Supervised Matrix Factorization for Survival Analysis: Likelihood & Estimation

Andrew Walther

December 12, 2025

1 Derivation of the Full Joint Likelihood

We define a generative Bayesian model where the latent structure L generates the genomics data Y and simultaneously predicts survival outcomes (t, δ).

1A Genomics Model (Bayesian Matrix Factorization

We model the observed gene expression matrix Y $(n \times p)$ using latent factors L $(n \times K)$ and factor loadings F $(p \times K)$.

Y come from known datasets (PDAC data - molecular subtypes (K))
-> use low rank approx for Y (LF')

1A.1 Likelihood

don't have to assume input matrix is non-negative, do require L (membership) to be non-negative

We assume Gaussian noise with precision τ_y:

$$Y_{ij} \sim \mathcal{N}\left(\left(LF^T\right)_{ij}, \tau_y^{-1}\right) \tag{1}$$

where the likelihood is

$$p\left(Y|L, F, \tau_y\right) \propto \exp\left(-\frac{\tau_y}{2}\|Y - LF^T\|_F^2\right) \tag{2}$$

don't have to constrain "F" to be non-negative

1A.2 Priors

Place sparsity-inducing priors on the factors and loadings (like Gaussian, Point-Normal, Exponential, Laplace, etc.)

flashier R package: includes good choices for priors

$$L_{ik} \sim g_L(\cdot), F_{jk} \sim g_F(\cdot) \tag{3}$$

1B Survival Model (Cox Proportional Hazards)

Model the hazard rate for patient i as a function of their latent factors L_i and regression coefficients $\beta_{(K\times 1)}$. The hazard function is

alternatives: piece-wise exponential

$$h(t|L_i) = h_0(t)\exp(L_i\beta) \tag{4}$$

1B.1 Partial Likelihood

Conditioned on observed events, the probability that patient i experiences an event at time t_i relative to the risk set $R(t_i)$ is:

$$L_{cox}\left(\beta, L\right) = \prod_{i:\delta_i=1} \frac{\exp\left(L_i\beta\right)}{\sum_{j\in R(t_i)} \exp\left(L_j\beta\right)} \tag{5}$$

1B.2 Prior

Place a regularization prior on the coefficients β like a Gaussian prior with precision τ_β:

$$\beta \sim \mathcal{N}\left(0, \tau_\beta^{-1} I\right) \tag{6}$$

updates: L, F, B (survival just includes L & B)

1C Full Posterior Distribution

We desire to estimate the posterior $p(L, F, \beta|Y, t, \delta)$. Then the log-posterior is proportional to the sum of the log-likelihoods and log-priors such that:

$$\log p(L, F, \beta|Y) \propto -\frac{\tau_y}{2}\|Y - LF^T\|_F^2 + \sum_{i=1}^{n} \delta_i \left(L_i\beta - \log \sum_{j \in R(t_i)} \exp(L_j\beta) \right) + \log P(L) + \log P(F) + \log P(\beta) \tag{7}$$

where

$$-\frac{\tau_y}{2}\|Y - LF^T\|_F^2 \tag{8}$$

is related to the genomics data matrix factorization

$$\sum_{i=1}^{n} \delta_i \left(L_i\beta - \log \sum_{j \in R(t_i)} \exp(L_j\beta) \right) \tag{9}$$

is related to the Cox Proportional Hazards survival model, and

$$\log P(L) + \log P(F) + \log P(\beta) \tag{10}$$

are penalty terms for the loadings, factors, and Cox regression components.

2 Objective Function (Quadratic Approximation)

Directly maximizing the Cox partial likelihood is computationally expensive within a matrix factorization loop because the log-sum-exp term couples all patients. We employ a Quadratic Approximation (Newton-Raphson) to create a tractable surrogate objective.

Approximate the Cox log-likelihood $l_{cox}(\eta)$ around the current linear predictor $\eta = L\beta$ using a second-order Taylor expansion. This creates a "Weighted Least Squares" pseudo-likelihood:

$$l_{cox} \approx -\frac{1}{2}(\mathbf{z} - L\beta)^T W (\mathbf{z} - L\beta) + C \tag{11}$$

where

similar to eta

- $\mathbf{z}$ (Pseudo-response): $z_i = \eta_i + u_i/W_{ii}$

approximations may affect convergence capability

- W (weights): A diagonal matrix where W_{ii} is the negative second derivative of the log-likelihood with respect to η_i

not actually diagonal, but treated as diagonal (ignore off-diag terms)

Final Objective Function: Define the surrogate objective function Q (negative log-posterior) to minimize for parameter estimation:

F

replace with full taylor expansion (eta) - focused on 2nd order term

$$Q(L, F, \beta) = \frac{\tau_y}{2}\|Y - LF^T\|_F^2 + \frac{1}{2}(\mathbf{z} - L\beta)^T W (\mathbf{z} - L\beta) + R(L) + R(F) + \frac{\tau_\beta}{2}\|\beta\|_2^2 \tag{12}$$

gaussian, quadratic in L

where $\frac{\tau_y}{2}\|Y - LF^T\|_F^2$ is the genomics MF fit, $\frac{1}{2}(\mathbf{z} - L\beta)^T W (\mathbf{z} - L\beta)$ is the survival fit, and $R(\cdot)$ are the negative log-prior (regularization penalty) for L and F. For standard Bayesian interpretation, these are squared L2 norms (Gaussian priors) or Elastic Net-type penalties.

focus on survival & quad approximation

-update L: condition on F & B (treat as fixed)
- update F: only condition on L (just MF term)
- update B: normal prior

3 Proof of Convexity

Prove that the objective function Q is convex with respect to the parameter block L when F and β are fixed. This ensures that each update step finds a unique global optimum.

taylor series expansion (2nd order) -> gets us close to posterior mode (bulk of density)

Claim: $Q(L|F, \beta, z, W)$ is convex with respect to L.

