Derivation of Empirical Bayes Matrix Factorization (EBMF) Models

Full Transcription of Notes (Pages 1–17)

November 21, 2025

1 1. Models (Pages 1–2)

1.1 Matrix Factorization Representation

We represent the $n \times p$ data matrix Y as:

$$Y = L^T F + E \quad (1.1) \tag{1}$$

where:

- L is a $k \times n$ matrix of **loadings**.
- F is a $k \times p$ matrix of **factors**.
- E is an $n \times p$ matrix of residuals (assumed normally distributed).

Simplest approaches to estimating L and F are based on Maximum Likelihood or Least Squares. These often include sparsity assumptions on L and F. Some methods have sparsity in loadings only, others in both.

1.2 The Empirical Bayes Approach

The Empirical Bayes approach exploits variational approximation methods to obtain simple algorithms that jointly estimate the **prior distributions** for both loadings and factors, as well as the loadings and factors themselves. Fitting EBMF with any prior family can be reduced to repeatedly solving the "Empirical Bayes Normal Means" (EBNM) problem with the same prior.

1.3 Summation Notation

The K-factor EBMF model has a bilinear structure. It is often helpful to define it as a sum of rank-1 matrices:

$$Y = \sum_{k=1}^{K} l_k f_k^T + E \tag{2}$$

- This represents a low-rank assumption. It supposes that high-dimensional observed data $Y_{n \times p}$ is composed of a relatively small number of underlying processes/components (K).
- If l_k is an $n \times 1$ vector (column) and f_k^T is a $1 \times p$ vector (row), then their product $l_k f_k^T$ is an $n \times p$ matrix.

This notation helps with clarity because the fitting algorithm updates the model one factor (k) at a time. The algorithm treats the contributions of other factors ($k' \neq k$) as fixed offsets.

2 2. Hierarchical Priors (Page 2)

We define the Empirical Bayes nature of the model via hierarchical priors:

$$\text{Loadings:} \quad l_{ik} \overset{iid}{\sim} g_l \in \mathcal{G}_l \quad \text{(Weights)} \tag{3}$$

$$\text{Factors:} \quad f_{jk} \overset{iid}{\sim} g_f \in \mathcal{G}_f \quad \text{(Latent features)} \tag{4}$$

The notation *iid* implies **exchangeability**. It assumes that, a priori, we have no reason to distinguish sample i and sample $i+1$ regarding their loading on factor k, other than what the distribution g_k dictates.

In classical Bayesian Factor Analysis, we might fix the prior $l_{ik} \sim \mathcal{N}(0, 1)$. In EBMF, the prior g_k is unknown. The derivation relies on **adaptive shrinkage**:

- If factor k is sparse, the estimated g_k might look like a "spike" at zero.
- If factor k is dense, g_k might be a broad Normal distribution.
- Different factors can have different priors (g_{l_k} vs g_{f_k}), allowing the model to handle a mix of sparse and dense signals simultaneously.

The user chooses a family $\mathcal{G}$ (e.g., mixture of normals, spike-and-slab), and the algorithm searches within this family to find the best g.

3 3. The Noise Model (Pages 2–4)

3.1 Gaussian Assumption

We define the residuals as:

$$E_{ij} \sim \mathcal{N}\left(0, \frac{1}{\tau_{ij}}\right) \quad \text{with } \tau = (\tau_{ij})_{n \times p} \tag{5}$$

Assuming the errors are Gaussian leads to the squared error loss function used in optimization. The probability density for a single point Y_{ij} is:

$$p(Y_{ij}|L, F, \tau) = \sqrt{\frac{\tau_{ij}}{2\pi}} \exp\left[-\frac{\tau_{ij}}{2}\left(Y_{ij} - \sum_k l_{ik} f_{jk}\right)^2\right] \tag{6}$$

3.2 Precision Structures

The paper allows flexibility by constraining the precision matrix τ.

- **Constant variance:** $\tau_{ij} = \tau$. (Simplest case).
- **Column-specific variance:** $\tau_{ij} = \tau_j$. (Standard in genomics, assumes some features are noisier than others).
- **Row-specific variance:** $\tau_{ij} = \tau_i$.

