Notes on Graphical Filtering

HT Attias
Golden Metallic, Inc

1 Graphical Filtering

Graphical filtering (GF) refers to inferring the value of a hidden variable from observed data in a probabilistic graphical model. GF is optimal and adaptive. It is optimal in the sense that it outputs the most probable value of the inferred variable. It is adaptive in the sense that its output depends on the statistical properties of the data, as captured by the graphical model. Hence, GF requires inferring the model (or part of it) from data, a process often termed learning.

2 Learning with Maximum Likelihood

This section discusses inferring graphical models from data by maximum likelihood (ML). The idea is that we have a probability distribution over observed data variables denoted by y_n, where $n = 1 : N$ denotes time. This distribution is controlled by a parameter θ. We assume the data are i.i.d. and denote them collectively by $y = \{y_n, n = 1 : N\}$. Then

$$p(y \mid \theta) = \prod_{n=1}^{N} p(y_n \mid \theta) \tag{1}$$

The likelihood function $\mathcal{L}$ is defined by

$$\mathcal{L} = \log p(y \mid \theta) = \sum_{n=1}^{N} \log p(y_n \mid \theta) \tag{2}$$

Learning by ML means computing the parameter value $\hat{\theta}$ that maximizes the likelihood,

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}(\theta) \tag{3}$$

2.1 The Gaussian model

A K-dim Gaussian distribution over a variable y_n is parametrized by a mean μ and a precision matrix ν,

$$\mathcal{N}(y_n \mid \mu, \nu) =\mid \frac{\nu}{2\pi} \mid^{1/2} \exp\left[-\frac{1}{2}(y_n - \mu)^T \nu (y_n - \mu)\right] \tag{4}$$

where the precision is defined as the inverse of the covariance.

To infer the parameters $\theta = (\mu, \nu)$ by ML from a dataset y, we write

$$\mathcal{L} = \frac{N}{2} \log \mid \nu \mid - \frac{1}{2} \sum_{n=1}^{N} (y_n - \mu)^T \nu (y_n - \mu) \tag{5}$$

(which is correct within a constant), and maximize it w.r.t. θ by setting the derivatives to zero,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mu} &= \nu \sum_{n=1}^{N} (y_n - \mu) = 0 \\ \frac{\partial \mathcal{L}}{\partial \nu} &= \frac{N}{2} \nu^{-1} - \frac{1}{2} \sum_{n=1}^{N} (y_n - \mu)(y_n - \mu)^T = 0 \end{aligned} \tag{6}$$

which yields

$$\begin{aligned} \mu &= \frac{1}{N} \sum_{n=1}^{N} y_n \\ \nu^{-1} &= \frac{1}{N} \sum_{n=1}^{N} (y_n - \mu)(y_n - \mu)^T = \frac{1}{N} \sum_{n=1}^{N} (y_n - \mu) y_n^T \end{aligned} \tag{7}$$

The Gaussian model is easy to infer from data. But at high dimension K it becomes hard to infer accurately, since the number of parameters in ν is $K(K+1)/2$ which grows quadratically with K. Generally, when a model has more parameters than data, the parameter estimates becomes very inaccurate. For instance, if ones wanted to use a Gaussian model to describe background brain activity using $K = 275$ MEG sensor data, one would have 37950 parameters to estimate, so $N = 1000$ as in averaged evoked data would be too small. Furthermore, for covariance matrices this problem is worse since the resulting estimate is ill-conditioned and they cannot be inverted. This can be fixed by using the factor analysis model or by adding probability distribution over the parameters.

2.2 The factor analysis model

The factor analysis (FA) model is a graphical model with hidden variables, termed factors. It is defined by describing the K-dim observed data y_n as generated by an unobserved L-dim factor vector x_n via a $K \times L$ mixing matrix A

$$y_n = A x_n + u_n \tag{8}$$

where u_n is a K-dim additive noise vector. A is termed the mixing matrix, or factor loading matrix. One can use this model to describe background brain activity, where x_n are brain source signals (or linear combination thereof) and u_n is sensor noise.

