\mathbf{x}_i; \mathbf{m}_{\mathbf{x}}^{(i)}, \boldsymbol{\Sigma}_{\mathbf{x}}\right) \tag{75}$$

where

$$\boldsymbol{\Sigma}_{\mathbf{x}} = \left(\mathbf{I} + \left\langle \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \right\rangle\right)^{-1} \tag{76}$$

$$\mathbf{m}_{\mathbf{x}}^{(i)} = \boldsymbol{\Sigma}_{\mathbf{x}} \left\langle \mathbf{W}^T \right\rangle \boldsymbol{\Psi}^{-1} \mathbf{t}_i \tag{77}$$

5.4 Calculation of the required expectations

Most of the moments required in the update equations can be obtained directly given the form of the distributions. For a Gaussian distribution for example, the expectation is simply the mean of the distribution. A slightly non-trivial expectation is $\left\langle \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \right\rangle$ which occurs in eq. (76). To compute this expectation we use the fact that $\boldsymbol{\Psi}^{-1}$ is a diagonal matrix as follows:

$$\left\langle \mathbf{W}^T \boldsymbol{\Psi}^{-1} \mathbf{W} \right\rangle = \left\langle \sum_k^d \frac{1}{p_k} \mathbf{w}_k \mathbf{w}_k^T \right\rangle$$

$$= \sum_k^d \frac{1}{p_k} \left\langle \mathbf{w}_k \mathbf{w}_k^T \right\rangle \tag{78}$$

Where $\mathbf{w}_k$ is the k^{th} row of $\mathbf{W}$, and p_k is the k^{th} diagonal element of $\boldsymbol{\Psi}$. Since we have shown that the distribution of each row of $\mathbf{W}$ is Gaussian with covariances and means as shown in eqs. (69) and (70) we can compute the required second order moments trivially.

5.5 Maximization equation for $\boldsymbol{\Psi}$

The noise covariance matrix is estimated using the standard EM algorithm. Differentiating the expectation of Eq. (56) w.r.t. $\boldsymbol{\Psi}^{-1}$, we get:

$$\boldsymbol{\Psi} = \frac{1}{N} \sum_i^N \left\langle (\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)(\mathbf{t}_i - \mathbf{W}\mathbf{x}_i)^T \right\rangle \tag{79}$$

$$= \frac{1}{N} \left[\sum_i^N \mathbf{t}_i \mathbf{t}_i^T - 2 \left\langle \mathbf{W} \right\rangle \left(\sum_i^N \left\langle \mathbf{x}_i \right\rangle \mathbf{t}_i^T \right) + \sum_i^N \left\langle \mathbf{W}\mathbf{x}_i \mathbf{x}_i^T \mathbf{W}^T \right\rangle \right] \tag{80}$$

Our only difficulty is in computing the term $\sum_i^N \left\langle \mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T \right\rangle$. We can work our way around this by noting that only the diagonal terms of $\boldsymbol{\Psi}$ are of interest. For this term, the k^{th} diagonal element can be written as follows:

$$\text{diag}_k\left[\sum_i^N \left\langle \mathbf{W}\mathbf{x}_i\mathbf{x}_i^T\mathbf{W}^T \right\rangle\right] = \sum_i^N \left\langle \mathbf{w}_k^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w}_k \right\rangle \tag{81}$$

$$= \text{Tr}\left[\left\langle \mathbf{w}_k\mathbf{w}_k^T \right\rangle \sum_i^N \left\langle \mathbf{x}_i\mathbf{x}_i \right\rangle\right] \tag{82}$$

$$\begin{aligned} = \text{Tr}&\left[\boldsymbol{\Sigma}_{\mathbf{w}}^k\left(N\boldsymbol{\Sigma}_{\mathbf{x}} + \sum_i^N \left\langle \mathbf{x}_i \right\rangle \left\langle \mathbf{x}_i \right\rangle^T\right)\right] \\ &+ \left\langle \mathbf{w}_k \right\rangle \left(N\boldsymbol{\Sigma}_{\mathbf{x}} + \sum_i^N \left\langle \mathbf{x}_i \right\rangle \left\langle \mathbf{x}_i \right\rangle^T\right) \left\langle \mathbf{w}_k \right\rangle^T \end{aligned} \tag{83}$$

5.6 Measuring success: High $\mathcal{F}(Q)$ or low $KL\left\{Q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathcal{D})\right\}$?

It must be pointed out that the variational framework is not inherently an approximation in itself. Free-form maximization of the functional $\mathcal{F}(Q)$ results in the solution that effectively reduces $KL\left\{Q(\boldsymbol{\theta})\|p(\boldsymbol{\theta}|\mathcal{D})\right\}$ to zero. However, it must be pointed out that *we* introduce the approximation, when we restrict Q to have a factorized form. Given this restriction, a high value of $\mathcal{F}(Q)$ does not necessarily mean that we have obtained a good approximation to the form of the posterior distribution. Conversely, a low value for the KL divergence between Q and the true posterior does not imply that we will have the tightest bound on the log-evidence of the data.

Which quantity then, is an appropriate measure of success? If we are doing Bayesian inference, then we are most likely interested in obtaining an approximation to the posterior that is as good as possible. In this case, the KL divergence is the true measure of how well we have inferred our model parameters from the data. If however, we are doing model comparison, then we would probably be more interested in achieving the tightest possible bound on the log-evidence of the data under each of the models we are comparing so that we can accurately judge the suitability of each model to the dataset at hand.

6 Modelling Nonlinear Manifolds — Mixture Models

Factor analysis is a linear model. This very fact makes it unsuitable for modelling non-linear manifolds. By using multiple factor analyzers however, we can create a soft partition of the data space such that each factor analyzer models a locally linear region of the manifold.

Using a mixture of factor analysis requires considerable extension to our graphical model as we can see in fig. 4. Each individual factor analyzer must automatically position itself within the input space (by adjusting its mean vector), and determine its appropriate factor loading matrix. In addition, each factor analyzer must also determine its own dimensionality. These

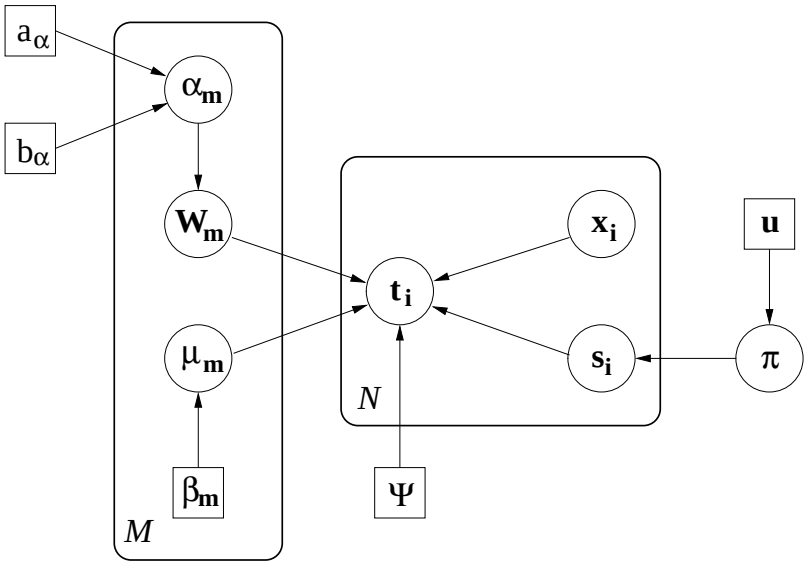

Figure 4: Graphical model for learning a mixture of factor analyzers with local dimensionality estimation

modifications are evident in the fact that we have now placed the $\boldsymbol{\mu}$, $\mathbf{W}$, and $\boldsymbol{\alpha}$ variables within a plate that indexes over the M factor analyzers in our mixture model.