3.3 Transformation to Loss Function

Assuming i.i.d. errors, the joint likelihood is the product of individual probabilities. Taking the log of the joint likelihood:

$$\log \mathcal{L}(\Theta) = \sum_{i,j}\left[\frac{1}{2}\log(\tau_{ij}) - \frac{1}{2}\log(2\pi) - \frac{\tau_{ij}}{2}\left(Y_{ij} - \sum_k l_{ik} f_{jk}\right)^2\right] \tag{7}$$

Maximizing likelihood is equivalent to minimizing the negative log-likelihood (Loss). Dropping terms constant with respect to L and F:

$$\text{Loss}(L, F) \approx \sum_{i,j} \frac{\tau_{ij}}{2}\left(Y_{ij} - \sum_k l_{ik} f_{jk}\right)^2 - \sum_{i,j} \frac{1}{2}\log(\tau_{ij}) \tag{8}$$

Holding precision constant, minimizing loss is equivalent to minimizing the weighted sum of squared errors.

4 4. Fitting the EBMF Model (Pages 4–6)

We need to estimate all of g_l, g_f, L, F, τ. We use a 2-step approach:

1. Estimate g_l, g_f, τ by maximizing the marginal likelihood (integrating out L and F).
2. Estimate L and F using the posterior distribution given those point estimates.

4.1 Marginal Likelihood

We want to find the hyperparameters that make the observed data Y most likely:

$$\mathcal{L}(g_l, g_f, \tau) = \iint p(Y|L, F, \tau) g_l(dL) \dots g_f(dF) \tag{9}$$

By the Law of Total Probability, since Y depends on latent variables L and F, we must integrate over all possible values they can take.

$$p(Y|g, \tau) = \iint p(Y, L, F|g, \tau)\, dL\, dF \tag{10}$$

This integration is computationally difficult, motivating the use of **Variational Approximation** to avoid solving the MLE directly.

5 5. Variational Approach (Pages 6–10)

EBMF is a "Type II Maximum Likelihood" method where we are interested in finding the distributions (priors) that generated L and F, rather than just the values of L and F themselves.

5.1 Derivation of the ELBO

We want to maximize $\log p(Y|\Theta)$. Let $q(Z)$ be an arbitrary distribution over the latent variables $Z = \{L, F\}$.

$$\log p(Y|\Theta) = \int q(Z) \log p(Y|\Theta)\, dZ \quad \text{(Expectation is constant)} \tag{11}$$

$$= \int q(Z) \log \left(\frac{p(Y, Z|\Theta)}{p(Z|Y, \Theta)} \right) dZ \quad \text{(Bayes Rule)} \tag{12}$$

$$= \int q(Z) \log \left(\frac{p(Y, Z|\Theta)}{q(Z)} \cdot \frac{q(Z)}{p(Z|Y, \Theta)} \right) dZ \quad \text{(Multiply by 1)} \tag{13}$$

$$= \int q(Z) \log \left(\frac{p(Y, Z|\Theta)}{q(Z)} \right) dZ + \int q(Z) \log \left(\frac{q(Z)}{p(Z|Y, \Theta)} \right) dZ \tag{14}$$

This splits the integral into two terms:

$$\log p(Y|\Theta) = \mathcal{F}(q, \Theta) + D_{KL}(q||p) \tag{15}$$

5.2 Variational Inequality (Proof via Jensen's Inequality)

To prove that $\mathcal{F}$ is a lower bound, we show $D_{KL} \geq 0$. Using the definition of KL divergence:

$$-D_{KL}(q||p) = \int q(Z) \log \left(\frac{p(Z|Y)}{q(Z)} \right) dZ \tag{16}$$

Let $f(x) = \log(x)$, which is concave. By Jensen's Inequality, $\mathbb{E}[\log(x)] \leq \log(\mathbb{E}[x])$.

$$\int q(Z) \log \left(\frac{p(Z|Y)}{q(Z)} \right) dZ \leq \log \left(\int q(Z) \frac{p(Z|Y)}{q(Z)}\, dZ \right) = \log \left(\int p(Z|Y)\, dZ \right) = \log(1) = 0 \tag{17}$$

Thus, $-D_{KL} \leq 0 \implies D_{KL} \geq 0$. Therefore:

$$\log p(Y|\Theta) \geq \mathcal{F}(q, \Theta) \tag{18}$$

Maximizing $\mathcal{F}(q)$ is equivalent to minimizing $D_{KL}(q||p)$, meaning we find the distribution q closest to the true posterior.