Proof:

outside loop: full updates

gradient = 0 (optimized) -> just consider 2nd order term (keep track of all terms)

inner loop: update L, F, B (ex: index on b)
- eta_hat(b) computed from (b-1) iteration

eta = LB (linear predictor) -> TTE outcome

eta_hat = current best estimate of LB

1. **Genomics Data MF Term**: $f_1(L) = \frac{\tau_y}{2}\|Y - LF^T\|_F^2$. This is a quadratic function of L. The Hessian with respect to vectorized L is $\tau_y\left(I_n \otimes F^T F\right)$, which is Positive Semi-Definite.
2. **Survival Model Term**: $f_2(L) = \frac{1}{2}\left(\mathbf{z} - L\beta\right)^T W\left(\mathbf{z} - L\beta\right)$. Since W is a diagonal matrix of non-negative weights (derived from the negative Hessian of a concave log-likelihood), this is a weighted quadratic form. Its Hessian is $W \otimes \left(\beta\beta^T\right)$, which is Positive Semi-Definite.
3. **Regularization**: If priors are log-concave (Gaussian, Laplace, Elastic Net), then $R(L)$ is convex.
4. **Conclusion**: The sum of convex functions is convex. Thus, the sub-problem for updating L is convex!

4 Parameter Update Algorithm

We use Block Coordinate Descent (BCD), effectively an Alternating Least Squares scene adapted for the Bayesian context.

Initialization: Initialize $L^{(0)}, F^{(0)}, \beta^{(0)}$

Then repeat the following until convergence (i.e. $\theta^{(b)} - \theta^{(b-1)} < \epsilon$ where ϵ is some arbitrarily small value:

4A Update Survival Approximation (z, W)

Fix L, β. Calculate linear predictors $\eta = L\beta$. Update the quadratic approximation parameters based on the current Cox model fit:

1. **Gradient**(u): $u_i = \delta_i - \sum_{j \in R(t_i)} \frac{\exp(\eta_i)}{\sum_{k \in R(t_j)} \exp(\eta_k)}$.
2. **HessianDiagonal**(H): Calculate the diagonal of the Hessian matrix H.
3. **Update**: Set weights $W_{ii} = -H_{ii}$ and pseudo-response $z_i = \eta_i + u_i / W_{ii}$.

4B Update Cox Coefficients (β)

Fix L, z, W. This becomes a Bayesian Weighted Linear Regression of $\mathbf{z}$ on L. Update the posterior distribution for β (assuming a Gaussian prior):

- **Precision**: $\Sigma_\beta^{-1} = L^T W L + \tau_\beta I$
- **Mean**: $\mu_\beta = \Sigma_\beta\left(L^T W \mathbf{z}\right)$
- **Set**: $\beta \leftarrow \mu_\beta$ (Posterior mean/mode)

keep track of what W, z mean

4C Update Latent Factors (L) - Supervised

Fix F, β, z, W. Update L by fusing information from Y and $\mathbf{z}$. For each subject i (row L_i), the posterior is Gaussian (assuming a Gaussian prior).

- Precision Fusion:
$$A_i = \tau_y\left(F^T F\right) + W_{ii}\left(\beta\beta^T\right) + \tau_L I \tag{13}$$
- Mean Fusion:
$$\mathbf{m}_i = A_i^{-1}\left(\tau_y F^T Y_i + W_{ii} z_i \beta\right) \tag{14}$$
- Set: $L_i \leftarrow \mathbf{m}_i$

4D Update Factor Loadings (F)

Fix L. Standard Bayesian Matrix Factorization update (row-wise for each gene j).

- Precision: $B_j = \tau_y \left(L^T L\right) + \tau_F I$
- Mean: $\mathbf{f}_j = B_j^{-1} \left(\tau_y L^T Y_{\cdot j}\right)$
- Set: $F_j \leftarrow \mathbf{f}_j$

5 Proof of Convergence

Proposition: The algorithm converges to a stationary point of the objective function Q.

Proof:

1. **Surrogate Function Construction**: The quadratic approximation $Q(\theta)$ is a local majorizer of the negative log-likelihood (due to the Taylor expansion of the concave Cox log-likelihood). Minimizing Q is guaranteed to improve (or maintain) the true likelihood l_{cox}.
2. **Block-wise Minimization**: In steps (2)-(4), solve for the exact global minimum of the convex surrogate function with respect to the active block of parameters. This ensures that $\mathcal{Q}\left(\theta^{(t+1)}\right) \leq \mathcal{Q}\left(\theta^{(t)}\right)$
3. **Boundedness**: The objective function is bounded below (log-likelihoods are bounded, and penalties are non-negative).
4. **Monotone Convergence**: A monotonically non-increasing sequence bounded below must converge to a limit. Therefore, the parameter estimates converge to a stationary point.