Scaleable Inference for Logistic Kalmans

October 15, 2019

1 Setup

1.1 Data

We're given a matrix $X \in \{0,1\}^{N_r \times N_c}$. We believe there may be some smoothness along the columns and/or rows. For example, we may believe that X_r is likely to be "similar" in some sense to X_{r+1}. We want to use this knowledge, together with a low-rank assumption, to find hidden structure in the matrix. To do this, we'll use a probabilistic matrix factorization model for X with a smoothness prior on the row and column loadings.

1.2 Model

- Hyperparameters
  - N_r – number of rows
  - N_c – number of columns
  - $L_U \in \mathbb{R}^{N_r}$ – a "location" for each row; we assume $L_{U,1} < L_{U,2} \cdots L_{U,N_r}$
  - $L_\alpha \in \mathbb{R}^{N_c}$ – a "location" for each column; we assume $L_{\alpha,1} < L_{\alpha,2} \cdots L_{\alpha,N_c}$
  - N_k – number of latent factors
  - $\rho_U \in \mathbb{R}^{N_k}$ – inverse smoothing factor for rows (large ρ means no smoothing)
  - $\rho_\alpha \in \mathbb{R}^{N_k}$ – inverse smoothing factor for columns
  - $\mu_U, \sigma_U \in \mathbb{R}^{N_k}$ – prior distribution for row loadings
  - $\mu_\alpha, \sigma_\alpha \in \mathbb{R}^{N_k}$ – prior distribution for columns loadings
- Random objects
  - $U \in \mathbb{R}^{N_r \times N_k}$, unobserved, the "row loadings." This random matrix is normally distributed, with

$$\begin{aligned}\mathbb{E}\left[U_{rk}\right] &= \mu_{U,k} \\ \operatorname{cov}\left(U_{rk}, U_{r'k'}\right) &= \mathbb{I}_{k=k'}\sigma^2_{U,k}\exp\left(-\frac{1}{2}\rho_{U,k}\left|L_{U,r} - L_{U,r'}\right|\right)\end{aligned}$$

  - $\alpha \in \mathbb{R}^{N_c \times N_k}$, unobserved, the "column loadings." Similarly,

$$\begin{aligned}\mathbb{E}\left[\alpha_{ck}\right] &= \mu_{\alpha,k} \\ \operatorname{cov}\left(\alpha_{ck}, \alpha_{c'k'}\right) &= \mathbb{I}_{k=k'}\sigma^2_{\alpha,k}\exp\left(-\frac{1}{2}\rho_{\alpha,k}\left|L_{\alpha,c} - L_{\alpha,c'}\right|\right)\end{aligned}$$

  - $Y \in \mathbb{R}^{N_r \times N_c}$, unobserved, with $Y_{rc} \sim \text{PolyaGamma}\left(1, \langle U_r, \alpha_c\rangle\right)$

- $X \in \mathbb{R}^{N_r \times N_c}$, observed, with $X_{rc} \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\langle U_r, \alpha_c \rangle}}\right)$.

Let $p(U, \alpha, Y, X; \rho, \mu, \sigma)$ denote the likelihood of these random variables under this model. Note this model does have some nonidentifiabilities (for example we can change the scale of U and inversely change the scale of α, or we could swap loading dimensions), so the meaning of these parameters should not be over-interpreted.

1.3 Held-out data

We will want to infer $U, \alpha, \rho, \mu, \sigma, N_k$ from X. We assume the matrix X is fully observed. However, in order to test model validity and overfitting, we will sometimes pretend that we cannot see all of X. Let

- $M_U \in \{0, 1\}^{N_r}$
- $M_\alpha \in \{0, 1\}^{N_c}$

When $M_{U,r} M_{\alpha,c} = 1$ we will sometimes pretend $X_{r,c}$ is unobserved. We will call these held-out entries the "test data."

1.4 Variational family

We will attempt to approximate the posterior $U, \alpha, Y | X$ using a mean-field variational family of the form

- $U_{rk} \sim \mathcal{N}\left(\hat{\mu}_{U,rk}, \hat{\sigma}^2_{U,rk}\right)$
- $\alpha_{ck} \sim \mathcal{N}\left(\hat{\mu}_{\alpha,ck}, \hat{\sigma}^2_{\alpha,ck}\right)$
- $Y_{rc} \sim \text{PolyaGamma}\left(Y_{rc}; 1, \hat{\Gamma}_{rc}\right)$

1.5 Some notation

Some remarks on notation:

- We will let $d_U \in \mathbb{R}^{N_r+1}$ denote

$$d_{U,r} = \begin{cases} \infty & r = 0 \\ |L_{U,r} - L_{U,r-1}| & r \in \{1 \cdots N_r - 1\} \\ \infty & r = N_r \end{cases}$$

- We will let $d_\alpha \in \mathbb{R}^{N_c+1}$ denote

$$d_{\alpha,c} = \begin{cases} \infty & c = 0 \\ |L_{\alpha,c} - L_{\alpha,c-1}| & c \in \{1 \cdots N_c - 1\} \\ \infty & c = N_c \end{cases}$$