The factors are modelled by a zero-mean Gaussian with unit precision, and the noise is modelled by a zero-mean Gaussian with a diagonal precision matrix λ,

$$\begin{aligned} p(x_n) &= \mathcal{N}(x_n \mid 0, I) \\ p(u_n) &= \mathcal{N}(u_n \mid 0, \lambda) \end{aligned} \tag{9}$$

Hence, the distribution of the data conditioned on the factor is

$$p(y_n \mid x_n) = \mathcal{N}(y_n \mid Ax_n, \lambda) \tag{10}$$

Denote the factors collectively by $x = \{x_n, x = 1 : N\}$, then the FA model is defined by the joint distribution over the observed and hidden variables

$$p(y, x \mid A, \lambda) = \prod_{n=1}^{N} p(y_n \mid x_n) p(x_n) \tag{11}$$

parameterized by the mixing matrix and noise precision, $\theta = \{A, \lambda\}$.

One could try inferring the model using ML by observing that the data distribution is Gaussian with mean zero and covariance $AA^T + \lambda^{-1}$. However, maximizing the likelihood function w.r.t. A, λ produces nonlinear equations which are difficult to solve. Instead, the FA model is inferred from data using an EM algorithm.

2.3 EM for inferring the FA model from data

Expectation maximization (EM) is an iterative algorithm that performs ML in graphical models with hidden variables. We start discussing EM by defining three quantities. (1) The posterior distribution in a graphical model is the probability distribution over the hidden variables conditioned on the observed data. In FA the hidden variables are the factors x, and the posterior is $p(x \mid y)$. It is computed using Bayes' rule,

$$p(x \mid y) = \frac{p(y, x)}{p(y)} \tag{12}$$

(2) The complete data likelihood l in a graphical model is the joint probability distribution over the observed and hidden variables. In FA

$$l(y, x \mid \theta) = \log p(y, x) \tag{13}$$

(3) The averaged complete data likelihood $\bar{l}$ in a graphical model is obtained by averaging the complete data likelihood over the hidden variables, w.r.t. the posterior distribution. In FA

$$\bar{l}(y \mid \theta) = El(y, x) = \int dx \, p(x \mid y) l(y, x) \tag{14}$$

The intuitive idea behind $\bar{l}$ is that, had we observed not just the data y but also the hidden variables x, we could have estimated the parameters by maximizing $l(y, x \mid \theta)$ w.r.t. θ. Since we do not, perhaps we could substitute for the true x in l some 'reasonable' value. However, having observed y, we actually know which values of x are reasonable – those that have high posterior probability $p(x \mid y)$. So one solution would be to substitute the x that maximizes the posterior, termed the MAP value. A better solution would be to substitute all possible values of x, and average over them w.r.t. to the posterior. This gives $\bar{l}$. There is also a rigorous argument (later).

Back to EM, each iteration has an E-step and an M-step. The E-step computes the posterior distribution $p(x \mid y)$ and from it the averaged complete data likelihood $\bar{l}$. The M-step maximizes $\bar{l}$ w.r.t. the parameters θ. It can be shown that each EM iteration increases the likelihood, and that the algorithm always converges. One usually starts EM by initializing the parameters to chosen values, and iterates until a convergence criterion is satisfied.

