Models
(1,1)
(2,1)-(2,4)
(3,1)-(3,8)
(3,11)-(3,17)

Derivation of EBMF Models

11/21/2025

- Factor analysis bayesian model (with TTE survival component)

(1.1)

- Matrix factorization represent $n \times p$ data matrix Y as:

$$Y = L^T F + E \quad (1.1)$$

where L is $k \times n$ matrix, F is $k \times p$ matrix, E is $n \times p$ matrix of residuals (assume normally distributed)

- L: "loadings"
- F: "factors"
- Simplest approaches to estimating L and F are based on maximum likelihood or least squares. Include sparsity assumptions on L and F.
- some methods have sparsity in loadings (L) only, others in both Loadings + Factors.

- 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.
  - ⇒ Empirical Bayes approach to matrix factorization
    - ↳ Fitting EBMF with any prior family can be reduced to repeatedly solving the "empirical bayes normal means" (EBNM) problem w/ same prior

- K-factor Empirical Bayes Matrix Factorization (EBMF) model

2.1 - linear structure

- Definition: $\boxed{Y = \sum_{k=1}^{K} l_k f_k^T + E}$ (instead of $Y = L^T F + E$ $\Rightarrow \sum_{k=1}^{K} l_k f_k^T$)
- caution: represents low-rank assumption. Supposes that high-dimensional observed data $Y_{n \times p}$ is composed of a relatively small number of underlying processes/components (k).

- Matrix vs. summation notation: Standard factor analysis is often written as $Y = LF^T$. This is explicitly decomposed into a sum of K rank-1 matrices. 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.

$$Y_{ij} \approx \underbrace{(l_1 f_1^T)}_{\text{effect of factor 1}} + \underbrace{(l_2 f_2^T)}_{\text{effect of factor 2}} + \cdots + \underbrace{(l_k f_k^T)}_{\text{effect of factor } k}$$

[summation of K components]

↳ cont: This summation notation helps w/ clarity because the fitting algorithm (algos 1 and 2) 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 & 2.3 - Hierarchical priors

(2.2) loadings: $l_{ik1}, \ldots, l_{ikn} \overset{iid}{\sim} g_{lk}, \; g_{lk} \in G_l$ (weights)

(2.3) factors: $f_{ik1}, \ldots, f_{ikn} \overset{iid}{\sim} g_{fk}, \; g_{fk} \in G_f$ (latent variables)
↳ selected "features"

- define Empirical Bayes nature of the model

- The notation $\overset{iid}{\sim}$ implies exchangeability. It assumes that, a priori, we have no reason to distinguish sample i and sample $i+1$ regarding their loading (weight) on factor k, other than what the distribution g_{fk} dictates.

- In classical Bayesian factor Analysis, we might fix the prior $l_{ik} \sim N(0,1)$. In EBMF, the prior g_{lk} is unknown. The "derivation" here relies on adaptive shrinkage.
  - If factor k is sparse, the estimated g_{lk} might look like a "spike" at zero.
  - If factor k is dense, g_{lk} might be a broad Normal distribution.
  - different factors (k) have different priors (g_{lk} vs. $g_{lk'}$) This allows the model to handle a mix of sparse and dense signals simultaneously.

- prior family (G): user chooses a family G ("all unimodal distributions at 0" or "spike-and-slab"). The algorithm then searches within this family to find the best g (prior).

2.4 - The Noise Model

$E_{ij} \sim N(0, \frac{1}{\tau_{ij}})$ with $\tau := (\tau_{ij}) \in T$

- defines the likelihood function

↳

- Gaussian Assumption: Assuming the errors E_{ij} are Gaussian leads to the squared error loss function used in the optimization.

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

- precision structure: paper derives flexibility by constraining the precision matrix τ. The term τ represents the set of allowed precision structures;
  - constant variance: $\tau_{ij} = \tau$. simplest case
  - column-specific variance: $\tau_{ij} = \tau_j$. standard in "flash", assumes some features are noisier than others
  - row-specific variance: $\tau_{ij} = \tau_i$

- Derivation of how (2.4) Gaussian assumption transforms into "Weighted Squared Error" loss function used in optimization.