In the previous sections we have glossed over the determination of the $\boldsymbol{\mu}$ parameter since we were dealing with a single factor analyser, and could subtract the sample mean from the data before performing our analysis. Indeed, if we had actually decided to perform computations for the $\boldsymbol{\mu}$ variable in the preceeding models we would have arrived at exactly the equations for the sample mean of the data (i.e. $\frac{1}{N}\sum_i^N \mathbf{t}_i$). In this section, we can no longer ignore the means since we are adapting the positions of our factor analyzers to account for local linearities within the data. Each mean is assigned a circular Gaussian prior with zero mean and precision (inverse covariance) β_m.

In addition to parameterizing each model, we introduce a hidden variable $\mathbf{s}_i$ for each data point that indicates which mixture component generated the data. The distribution of $\mathbf{s}_i$ is multinomial with a probability vector $\boldsymbol{\pi}$, which in turn is modelled by a conjugate Dirichlet distribution parameterized by $\mathbf{u}$.

As with the case of the single factor analyzer, we shall first try and determine an expression for $p(\mathcal{D}, \boldsymbol{\theta})$.

$$p(\mathcal{D}, \boldsymbol{\theta}) = \left[\prod_i^N p(\mathbf{t}_i|\mathbf{x}_i, \mathbf{s}_i, \mathbf{W}, \boldsymbol{\mu}) p(\mathbf{x}_i) p(\mathbf{s}_i|\boldsymbol{\pi})\right] p(\boldsymbol{\pi}) \left[\prod_m^M p(\mathbf{W}_m|\boldsymbol{\alpha}_m) p(\boldsymbol{\alpha}_m) p(\boldsymbol{\mu}_m)\right] \tag{84}$$

$$= \left[\prod_i^N \prod_m^M \left[p(\mathbf{t}_i|\mathbf{x}_i, s_{im}, \mathbf{W}_m, \boldsymbol{\mu}_m) p(\mathbf{x}_i) p(s_{im}|\boldsymbol{\pi}) \right]^{s_{im}} \right] p(\boldsymbol{\pi}) \tag{85}$$

$$\cdot \left[\prod_m^M p(\mathbf{W}_m|\boldsymbol{\alpha}_m) p(\boldsymbol{\alpha}_m) p(\boldsymbol{\mu}_m) \right] \tag{86}$$

And hence

$$\begin{aligned} \log p(\mathcal{D}, \boldsymbol{\theta}) = & \sum_i^N \sum_m^M s_{im} \left[\log p(\mathbf{t}_i|\mathbf{x}_i, s_{im}, \mathbf{W}_m, \boldsymbol{\mu}_m) + \log p(\mathbf{x}_i) + \log p(s_{im}|\boldsymbol{\pi}) \right] \\ & + \log p(\boldsymbol{\pi}) + \sum_m^M \left[\log p(\mathbf{W}_m|\boldsymbol{\alpha}_m) + \log p(\boldsymbol{\alpha}_m) + \log p(\boldsymbol{\mu}_m) \right] \end{aligned} \tag{87}$$

$$\begin{aligned} = & -\frac{1}{2} \sum_i^N \sum_m^M s_{im} (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m) \\ & - \frac{1}{2} \sum_i^N \sum_m^M s_{im} \mathbf{x}_i^T \mathbf{x}_i + \sum_i^N \sum_m^M s_{im} \log \boldsymbol{\pi}_m \\ & + \sum_m^M (u_m - 1) \log \boldsymbol{\pi}_m \\ & + \sum_m^M \sum_i^q \frac{d}{2} \log \alpha_{im} - \frac{1}{2} \sum_m^M \sum_i^q \alpha_{im} \mathbf{w}_{im}^T \mathbf{w}_{im} \\ & + \sum_m^M \sum_i^q (a_\alpha - 1) \log \alpha_{im} - \sum_m^M \sum_i^q b_\alpha \alpha_{im} \\ & - \frac{1}{2} \sum_m^M \beta_m \boldsymbol{\mu}_m^T \boldsymbol{\mu}_m + const \end{aligned} \tag{88}$$

We assume the factorization

$$\begin{aligned} Q(\mathbf{X}, \mathbf{S}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\mu}) &= Q(\mathbf{X}, \mathbf{S}) Q(\mathbf{W}) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\pi}) Q(\boldsymbol{\mu}) \\ &= Q(\mathbf{X}|\mathbf{S}) Q(\mathbf{S}) Q(\mathbf{W}) Q(\boldsymbol{\alpha}) Q(\boldsymbol{\pi}) Q(\boldsymbol{\mu}) \end{aligned} \tag{89}$$

6.1 Estimation of $Q(\boldsymbol{\pi})$

Taking expectations of eq. (88) w.r.t. all distributions except $Q(\boldsymbol{\pi})$ and retaining only the terms containing $\boldsymbol{\pi}$ we get:

$$\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}-\{\boldsymbol{\pi}\}}} = \sum_i^N \sum_m^M \langle s_{im} \rangle \log \boldsymbol{\pi}_m + \sum_m^M (u_m - 1) \log \boldsymbol{\pi}_m + const \tag{90}$$

$$= \sum_m^M \left(u_m + \sum_i^N \langle s_{im} \rangle - 1 \right) \log \boldsymbol{\pi}_m + const \tag{91}$$

We can infer:

$$Q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}; \hat{\mathbf{u}})^{\text{XIII}} \tag{92}$$

where

$$\hat{u}_m = u_m + \sum_i^N \langle s_{im} \rangle \tag{93}$$

6.2 Estimation of $Q(\boldsymbol{\alpha})$

Taking expectations of eq. (88) w.r.t. all distributions except $Q(\boldsymbol{\alpha})$ and retaining only the terms containing $\boldsymbol{\alpha}$ we get:

$$\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}-\{\boldsymbol{\alpha}\}}} = \sum_m^M \sum_i^q \frac{d}{2} \log \alpha_{im} - \frac{1}{2} \sum_m^M \sum_i^q \alpha_{im} \left\langle \|\mathbf{w}_{im}\|^2 \right\rangle$$
$$+ \sum_m^M \sum_i^q (a_\alpha - 1) \log \alpha_{im} - \sum_m^M \sum_i^q b_\alpha \alpha_{im} + const \tag{94}$$

$$= \sum_m^M \sum_i^q \left(a_\alpha + \frac{d}{2} - 1 \right) \log \alpha_{im} - \sum_m^M \sum_i^q \left(b_\alpha + \frac{\left\langle \|\mathbf{w}_{im}\|^2 \right\rangle}{2} \right) \alpha_{im} + const \tag{95}$$

$$Q(\boldsymbol{\alpha}) = \prod_m^M \prod_i^q Q(\alpha_{im}) \tag{96}$$

$$= \prod_m^M \prod_i^q \mathcal{G}\left(\alpha_{im}; \hat{a}_\alpha, \hat{b}_\alpha^{(im)} \right) \tag{97}$$

[XIII] We use the notation $\mathcal{D}(\boldsymbol{\pi}; \mathbf{u})$ to denote the Dirichlet distribution which is mathematically defined as:

$$\mathcal{D}(\boldsymbol{\pi}; \mathbf{u}) = \frac{\Gamma\left(\sum_m^M u_m\right)}{\prod_m^M \Gamma(u_m)} \prod_m^M \boldsymbol{\pi}_m^{u_m - 1}$$

where

$$\hat{a}_\alpha = a_\alpha + \frac{d}{2} \tag{98}$$

$$\hat{b}_\alpha^{(im)} = b_\alpha + \frac{\left\langle \|\mathbf{w}_{im}\|^2 \right\rangle}{2} \tag{99}$$