- We will let $\tilde{X} \in \mathbb{R}^{N_r \times N_c}$ denotes

$$\tilde{X}_{rc} = \left(X_{rc} - \frac{1}{2}\right)(1 - M_{U,r} M_{\alpha,c})$$

- Throughout, we will use zero-based indexing. For example U is indexed as

$$\{U_{rk}\}_{r \in \{0 \cdots N_r - 1\}, k \in \{0 \cdots N_k - 1\}}$$

  and d_U is indexed as

$$\{d_{U,r}\}_{r \in \{0 \cdots N_r\}}$$

1.6 ELBO

The ELBO corresponding to this family (after dropping terms for the test data) is given by

$$\begin{aligned}\mathcal{L}\left(\rho,\mu,\sigma,\hat{\mu},\hat{\sigma},\hat{\Gamma}\right) = \mathbb{E}_q&\left[\log\frac{p(U;\rho,\mu,\sigma)}{q(U;\rho,\mu,\sigma)}\right]\\ &+\mathbb{E}_q\left[\log\frac{p(\alpha;\rho,\mu,\sigma)}{q(\alpha;\rho,\mu,\sigma)}\right]\\ &+\sum_{r=0}^{N_r-1}\sum_{c=0}^{N_c-1}(1-M_{U,r}M_{\alpha,c})\,\mathbb{E}_q\left[\log\frac{p(X_{rc},Y_{rc}|U,\alpha;\rho,\mu,\sigma)}{q(Y_{rc}|U,\alpha;\rho,\mu,\sigma)}\right]\end{aligned}$$

1.7 Tasks

Our main goal is to find a good choice for the parameters $\left(N_k,\rho,\mu,\sigma,\hat{\mu},\hat{\sigma},\hat{\Gamma}\right)$. Our approach is to first select some value of N_k, use many rounds of coordinate ascent on the ELBO to optimize the other parameters, and check final performance by looking at predictive capability on the held-out test data. In light of this, these are the main things we need to be able to do:

- Evaluate the ELBO
- Perform coordinate updates for the ELBO
- Evaluate predictive capability on the test data

In the sequel we build up the machinery to do all of these things efficiently on the GPU. We start by making a careful study of the prior covariance for U,α.

2 Covariances of U,α

We here focus on U (the equations for α are much the same).

The random variable U is understood as an $N_r\times N_k$ matrix, rather than a vector. This makes talking about covariances slightly tricky. One point of view is that we can unravel the variable U into a big vector in $\mathbb{R}^{N_rN_k}$. Traces, inverse, and determinants of the covariance are then all defined as usual. Another point of view is that the covariance is just a linear operator on matrix-space. Here you have to think about traces, inverses, dot products, and determinants in matrix-space; this is a little confusing but it's really not so bad because you know in your head you can always figure out what you're supposed to do by just unraveling U into a big vector. We will mostly adopt this latter perspective.

We can thus define the prior covariance for U as an operator defined by the fact that

$$(\Sigma_U\xi)_{rk}=\sum_{r'}\sigma_{U,k}^2\exp\left(-\frac{1}{2}\rho_{U,k}\left|L_{U,r}-L_{U,r'}\right|\right)\xi_{r'k'}$$

Note this operator has a very special structure. It enables us to get:

- **An inverse.** Let
  - $\Phi_{U.k}$ be the matrix defined by

$$\left(\Phi_{U,k}\right)_{rr'}\triangleq\mathbb{I}_{r=r'}\sigma_{U,k}^{-2}\left(\frac{\left(1-e^{-\rho_{U,k}d_{U,r}}e^{-\rho_{U,k}d_{U,r+1}}\right)}{\left(1-e^{-\rho_{U,k}d_{U,r}}\right)\left(1-e^{-\rho_{U,k}d_{U,r+1}}\right)}\right)$$

– $D_{U,k}$ be the matrix defined by

$$(D_{U,k}\xi)_r \triangleq \sigma_{U,k}^{-2} \begin{cases} \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r+1}}}{1-e^{-\rho_{U,k}d_{U,r+1}}}\right)\xi_{r+1} & \text{if } r = 0 \\ \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r}}}{1-e^{-\rho_{U,k}d_{U,r}}}\right)\xi_{r-1} & \text{if } r = N_r - 1 \\ \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r}}}{1-e^{-\rho_{U,k}d_{U,r}}}\right)\xi_{r-1} + \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r+1}}}{1-e^{-\rho_{U,k}d_{U,r+1}}}\right)\xi_{r+1} & \text{else} \end{cases}$$

Finally, let

$$\Phi_U = \bigoplus_k \Phi_{U,k}$$
$$D_U = \bigoplus_k D_{U,k}$$

These big operators act on the space of matrices of size $N_c \times N_k$ by acting on each column indepenently using the corresponding matrix, defined above. It is straightforward to verify that $\Sigma_U^{-1} = \Phi - D$. Note that these operators can be computed in linear time.

- **A determinant.** We have that

$$\log|\Sigma_U| = \sum_{r=0}^{N_r-1}\sum_{k=0}^{N_k} \log\sigma_{U,k}^2 + \sum_{r=0}^{N_r-2}\sum_{k=0}^{N_k} \log\left(1 - e^{-\rho_{U,k}d_{U,r+1}}\right)$$

These facts will be handy later on.