E-step. For the E-step in FA, the posterior factorizes over time

$$p(x \mid y) = \prod_{n=1}^{N} p(x_n \mid y_n) \tag{15}$$

Next, consider the log and observe that, within a constant,

$$\begin{aligned} \log p(x_n \mid y_n) &= \log p(y_n \mid x_n) + \log p(x_n) \\ &= -\frac{1}{2}(y_n - Ax_n)^T \lambda (y_n - Ax_n) - \frac{1}{2} x_n^T x_n \end{aligned} \tag{16}$$

This implies that the posterior is Gaussian. To find its mean and precision, compute the gradient; the mean is the value of x_n that makes it vanish, and the precision is the coefficient of the linear term. The gradient is

$$\frac{\partial}{\partial x_n} \log p(y_n \mid x_n) = A^T \lambda y_n - (A^T \lambda A + I) x_n \tag{17}$$

Hence

$$\begin{aligned} p(x_n \mid y_n) &= \mathcal{N}(x_n \mid \bar{x}_n, \Gamma) \\ \bar{x}_n &= \Gamma^{-1} A^T \lambda y_n \\ \Gamma &= A^T \lambda A + I \end{aligned} \tag{18}$$

Sufficient statistics. We now define quantities that are required for the M-step. They are termed sufficient statistics (SS), and are obtained from moments of the posterior. Since the posterior $p(x_n \mid y_n)$ is Gaussian, its only non-zero moments are the first and second,

$$\begin{aligned} E x_n &= \bar{x}_n \\ E x_n x_n^T &= \bar{x}_n \bar{x}_n^T + \Gamma^{-1} \end{aligned} \tag{19}$$

The SS, R_{yx} and R_{xx}, are the data-factor and the factor-factor correlations

$$\begin{aligned} R_{yx} &= E \sum_{n=1}^{N} y_n x_n^T = \sum_{n=1}^{N} y_n \bar{x}_n^T \\ R_{xx} &= E \sum_{n=1}^{N} x_n x_n^T = \sum_{n=1}^{N} \bar{x}_n \bar{x}_n^T + N\Gamma^{-1} \end{aligned} \tag{20}$$

plus $R_{xy} = R_{yx}^T$. We also define the data-data correlation

$$R_{yy} = \sum_{n=1}^{N} y_n y_n^T \tag{21}$$

M-step. For the M-step we have, within a constant

$$\begin{aligned}\bar{l} &= E\sum_{n=1}^{N}[\log p(y_n \mid x_n) + \log p(x_n)] \\ &= \frac{N}{2}\log \mid \lambda \mid -\frac{1}{2}E\sum_{n=1}^{N}(y_n - Ax_n)^T\lambda(y_n - Ax_n) \qquad (22)\end{aligned}$$

To maximize $\bar{l}$ w.r.t. the parameter we set its gradient to zero,

$$\begin{aligned}\frac{\partial \bar{l}}{\partial A} &= E\sum_{n=1}^{N}(y_n - Ax_n)x_n^T = R_{yx} - AR_{xx} = 0 \\ \frac{\partial \bar{l}}{\partial \lambda} &= \frac{N}{2}\lambda^{-1} - \frac{1}{2}E\sum_{n=1}^{N}(y_n - Ax_n)(y_n - Ax_n)^T = \frac{N}{2}\lambda^{-1} - \frac{1}{2}(R_{yy} - AR_{xy}) = 0 \qquad (23)\end{aligned}$$

Hence, we obtain

$$\begin{aligned}A &= R_{yx}R_{xx}^{-1} \\ \lambda^{-1} &= \frac{1}{N}\mathrm{diag}(R_{yy} - AR_{xy}) \qquad (24)\end{aligned}$$

Reconstruction by GF. After convergence, the factors can be reconstructed from data using their MAP value, i.e., by the spatial graphical filter

$$\bar{x}_n = \Gamma^{-1}A^T\lambda y_n \qquad (25)$$

This is the most likely value of the factor given the observed data, since it maximizes their posterior distribution $p(x_n \mid y_n)$.

Likelihood. Since $p(y_n) = \mathcal{N}(y_n \mid 0, \Sigma^{-1})$ with $\Sigma = AA^T + \lambda^{-1}$, the likelihood is

$$\mathcal{L} = \sum_{n=1}^{N}\log p(y_n) = \frac{N}{2}\log \mid \Sigma^{-1} \mid -\frac{1}{2}\sum_{n=1}^{N}y_n^T\Sigma^{-1}y_n \qquad (26)$$

within a constant. $\mathcal{L}$ must not decrease as EM iterates.