1) model assumes errors are Normally distributed

$$E_{ij} = Y_{ij} - \left(\sum_k l_{ik} f_{jk}\right) \sim N\left(0, \frac{1}{\tau_{ij}}\right)$$

↳ observed data Y_{ij}, given the latent factors, follows a Gaussian distribution with mean $\mu_{ij} = \sum_k l_{ik} f_{jk}$ and variance $\sigma^2_{ij} = 1/\tau_{ij}$

2) probability density function (PDF)

- probability density for a single point Y_{ij} is:

$$P(Y_{ij} | L, F, \tau) = \frac{1}{\sqrt{2\pi\sigma^2_{ij}}} \exp\left(\frac{-(Y_{ij} - \mu_{ij})^2}{2\sigma^2_{ij}}\right)$$

- substitute precision $\tau_{ij} = 1/\sigma^2_{ij}$

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

3) Joint likelihood

- assuming iid errors, probability of entire data matrix Y is product of individual probabilities

$$L(\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{p} \left[\sqrt{\frac{\tau_{ij}}{2\pi}} \exp\left(-\frac{\tau_{ij}}{2}\left(Y_{ij} - \sum_k l_{ik} f_{jk}\right)^2\right)\right]$$

4) log-likelihood

- take a log of both sides (product → sum)

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

↳

5) re-write as "loss"

- maximizing likelihood is equivalent to minimizing negative log-likelihood,
- consider NLL, drop terms that are constant wrt L and F

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

- holding τ (precision) constant, minimizing loss is equivalent to minimizing:

$$SSE = \sum_{i,j} \tau_{ij}\left(y_{ij} - \hat{y}_{ij}\right)^2$$

- weighted sum of squared errors,
- $\tau_{ij}=1$ (constant noise) → standard least squares
- $\tau_{ij}=0$ (infinite variance) → ignore data point (missing data imputation)

→ connect to (A.6) where a function $F(\tau)$ is defined to update the precision parameters:

$$F(\tau) = \frac{1}{2}\sum_{i,j}\left[\log(\tau_{ij}) + \tau_{ij}\overline{R^2}_{ij}\right]$$

$\overline{R^2}_{ij}$: expected squared residuals

↖ + or − ? typo?

Notes

- Y: observed data matrix w/ dimension $n \times p$
- l_k: kth set of "loadings" (n-vector)
- f_k: kth "factor" (p-vector)
- G_l, G_f: pre-specified families of distributions
- g_{lk}, g_{fk}: unknown "prior" distributions to be estimated
- E: $n \times p$ matrix of independent error terms
- τ: unknown $n \times p$ matrix of precisions (τ_{ij}) (constant / column-specific / row-specific)

- "Empirical Bayes" in EBMF means fitting (2.11)-(2.14) by obtaining point estimates for the priors g_{lk}, g_{fk}, $(k=1,\dots,K)$ and approximate the posterior distributions for the parameters l_k, f_k given those point estimates.

3- Fitting the EBMF Model

- need to estimate all of g_l, g_f, l, f, τ
- 2 step approach:
  - estimate g_l, g_f, τ by maximizing likelihood (product of all g_l and g_f)

$$L(g_l, g_f, \tau) := \int\int p(Y \mid l, f, \tau)\, g_l(dl_1)\cdots g_l(dl_n)\, g_f(df_1)\cdots g_f(df_p) \quad \text{over } g_l, g_f, \tau$$

- estimate l and f using posterior distribution: $p(l, f \mid Y, \tilde{g}_l, \tilde{g}_f, \tilde{\tau})$

* Derivation ↓

- marginal likelihood:

$$L(g_l, g_f, \tau) := \int\int p(Y \mid l, f, \tau)\, g_l(dl) \cdots g_f(df)$$

- integrate out unknown latent variables l and f
- find how "likely" observed data Y is given hyperparameters (g_l, g_f, τ)

1) probability of data given hyperparameters
- find estimated priors $\tilde{g}_l, \tilde{g}_f$, and precision $\tilde{\tau}$ that make observed data most likely. Maximize

$$p(Y \mid g_l, g_f, \tau)$$

2) Law of Total Probability
- data Y depends on latent factors l (loadings) and f (factors) (not observed), can't directly compute $p(Y)$. Need to integrate over all possible values that l and f take.

$$p(Y \mid g_l, g_f, \tau) = \int_f \int_l p(Y, l, f \mid g_l, g_f, \tau)\, dl\, df$$

3) factorize joint distribution
- factorize joint probability inside integral

$$p(Y, l, f) = \underbrace{p(Y \mid l, f, \tau)}_{\text{likelihood}} \cdot \underbrace{p(l \mid g_l)}_{\text{prior on } l} \cdot \underbrace{p(f \mid g_f)}_{\text{prior on } f}$$

4) expand priors
- consider rank 1 case ($k=1$). l and f vectors consist of n and p independent draws:

$$\cdot\, p(l \mid g_l) = \prod_{i=1}^{n} g_l(l_i)$$

$$\cdot\, p(f \mid g_f) = \prod_{j=1}^{p} g_f(f_j)$$

· Note: $g(dl)$ allows for distributions that are partly continuous (slab) and partly discrete (spike at zero)

↳ computational difficulty motivates use of variational approximation to avoid solving MLE (3.1) directly!