6.3 Estimation of $Q(\boldsymbol{\mu})$

$$\begin{aligned}
\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}-\{\boldsymbol{\mu}\}}} &= -\frac{1}{2} \sum_i^N \sum_m^M \langle s_{im} \rangle \left\langle (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m) \right\rangle \\
&\quad - \frac{1}{2} \sum_m^M \beta_m \boldsymbol{\mu}_m^T \boldsymbol{\mu}_m + const \qquad (100) \\
&= -\frac{1}{2} \sum_i^N \sum_m^M \langle s_{im} \rangle \left[\boldsymbol{\mu}_m^T \boldsymbol{\Psi}^{-1} \boldsymbol{\mu}_m - 2 \boldsymbol{\mu}_m^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle) \right. \\
&\quad \left. + (\mathbf{t}_i - \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle) \right] \qquad (101) \\
&\quad - \frac{1}{2} \sum_m^M \beta_m \boldsymbol{\mu}_m^T \boldsymbol{\mu}_m + const \\
&= -\frac{1}{2} \sum_m^M \left[\boldsymbol{\mu}_m^T \left(\beta_m \mathbf{I} + \boldsymbol{\Psi}^{-1} \sum_i^N \langle s_{im} \rangle \right) \boldsymbol{\mu}_m \right. \\
&\quad - 2 \boldsymbol{\mu}_m^T \boldsymbol{\Psi}^{-1} \sum_i^N \langle s_{im} \rangle (\mathbf{t}_i - \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle) \\
&\quad \left. + \sum_i^N \langle s_{im} \rangle (\mathbf{t}_i - \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle) \right] + const \qquad (102)
\end{aligned}$$

We can deduce

$$\begin{aligned}
Q(\boldsymbol{\mu}) &= \prod_m^M Q(\boldsymbol{\mu}_m) \\
&= \prod_m^M \mathcal{N}\left(\boldsymbol{\mu}_m; \mathbf{m}_{\boldsymbol{\mu}}^m, \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^m\right) \qquad (103)
\end{aligned}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{m} = \left(\beta_m \mathbf{I} + \boldsymbol{\Psi}^{-1} \sum_{i}^{N} \langle s_{im} \rangle \right)^{-1} \tag{104}$$

$$\mathbf{m}_{\boldsymbol{\mu}}^{m} = \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{m} \boldsymbol{\Psi}^{-1} \sum_{i}^{N} \langle s_{im} \rangle \left(\mathbf{t}_i - \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle \right) \tag{105}$$

6.4 Estimation of $Q(\mathbf{W})$

$$\begin{aligned}
\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}-\{\mathbf{W}\}}} &= -\frac{1}{2} \sum_{i}^{N} \sum_{m}^{M} \langle s_{im} \rangle \left\langle (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m) \right\rangle \\
&\quad - \frac{1}{2} \sum_{m}^{M} \sum_{i}^{q} \langle \alpha_{im} \rangle \, \mathbf{w}_{im}^T \mathbf{w}_{im} + const \\
&= -\frac{1}{2} \sum_{i}^{N} \sum_{m}^{M} \langle s_{im} \rangle \sum_{k}^{d} p_k \left\langle \left(t_{ik} - \mathbf{w}_{mk}^T \mathbf{x}_i - \mu_{mk} \right)^2 \right\rangle \qquad (106) \\
&\quad - \frac{1}{2} \sum_{m}^{M} \sum_{k}^{d} \mathbf{w}_{mk}^T \langle \mathbf{A}_m \rangle \, \mathbf{w}_{mk} + const \\
&= -\frac{1}{2} \sum_{m}^{M} \sum_{k}^{d} \left[p_k \sum_{i}^{N} \langle s_{im} \rangle \left\langle \left(t_{ik} - \mathbf{w}_{mk}^T \mathbf{x}_i - \mu_{mk} \right)^2 \right\rangle + \mathbf{w}_{mk}^T \langle \mathbf{A}_m \rangle \, \mathbf{w}_{mk} \right] \\
&\quad + const \qquad (107)
\end{aligned}$$

We can deduce

$$Q(\mathbf{W}) = \prod_{m}^{M} \prod_{k}^{d} Q(\mathbf{w}_{mk}) \tag{108}$$

$$= \prod_{m}^{M} \prod_{k}^{d} \mathcal{N} \left(\mathbf{w}_{mk}; \mathbf{m}_{\mathbf{w}}^{m(k)}, \boldsymbol{\Sigma}_{\mathbf{w}}^{m(k)} \right) \tag{109}$$

where

$$\boldsymbol{\Sigma}_{\mathbf{w}}^{m(k)} = \left(p_k \sum_{i}^{N} \langle s_{im} \rangle \left\langle \mathbf{x}_i \mathbf{x}_i^T | m \right\rangle + \langle \mathbf{A}_m \rangle \right)^{-1} \tag{110}$$

$$\mathbf{m}_{\mathbf{w}}^{m(k)} = p_k \boldsymbol{\Sigma}_{\mathbf{w}}^{m(k)} \left(\sum_{i}^{N} \langle s_{im} \rangle \langle \mathbf{x}_i | m \rangle \left(t_{ik} - \langle \mu_{mk} \rangle \right) \right) \tag{111}$$

6.5 Estimation of $Q(\mathbf{X}|\mathbf{S})$

Since the distribution of $\mathbf{X}$ is conditioned on $\mathbf{S}$ we must not take expectations w.r.t. $Q(\mathbf{S})$.

$$\begin{aligned}\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}-\{\mathbf{X},\mathbf{S}\}}} = &-\frac{1}{2}\sum_i^N \sum_m^M s_{im} \left\langle (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m) \right\rangle \\ &- \frac{1}{2}\sum_i^N \sum_m^M s_{im} \mathbf{x}_i^T \mathbf{x}_i + const\end{aligned}$$

From which we can infer that

$$\log Q(\mathbf{X}|\mathbf{S}) = \log \prod_i^N Q(\mathbf{x}_i|\mathbf{s}_i) \tag{112}$$

$$= \log \prod_i^N \prod_m^M Q(\mathbf{x}_i|m)^{s_{im}} \tag{113}$$

$$= \sum_i^N \sum_m^M s_{im} \log Q(\mathbf{x}_i|m) \tag{114}$$

where

$$Q(\mathbf{x}_i|m) = \mathcal{N}\left(\mathbf{x}_i; \mathbf{m}_{\mathbf{x}}^{m(i)}, \boldsymbol{\Sigma}_{\mathbf{x}}^m\right) \tag{115}$$

and

$$\boldsymbol{\Sigma}_{\mathbf{x}}^m = \left(\mathbf{I} + \left\langle \mathbf{W}_m^T \boldsymbol{\Psi}^{-1} \mathbf{W}_m \right\rangle\right)^{-1} \tag{116}$$

$$\mathbf{m}_{\mathbf{x}}^{m(i)} = \boldsymbol{\Sigma}_{\mathbf{x}}^m \left\langle \mathbf{W}_m \right\rangle^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \langle \boldsymbol{\mu}_m \rangle) \tag{117}$$