3 The ELBO

We will break the ELBO down in different ways in this document. For computing the ELBO itself, we break it down by defining

$$\mathcal{L}_U(\rho_U, \mu_U, \sigma_U, \hat{\mu}_U, \hat{\sigma}_U) = \mathbb{E}_q\left[\log\frac{p(U;\rho,\mu,\sigma)}{q(U;\rho,\mu,\sigma)}\right]$$
$$\mathcal{L}_\alpha(\rho_\alpha, \mu_\alpha, \sigma_\alpha, \hat{\mu}_\alpha, \hat{\sigma}_\alpha) \triangleq \mathbb{E}_q\left[\log\frac{p(\alpha;\rho,\mu,\sigma)}{q(\alpha;\rho,\mu,\sigma)}\right]$$
$$\mathcal{L}_{X,rc}\left(\hat{\mu}_{U,r}, \hat{\sigma}_{U,r}, \hat{\mu}_{\alpha,c}, \hat{\sigma}_{\alpha,c}, \hat{\Gamma}_{rc}\right) \triangleq \mathbb{E}_q\left[\log\frac{p(X_{rc}, Y_{rc}|U,\alpha;\rho,\mu,\sigma)}{q(Y_{rc}|U,\alpha;\rho,\mu,\sigma)}\right]$$

so that

$$\begin{aligned}\mathcal{L}\left(\rho,\mu,\sigma,\hat{\mu},\hat{\sigma},\hat{\Gamma}\right) = &\ \mathcal{L}_U(\rho_U, \mu_U, \sigma_U, \hat{\mu}_U, \hat{\sigma}_U) \\ &+ \mathcal{L}_\alpha(\rho_\alpha, \mu_\alpha, \sigma_\alpha, \hat{\mu}_\alpha, \hat{\sigma}_\alpha) \\ &+ \sum_{rc}(1 - M_{U,r}M_{\alpha,c})\,\mathcal{L}_{X,rc}\left(\hat{\mu}_{U,r}, \hat{\sigma}_{U,r}, \hat{\mu}_{\alpha,c}, \hat{\sigma}_{\alpha,c}, \hat{\Gamma}_{rc}\right)\end{aligned}$$

We compute each kind of term separately.

3.1 $\mathcal{L}_U, \mathcal{L}_\alpha$

We here focus on $\mathcal{L}_U$ (the equations for $\mathcal{L}_\alpha$ are much the same). Prior and variational family are both gaussians, so its just the negative KL of gaussians:

$$\mathcal{L}_U = -\frac{1}{2}\left(\mathrm{tr}\left(\Sigma_U^{-1}\hat{\Sigma}_U\right) + \|\mu_U - \hat{\mu}_U\|^2_{\Sigma_U^{-1}} - N_r N_k + \log\frac{|\Sigma_U|}{\left|\hat{\Sigma}_U\right|}\right)$$

Two remarks

- The parameter $\mu_U \in \mathbb{R}^{N_k}$ isn't the same shape as $\hat{\mu}_U \in \mathbb{R}^{N_r \times N_k}$. However, we will sometimes treat them as the same shape, an abuse of notation supported by broadcasting; $\mu_{U,rk} = \mu_{U,k}$.where μ, Σ specify the prior model and $\hat{\mu}, \hat{\Sigma}$ specify the variational family.
- The $\|(\cdot)\|^2_{(\cdot)}$ indicates the usual mahalanobis norm.

We have already made an extensive study of the inverse covariance and the variational family is completely mean-field, so we are now in a pretty good position to compute this object:

$$\mathcal{L}_U = -\frac{1}{2}\left(\|\hat{\sigma}_U\|^2_{\Phi_U} + \|\mu_U - \hat{\mu}_U\|^2_{\Phi_U - D_U} - N_r N_k - \sum_{rk}\log\hat{\sigma}^2_{U,rk} + \log|\Sigma_U|\right)$$

3.2 $\mathcal{L}_{X,rc}$

We find that

$$\begin{aligned}
\mathcal{L}_{X,rc} = &\left(X_{rc} - \frac{1}{2}\right)\langle\hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c}\rangle \\
&- \log 2\cosh\frac{\hat{\Gamma}_{rc}}{2} \\
&- \frac{1}{2}\left(\mathbb{E}_q\left[\langle U_r, \alpha_c\rangle^2\right] - \hat{\Gamma}^2_{rc}\right)\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}
\end{aligned}$$

We can compute that

$$\mathbb{E}_q\left[\langle U_r, \alpha_c\rangle^2\right] = \langle\hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c}\rangle^2 + \sum_k\left(\hat{\sigma}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\sigma}^2_{U,rk}\hat{\mu}^2_{\alpha,ck}\right)$$

As we will show later, it turns out that for any fixed $\hat{\mu}, \hat{\sigma}$ the optimal value of $\hat{\Gamma}^2_{rc}$ is given by $\mathbb{E}_q\left[\langle U_r, \alpha_c\rangle^2\right]$. In practice we will only evaluate the ELBO for this optimal value of $\hat{\Gamma}^2$. This yields a cancellation. In this case we only really need to compute:

$$\mathcal{L}_{X,rc} = \left(X_{rc} - \frac{1}{2}\right)\langle\hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c}\rangle - \log 2\cosh\frac{1}{2}\sqrt{\mathbb{E}_q\left[\langle U_r, \alpha_c\rangle^2\right]}$$