Next...: use variational approximation to approximate estimates for g_ℓ, g_f, and τ by maximizing likelihood (3.1).
- goal: transform (3.1) "impossible integration" → "solvable optimization"!

Variational Approach

- 3.2: write log of likelihood (3.1) as:

$$l(g_\ell, g_f, \tau) := \log L(g_\ell, g_f, \tau)$$

- In factor analysis (maximum likelihood), find loadings (L) and factors (F) that minimize the error: $\min \|Y - LF\|^2$

- In 3.2, L and F are missing from likelihood parameters ($l(g_\ell, g_f, \tau)$). EBMF is a "Type II Maximum Likelihood" method where we don't care about the specific values of L and F. We are interested in finding the distributions (priors g_ℓ, g_f) that generated them (L and F).
- Do this by integrating out the latent variables L and F.

$$L(g_\ell, g_f, \tau) = \int p(Y, L, F \mid g_\ell, g_f, \tau)\, dL\, dF$$

L> (3.2) says: "find the shape of the priors (g) that makes the observed data (Y) most probable, averaging over all possible factor configurations"
- Take log to turn product of probabilities into a sum of expectations (separate L and F terms)

- 3.3
  - Target: $l(g_\ell, g_f, \tau) = F(q, g_\ell, g_f, \tau) + D_{KL}(q \| p)$

1) marginal likelihood

Let $\Theta = \{g_\ell, g_f, \tau\}$ by hyperparameters. Let $Z = \{L, F\}$ be the latent variables, the log of the marginal likelihood is: $\log[p(Y \mid \Theta)]$

2) Expectation

$\log p(Y \mid \Theta)$ is constant wrt Z. Thus, if we take expectation wrt any distribution $q(z)$, the value remains unchanged.

$$\log p(Y \mid \Theta) = \int q(z) \log p(Y \mid \Theta)\, dz$$

Note: $\int q(z) \cdot c\, dz = c \underbrace{\int q(z)\, dz}_{=1} = c \cdot 1 = c$

3) Bayes Rule Expansion

- consider conditional probability:

$$P(z|Y,\mathcal{D}) = \frac{P(Y,z|\mathcal{D})}{P(Y|\mathcal{D})}$$

- rearrange:

$$P(Y|\mathcal{D}) = \frac{P(Y,z|\mathcal{D})}{P(z|Y,\mathcal{D})}$$

- substitution into integral from (2) ↓ Plug in

$$\log P(Y|\mathcal{D}) = \int q(z) \log\left(\frac{P(Y,z|\mathcal{D})}{P(z|Y,\mathcal{D})}\right) dz$$

4) variational distribution q

- multiply and divide by approximation $q(z)$

$$\log(P(Y|\mathcal{D})) = \int q(z) \log\left[\frac{P(Y,z|\mathcal{D})}{P(z|Y,\mathcal{D})} \cdot \frac{q(z)}{q(z)}\right] dz$$

5) rearrange terms

- group numerators + denominators to match definitions of ELBO and KL divergence.
  - ELBO: $\frac{P(Y,z|\mathcal{D})}{q(z)}$ • KL: $\frac{q(z)}{P(z|Y,\mathcal{D})}$

↳ $$\log P(Y|\mathcal{D}) = \int q(z) \log\left[\underbrace{\frac{P(Y,z|\mathcal{D})}{q(z)}}_{\text{ELBO}} \cdot \underbrace{\frac{q(z)}{P(z|Y,\mathcal{D})}}_{\text{KL}}\right] dz$$

6) Split up logarithm

- $\log(A \cdot B) = \log(A) + \log(B)$

↳ $$\log P(Y|\mathcal{D}) = \int q(z) \left[\log\left(\frac{P(Y,z|\mathcal{D})}{q(z)}\right) + \log\left(\frac{q(z)}{P(z|Y,\mathcal{D})}\right)\right] dz$$

(ELBO, KL)

7) Separate integral (ELBO + KL components)

- definition of $F(q,\mathcal{D})$ (ELBO) from (3,4)

$$\int q(z) \log \frac{P(Y,z|\mathcal{D})}{q(z)} dz$$

- definition of $D_{KL}(q||p)$ (KL divergence) from (3,5)

$$\int q(z) \log \frac{q(z)}{P(z|Y,\mathcal{D})} dz$$ Note: $D_{KL}(q||p) = \int q \log\left(\frac{q}{p}\right) = -\int q \log \frac{p}{q}$.