Using the matrix inversion lemma

$$\Sigma^{-1} = \lambda - \lambda A(A^T\lambda A + I)^{-1}A^T\lambda = \lambda - \lambda A\Gamma^{-1}A^T\lambda \qquad (27)$$

it can be shown that

$$\begin{aligned}\mathcal{L} &= \frac{N}{2}\log\frac{\mid \lambda \mid}{\mid \Gamma \mid} - \frac{1}{2}\sum_{n=1}^{N}y_n^T\lambda y_n + \frac{1}{2}\sum_{n=1}^{N}\bar{x}_n^T\Gamma\bar{x}_n \\ &= \frac{N}{2}\log\frac{\mid \lambda \mid}{\mid \Gamma \mid} - \frac{1}{2}\mathrm{Tr}\lambda R_{yy} + \frac{1}{2}\mathrm{Tr}\Gamma R_{xx} \qquad (28)\end{aligned}$$

(using also that $\mid \Gamma^{-1} \mid = \mid I - \lambda A\Gamma^{-1}A^T \mid$).

2.4 A derivation of EM

The previous section presented the EM algorithm for learning the FA model from data. This section provides a mathematical proof that EM performs maximum likelihood, and that its convergence is guaranteed. The proof uses a variational technique.

Whereas the proof uses the notation of the FA model, it is valid for any graphical model, if we let x and y denote the hidden and observed variables, respectively, in that model.

We begin by defining a new objective function $\mathcal{F}$, which is sometime referred to as 'free energy',

$$\mathcal{F}(\theta, q) = \int dx\; q(x \mid y) \left[\log p(y, x \mid \theta) - \log q(x \mid y)\right] \tag{29}$$

It is a function of two independent quantities: the model parameters θ, and of an arbitrary conditional distribution $q(x \mid y)$ over the hidden variables. Next, we show that maximizing $\mathcal{F}$ w.r.t. θ and q alternately results in the EM algorithm of the previous section, where the E-step maximizes w.r.t. q and the M-step w.r.t. θ. Moreover, after each maximization w.r.t. q, $\mathcal{F}$ equals the likelihood.

To maximize w.r.t. q, we first modify $\mathcal{F}$ by adding a Lagrange multiplier term, $\mathcal{F} \to \mathcal{F} + \alpha\left[\int dx\; q(x \mid y) - 1\right]$, that enforces the normalization of q. Then we set the gradient w.r.t. q to zero while keeping the parameters θ fixed,

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial q(x \mid y)} &= \log p(y, x) - \log q(x \mid y) - 1 + \alpha = 0 \\ \frac{\partial \mathcal{F}}{\partial \alpha} &= \int dx\; q(x \mid y) - 1 = 0 \end{aligned} \tag{30}$$

and obtain $q(x \mid y) = p(y, x)/e^{1-\alpha}$ and $\int dx\; q(x \mid y) = 1$, hence $e^{1-\alpha} = p(y)$ and

$$q(x \mid y) = \frac{p(y, x)}{p(y)} = p(x \mid y) \tag{31}$$

Therefore, the distribution q that maximizes $\mathcal{F}$ is the posterior distribution over the hidden variables, computed for the fixed value of θ. Notice that we have obtained this result without invoking Bayes' rule of Eq. (12). Hence, maximizing $\mathcal{F}$ w.r.t. q yields the E-step of EM.

To maximize w.r.t. θ, observe that after the previous step we have $q(x \mid y) = p(x \mid y)$, hence

$$\mathcal{F} = \bar{l}(y \mid \theta) + H_q \tag{32}$$

where $\bar{l}$ is the averaged complete data likelihood of Eq. (14), and $H_q = -\int dx\; q(x \mid y) \log q(x \mid y)$ is the entropy of q. Since the latter is independent of θ, maximizing $\mathcal{F}$ w.r.t. θ is equivalent to the maximization of $\bar{l}$, which is the M-step of EM.