6.6 Estimation of $Q(\mathbf{S})$

Since the distribution $Q(\mathbf{X}|\mathbf{S})$ is conditioned on $\mathbf{S}$ we can derive (see appendix B.1):

$$\log Q(\mathbf{S}) = \langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}-\{\mathbf{S}\}}} + \text{entropy}\left\{Q(\mathbf{X}|\mathbf{S})\right\} + const \tag{118}$$

Now we have

$$\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_{\boldsymbol{\theta}-\{\mathbf{S}\}}} = -\frac{1}{2} \sum_i^N \sum_m^M s_{im} \left\langle (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m)^T \boldsymbol{\Psi}^{-1} (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m) \right\rangle$$
$$- \frac{1}{2} \sum_i^N \sum_m^M s_{im} \left\langle \mathbf{x}_i^T \mathbf{x}_i | m \right\rangle + \sum_i^N \sum_m^M s_{im} \langle \log \boldsymbol{\pi}_m \rangle + const \tag{119}$$

also

$$\text{entropy} \{Q(\mathbf{X}|\mathbf{S})\} = -\int Q(\mathbf{X}|\mathbf{S}) \log Q(\mathbf{X}|\mathbf{S}) d\mathbf{X} \tag{120}$$
$$= -\int Q(\mathbf{X}|\mathbf{S}) \log \prod_i^N \prod_m^M Q(\mathbf{x}_i|m)^{s_{im}} d\mathbf{X} \tag{121}$$
$$= -\int Q(\mathbf{X}|\mathbf{S}) \sum_i^N \sum_m^M s_{im} \log Q(\mathbf{x}_i|m) d\mathbf{X} \tag{122}$$

Given that the posterior distributions of the $\mathbf{X}$ variables are independent Gaussian (see eqs. (114) and (115)) we can write:

$$\text{entropy} \{Q(\mathbf{X}|\mathbf{S})\} = -\sum_i^N \sum_m^M s_{im} \int Q(\mathbf{x}_i|m) \log Q(\mathbf{x}_i|m) d\mathbf{x}_i \tag{123}$$
$$= -\sum_i^N \sum_m^M s_{im} \text{entropy} \{Q(\mathbf{x}_i|m)\} \tag{124}$$
$$= \frac{1}{2} \sum_i^N \sum_m^M s_{im} \log |\boldsymbol{\Sigma}_{\mathbf{x}}^m| + const \tag{125}$$

From eqs. (118), (119) and (125), we can infer:

$$Q(\mathbf{S}) = \prod_i^N \prod_m^M Q(s_{im} = 1)^{s_{im}} \tag{126}$$

where

$$\begin{aligned} \log Q(s_{im} = 1) = -\frac{1}{2} \{ & \mathbf{t}_i^T \boldsymbol{\Psi}^{-1} \mathbf{t}_i - 2 \mathbf{t}_i^T \boldsymbol{\Psi}^{-1} \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle - 2 \mathbf{t}_i^T \boldsymbol{\Psi}^{-1} \langle \boldsymbol{\mu}_m \rangle \\ & + 2 \langle \boldsymbol{\mu}_m^T \rangle \boldsymbol{\Psi}^{-1} \langle \mathbf{W}_m \rangle \langle \mathbf{x}_i | m \rangle + \text{Tr} \left[\langle \mathbf{W}_m^T \boldsymbol{\Psi}^{-1} \mathbf{W}_m \rangle \langle \mathbf{x}_i \mathbf{x}_i^T | m \rangle \right] \\ & + \text{Tr} \left[\boldsymbol{\Psi}^{-1} \langle \boldsymbol{\mu}_m \boldsymbol{\mu}_m^T \rangle \right] \} + \langle \log \boldsymbol{\pi}_m \rangle - \frac{1}{2} \langle \mathbf{x}_i^T \mathbf{x}_i | m \rangle + \frac{1}{2} \log |\boldsymbol{\Sigma}_{\mathbf{x}}^m| \\ & + const \end{aligned} \tag{127}$$

The constant term in the above equation can be determined by normalizing. We have encountered all the above sufficient statistics before except for $\langle \log \boldsymbol{\pi}_m \rangle$. Given that the distribution $Q(\boldsymbol{\pi})$ is Dirichlet, we can use the result:

$$\langle \log \boldsymbol{\pi}_m \rangle = \psi(u_m) - \psi\left(\sum_j^M u_j\right) \tag{128}$$

where $\psi(\cdot)$ is the Digamma function defined as follows:

$$\psi(x) = \frac{\partial}{\partial x} \log \Gamma(x) \tag{129}$$

6.7 Maximization equation for $\boldsymbol{\Psi}$

The noise covariance matrix is estimated using the standard EM algorithm. Differentiating the expectation of Eq. (88) w.r.t. $\boldsymbol{\Psi}^{-1}$, we get:

$$\boldsymbol{\Psi} = \frac{1}{N}\sum_i^N \sum_m^M \langle s_{im} \rangle \left\langle (\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m)(\mathbf{t}_i - \mathbf{W}_m \mathbf{x}_i - \boldsymbol{\mu}_m)^T \right\rangle \tag{130}$$

$$= \frac{1}{N}\left[\sum_i^N \mathbf{t}_i \mathbf{t}_i^T - 2\sum_m^M \left(\sum_i^N \langle s_{im} \rangle \mathbf{t}_i\right) \langle \boldsymbol{\mu}_m \rangle^T + \sum_m^M \left(\sum_i^N \langle s_{im} \rangle\right)\left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^m + \langle \boldsymbol{\mu}_m \rangle \langle \boldsymbol{\mu}_m \rangle^T\right)\right.$$
$$\left. - 2\sum_m^M \langle \mathbf{W}_m \rangle \left(\sum_i^N \langle s_{im} \rangle \mathbf{x}_i \big(\mathbf{t}_i - \langle \boldsymbol{\mu}_m \rangle\big)^T\right) + \sum_i^N \sum_m^M \langle s_{im} \rangle \left\langle \mathbf{W}_m \mathbf{x}_i \mathbf{x}_i^T \mathbf{W}_m^T \right\rangle\right] \tag{131}$$

Our only difficulty is in computing the term $\sum_i^N \sum_m^M \langle s_{im} \rangle \left\langle \mathbf{W}_m \mathbf{x}_i \mathbf{x}_i^T \mathbf{W}_m^T \right\rangle$. We can work our way around this by noting that only the diagonal terms of $\boldsymbol{\Psi}$ are of interest. For this term, the k^{th} diagonal element can be written as follows:

$$\text{diag}_k\left[\sum_i^N \sum_m^M \langle s_{im} \rangle \left\langle \mathbf{W}_m \mathbf{x}_i \mathbf{x}_i^T \mathbf{W}_m^T \right\rangle\right] = \sum_i^N \sum_m^M \langle s_{im} \rangle \left\langle \mathbf{w}_{mk}^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{w}_{mk} \right\rangle \tag{132}$$

$$= \sum_m^M \text{Tr}\left[\left\langle \mathbf{w}_{mk} \mathbf{w}_{mk}^T \right\rangle \sum_i^N \langle s_{im} \rangle \langle \mathbf{x}_i \mathbf{x}_i | m \rangle\right] \tag{133}$$

Typically the matrix $\boldsymbol{\Psi}$ in Eq. (131) is diagonalized after it is computed, but by realizing that only its diagonal terms are required, we may be able to write the expression simply for the k^{th}

diagonal term in a manner which is numerically much more efficient.