3.3 Summary

We have

$$\mathcal{L}_U = -\frac{1}{2}\left(\langle\hat{\sigma}_U|\Phi_U|\hat{\sigma}_U\rangle + \langle\mu_U - \hat{\mu}_U|\left(\Phi_U - D_U\right)\left(\mu_U - \hat{\mu}_U\right)\rangle - N_r N_k - \sum_{rk}\log\hat{\sigma}^2_{U,rk} + \log|\Sigma_U|\right)$$
$$\mathcal{L}_\alpha = -\frac{1}{2}\left(\langle\hat{\sigma}_\alpha|\Phi_\alpha|\hat{\sigma}_\alpha\rangle + \langle\mu_\alpha - \hat{\mu}_\alpha|\left(\Phi_\alpha - D_\alpha\right)\left(\mu_\alpha - \hat{\mu}_\alpha\right)\rangle - N_r N_k + \sum_{rk}\log\hat{\sigma}^2_{\alpha,rk} - \log\left|\Sigma_\alpha^{-1}\right|\right)$$
$$\mathcal{L}_{X,rc} = \left(X_{rc} - \frac{1}{2}\right)\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\rangle - \log 2\cosh\frac{\hat{\Gamma}_{rc}}{2} - \frac{1}{2}\left(\mathbb{E}_q\left[\langle U_r,\alpha_c\rangle^2\right] - \hat{\Gamma}^2_{rc}\right)\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}$$

where

$$\mathbb{E}_q\left[\langle U_r,\alpha_c\rangle^2\right] = \langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\rangle^2 + \sum_k\left(\hat{\sigma}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\sigma}^2_{U,rk}\hat{\mu}^2_{\alpha,ck}\right)$$

If you evaluate this at the optimal $\hat{\Gamma}^2$ with everything else fixed, you get a different expression for the last term:

$$\mathcal{L}_{X,rc} = \left(X_{rc} - \frac{1}{2}\right)\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\rangle - \log 2\cosh\frac{1}{2}\sqrt{\mathbb{E}_q\left[\langle U_r,\alpha_c\rangle^2\right]}$$

4 Coordinate updates for $\hat{\mu}_U, \hat{\mu}_\alpha, \hat{\sigma}^2_U, \hat{\sigma}^2_\alpha$

We here focus on U (the equations for α are much the same).

4.1 $\hat{\mu}$

Our first task is to collect all the terms from the ELBO that pertain to $\hat{\mu}_U$, i.e.

$$\begin{aligned}\mathcal{L}_{\hat{\mu}_U} &= -\frac{1}{2}\langle\mu_U - \hat{\mu}_U|\Phi_U - D_U|\mu_U - \hat{\mu}_U\rangle + \sum_{rc}\tilde{X}_{rc}\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\rangle \\ &\quad - \frac{1}{2}\sum_{rc}\left(1 - M_{U,r}M_{\alpha,c}\right)\left(\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\rangle^2 + \sum_k\hat{\mu}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck}\right)\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}\end{aligned}$$

Notice that this loss is wholly independent of $\hat{\mu}_U$. It should be computed at the same time.

We start by rewriting this loss using a new linear operator:

$$(\Psi\xi)_{rk} = \sum_{c=0}^{N_c-1}\left(1 - M_{U,r}M_{\alpha,c}\right)\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}\left[\left(\sum_{k'=0}^{N_k-1}\hat{\mu}_{\alpha,ck}\hat{\mu}_{\alpha,ck'}\xi_{rk'}\right) + \left(\hat{\sigma}^2_{\alpha,ck}\xi_{rk}\right)\right]$$

So that we may write

$$\begin{aligned}\mathcal{L}_{\hat{\mu}_U} &= -\frac{1}{2}\|\hat{\mu}_U\|^2_{\Phi+\Psi-D} + \langle\hat{\mu}_U, b\rangle \\ b &\triangleq \tilde{X}\hat{\mu}_\alpha + (\Phi - D)\,\mu_U\end{aligned}$$

The optimum of this quadratic form is found by solving

$$(\Phi + \Psi - D)\,\hat{\mu}_U = b$$

However, inverting this operator isn't quick on a GPU, despite its tridiagonal flavor. We instead apply a damped Jacobi update, i.e. we note that the true solution should satisfy

$$\begin{aligned}(\Phi + \Psi)\,\hat{\mu}_U &= \xi \\ \xi &\triangleq b + D\hat{\mu}_U\end{aligned}$$

This suggests (and it turns out to be true) that a good search direction can be found by computing

$$\nu = (\Phi + \Psi)^{-1}\xi$$

and searching in the direction of $\Delta = \nu - \hat{\mu}_U$. We can even figure out exactly how far we ought to go in this direction by optimizing

$$\ell(c) = -\frac{1}{2}\left\|\hat{\mu}_U + c\Delta\right\|^2_{\Phi+\Psi-D} + \langle \hat{\mu}_U + c\Delta, b\rangle$$

which yields

$$c = \frac{\langle \Delta, b\rangle - \langle \hat{\mu}_U, (\Phi + \Psi - D)\,\Delta\rangle}{\|\Delta\|^2_{\Phi+\Psi-D}}$$

We can also just take $c = 2/3$, which is a "popular choice" according to Wikipedia (though its not guaranteed to actually converge or even guaranteed to go uphill).