5.3 Mean Field Assumption

To make the maximization feasible, we restrict q to the family of distributions that fully factorize:

$$q(L, F) = \prod_{i=1}^{n} q_{l_i}(l_i) \prod_{j=1}^{p} q_{f_j}(f_j) \implies q(L, F) = q(L)q(F) \tag{19}$$

This independence assumption simplifies the expectation of the likelihood term. Specifically, for the cross-term in the Gaussian expansion:

$$\mathbb{E}_q[l^T f] = \mathbb{E}_q[l]^T \mathbb{E}_q[f] \tag{20}$$

This separation means that when updating loadings $q(l)$, the factors f look like constants (their expectations). This converts the complex optimization into an iterative process.

6 6. Expected Squared Residuals (Pages 11–12)

To update the model, we minimize the Expected Squared Error. We cannot compute $(Y - LF)^2$ directly because L and F are distributions.

1. Expand the quadratic for a single entry $Y_{ij} \approx l_i f_j$:

$$(Y_{ij} - l_i f_j)^2 = Y_{ij}^2 - 2Y_{ij} l_i f_j + l_i^2 f_j^2 \tag{21}$$

2. Take the expectation using the independence assumption ($q(l, f) = q(l)q(f)$):

$$\mathbb{E}[(Y_{ij} - l_i f_j)^2] = Y_{ij}^2 - 2Y_{ij}\mathbb{E}[l_i]\mathbb{E}[f_j] + \mathbb{E}[l_i^2]\mathbb{E}[f_j^2] \tag{22}$$

3. Rearrange to align with the residual of means ($\mu_l = \mathbb{E}[l], \mu_f = \mathbb{E}[f]$):

$$\bar{R}_{ij}^2 = (Y_{ij} - \mu_l \mu_f)^2 - \mu_l^2 \mu_f^2 + \mathbb{E}[l^2]\mathbb{E}[f^2] \tag{23}$$

This accounts for the variance of the factors. If the model is uncertain about a loading (large $\mathbb{E}[l^2]$), the expected residual grows, forcing the model to increase estimated noise variance or shrink the priors.

7 7. The EBNM Problem (Pages 13–14)

The complex matrix factorization reduces to the "Empirical Bayes Normal Means" (EBNM) problem.

- **Observations:** $x = (x_1, \ldots, x_n)$
- **Observation Model:** $x_j | \theta_j, s_j \sim \mathcal{N}(\theta_j, s_j^2)$ (Gaussian errors with known sd s_j)
- **Prior Model:** $\theta_j \sim g \in \mathcal{G}$ (Parameters drawn from unknown prior)

7.1 Solving the EBNM Problem

1. **Estimate Prior:** Estimate $\hat{g}$ by maximizing the marginal likelihood:

$$\hat{g} = \arg\max_{g \in \mathcal{G}} \prod_j \int p(x_j | \theta_j, s_j) g(d\theta_j) \tag{24}$$

2. **Posterior:** Treat $\hat{g}$ as the true prior and compute posterior moments:

$$\text{Mean (1st moment):} \quad \bar{\theta}_j := \mathbb{E}[\theta_j | x, s, \hat{g}] \tag{25}$$

$$\text{2nd moment:} \quad \bar{\theta^2}_j := \mathbb{E}[\theta_j^2 | x, s, \hat{g}] \tag{26}$$

Note: $E[\theta^2] = E[\theta]^2 + \text{Var}(\theta)$.