$$\text{diag}_k[\boldsymbol{\Psi}] = \frac{1}{N}\sum_i^N\sum_m^M \langle s_{im}\rangle \left\langle (t_{ik} - \mathbf{w}_{mk}^T\mathbf{x}_i - \mu_{mk})^2 \right\rangle \tag{134}$$

$$\begin{aligned} &= \frac{1}{N}\Bigg[\sum_i^N t_{ik}^2 - 2\sum_m^M\left(\sum_i^N \langle s_{im}\rangle\, t_{ik}\right)\langle \mu_{mk}\rangle + \sum_m^M\left(\sum_i^N \langle s_{im}\rangle\right)\left(\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^m(k,k) + \langle\mu_{mk}\rangle^2\right) \\ &\quad - 2\sum_m^M \langle \mathbf{w}_{mk}\rangle^T\left(\sum_i^N \langle s_{im}\rangle\, \mathbf{x}_i\big(t_{ik} - \langle\mu_{mk}\rangle\big)\right) \\ &\quad + \sum_m^M \text{Tr}\left[\left\langle \mathbf{w}_{mk}\mathbf{w}_{mk}^T\right\rangle \sum_i^N \langle s_{im}\rangle \left\langle \mathbf{x}_i\mathbf{x}_i|m\right\rangle\right]\Bigg] \end{aligned} \tag{135}$$

6.8 Functional monitoring

From our definition of the functional $\mathcal{F}(Q)$ in Eq. (50), we can write:

$$\mathcal{F}(Q) = \int Q(\boldsymbol{\theta})\log\frac{p(\mathcal{D},\boldsymbol{\theta})}{Q(\boldsymbol{\theta})}d\boldsymbol{\theta} \tag{136}$$

$$= \langle \log p(\mathcal{D},\boldsymbol{\theta})\rangle_{Q(\boldsymbol{\theta})} + \text{Entropy}[Q(\boldsymbol{\theta})] \tag{137}$$

From Eq. (87) we have:

$$\begin{aligned} \langle \log p(\mathcal{D},\boldsymbol{\theta})\rangle = &\sum_i^N\sum_m^M \langle s_{im}\rangle \left[\langle \log p(\mathbf{t}_i|\mathbf{x}_i, s_{im}, \mathbf{W}_m, \boldsymbol{\mu}_m)\rangle + \langle \log p(\mathbf{x}_i)\rangle + \langle \log p(s_{im}|\boldsymbol{\pi})\rangle\right] \\ &+ \langle \log p(\boldsymbol{\pi})\rangle + \sum_m^M\left[\langle \log p(\mathbf{W}_m|\boldsymbol{\alpha}_m)\rangle + \langle \log p(\boldsymbol{\alpha}_m)\rangle + \langle \log p(\boldsymbol{\mu}_m)\rangle\right] \end{aligned} \tag{138}$$

Consider each of the above terms individually:

$$\sum_i^N\sum_m^M \langle s_{im}\rangle \langle \log p(\mathbf{t}_i|\mathbf{x}_i, s_{im}, \mathbf{W}_m, \boldsymbol{\mu}_m)\rangle \tag{139}$$

$$= -\frac{N}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_i^N\sum_m^M \langle s_{im}\rangle \left\langle (\mathbf{t}_i - \mathbf{W}_m\mathbf{x}_i - \boldsymbol{\mu}_m)^T\boldsymbol{\Psi}^{-1}(\mathbf{t}_i - \mathbf{W}_m\mathbf{x}_i - \boldsymbol{\mu}_m)\right\rangle + const \tag{140}$$

$$= -\frac{N}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_k^d \frac{1}{\text{diag}_k[\boldsymbol{\Psi}]}\underbrace{\sum_i^N\sum_m^M \langle s_{im}\rangle \left\langle (t_{ik} - \mathbf{w}_{mk}^T\mathbf{x}_i - \mu_{mk})^2\right\rangle}_{=N\text{diag}_k[\boldsymbol{\Psi}]\text{ (see Eq. (134))}} + const \tag{141}$$

$$= -\frac{N}{2}\log|\mathbf{\Psi}| - \frac{1}{2}Nd + const \tag{142}$$

$$= -\frac{N}{2}\log|\mathbf{\Psi}| + const \tag{143}$$

$$\sum_i^N \sum_m^M \langle s_{im}\rangle \langle \log p(\mathbf{x}_i)\rangle = -\frac{1}{2}\sum_i^N \sum_m^M \langle s_{im}\rangle \left\langle \mathbf{x}_i^T \mathbf{x}_i | m\right\rangle + const \tag{144}$$

$$= -\frac{1}{2}\sum_i^N \sum_m^M \langle s_{im}\rangle \operatorname{Tr}\left[\left\langle \mathbf{x}_i \mathbf{x}_i^T | m\right\rangle\right] + const \tag{145}$$

$$= -\frac{1}{2}\sum_m^M \left(\sum_i^N \langle s_{im}\rangle\right) \operatorname{Tr}\left[\mathbf{\Sigma}_{\mathbf{x}}^m\right] - \frac{1}{2}\sum_m^M \sum_i^N \langle s_{im}\rangle \langle \mathbf{x}_i | m\rangle^T \langle \mathbf{x}_i | m\rangle + const \tag{146}$$

$$\sum_i^N \sum_m^M \langle s_{im}\rangle \langle \log p(s_{im}|\boldsymbol{\pi})\rangle = \sum_i^N \sum_m^M \langle s_{im}\rangle \langle \log \pi_m\rangle \tag{147}$$

$$= \sum_i^N \sum_m^M \langle s_{im}\rangle \big(\psi(u_m) - \psi(u_0)\big) \tag{148}$$

$$= \sum_m^M \left(\sum_i^N \langle s_{im}\rangle\right) \psi(u_m) - N\psi(u_0) \tag{149}$$

$$\langle \log p(\boldsymbol{\pi})\rangle = \log\Gamma(u_0) - \sum_m^M \log\Gamma(u_m) + \sum_m^M (u_m - 1)\langle \log \pi_m\rangle \tag{150}$$

$$= \log\Gamma(u_0) - \sum_m^M \log\Gamma(u_m) + \sum_m^M (u_m - 1)\big(\psi(u_m) - \psi(u_0)\big) \tag{151}$$

$$= \log\Gamma(u_0) - \sum_m^M \log\Gamma(u_m) + \sum_m^M (u_m - 1)\psi(u_m) - \psi(u_0)\sum_m^M (u_m - 1) \tag{152}$$

$$\sum_m^M \langle \log p(\mathbf{W}_m|\boldsymbol{\alpha}_m)\rangle = \frac{d}{2}\sum_m^M \sum_i^q \langle \log \alpha_{mi}\rangle - \frac{1}{2}\sum_m^M \sum_i^q \langle \alpha_{mi}\rangle \left\langle \mathbf{w}_{mi}^T \mathbf{w}_{mi}\right\rangle \tag{153}$$

$$= \frac{d}{2}\sum_m^M \sum_i^q \Big(\psi(\hat{a}_\alpha) - \log \hat{b}_\alpha^{(mi)}\Big) - \frac{1}{2}\sum_m^M \sum_k^d \left\langle \mathbf{w}_{mk}^T \mathbf{A}_m \mathbf{w}_{mk}\right\rangle \tag{154}$$

$$= \frac{d}{2}\sum_m^M \sum_i^q \Big(\psi(\hat{a}_\alpha) - \log \hat{b}_\alpha^{(mi)}\Big) - \frac{1}{2}\sum_m^M \sum_k^d \operatorname{Tr}\left[\left\langle \mathbf{w}_{mk}\mathbf{w}_{mk}^T\right\rangle \langle \mathbf{A}_m\rangle\right] \tag{155}$$