Finally, we show that the E-step renders $\mathcal{F}$ equal the likelihood. It can be shown that $\mathcal{F}$ is related to the likelihood $\mathcal{L} = \log p(y)$ by

$$\mathcal{F}(\theta, q) = \mathcal{L}(\theta) - KL[q(x \mid y) \mid\mid p(x \mid y)] \tag{33}$$

where the second term on the r.h.s. is the KL (Kullback-Leibler, or information theory) distance between the arbitrary distribution q and the posterior distribution. Generally, the KL distance between $q(x)$ and $p(x)$ is defined by $KL[q(x) \mid\mid p(x)] = \int dx\; q(x)[\log q(x) - \log p(x)]$. It is always nonnegative, and vanishes when the two distributions are identical. Hence, for an arbitrary q we have the inequality $\mathcal{F} \leq \mathcal{L}$, i.e., $\mathcal{F}$ is upper bounded by $\mathcal{L}$, and following the E-step which sets $q(x \mid y) = p(x \mid y)$, we have $\mathcal{F} = \mathcal{L}$.

To summarize, we have shown that (1) both the M-step and the E-step increase $\mathcal{F}$ or leave is unchanged, and (2) after the E-step $\mathcal{F}$ equals the likelihood. Hence, each EM iteration increases the likelihood.

It remains to show that EM always converges. Let $p_0(y)$ denote the distribution of the observed data

$$p_0(y) = \frac{1}{N} \sum_{n=1}^{N} \delta(y - y_n) \tag{34}$$

Consider the KL distance between p_0 and the model distribution $p(y \mid \theta)$

$$\begin{aligned} KL(p_0 \,||\, p) &= \int dy\, p_0(y) \left[\log p_0(y) - \log p(y \mid \theta)\right] \\ &= -H_{p_0} - \frac{1}{N}\mathcal{L}(\theta) \end{aligned} \tag{35}$$

since $\int dy\, p_0(y) \log p(y \mid \theta) = (1/N) \sum_n \log p(y_n \mid \theta) = \mathcal{L}/N$. Since the entropy of p_0 is independent of θ, it follows that maximizing the likelihood w.r.t. θ is equivalent to minimizing the KL distance. Next, since the distance is nonnegative, we have the inequality

$$\mathcal{L} \leq -NH_{p_0} \tag{36}$$

i.e., the likelihood is upper bounded in the negative entropy of the observed data. Therefore, any algorithm that maximizes the likelihood, including EM, is guaranteed to converge.

However, the likelihood may have several local maxima, and generally there is no guarantee for convergence to a global maximum.

2.4.1 Exercise

Using the definition of $\mathcal{F}$ in Eq. (29), derive the likelihood of the FA model in Eq. (28).

3 Bayesian inference of models from data

The ML approach to learning computes an estimate for the model parameters from data. To make a prediction for a new data point, such as its class or the next element in its time series, one uses that parameter value.

In contrast, the Bayesian approach computes not just a single estimate, but a full probability distribution over the model parameters. This means that given the data y, one considers all possible values of the parameters θ, and for each value computes how probable it is given the data, i.e., its posterior distribution $p(\theta \mid y)$. To make a prediction for a new data point, one computes the prediction for all parameter values, then weighted-averages them w.r.t. the posterior. This approach is more robust to overfitting.

One well-known feature of the Bayesian approach is the use of priors. Since Bayes' rule implies that $p(\theta \mid y) = p(y \mid \theta)p(\theta)/p(y)$, one must specify a prior distribution $p(\theta)$ over the parameters.