Now so far we have talked about how to update $\hat{\mu}_U$ if all the other variables are fixed and known, including $\hat{\Gamma}^2$. In practice, right before we make our $\hat{\mu}$ update we'll want to update $\hat{\Gamma}$ to its optimal value. So Ψ will need to know the best value of $\hat{\Gamma}$ given our initial conditions, namely

$$\hat{\Gamma}_{rc} \leftarrow \sqrt{\langle \hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c}\rangle^2 + \sum_k \left(\hat{\sigma}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\sigma}^2_{U,rk}\hat{\mu}^2_{\alpha,ck}\right)}$$

We do this partially because it makes us go uphill faster, and partially just because we never want to explicitly store all of $\hat{\Gamma}_{rc}$ at any one time.

4.2 $\hat{\sigma}$

Our first task is to collect all the terms from the ELBO that pertain to $\hat{\sigma}_U$, i.e.

$$\begin{aligned}\mathcal{L}_{\hat{\sigma}_U} = &-\frac{1}{2}\left(\langle \hat{\sigma}_U|\Phi_U|\hat{\sigma}_U\rangle - 2\sum_{rk}\log\hat{\sigma}_{U,rk}\right)\\ &-\frac{1}{2}\sum_{rck}(1 - M_{U,r}M_{\alpha,c})\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}\hat{\sigma}^2_{U,rk}\left(\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{\alpha,ck}\right)\end{aligned}$$

Notice that this loss is wholly independent of $\hat{\mu}_U$. It should be computed at the same time.

Take derivative w.r.t. $\hat{\sigma}_{U,rk}$:

$$\begin{aligned}\frac{\partial\mathcal{L}_{\hat{\sigma}_U}}{\partial\hat{\sigma}_{U,rk}} = &-\hat{\sigma}_{U,rk}\sigma^{-2}_{U,k}\left(\frac{\left(1 - e^{-\rho_{U,k}d_{U,r}}e^{-\rho_{U,k}d_{U,r+1}}\right)}{\left(1 - e^{-\rho_{U,k}d_{U,r}}\right)\left(1 - e^{-\rho_{U,k}d_{U,r+1}}\right)}\right) + \frac{1}{\hat{\sigma}_{U,rk}}\\ &-\hat{\sigma}_{U,rk}\sum_c(1 - M_{U,r}M_{\alpha,c})\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}\left(\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{\alpha,ck}\right)\end{aligned}$$

ergo

$$\hat{\sigma}^{-2}_{U,rk} = \sigma^{-2}_{U,k}\left(\frac{\left(1 - e^{-\rho_{U,k}d_{U,r}}e^{-\rho_{U,k}d_{U,r+1}}\right)}{\left(1 - e^{-\rho_{U,k}d_{U,r}}\right)\left(1 - e^{-\rho_{U,k}d_{U,r+1}}\right)}\right) + \sum_c(1 - M_{U,r}M_{\alpha,c})\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}\left(\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{\alpha,ck}\right)$$

As above, we'll recompute $\hat{\Gamma}_{rc}$ right before we make this update.

4.3 Summary

We get updates for $\hat{\mu}_U, \hat{\sigma}_U$ by defining

$$(\Psi\xi)_{rk} = \sum_{c=0}^{N_c-1} (1 - M_{U,r} M_{\alpha,c}) \frac{\tanh \hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}} \left[\left(\sum_{k'=0}^{N_k-1} \hat{\mu}_{\alpha,ck} \hat{\mu}_{\alpha,ck'} \xi_{rk'} \right) + \left(\hat{\sigma}^2_{\alpha,ck} \xi_{rk} \right) \right]$$

and taking

$$\begin{aligned}
b &\leftarrow \tilde{X} \hat{\mu}_\alpha + (\Phi_U - D_U)\, \mu_U \\
\xi &\leftarrow b + D_U \hat{\mu}_U \\
\nu &\leftarrow (\Phi_U + \Psi_U)^{-1} \xi \\
\hat{\sigma}^{-2}_{U,rk} &\leftarrow \sigma^{-2}_{U,k} \left(\frac{\left(1 - e^{-\rho_{U,k} d_{U,r}} e^{-\rho_{U,k} d_{U,r+1}}\right)}{\left(1 - e^{-\rho_{U,k} d_{U,r}}\right)\left(1 - e^{-\rho_{U,k} d_{U,r+1}}\right)} \right) + \sum_c (1 - M_{U,r} M_{\alpha,c}) \frac{\tanh \hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}} \left(\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{\alpha,ck} \right) \\
\Delta &\leftarrow \nu - \hat{\mu}_U \\
c &\leftarrow \begin{cases} 2/3 & \text{if you're bored} \\ \frac{\langle \Delta, b \rangle - \langle \hat{\mu}_U, (\Phi + \Psi - D)\Delta \rangle}{\|\Delta\|^2_{\Phi+\Psi-D}} & \text{otherwise} \end{cases} \\
\hat{\mu}_U &\leftarrow \hat{\mu}_U + c\Delta
\end{aligned}$$