$$= \frac{d}{2}\sum_m^M\sum_i^q\Big(\psi(\hat{a}_\alpha) - \log\hat{b}_\alpha^{(mi)}\Big) - \frac{1}{2}\sum_m^M\Big(\sum_k^d \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\mathbf{w}}^{m(k)}\left\langle\mathbf{A}_m\right\rangle\right]\Big) - \frac{1}{2}\sum_m^M\sum_k^d \left\langle\mathbf{w}_{mk}\right\rangle^T\left\langle\mathbf{A}_m\right\rangle\left\langle\mathbf{w}_{mk}\right\rangle \tag{156}$$

$$\sum_m^M\left\langle\log p(\boldsymbol{\alpha}_m)\right\rangle = \sum_m^M\sum_i^q\big(a_\alpha\log b_\alpha - \log\Gamma(a_\alpha)\big) + (a_\alpha-1)\sum_m^M\sum_i^q\left\langle\log\alpha_{mi}\right\rangle - b_\alpha\sum_m^M\sum_i^q\left\langle\alpha_{mi}\right\rangle \tag{157}$$

$$= Mq\Big(a_\alpha\log b_\alpha - \log\Gamma(a_\alpha)\Big) + (a_\alpha-1)\sum_m^M\sum_i^q\Big(\psi(\hat{a}_\alpha) - \log\hat{b}_\alpha^{(mi)}\Big) - b_\alpha\sum_m^M\sum_i^q\left\langle\alpha_{mi}\right\rangle \tag{158}$$

$$\sum_m^M\left\langle\log p(\boldsymbol{\mu}_m)\right\rangle = \frac{d}{2}\sum_m^M\log\beta_m - \frac{1}{2}\sum_m^M\beta_m\left\langle\boldsymbol{\mu}_m^T\boldsymbol{\mu}_m\right\rangle \tag{159}$$

$$= \frac{d}{2}\sum_m^M\log\beta_m - \frac{1}{2}\sum_m^M\beta_m\Big(\mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^m\right] + \left\langle\boldsymbol{\mu}_m\right\rangle^T\left\langle\boldsymbol{\mu}_m\right\rangle\Big) \tag{160}$$

We can compute the entropies[XIV] of the various distributions as follows:

$$\mathrm{Entropy}[Q(\boldsymbol{\mu})] = \sum_m^M\mathrm{Entropy}[Q(\boldsymbol{\mu}_m)] \tag{161}$$

$$= \frac{1}{2}\sum_m^M\log\left((2\pi e)^d\left|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^m\right|\right) \tag{162}$$

$$= \frac{1}{2}\sum_m^M\log\left|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^m\right| + const \tag{163}$$

[XIV] Entropies of some standard distributions:

$$\mathrm{Entropy}\left[\mathcal{N}\left(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}\right)\right] = \frac{1}{2}\log\left((2\pi e)^d\left|\boldsymbol{\Sigma}\right|\right)$$

$$\mathrm{Entropy}\left[\mathcal{G}\left(x;a,b\right)\right] = \log\Gamma(a) - (a-1)\psi(a) - \log b + a$$

$$\mathrm{Entropy}\left[\mathcal{D}\left(\mathbf{x};\mathbf{u}\right)\right] = -\log\Gamma(u_0) + \sum_m^M\log\Gamma(u_m) - \sum_m^M(u_m-1)\psi(u_m) + \psi(u_0)\sum_m^M(u_m-1)$$

$$\text{Entropy}[Q(\mathbf{W})] = \sum_m^M \sum_k^d \text{Entropy}[Q(\mathbf{w}_{mk})] \tag{164}$$

$$= \frac{1}{2} \sum_m^M \sum_k^d \log\left|\boldsymbol{\Sigma}_{\mathbf{w}}^{m(k)}\right| + const \tag{165}$$

$$\text{Entropy}[Q(\mathbf{X}, \mathbf{S})] = -\int Q(\mathbf{X}|\mathbf{S})Q(\mathbf{S})\big(\log Q(\mathbf{X}|\mathbf{S}) + \log Q(\mathbf{S})\big)d\mathbf{X}d\mathbf{S} \tag{166}$$

$$= -\int Q(\mathbf{X}|\mathbf{S})Q(\mathbf{S}) \log Q(\mathbf{X}|\mathbf{S})d\mathbf{X}d\mathbf{S} - \int Q(\mathbf{X}|\mathbf{S})Q(\mathbf{S}) \log Q(\mathbf{S})d\mathbf{X}d\mathbf{S} \tag{167}$$

$$= \langle \text{Entropy}[Q(\mathbf{X}|\mathbf{S})]\rangle_{Q(\mathbf{S})} + \text{Entropy}[Q(\mathbf{S})] \tag{168}$$

which from Eq. (125) gives us:

$$= \frac{1}{2} \sum_i^N \sum_m^M \langle s_{im}\rangle \log|\boldsymbol{\Sigma}_{\mathbf{x}}^m| - \sum_i^N \sum_m^M \langle s_{im}\rangle \log\langle s_{im}\rangle \tag{169}$$

$$= \frac{1}{2} \sum_m^M \Big(\sum_i^N \langle s_{im}\rangle\Big) \log|\boldsymbol{\Sigma}_{\mathbf{x}}^m| - \sum_i^N \sum_m^M \langle s_{im}\rangle \log\langle s_{im}\rangle \tag{170}$$

$$\text{Entropy}[Q(\boldsymbol{\alpha})] = \sum_m^M \sum_i^q \text{Entropy}[Q(\boldsymbol{\alpha})] \tag{171}$$

$$= \sum_m^M \sum_i^q \log\Gamma(\hat{a}_\alpha) - (\hat{a}_\alpha - 1)\psi(\hat{a}_\alpha) - \log\hat{b}_\alpha^{(mi)} + \hat{a}_\alpha \tag{172}$$

$$= Mq\Big(\log\Gamma(\hat{a}_\alpha) - (\hat{a}_\alpha - 1)\psi(\hat{a}_\alpha) + \hat{a}_\alpha\Big) - \sum_m^M \sum_i^q \log\hat{b}_\alpha^{(mi)} \tag{173}$$

$$\text{Entropy}[Q(\boldsymbol{\pi})] = -\log\Gamma(u_0) + \sum_m^M \log\Gamma(u_m) - \sum_m^M (u_m - 1)\psi(u_m) + \psi(u_0)\sum_m^M (u_m - 1) \tag{174}$$

6.8.1 Functional contribution of each model

In order to assess when a model needs to be split, we need to estimate the contribution of each model to the overall functional. The functional is written as follows:

$$\mathcal{F}(Q) = \int d\boldsymbol{\theta} Q(\boldsymbol{\theta}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} \tag{175}$$

$$\begin{aligned} &= \int d\boldsymbol{\pi} Q(\boldsymbol{\pi}) \log \frac{p(\boldsymbol{\pi})}{Q(\boldsymbol{\pi})} + \sum_m^M \int d\boldsymbol{\mu}_m Q(\boldsymbol{\mu}_m) \log \frac{p(\boldsymbol{\mu}_m)}{Q(\boldsymbol{\mu}_m)} \\ &\quad + \sum_m^M \int d\boldsymbol{\alpha}_m Q(\boldsymbol{\alpha}_m) \left[\log \frac{p(\boldsymbol{\alpha}_m)}{Q(\boldsymbol{\alpha})} + \int d\mathbf{W}_m Q(\mathbf{W}_m) \log \frac{p(\mathbf{W}_m|\boldsymbol{\alpha}_m)}{Q(\mathbf{W}_m)} \right] \\ &\quad + \sum_i^N \sum_m^M \langle s_{im} \rangle \left[\int d\boldsymbol{\pi} Q(\boldsymbol{\pi}) \log \frac{p(m|\boldsymbol{\pi})}{Q(m)} + \int d\mathbf{x}_i Q(\mathbf{x}_i|m) \log \frac{p(\mathbf{x}_i)}{Q(\mathbf{x}_i|m)} \right. \\ &\quad \left. + \int d\boldsymbol{\mu}_m Q(\boldsymbol{\mu}_m) \int d\mathbf{W}_m Q(\mathbf{W}_m) \int d\mathbf{x}_i Q(\mathbf{x}_i|m) p(\mathbf{t}_i|\mathbf{x}_i, m, \mathbf{W}_m, \boldsymbol{\mu}_m) \right] \end{aligned} \tag{176}$$

The last term in the expression (normalized by $\sum_i^N \langle s_{im} \rangle$) is the contribution of each model to the functional.