8 8. Connecting EBMF and EBNM (Pages 15–17)

8.1 Main Result

Updating the posterior q_l and prior g_l for the loadings vector $l_{\cdot k}$, while holding factors f and noise τ fixed, is identical to solving an EBNM problem on pseudo-observations.

$$(\hat{g}_l, \hat{q}_l) = \text{EBNM}(\tilde{x}, s) \tag{27}$$

8.2 Deriving Pseudo-Parameters

We match terms in the ELBO with the Gaussian log-likelihood. 1. Isolate terms in ELBO:

$$\text{Likelihood} \propto -\frac{1}{2}\sum_{i,j} \tau_{ij}\mathbb{E}_q[(R_{ij}^{-k} - l_{ik}f_{jk})^2] \tag{28}$$

where R_{ij}^{-k} is the residual excluding factor k. 2. Expand square with respect to l_{ik}:

$$\propto \sum_j \tau_{ij}\left[l_{ik}R_{ij}^{-k}\mathbb{E}[f_{jk}] - \frac{1}{2}l_{ik}^2\mathbb{E}[f_{jk}^2]\right] \tag{29}$$

3. Compare to the log-likelihood of a single data point x drawn from $\mathcal{N}(l_{ik}, s^2)$:

$$\log\mathcal{N}(x|l_{ik}, s^2) \propto -\frac{1}{2s^2}l_{ik}^2 + \frac{x}{s^2}l_{ik} \tag{30}$$

4. Match Coefficients:

$$\text{For } l_{ik}^2: \quad \frac{1}{s^2} = \sum_j \tau_{ij}\mathbb{E}[f_{jk}^2] \implies s_{ik} = \left(\sum \tau_{ij}\mathbb{E}[f_{jk}^2]\right)^{-1/2} \tag{31}$$

$$\text{For } l_{ik}: \quad \frac{\tilde{x}}{s^2} = \sum_j \tau_{ij}R_{ij}^{-k}\mathbb{E}[f_{jk}] \implies \tilde{x}_{ik} = \frac{\sum \tau_{ij}R_{ij}^{-k}\mathbb{E}[f_{jk}]}{\sum \tau_{ij}\mathbb{E}[f_{jk}^2]} \tag{32}$$

8.3 Proof of Unbiasedness and Variance (Pages 16–17)

What if we knew F and τ? The estimator for l_i would be a Weighted Least Squares (WLS) regression of the data row onto the factor. The WLS solution is $\hat{\beta} = \frac{\sum wxy}{\sum wx^2}$. Here, our estimator is:

$$\hat{l}_i = \frac{\sum_j \tau_{ij}Y_{ij}f_j}{\sum_j \tau_{ij}f_j^2} \tag{33}$$

Unbiased Mean: Assume true model $Y_{ij} = l_i f_j + E_{ij}$.

$$\hat{l}_i = \frac{\sum \tau_{ij}(l_i f_j + E_{ij})f_j}{D} = l_i\frac{\sum \tau_{ij}f_j^2}{D} + \frac{\sum \tau_{ij}E_{ij}f_j}{D} \tag{34}$$

Since $\mathbb{E}[E_{ij}] = 0$, $\mathbb{E}[\hat{l}_i] = l_i$.

Variance: Let $D = \sum \tau_{ij}f_j^2$.

$$\text{Var}(\hat{l}_i) = \text{Var}\left(\frac{\sum \tau_{ij}f_jY_{ij}}{D}\right) \tag{35}$$

$$= \frac{1}{D^2}\sum_j (\tau_{ij}f_j)^2\text{Var}(Y_{ij}) \quad \text{(Independence)} \tag{36}$$

$$= \frac{1}{D^2}\sum_j \tau_{ij}^2 f_j^2\left(\frac{1}{\tau_{ij}}\right) \tag{37}$$

$$= \frac{1}{D^2}\sum_j \tau_{ij}f_j^2 = \frac{1}{D^2}\cdot D = \frac{1}{D} \tag{38}$$

Thus, $\text{Var}(\hat{l}_i) = \frac{1}{\sum_j \tau_{ij}f_j^2}$. This confirms the standard error s derived in the EBNM mapping.