In practice, we should compute ν and $\hat{\sigma}^{-2}_{U,rk}$ at the same time. Both of them involve the expensive computation of the optimal $\hat{\Gamma}$. In code, the computation of $\nu, \hat{\sigma}^{-2}_{U,rk}$ will look something like this

```
erd=exp(-rhoU[None,:] * dU[:,None]) # N_{r+1} × N_k
inners = (1/sigUsq[None,:]) * (1-erd[1:]*erd[:-1]) / ((1-erd[1:])*(1-erd[:-1])) # N_r ×
N_k
prepalpha = einsum(ck,cℓ → ckℓ,μ̂_α,μ̂_α) + diag(sighatalphasq) # N_c × N_k × N_k
for (st,en) in rowbatches:
  G1 = (muhatU[st:en] @ muhatalpha.T)**2
  G2 = sighatUsq[st:en] @ sighatalphasq.T
  G3 = muhatUsq[st:en] @ sighatalphasq.T
  G4 = sighatUsq[st:en] @ muhatalphasq.T
  G = sqrt(G1+G2+G3+G4)
  M = tanh(G/2)/(2*G) * (1-MU[st:en,None]*Malpha[None,:])
  Psi_plus_Phi = einsum(rc,ckℓ → rkℓ,M,prepalpha) + diag(inners[st:en])
  nu[st:en] = solve(Psi_plus_Phi,xi[st:en])
  newsighatsqi[st:en] = inners[st:en] + M @ (sighatalphasq + muhatalpha**2)
```

Note that we have to batch over rows so our RAM doesn't explode.

5 Coordinate updates for μ, σ, ρ

We here focus on U (the equations for α are much the same).

5.1 μ

We here focus on U (the equations for α are much the same). Relevant terms:

$$\mathcal{L}_{\mu,U} = -\frac{1}{2} \left(\langle \mu_U - \hat{\mu}_U | \Sigma_U^{-1} | \mu_U - \hat{\mu}_U \rangle \right)$$

So we'd want to set $\mu = \hat{\mu}$, but recall that μ is constrained to be the same for every row, i.e. the problem is actually

$$\begin{aligned}\mathcal{L}_{\mu,U} &= -\frac{1}{2}\left(\left\langle \mathbf{1}\mu_U^T - \hat{\mu}_U \middle| \Sigma_U^{-1} \middle| \left(\mathbf{1}\mu_U^T - \hat{\mu}_U\right)\right\rangle\right) \\ &= -\frac{1}{2}\left\|\mathbf{1}\mu_U^T\right\|_{\Sigma_U^{-1}}^2 + \left\langle \mathbf{1}\mu_U \middle| \Sigma_U^{-1}\hat{\mu}_U\right\rangle\end{aligned}$$

This separates by loadings into

$$\mathcal{L}_{\mu,U} = -\frac{1}{2}\mu_{U.k}^2\left\|\mathbf{1}\right\|_{\Sigma_U^{-1}}^2 + \mu_U\left\langle \mathbf{1}\middle|\Sigma_U^{-1}\hat{\mu}_{U,k}\right\rangle$$

which leads to

$$\mu_{U,k} \leftarrow \frac{\left\langle \hat{\mu}_{U,k}\middle|\Sigma_U^{-1}\middle|\mathbf{1}\right\rangle}{\left\|\mathbf{1}\right\|_{\Sigma_U^{-1}}^2}$$

5.2 σ

$$\mathcal{L}_{\sigma,U} = -\frac{1}{2}\left\|\hat{\sigma}_U\right\|_{\Phi_U}^2 - \frac{1}{2}\left\|\mu_U - \hat{\mu}_U\right\|_{\Sigma_U^{-1}}^2 - \frac{1}{2}N_r\sum_k \log\sigma_{U,k}^2$$

As above, this is separable in loadings:

$$\mathcal{L}_{\sigma,U,k} = -\frac{1}{2}\sigma_{U,k}^{-2}\left(\left\|\hat{\sigma}_U\right\|_{\sigma_{U,k}^2\Phi_{U.k}}^2 + \left\|\mu_{U,k} - \hat{\mu}_{U,k}\right\|_{\sigma_{U,k}^2\Sigma_{U,k}^{-1}}^2\right) - \frac{1}{2}N_r\log\sigma_{U,k}^2$$

Note that $\sigma_{U,k}^2\Sigma_{U,k}^{-1}$ is actually not a function of $\sigma_{U,k}$ (there's a cancellation). So when we take derivatives we get that

$$\sigma_{U,k}^2 = \frac{1}{N_r}\left(\left\|\hat{\sigma}_U\right\|_{\sigma_{U,k}^2\Phi_{U.k}}^2 + \left\|\mu_{U,k} - \hat{\mu}_{U,k}\right\|_{\sigma_{U,k}^2\Sigma_{U,k}^{-1}}^2\right)$$

6 ρ

Line search, independent in k. The relevant terms we need to collect are

$$\mathcal{L}_{\rho,U,k} = -\frac{1}{2}\left(\left\langle\hat{\sigma}_{U,k}\middle|\Phi_{U,k}\middle|\hat{\sigma}_{U,k}\right\rangle + \left\langle \mu_{U,k} - \hat{\mu}_{U,k}\middle|\Sigma_{U,k}^{-1}\middle|\mu_{U,k} - \hat{\mu}_{U,k}\right\rangle + \sum_{r=0}^{N_r-2}\log\left(1 - e^{-\rho_{U,k}d_{U,r+1}}\right)\right)$$