Another feature of the Bayesian approach is that it makes no distinction between parameters and hidden variables. The model parameters are simply added to hidden variables of the model and treated on equal footing. In the FA model, for instance, the Bayesian approach infers from data y both the parameters A, λ and the factors x by computing the joint posterior distribution $p(x, A, \lambda \mid y)$. The parameters posterior $p(A, \lambda \mid y)$ can be obtained from it by integrating over x.

It is wasy to show that ML is actually an approximation to the Bayesian approach. This approximation, termed the maximum a-posteriori (MAP) estimate, computes from data the most probable parameter value using a flat prior. To see this, consider the log of the posterior

$$\log p(\theta \mid y) = \mathcal{L}(\theta) + \log p(\theta) + \text{const.} \tag{37}$$

where $\mathcal{L}(\theta) = \log p(y \mid \theta)$ is the likelihood, and the θ-independent term $\log p(y)$ has been replaced by a constant. The MAP estimate maximizes the posterior w.r.t. θ. When the prior is flat, $p(\theta) = \text{const.}$, this maximization reduces precisely to ML.

For a non-flat prior, the MAP estimate differs from ML. The prior is often referred to as a regularizing term, since it can regularize ML parameter estimates. For instance, whereas the ML estimate of a covariance matrix in the Gaussian may result in an ill-conditioned matrix, a suitable prior on the covariance leads to a covariance matrix that is always well-conditioned.

In simple graphical models, such as the Gaussian model

$$p(y \mid \theta) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \mu, \nu) \tag{38}$$

the posterior $p(\theta \mid y)$ can be computed analytically. However, models of interest are usually more complex, and the posterior becomes computationally intractable. Hence, approximations must be made. Sampling (Monte Carlo) techniques provide one avenue for computing approximate posteriors. Here we focus on the more recent approach suggested by variational techniques.

4 Variational Bayesian learning of the FA model

A Bayesian approach to the FA model requires prior distributions over the model parameters $\theta = (A, \lambda)$. Here we use a Gaussian prior on the mixing matrix A, and a flat prior on the diagonal noise precision matrix λ.

For A we choose a prior that factorizes over the rows. Define a^i as ith row of A. It is a column vectors whose jth element is the ij element of A, i.e.,

$$a^i_j = A_{ij} \tag{39}$$

The prior over row i is Gaussian with mean zero and precision $\lambda_i \alpha$, where α is a diagonal $L \times L$ matrix. Hence

$$p(A \mid \lambda, \alpha) = p(a^1, ..., a^K) = \prod_{i=1}^{K} \mathcal{N}(a^i \mid 0, \lambda_i \alpha) \tag{40}$$

Since α is diagonal, this prior actually describes all matrix elements A_{ij} as independent, $p(A) = \prod_{ij} \mathcal{N}(A_{ij} \mid 0, \lambda_i \alpha_j)$. The parameters α_j which parametrize the prior distribution are termed hyperparameters.

The distribution $p(A)$ is a conjugate prior, as will be shown below. A prior is termed conjugate w.r.t. a model if the corresponding posterior has the same functional form. The approximate posterior will have the form $q(A \mid y) = \prod_i \mathcal{N}(a^i \mid \bar{a}^i, \lambda_i \Psi)$, which is identical to the prior's form with the diagonal α replaced by the non-diagonal Ψ computed from the data. One motivation for using conjugate priors is that it often significantly simplifies some of the required calculations.

Next, the Bayesian approach computes the joint posterior over the factors x and parameters A, λ. For simplicity, we consider the joint posterior over x, A only and infer λ by the MAP approximation. However, $p(x, A \mid y)$ cannot be computed analytically. Specifically, it can be computed only within normalization. To see this, use Bayes' rule to write

$$\begin{aligned} p(x, A \mid y) &= \frac{1}{Z} p(y \mid x, A) p(x) p(A) \\ Z &= \int dx dA \, p(y \mid x, A) p(x) p(A) \end{aligned} \tag{41}$$

and observe that the integral which defines the normalization constant Z has no closed form solution.