Based on the separation of the integral we get:

$$\underbrace{l(\vartheta_\ell, \vartheta_f, \tau)}_{\text{likelihood}} = \underbrace{F(q, \vartheta_\ell, \vartheta_f, \tau)}_{\text{ELBO}} + \underbrace{D_{KL}(q\|p)}_{\text{KL divergence}} \rightarrow \text{this is (3,3)}$$

3,4 – Evidence Lower Bound

- derived for 3,2 → 3,3

$$F(q, \vartheta_\ell, \vartheta_f, \tau) = \int q(l,f) \log \frac{p(Y, l, f \mid \vartheta_\ell, \vartheta_f, \tau)}{q(l,f)} \, dl \, df$$

- defined directly from first term of log-likelihood splitting derived for (3,3)
  1) expanded ~~marginal~~ log-likelihood:
  $$l(\vartheta) = \int q(z) \log\left(\frac{p(Y,z|\vartheta)}{p(z)}\right) dz + \int q(z) \log\left(\frac{q(z)}{p(z|Y,\vartheta)}\right) dz$$

  2) F is defined to be exactly the first integral
  $$F(q,\vartheta) := E_q \underbrace{[\log p(Y,z|\vartheta)}_{a} \underbrace{- \log(q(z))]}_{b}$$

  a) q places probability mass where complete data likelihood is high (good fit)
  b) Shannon entropy of q, encourages q to be spread out, preventing the model from collapsing to a single point estimate.

3,5 - KL divergence

- derived for 3,2 → 3,3

$$D_{KL}(q\|p) = -\int q(l,f) \log \frac{p(l,f \mid Y, \vartheta_\ell, \vartheta_f, \tau)}{q(l,f)} \, dl \, df$$

- 2nd term of split integral in (3,3)

$$\int q(z) \log\left(\frac{q(z)}{p(z|Y,\vartheta)}\right) dz \rightarrow \text{definition of KL divergence from distribution } q \text{ to distribution } p(\cdot|Y)$$

- negative sign on integral is identical to $\int q \log\left(\frac{q}{p}\right)$ → maximizing term inside log, => minimizing divergence
- KL divergence is non-negative so $F(q, \vartheta_\ell, \vartheta_f, \tau)$ is lower bound for log-likelihood (3,6) → proof

3.5 - Variational Inequality

- $F(q, g_\ell, g_f, Y)$ is a lower bound for the log-likelihood:

$$l(g_\ell, g_f, Y) \geq F(q, g_\ell, g_f, Y)$$

With equality when $q(\ell, f) = p(\ell, f \mid Y, g_\ell, g_f, Y)$

- **Proof** (via Jensen's Inequality)
  - to prove that F is a lower bound, prove that the term D_{KL} is non-negative
  - definition of K-L Divergence,

$$-D_{KL}(q\|p) = \int q(z) \log\left(\frac{p(z|Y)}{q(z)}\right) dz$$

  - let $f(x) = \log(x)$ (concave)
  - apply Jensen's inequality: $E[\log(x)] \leq \log(E(x))$

$$\Rightarrow \int q(z) \log\left[\frac{p(z|Y)}{q(z)}\right] dz \leq \log\left(\int q(z) \frac{p(z|Y)}{q(z)} dz\right)$$

  - simplify integral inside log

$$\int q(z) \frac{p(z|Y)}{q(z)} dz = \underbrace{\int p(z|Y) dz}_{\text{PDF integrates to 1}} = 1 \quad (p(z|Y) \text{ integrates to 1!})$$

  - plug into log( )

$$\log\left(\int q(z) \frac{p(z|Y)}{q(z)} dz\right) = \log\left(\int p(z|Y) dz\right) = \log(1) = 0$$

- Therefore (consider inequality)

$$-D_{KL}(q\|p) = \int q(z) \log\left(\frac{p(z|Y)}{q(z)}\right) dz \leq 0$$

$$\text{So, } -D_{KL}(q\|p) \leq 0 \Rightarrow D_{KL}(q\|p) \geq 0$$

- Now consider (3.3) again,

$$l(\theta) = F(q, \theta) + D_{KL} \quad (\text{and } D_{KL}(q\|p) \geq 0$$

Such that it must be true that

$$l(\theta) \geq F(q, \theta) \quad (3.6)$$

3.7 - maximize objective

Based on (3.6), $l(q_\ell, q_f, \tau) = \max_q F(q, q_\ell, q_f, \tau)$ with maximization over all possible distributions $q(\ell, f)$.

- Can't minimize $D_{KL}(q\|p)$ directly because calculating p (posterior) requires the integral $p(y)$.
- $l(\theta)$ is fixed wrt $q \Rightarrow l(\theta) = F(q) + D_{KL}(q\|p) \Rightarrow D_{KL}(q\|p) = l(\theta) - F(q)$
- So maximizing $F(q)$ is identical to minimizing $l(\theta) - D_{KL}(q\|p)$. $l(\theta)$ is constant so this is equivalent to minimizing $D_{KL}(q\|p)$ $\Rightarrow F(q) = l(\theta) - D_{KL}(q\|p)$ (D_{KL}: make as small as possible.)
- So finding q^* that maximizes F means that q^* is the distribution closest to the true posterior.