A Useful matrix algebra results

A.1 Matrix inversion theorem

The famous Sherman-Morrison-Woodbury theorem:

$$(\mathbf{A} + \mathbf{XRY})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{X}(\mathbf{R}^{-1} + \mathbf{YA}^{-1}\mathbf{X})^{-1}\mathbf{YA}^{-1}$$

A.2 Another useful result

$$\begin{aligned} \mathbf{A} &= \mathbf{W}^T(\boldsymbol{\Psi} + \mathbf{W}\mathbf{W}^T)^{-1} \\ &= \mathbf{W}^T \left[\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\mathbf{W}(\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{\Psi}^{-1} \right] \\ &= \left[\mathbf{I} - \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W}(\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1} \right] \mathbf{W}^T\boldsymbol{\Psi}^{-1} \\ &= \left[\mathbf{I} + (\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1} - (\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W})(\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1} \right] \mathbf{W}^T\boldsymbol{\Psi}^{-1} \\ &= (\mathbf{I} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{\Psi}^{-1} \end{aligned}$$

In a simliar fashion, we can also prove the more general result

$$\mathbf{W}^T(\boldsymbol{\Psi} + \mathbf{WAW}^T)^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{W}^T\boldsymbol{\Psi}^{-1}\mathbf{W})^{-1}\mathbf{W}^T\boldsymbol{\Psi}^{-1}$$

B Use of the factorial variational approximation

We lower bound the log evidence using Jensen's inequality as follows:

$$\log p(\mathcal{D}) = \log \int p(\mathcal{D}, \boldsymbol{\theta}) d\boldsymbol{\theta} \tag{177}$$

$$= \log \int Q(\boldsymbol{\theta}) \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} \tag{178}$$

$$\geq \int Q(\boldsymbol{\theta}) \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q(\boldsymbol{\theta})} d\boldsymbol{\theta} = \mathcal{F}(Q) \tag{179}$$

Maximizing the lower bound implies maximizing the functional $\mathcal{F}(Q)$ over the space of probability distributions $Q(\boldsymbol{\theta})$. If we assume that $Q(\boldsymbol{\theta})$ factors over the individual variables θ_i, then we can write $Q(\boldsymbol{\theta}) = \prod_i Q_i(\theta_i)$. Let us consider a simple example in which $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3\}$, and $Q(\boldsymbol{\theta}) = Q_1(\theta_1)Q_2(\theta_2)Q_3(\theta_3)$. For the ease of notation, we shall use the symbol Q_i to denote $Q_i(\theta_i)$. Hence for our current example, our functional $\mathcal{F}(Q)$ is of the form:

$$\mathcal{F}(Q) = \int Q_1 Q_2 Q_3 \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q_1 Q_2 Q_3} d\theta_1 d\theta_2 d\theta_3 \tag{180}$$

From the calculus of variations we know that maximizing $\mathcal{F}(Q)$ is actually a constrained maximization since we must ensure that $\int Q(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$. This constraint can be incorporated into the integrand by the use of Lagrange multipliers. To this end we define a new function $z(\boldsymbol{\theta})$ as follows:

$$z(\boldsymbol{\theta}) = \int_{-\infty}^{\boldsymbol{\theta}} Q_1(\theta_1')Q_2(\theta_2')Q_3(\theta_3') d\theta_1' d\theta_2' d\theta_3' \tag{181}$$

Giving us the differential constraint:

$$\dot{z} - Q_1 Q_2 Q_3 = 0 \tag{182}$$

with the end point constraints being $z(-\infty) = 0$ and $z(\infty) = 1$. If we let $g(Q_1, Q_2, Q_3, \boldsymbol{\theta})$ represent the integrand:

$$g(Q_1, Q_2, Q_3, \boldsymbol{\theta}) = Q_1 Q_2 Q_3 \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q_1 Q_2 Q_3} \tag{183}$$

The we can incorporate the constraint into the integral by augmenting the integrand with the help of the Lagrange multiplier λ as follows:

$$g_a(Q_1, Q_2, Q_3, \boldsymbol{\theta}, z, \lambda) = Q_1 Q_2 Q_3 \log \frac{p(\mathcal{D}, \boldsymbol{\theta})}{Q_1 Q_2 Q_3} + \lambda(\dot{z} - Q_1 Q_2 Q_3) \tag{184}$$

Maximizing the functional $\mathcal{F}(Q)$ w.r.t. each of the distributions Q_i involves solving the Euler equations:

$$\frac{\partial g_a}{\partial Q_i} - \frac{d}{d\boldsymbol{\theta}}\left(\frac{\partial g_a}{\partial \dot{Q}_i}\right) = 0 \tag{185}$$

$$\frac{\partial g_a}{\partial z} - \frac{d}{d\boldsymbol{\theta}}\left(\frac{\partial g_a}{\partial \dot{z}}\right) = 0 \tag{186}$$

Where $\dot{Q}_i = dQ_i/d\boldsymbol{\theta}$

Substituting from eq. (184) in eq. (186) we get:

$$\frac{d\lambda}{d\boldsymbol{\theta}} = \mathbf{0} \tag{187}$$

Which implies that λ is independent of $\boldsymbol{\theta} = \{\theta_1, \theta_2, \theta_3\}$. This is an important result that will be used in the following steps. Similarly substituting from eq. (184) in eq. (185) and performing the differentiation with $Q_i = Q_1$ we get:

$$Q_2Q_3\left[\log p(\mathcal{D}, \boldsymbol{\theta}) - \log Q_1 - \log Q_2Q_3\right] - Q_2Q_3 - \lambda Q_2Q_3 = 0 \tag{188}$$

Integrating the above equation w.r.t. θ_2 and θ_3 we get:

$$\langle \log p(\mathcal{D}, \boldsymbol{\theta})\rangle_{Q_2Q_3} - \log Q_1 - \int Q_2Q_3 \log Q_2Q_3 d\theta_2 d\theta_3 - 1 - \lambda = 0 \tag{189}$$

Solving for Q_1 we get:

$$Q_1 = \frac{\exp\langle \log p(\mathcal{D}, \boldsymbol{\theta})\rangle_{Q_2Q_3}}{\exp\left(1 + \lambda + \int Q_2Q_3 \log Q_2Q_3 d\theta_2 d\theta_3\right)} \tag{190}$$

From eq. (187) and the assumed factorization $Q(\boldsymbol{\theta}) = \prod_i Q_i(\theta_i)$ we know that the denominator is independent of θ_1 and can be treated as a normalizing constant. Hence in general we can express the solution for the individual Q_i that maximizes the functional $\mathcal{F}(Q)$ under the assumed factorization as:

$$Q_i(\theta_i) = \frac{\exp\langle \log p(\mathcal{D}, \boldsymbol{\theta})\rangle_{Q_{k\neq i}}}{\int \exp\langle \log p(\mathcal{D}, \boldsymbol{\theta})\rangle_{Q_{k\neq i}} d\theta_i} \tag{191}$$

B.1 Solution for partial factorization

Let us drop the assumption of complete factorization for now and examine how our solution changes when the factorization is partial. Suppose our current example had the partial factorization $Q(\boldsymbol{\theta}) = Q_{12}(\theta_1, \theta_2)Q_3(\theta_3) = Q_1(\theta_1|\theta_2)Q_2(\theta_2)Q(\theta_3)$, then if we were trying to find

a solution for Q_1 we would not be able to separate the $\log Q_1$ term out of the integral as we have done in eq. (189), due to the fact that $Q_1 = Q_1(\theta_1|\theta_2)$ and has a dependency on θ_2. Our only way out of the problem is to infer the joint distribution $Q_{12}(\theta_1, \theta_2)$, and hope to be able to factor the resulting distribution into $Q_1(\theta_1|\theta_2)Q_2(\theta_2)$.