We use a variational technique to approximate the posterior. The idea is to restrict its form to a product of factor distributions, where each factor is termed a variational posterior. Those factors are then computing by maximizing the objective function $\mathcal{F}$, discussed above. Here we restrict the posterior to a product of two variational posteriors, one over x and the other over A,

$$p(x, A \mid y) \approx q(x, A \mid y) = q(x \mid y) q(A \mid y) \tag{42}$$

i.e., the factors are independent of the mixing matrix given the data. This is an approximation, since truly x and A are correlated given y. The advantage is that the approximate posterior can be computed analytically.

To obtain the posteriors $q(x \mid y)$ and $q(A \mid y)$ and the MAP estimates of λ, α, consider

$$\mathcal{F}[q, \lambda, \alpha] = \int dx dA \, q(x \mid y) q(A \mid y) \left[\log p(y, x, A \mid \lambda, \alpha) - \log q(x \mid y) - \log q(A \mid y)\right] \tag{43}$$

and maximize it w.r.t. the qs and λ, α.

Maximization w.r.t. $q(x \mid y)$ yields

$$q(x \mid y) = \prod_{n=1}^{N} q(x_n \mid y_n) \tag{44}$$

where

$$\begin{aligned} q(x_n \mid y_n) &= \mathcal{N}(x_n \mid \bar{x}_n, \Gamma) \\ \bar{x}_n &= \Gamma^{-1} \bar{A}^T \lambda y_n \\ \Gamma &= \bar{A}^T \lambda \bar{A} + K \Psi^{-1} + I \end{aligned} \tag{45}$$

The sufficient statistics of the mixing matrix, $\bar{A}$ and Ψ, are computed below.

Maximization w.r.t. $q(A \mid y)$ yields

$$q(A \mid y) = q(a^1, ..., a^K \mid y) = \prod_{i=1}^{K} \mathcal{N}(a^i \mid \bar{a}^i, \lambda_i \Psi) \tag{46}$$

where the mean $\bar{a}^i = \bar{A}_{ij}$ and precision matrix Ψ are obtained by

$$\begin{aligned}\bar{A} &= R_{yx}\Psi^{-1} \\ \Psi &= R_{xx} + \alpha \qquad (47)\end{aligned}$$

The sufficient statistics of the factors, R_{yx} and R_{xx}, and the data correlation matrix R_{yy}, are given by

$$\begin{aligned}R_{yx} &= \sum_{n=1}^{N} y_n \bar{x}_n^T \\ R_{xx} &= \sum_{n=1}^{N} \bar{x}_n \bar{x}_n^T + N\Gamma \\ R_{yy} &= \sum_{n=1}^{N} y_n y_n^T \qquad (48)\end{aligned}$$

Maximization w.r.t. α and λ yields

$$\begin{aligned}\alpha^{-1} &= \mathrm{diag}\left(\frac{1}{K}\bar{A}^T\lambda\bar{A} + \Psi^{-1}\right) \\ \lambda^{-1} &= \frac{1}{N+L}\mathrm{diag}\left(R_{yy} - \bar{A}R_{yx}^T\right) \qquad (49)\end{aligned}$$

This completes the variational Bayes (VB) EM algorithm for the FA model. Finally, the objective function that this algorithm maximizes is

$$\begin{aligned}\mathcal{F} &= \frac{N+L}{2}\log \mid \lambda \mid -\frac{1}{2}\sum_{n=1}^{N} y_n^T \lambda y_n - \frac{N}{2}\log \mid \Gamma \mid +\frac{1}{2}\sum_{n=1}^{N} \bar{x}_n^T \Gamma \bar{x}_n \\ &+ \frac{K}{2}\log \mid \alpha \mid -\frac{1}{2}\mathrm{Tr}(\bar{A}^T\lambda\bar{A}\alpha) \qquad (50)\end{aligned}$$