Note: variational approach simplifies problem by maximizing F but restricting the family of distributions for q.

3.8 - mean field assumption

- restrict q to family Q of distributions that "fully-factorize":

$$Q = \left\{ q : q(\ell, f) = \prod_{i=1}^{n} q_{\ell,i}(\ell_i) \prod_{j=1}^{p} q_{f,j}(f_j) \right\}$$

- constraint to make maximization for (3.7) feasible

1) factorization indices:
   - $\prod_{i=1}^{n} q_{\ell,i}(\ell_i)$: assumes loading for sample 1 is independent of the loading for sample 2
   - $\prod_{j=1}^{p} q_{f,j}(f_j)$: assumes factor value for gene 1 is independent of gene 2.

   L> implication: $q(\ell, f) = q(\ell)\, q(f)$ (can factor)

2) independence assumption
   - expectation of the likelihood term in F simplifies. Note the cross-term in the Gaussian likelihood: $\ell_i^T f_j$.

$$E_q[\ell_i^T f_j] = \int\int q(\ell, f)(\ell_i^T f_j)\, d\ell\, df \Rightarrow$$

↙ assume loadings + factors are independent

Assuming, $q(l,f) = q(l)\,q(f)$, the integral separates to be:

$$\left(\int q(l) l_i \, dl\right)^T \left(\int q(f) f_j \, df\right) = E_q[l_i]^T E_q[f_j]$$

- separation means that when $q(\ell)$ is updated, the factors (f) look like constants (expectations). This converts the complex optimization into an iterative EBMA problem

↳ So optimize $F(q_l, q_f, g_l, g_f, \tau)$ by alternating between optimizing over variables related to $l\,[(q_l, g_l)]$, over variables related to $f\,[(q_f, g_f)]$, and over τ.
↳ each step is guaranteed to increase F.

EBMF Algo 1

- initial values: $q_l^{(0)}, q_f^{(0)}, g_l^{(0)}, g_f^{(0)}$

1) $t \leftarrow 0$
2) repeat
3) $t \leftarrow t+1$
4) $\tau^{(t)} \leftarrow \arg\max_\tau F(q_l^{(t-1)}, q_f^{(t-1)}, g_l^{(t-1)}, g_f^{(t-1)}, \tau)$
5) $q_l^{(t)}, g_l^{(t)} \leftarrow \arg\max_{q_l, g_l} F(q_l, q_f^{(t-1)}, g_l, g_f^{(t-1)}, \tau^{(t)})$
6) $q_f^{(t)}, g_f^{(t)} \leftarrow \arg\max_{q_f, g_f} F(q_l^{(t)}, q_f, g_l^{(t)}, g_f, \tau^{(t)})$

maximization steps

7) until converged
8) return $q_l^{(t)}, q_f^{(t)}, g_l^{(t)}, g_f^{(t)}, \tau^{(t)}$

- step 4, update of τ, involves computing expected squared residuals:

$$\overline{R^2}_{ij} := E_{q_l, q_f}[(Y_{ij} - l_i f_j)^2] = [Y_{ij} - E_{q_l}(l_i) E_{q_f}(f_j)]^2 - E_{q_l}(l_i)^2 E_{q_f}(f_j)^2 + E_{q_l}(l_i^2) E_{q_f}(f_j^2)$$

3,9- definition

- ELBO contains a term for the expected log-likelihood of the data. Noise (error) is Gaussian, so maximizing likelihood is equivalent to minimizing the Expected Squared Error.

$$\overline{R^2}_{ij} := E_{q_l, q_f}[(Y_{ij} - l_i f_j)^2] \hookrightarrow$$

3.10 -

- We can't compute $(Y-\ell f)^2$ directly because ℓ and f are distributions!

1) expand quadratic: Let $x = \ell_i f_j$:

$$(Y_{ij} - x)^2 = Y_{ij}^2 - 2Y_{ij}x + x^2$$

2) expectation

$$E[(Y_{ij} - \ell_i f_j)^2] = Y_{ij}^2 - 2Y_{ij}E[\ell_i f_j] + E[(\ell_i f_j)^2]$$

3) apply independence assumption ($a(\ell, f) = g(\ell)\,g(f)$)

- $E[\ell_i f_j] = E[\ell_i]\,E[f_j]$
- $E[(\ell_i f_j)^2] = E[\ell_i^2 f_j^2] = E[\ell_i^2]\,E[f_j^2]$

Sub in:

$$\overline{R^2}_{ij} = Y_{ij}^2 - 2Y_{ij}E[\ell_i]E[f_j] + E[\ell_i^2]E[f_j^2]$$