7 Initialization

- Get $\hat{\mu}$. Setting $\hat{\Gamma}_{ct} = \hat{\sigma} = 0$ and $\sigma = \infty$, we obtain the following objective in $\hat{\mu}$:

$$\mathcal{L}_{\hat{\mu}} = \sum_{rc}\left(1 - M_{U,r}M_{\alpha,c}\right)\left[4\left(X_{rc} - \frac{1}{2}\right)\left\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\right\rangle - \frac{1}{2}\left\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\right\rangle^2\right]$$

  We approach this by first solving the simpler problem

$$\mathcal{L}_{\hat{\mu}} = \sum_{rc}\left(1 - M_{U,r}\right)\left(1 - M_{\alpha,c}\right)\left[4\left(X_{rc} - \frac{1}{2}\right)\left\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\right\rangle - \frac{1}{2}\left\langle\hat{\mu}_{U,r},\hat{\mu}_{\alpha,c}\right\rangle^2\right]$$

which is equivalent to the SVD objective on $4\left(X_{rc}-\frac{1}{2}\right)$ for the submatrix where $(1-M_U)(1-M_\alpha)^T = 1$. We use SVD to grab this. But this only gives us values for the observed rows and columns. To get the sometimes-unobserved columns, we fix the $\hat{\mu}_{U,r}$ we already found. Then the $\hat{\mu}_\alpha$ updates are just

$$\hat{\mu}_{\alpha,r} = 4\left(\sum_r (1-M_{U,r})\,\hat{\mu}_{U,r}\hat{\mu}_{U,r}^T\right)^{-1}\left(\sum_r (1-M_{U,r})\left(X_{rc}-\frac{1}{2}\right)\hat{\mu}_{U,r}\right)$$

which we can get because we already have the values of $\hat{\mu}_U$ that we need for that. Likewise for $\hat{\mu}_\alpha$.

- We take a large initial posterior variance:

$$\hat{\sigma}_{U,rk} = \frac{1}{2}\left|\hat{\mu}_{U,rk}\right|$$

- We initialize μ, σ by the empirical distributions of $\hat{\mu}^2$ within each loading.
- We initialize with ρ_U essentially infinite (no smoothing).

8 Cheat Sheet

- Distances:

$$d_{U,r} = \begin{cases} \infty & r = 0 \\ |L_{U,r} - L_{U,r-1}| & r \in \{1 \cdots N_r - 1\} \\ \infty & r = N_r \end{cases}$$

- Masked data minus one-half has a name:

$$\tilde{X}_{rc} = \left(X_{rc} - \frac{1}{2}\right)(1 - M_{U,r}M_{\alpha,c})$$

- We made some operators:

$$(\Phi_U\xi)_{rk} \triangleq \sigma_{U,k}^{-2}\left(\frac{\left(1-e^{-\rho_{U,k}d_{U,r}}e^{-\rho_{U,k}d_{U,r+1}}\right)}{\left(1-e^{-\rho_{U,k}d_{U,r}}\right)\left(1-e^{-\rho_{U,k}d_{U,r+1}}\right)}\right)\xi_{rk}$$

$$(D_U\xi)_{rk} \triangleq \sigma_{U,k}^{-2}\begin{cases} \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r+1}}}{1-e^{-\rho_{U,k}d_{U,r+1}}}\right)\xi_{r+1} & \text{if } r = 0 \\ \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r}}}{1-e^{-\rho_{U,k}d_{U,r}}}\right)\xi_{r-1} & \text{if } r = N_r - 1 \\ \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r}}}{1-e^{-\rho_{U,k}d_{U,r}}}\right)\xi_{r-1} + \left(\frac{e^{-\frac{1}{2}\rho_{U,k}d_{U,r+1}}}{1-e^{-\rho_{U,k}d_{U,r+1}}}\right)\xi_{r+1} & \text{else} \end{cases}$$

$$(\Psi_U\xi)_{rk} \triangleq \sum_{c=0}^{N_c-1}(1-M_{U,r}M_{\alpha,c})\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}}\left[\left(\sum_{k'=0}^{N_k-1}\hat{\mu}_{\alpha,ck}\hat{\mu}_{\alpha,ck'}\xi_{rk'}\right) + \left(\hat{\sigma}_{\alpha,ck}^2\xi_{rk}\right)\right]$$

- Inverse and determinant of covariance:

$$|\Sigma_U| = \sum_{r=0}^{N_r-1}\sum_{k=0}^{N_k}\log\sigma_{U,k}^2 + \sum_{r=0}^{N_r-2}\sum_{k=0}^{N_k}\log\left(1-e^{-\rho_{U,k}d_{U,r+1}}\right)$$

$$\Sigma_U^{-1} = \Phi_U - D_U$$

- Expected logits squared:

$$\mathbb{E}_q\left[\langle U_r, \alpha_c\rangle^2\right] = \langle \hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c}\rangle^2 + \sum_k \left(\hat{\sigma}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\mu}^2_{U,rk}\hat{\sigma}^2_{\alpha,ck} + \hat{\sigma}^2_{U,rk}\hat{\mu}^2_{\alpha,ck}\right)$$