The final solution also changes if we were trying to maximize the functional w.r.t. Q_2. We would proceed as if we assumed full factorization as before and arrive at the following equation which is analogous to eq. (189)

$$\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_1 Q_3} - \log Q_2 - \int Q_1 Q_3 \log Q_1 Q_3 d\theta_1 d\theta_3 - 1 - \lambda = 0 \tag{192}$$

In this situation however, we must keep in mind that Q_1 is actually $Q_1(\theta_1|\theta_2)$ and hence has a dependency on θ_2. Now when we solve for Q_2 we should take care to place all of the terms in the equation that have a dependency on θ_2 in the numerator. We then arrive at the equation:

$$Q_2(\theta_2) \propto \exp\left(\langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_1 Q_3} - \int Q_1(\theta_1|\theta_2) \log Q_1(\theta_1|\theta_2) d\theta_1 \right) \tag{193}$$

or equivalently

$$\log Q_2 = \langle \log p(\mathcal{D}, \boldsymbol{\theta}) \rangle_{Q_1 Q_3} + \text{entropy}\left\{ Q_1(\theta_1|\theta_2) \right\} + const \tag{194}$$

C Computing $\langle \log x \rangle$ and $\langle \log |\mathbf{X}| \rangle$ expectations from the Gamma & Wishart distributions

(Thanks to Matt Beal for the hint about the derivation)

The Gamma distribution is frequently used as a conjugate prior to a precision (inverse variance) variable. The multivariate extension of this conjugate prior to precision (inverse covariance) *matrices* is the Wishart prior. These two distributions have the following form.

$$\mathcal{G}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{(a-1)} \exp(-bx) \tag{195}$$

$$\mathcal{W}(\mathbf{X}; \nu, \mathbf{S}) = \frac{|\mathbf{S}|^{-\nu/2}}{2^{\nu d/2} \pi^{d(d-1)/4} \prod_i^d \Gamma\left(\frac{\nu+1-i}{2}\right)} |\mathbf{X}|^{(\nu-d-1)/2} \exp\left\{ -\frac{1}{2} \text{Tr}\left[\mathbf{S}^{-1}\mathbf{X}\right] \right\} \tag{196}$$

The Gamma distribution is valid over $x \geq 0$, and the Wishart is valid over all positive-definite $\mathbf{X}$. When working with these distributions, it is sometimes required to compute the expectations $\langle \log x \rangle_{\mathcal{G}}$ and $\langle \log |\mathbf{X}| \rangle_{\mathcal{W}}$.

C.1 For the Gamma distribution

Consider:

$$\mathcal{G}(x; a, b) = \frac{b^a}{\Gamma(a)} x^{(a-1)} \exp(-bx) \tag{197}$$

$$= \frac{1}{Z} x^{(a-1)} \exp(-bx) \tag{198}$$

where:

$$Z = \frac{\Gamma(a)}{b^a} \tag{199}$$

$$= \int_0^\infty x^{(a-1)} \exp(-bx) dx \tag{200}$$

Differentiating w.r.t. the parameter a we get:

$$\frac{d}{da} Z = \frac{d}{da} \int_0^\infty x^{(a-1)} \exp(-bx) dx \tag{201}$$

$$= \int_0^\infty (\log x) x^{(a-1)} \exp(-bx) dx \tag{202}$$

$$= Z \langle \log x \rangle_{\mathcal{G}} \tag{203}$$

Hence:

$$\langle \log x \rangle_{\mathcal{G}} = \frac{1}{Z} \frac{d}{da} Z \tag{204}$$

$$= \frac{d}{da} \log Z \tag{205}$$

$$= \frac{d}{da} \big(\log \Gamma(a) - a \log b \big) \tag{206}$$

$$= \psi(a) - \log b \tag{207}$$

C.2 For the Wishart distribution

Consider:

$$\mathcal{W}(\mathbf{X}; \nu, \mathbf{S}) = \frac{|\mathbf{S}|^{-\nu/2}}{2^{\nu d/2} \pi^{d(d-1)/4} \prod_i^d \Gamma\left(\frac{\nu+1-i}{2}\right)} |\mathbf{X}|^{(\nu-d-1)/2} \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left[\mathbf{S}^{-1}\mathbf{X}\right] \right\} \tag{208}$$

$$= \frac{1}{Z} |\mathbf{X}|^{(\nu-d-1)/2} \exp\left\{ -\frac{1}{2} \operatorname{Tr}\left[\mathbf{S}^{-1}\mathbf{X}\right] \right\} \tag{209}$$

where:

$$Z = \frac{2^{\nu d/2}\pi^{d(d-1)/4}\prod_i^d \Gamma\left(\frac{\nu+1-i}{2}\right)}{|\mathbf{S}|^{-\nu/2}} \tag{210}$$

$$= \int_{\mathbf{O}}^{\infty} |\mathbf{X}|^{(\nu-d-1)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\mathbf{S}^{-1}\mathbf{X}\right]\right\} d\mathbf{X} \tag{211}$$

Differentiating w.r.t. the parameter ν we get:

$$\frac{d}{d\nu}Z = \frac{d}{d\nu}\int_{\mathbf{O}}^{\infty} |\mathbf{X}|^{(\nu-d-1)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\mathbf{S}^{-1}\mathbf{X}\right]\right\} d\mathbf{X} \tag{212}$$

$$= \int_{\mathbf{O}}^{\infty} (\log|\mathbf{X}|)\, |\mathbf{X}|^{(\nu-d-1)/2} \exp\left\{-\frac{1}{2}\operatorname{Tr}\left[\mathbf{S}^{-1}\mathbf{X}\right]\right\} d\mathbf{X} \tag{213}$$

$$= Z\left\langle \log|\mathbf{X}|\right\rangle_{\mathcal{W}} \tag{214}$$

Hence:

$$\left\langle \log|\mathbf{X}|\right\rangle_{\mathcal{W}} = \frac{1}{Z}\frac{d}{d\nu}Z \tag{215}$$

$$= \frac{d}{d\nu}\log Z \tag{216}$$

$$= \frac{d}{d\nu}\left(\frac{\nu d}{2}\log 2 + \frac{d(d-1)}{4}\log\pi + \sum_i^d \log\Gamma\left(\frac{\nu+1-i}{2}\right) + \frac{\nu}{2}\log|\mathbf{S}|\right) \tag{217}$$

$$= \sum_i^d \psi\left(\frac{\nu+1-i}{2}\right) + \frac{d}{2}\log 2 + \frac{1}{2}\log|\mathbf{S}| \tag{218}$$