4) rearrange to align w/ (3.10)

let $\mu_\ell = E[\ell_i]$, $\mu_f = E[f_i]$. Consider the squared error using just the means:

$$(Y_{ij} - \mu_\ell \mu_f)^2 = Y_{ij}^2 - 2Y_{ij}\mu_\ell\mu_f + \mu_\ell^2\mu_f^2$$

true expected error (step 3)

$$\text{Error}_{\text{true}} = Y_{ij}^2 - 2Y_{ij}\mu_\ell\mu_f + E[\ell^2]\,E[f^2]$$

now write:

$$\overline{R^2}_{ij} = (Y_{ij} - \mu_\ell\mu_f)^2 - \mu_\ell^2\mu_f^2 + E[\ell^2]\,E[f^2]$$

↳ accounts for the variance of the factors. If the model is very uncertain about a loading ℓ_i, the term $E[\ell_i^2]$ grows, increasing expected residual $\overline{R^2}$.

→ model increases estimated noise variance ($1/\tau$) or shrink the priors to reduce uncertainty.

3.3 - The EBMM Problem

- observations $x = (x_1, \ldots, x_n)$
- underlying quantities $\theta = (\theta_1, \ldots, \theta_n)$
- independent Gaussian errors w/ known sd $s = (s_1, \ldots, s_n)$
- elements of θ are iid from some distribution, $g \in G$

3.11 - observation model

$$x/\theta \sim N_n(\theta, \text{diag}(s_1^2, \ldots, s_n^2))$$

- Assume: error around the true parameter is Gaussian and independent for each data point.

3.12 - prior model

$\theta_1, \ldots, \theta_n \overset{iid}{\sim} g, \quad g \in \mathcal{G}$

- assume all n parameters (θ_i) are drawn from the same underlying distribution
- g belongs to family $\mathcal{G}$, must estimate the specific shape of g based on data X.

Now solve the EBNM problem by fitting (3.11)-(3.12) by the following:

3.13 - solve for prior

- estimate prior g from data (Empirical Bayes)

$$\hat{g} = \underset{g \in \mathcal{G}}{\operatorname{argmax}} \sum_j \int p(x_j | \theta_j, s_j) g(d\theta_j)$$

derivation

1) likelihood: maximize probability of observing the data X. (depends on unobserved θ)
2) Marginalize θ: integrate out unknown θ.

$$p(x_j | g) = \int p(x_j | \theta_j, s_j) \underbrace{p(\theta_j | g) d\theta_j}_{\text{write as measure } g(d\theta_j) \text{ for point masses at zero}}$$

3) Independence: Since θ_j are iid, probability of full vector x is product of individual probabilities

$$L(g) = \prod_{j=1}^{n} p(x_j | g)$$

↳ how?

4) optimization: find $\hat{g}$ (in $\mathcal{G}$) that maximizes $L(g)$

- Two approaches for maximizing $L(g)$:
  1) parametric optimization → numerical optimization algos
  2) convex optimization

3.14 - posterior distribution

- estimate prior $\hat{g}$ → treat it as true prior and proceed w/ standard Bayesian inference.

$$\cdot \underbrace{p(\theta | x, s, \hat{g})}_{\text{posterior}} \propto \prod_j \underbrace{\hat{g}(\theta_j)}_{\text{prior}} \underbrace{p(x_j | \theta_j, s_j)}_{\text{likelihood}}$$

- comes from Bayes' Theorem: posterior ∝ prior × likelihood

- prior: $\hat{g}(\theta_j)$
- likelihood: $p(x_j | \theta_j, s_j)$ → normal dist. from 3.11
- posterior: $p(\theta | x, s, \hat{g})$

• posterior: belief about θ_j after viewing data x_j.

3.15 + 3.16 – Moments

- EBMF doesn't need full posterior distribution object (only 1st + 2nd moments)

(3.15) posterior mean (1st moment)

$$\bar{\theta}_j := E(\theta_j | x, s, \hat{g}) = \int \underbrace{\theta_j}_{E[\,]} \underbrace{p(\theta_j | x, s, \hat{g})}_{\text{posterior}} d\theta_j$$

(3.16) 2nd moment

$$\bar{\theta}^2_j := E(\theta^2_j | x, s, \hat{g}) = \int \underbrace{\theta^2_j}_{E[\cdot^2]} \underbrace{p(\theta_j | x, s, \hat{g})}_{\text{posterior}} d\theta_j$$

↳ (3.15) had $E[(\hat{\mu} - \mu)^2]$, need $E[\hat{\mu}^2]$, $E[\mu^2]$ to compute residual error.
↳ $E[\hat{\theta}^2] = E[\theta]^2 + \text{Var}(\theta)$