- ELBO:

$$\begin{aligned}
\mathcal{L}_U &= -\frac{1}{2}\left(\langle\hat{\sigma}_U|\Phi_U|\hat{\sigma}_U\rangle + \langle\mu_U - \hat{\mu}_U|\left(\Phi_U - D_U\right)\left(\mu_U - \hat{\mu}_U\right)\rangle - N_r N_k - \sum_{rk}\log\hat{\sigma}^2_{U,rk} + \log|\Sigma_U|\right) \\
\mathcal{L}_\alpha &= -\frac{1}{2}\left(\langle\hat{\sigma}_\alpha|\Phi_\alpha|\hat{\sigma}_\alpha\rangle + \langle\mu_\alpha - \hat{\mu}_\alpha|\left(\Phi_\alpha - D_\alpha\right)\left(\mu_\alpha - \hat{\mu}_\alpha\right)\rangle - N_r N_k + \sum_{rk}\log\hat{\sigma}^2_{\alpha,rk} - \log\left|\Sigma_\alpha^{-1}\right|\right) \\
\mathcal{L}_{X,rc} &= \left(X_{rc} - \frac{1}{2}\right)\langle\hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c}\rangle - \log 2\cosh\frac{\hat{\Gamma}_{rc}}{2} - \frac{1}{2}\left(\mathbb{E}_q\left[\langle U_r, \alpha_c\rangle^2\right] - \hat{\Gamma}^2_{rc}\right)\frac{\tanh\hat{\Gamma}_{rc}/2}{2\hat{\Gamma}_{rc}} \\
\mathcal{L}_{X,rc}\left(\hat{\Gamma}^*\right) &= \left(X_{rc} - \frac{1}{2}\right)\langle\hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c}\rangle - \log 2\cosh\frac{1}{2}\sqrt{\mathbb{E}_q\left[\langle U_r, \alpha_c\rangle^2\right]}
\end{aligned}$$

- Loss for $\hat{\mu}_U$:

$$\mathcal{L}_{\hat{\mu}_U} = -\frac{1}{2}\|\hat{\mu}_U\|^2_{\Phi+\Psi-D} + \left\langle\hat{\mu}_U, \tilde{X}\hat{\mu}_\alpha + \left(\Phi - D\right)\mu_U\right\rangle$$

- If you want to solve $Ax = b$ and already have a seach direction Δ in mind, the best update is

$$x \leftarrow x + \frac{\langle\Delta|b\rangle - \langle x|A|\Delta\rangle}{\|\Delta\|^2_A}\Delta$$

9 Low-level API

Key elements of the low-level API:

- Classes:

  KalmanParameters A class to store $N_k, \rho_U, \mu_U, \sigma_\alpha, \hat{\mu}_\alpha, \hat{\sigma}_\alpha, d_U, M_U$ or the same for α. Functionality includes:

  $\sigma^2_U, \hat{\sigma}^2_U$ is precomputed

  $\pi_{U,rk} = e^{-\frac{1}{2}\rho_{U,k}d_{U,r+1}}$ is precomputed

  $\pi^2_{U,rk} = e^{-\rho_{U,k}d_{U,r+1}}$ is precomputed

  SilkParameters A class to store a pair of kalman parameters – $(\mathtt{kp}_U, \mathtt{kp}_\alpha)$. Functionality includes:

  transpose Returns a new version that swaps U and α

  BinaryMatrixMinusOneHalfWithCornerCut A class which acts as a matrix with entries in $\left\{\frac{1}{2}, -\frac{1}{2}, 0\right\}$. The held-out data gets zeros. The other data gets the $\pm\frac{1}{2}$.Functionality includes:

  transpose Return a new version with rows and columns swapped

  BinaryMatrixMinusOneHalfJustACorner A class which acts as a matrix with entries in $\left\{\frac{1}{2}, -\frac{1}{2}, 0\right\}$. The training data gets zeros. The other data gets the $\pm\frac{1}{2}$. Functionality includes:

  transpose Return a new version with rows and columns swapped

- Functions:

  mult_Phi Left-multiply by Φ

mult_D Left-multiply by D

mult_Psi Left-multiply by Ψ_U (at optimal value of $\hat{\Gamma}$)

mult_Sigma Left-multiply of $\Phi - D$

solve_Psi_Phi Solves $(\Phi_U + \Psi_U)\, x = \xi$ and also returns a new-and-improved value for $\hat{\sigma}_U^2$ as a bonus prize

kalman_KL Evaluate KL from mean field model to true model

data_loss Returns $\sum_{rc} (1 - M_{U,r} M_{\alpha,c})\, \mathcal{L}_{X,rc}$

test_loss Returns predictive likelihood on held-out data,

$$\sum_{rc} \left[M_{U,r} M_{\alpha,c} \left(X_{rc} - \frac{1}{2} \right) \langle \hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c} \rangle - \log 2 \cosh \frac{1}{2} \langle \hat{\mu}_{U,r}, \hat{\mu}_{\alpha,c} \rangle \right]$$

variational_update Updates $\hat{\mu}_U, \hat{\sigma}_U$

prior_gaussian_update Updates μ, σ

prior_smoothing_update Updates ρ_U

Note that many of these functions are focused on the U variables. To do the corresponding things for the α variables you pass in the transposed silk parameters and transposed matrix.