3.17 – mapping

- $\text{EBNM}(x, s) = (\hat{g}, p)$
- inputs
  - x: noisy data from matrix residuals
  - s: noise standard deviation (from precision τ)
- outputs
  - $\hat{g}$: fitted prior distribution
  - p: posterior (moments $\bar{\theta}$ and $\bar{\theta}^2$)

*Note: calculation of x, s in 3.4

3.4 – Connecting EBMF and EBNM Problems

- prove task (complexity) of updating a single factor (f) in Matrix Factorization is identical to solving the simple EBNM problem (section 3.3)
- To update new model, need to calculate two vectors – pseudo data ($\hat{l}$) and pseudo standard errors (s_e) → plug into solver

3.18 – main result: $\underset{q_e, g_e}{\text{argmax}} \; F(q, g, \tau) = \text{EBNM}(\hat{l}, s_e)$

– update posterior (q_e) and prior (g_e) for the loadings vector l, holding factors (f) and noise (τ) fixed. Maximizing ELBO (F) to do this is the same as doing EBNM on pseudo observations

3.19 + 3.20 - derive pseudo parameters

- match terms in ELBO w/ Gaussian log-likelihood.

1) isolate terms in the ELBO w/ l_{ik}

- likelihood = $E_{q_\ell, q_f}[\log p(Y|\ell, f, \tau)]$
- sub in likelihood

$$= E_{q_\ell, q_f}\left[\frac{-1}{2}\sum_{ij}\tau_{ij}(Y_{ij} - \sum_k \ell_{ik} f_{jk})^2\right]$$

2) consider single factor (k)

- update one factor at a time (update kth loading on ith person (ℓ_{ik}))

$$R_{ij}^{-k} = Y_{ij} - \sum_{k' \neq k} \ell_{ik'} f_{jk} \Rightarrow Y_{ij} \propto R_{ij}^{-k} + \ell_{ik} f_{jk}$$

expand square w ℓ_{ik} terms:

$$\propto -\frac{1}{2}\sum_j \tau_{ij} E_{q_f}\left[(R_{ij}^{-k} - \ell_{ik} f_{jk})^2\right]$$

$$\propto -\frac{1}{2}\sum_j \tau_{ij} E_{q_f}\left[(R_{ij}^{-k})^2 - 2\ell_{ik} f_{jk} R_{ij}^{-k} + \ell_{ik}^2 f_{jk}^2\right]$$

3) expectation and group by ℓ_{ik}

- independence $\Rightarrow E[\ell f] = E[\ell]E[f]$

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

arrange as quadratic equation in ℓ_{ik}:

$$\text{objective}(\ell_{ik}) \propto -\frac{1}{2}\ell_{ik}^2\left(\sum_j \tau_{ij} E(f_{jk}^2)\right) + \ell_{ik}\left(\sum_j \tau_{ij} R_{ij}^{-k} E[f_{jk}]\right)$$

4) align w/ NM problem

- log-likelihood of single data point x from $N(\ell_{ik}, s^2)$

$$\log N(x|\ell_{ik}, s^2) \propto -\frac{1}{2s^2}(x - \ell_{ik})^2 \Rightarrow -\frac{1}{2s^2}\ell_{ik}^2 + \frac{x}{s^2}\ell_{ik} - \frac{x^2}{2s^2}$$

5) compare coefficients (ℓ_{ik}^2 and ℓ_{ik})

- ℓ_{ik}^2: precision: $\frac{1}{s^2} = \sum_j \tau_{ij} E[f_{jk}^2] \Rightarrow s_{\ell(\cdot)i} = \left(\sum_j \tau_{ij} E[f_{jk}^2]\right)^{-\frac{1}{2}}$ (3.20)
- ℓ_{ik}: scaled mean: $\frac{x}{s^2} = \sum_j \tau_{ij} R_{ij}^{-k} E[f_{jk}]$ $\xrightarrow{\text{solve for } x}$ $\hat{\ell}_i = s^2 x B = \frac{\sum_j \tau_{ij} R_{ij}^{-k} E(f_{jk})}{\sum_j \tau_{ij} E[f_{jk}^2]} \Rightarrow \hat{\ell}(k) = \frac{\sum_j \tau_{ij} Y_{ij} E[f_{jk}]}{\sum_j \tau_{ij} E[f_{jk}^2]}$ (3.19)

3.21+3.22 – Moments

- need 1st + 2nd moments of factors (F), defined as:
  - $\bar{f} := (E_{q_f}[f_j])$
  - $\bar{f^2} := (E_{q_f}[f_j^2])$
- Note: (3.19) uses $\bar{f^2}$, not $(\bar{f})^2$
  - $\bar{f^2} = (\bar{f})^2 + Var(f)$
  - If f is uncertain -> $\bar{f^2}$ is large
  - large denominator -> loading ($\hat{l}$) shrinks -> 0

3.23–3.26 – known factor intuition

– what if we knew f and Y?

↳ model becomes n independent regressions of the rows of Y on f, and the maximum likelihood estimate for l has elements:

(3.23) Regression estimate if f is fixed constants, estimating l_i is a weighted least squares (WLS) regression of the data row Y_i onto the factor f.

WLS solution is $\hat{\beta} = \frac{\Sigma wxy}{\Sigma wx2}$ where $w=\tau, x=f, y=Y$ $\Rightarrow$

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

(3.24) Regression Standard Error

– SE of WLS estimate is $(X^T W X)^{-\frac{1}{2}}$

$$\Rightarrow s_i = \left(\sum_j \tau_{ij} f_j^2\right)^{-1/2}$$

(3.25)+(3.26)

- goal: prove $\hat{l}_i \sim N(l_i, s_i^2)$
- estimator from WLS solution is $\hat{l}_i = \frac{\sum_j \tau_{ij} Y_{ij} f_j}{\sum_j \tau_{ij} f_j^2}$
- distribution shape: Y_{ij} is drawn from a Normal distribution, and $\hat{l}_i$ is a linear combination of the Y_{ij} values so $\hat{l}_i \sim$ Normal (find Mean and Variance)
- prove mean is unbiased ($E(\hat{l}_i) = l_i$) true model: $Y_{ij} = l_i f_j + \varepsilon_{ij}$

$$\hat{l}_i = \frac{\sum_j \tau_{ij}(l_i f_j + \varepsilon_{ij}) f_j}{\sum_j \tau_{ij} f_j^2} \Rightarrow \hat{l}_i = \frac{\sum_j \tau_{ij} l_i f_j^2}{\sum_j \tau_{ij} f_j^2} + \frac{\sum_j \tau_{ij} \varepsilon_{ij} f_j}{\sum_j \tau_{ij} f_j^2}$$

- factor out constant $l_i \Rightarrow \hat{l}_i = l_i \left[\frac{\sum_j \tau_{ij} f_j^2}{\sum_j \tau_{ij} f_j^2}\right] + \text{"}\varepsilon\text{"} \Rightarrow \hat{l}_i = l_i + \text{"}\varepsilon\text{"}$

- consider expectation:

$$E[\hat{\lambda}_i] = E[\lambda_i] + E\left[\frac{\sum_j \gamma_{ij} f_j \varepsilon_{ij}}{\sum_j \gamma_{ij} f_j^2}\right] \qquad E[\varepsilon_{ij}] = 0$$

$$= \lambda_i + 0$$

So, $E[\hat{\lambda}_i] = \lambda_i$

- Show $\text{var}(\hat{\lambda}_i) = s_i^2$; Note $\text{Var}(cX) = c^2\text{Var}(X)$ w/ Y_{ij} independent errors,
  - consider denominator $D = \sum_j \gamma_{ij} f_j^2$

$$\text{Var}(\hat{\lambda}_i) = \text{var}\left(\frac{\sum_j \gamma_{ij} f_j Y_{ij}}{D}\right) = \frac{1}{D^2}\sum_j (\gamma_{ij} f_j)^2 \text{Var}(Y_{ij})$$

- Subin $\text{Var}(Y_{ij}) = \frac{1}{\gamma_{ij}}$:

$$= \frac{1}{D^2}\sum_j \gamma_{ij}^2 f_j^2 (1/\gamma_{ij})$$

$$= \frac{1}{D^2}\sum_j \gamma_{ij} f_j^2 \quad \text{where } D = \sum_j \gamma_{ij} f_j^2$$

So,

$$\text{Var}(\hat{\lambda}_i) = \frac{1}{D^2} \cdot D = \frac{1}{D} \Rightarrow \text{Var}(\hat{\lambda}_i) = \frac{1}{\sum_j \gamma_{ij} f_j^2}$$

such that $SD(\hat{\lambda}_i) = (\sum_j \gamma_{ij} f_j^2)^{-1/2}$ ✓

Therefore, we showed mean + variance so we have $\hat{\lambda}_i \sim N(\lambda_i, s_i^2)$ (3.25)

(3.26) - prior

- $\lambda_1, \ldots, \lambda_n \overset{iid}{\sim} g_\ell, \; g \in \mathcal{G}$ (3.26)

So if we combine (3.26) and (3.25) (prior and likelihood), we get the structure for each loading:

1) observation: $\hat{\lambda}_i | \lambda_i \sim N(\lambda_i, s_i^2)$

2) prior: $\lambda_i \sim g_\ell$

which is the definition of the EBMM problem