Notes

Jackson Loper

January 2019

1 Optimization perspective

1.1 Objective

For any fixed matrix X with entries $X_{cg} \in \{-.5, .5\}$, consider the problem of optimizing

$$L(Z,\alpha) = \sum_{c,g} \left(X_{c,g} \left(\sum_k Z_{ck}\alpha_{gk} \right) - \log 2 \cosh \frac{1}{2} \sum_k Z_{ck}\alpha_{gk} \right)$$

1.2 Minorization

Observe that for any initial condition, $\tilde{Z}, \tilde{\alpha}$, we may obtain a simple minorizaton for this problem. Indeed, let

$$\begin{aligned}
M_{cg} = M_{cg}(\tilde{Z},\tilde{\alpha}) &= \frac{\tanh\left(\frac{1}{2}\sum_k \tilde{Z}_{ck}\tilde{\alpha}_{gk}\right)}{2\sum_k \tilde{Z}_{ck}\tilde{\alpha}_{gk}} \\
\kappa_{cg} = \kappa_{cg}(\tilde{Z},\tilde{\alpha}) &= \frac{1}{2} M_{cg} \left(\sum_k \tilde{Z}_{ck}\tilde{\alpha}_{gk} \right)^2 - \log 2 \cosh \frac{1}{2} \sum_k \tilde{Z}_{ck}\tilde{\alpha}_{gk} \\
\tilde{L}_{M,k}(Z,\alpha) &= \sum_{c,g} \left(X_{c,g} \left(\sum_k Z_{ck}\alpha_{gk} \right) + \kappa_{cg} - \frac{1}{2} M_{cg} \left(\sum_k Z_{ck}\alpha_{gk} \right)^2 \right)
\end{aligned}$$

Then observe that

$$\tilde{L}_{M,k}(\tilde{Z},\tilde{\alpha}) = L(\tilde{Z},\tilde{\alpha})$$

Furthermore, it is well-known that

$$\tilde{L}_{M,k}(Z,\alpha) \leq L(Z,\alpha) \qquad \forall Z, \alpha$$

Thus $\tilde{L}$ is a so-called "minorizer" for L from the initial condition $\tilde{Z}, \tilde{\alpha}$. We can therefore be guaranteed that if we can find Z, α that improves $\tilde{L}$, it will also improve our value of L. That is, if we can find Z, α such that $\tilde{L}_{M,k}(Z,\alpha) > \tilde{L}_{M,k}(\tilde{Z},\tilde{\alpha})$, then we will also have $L(Z,\alpha) > L(\tilde{Z},\tilde{\alpha})$. This suggests the following iterative process:

- Start with some initial condition $\tilde{Z}, \tilde{\alpha}$.
- Calculate $M(\tilde{Z}, \tilde{\alpha}), k(\tilde{Z}, \tilde{\alpha})$
- Find Z, α such that $\tilde{L}_{M,k}(Z, \alpha) > \tilde{L}_{M,k}(\tilde{Z}, \tilde{\alpha})$
- Set $\tilde{Z} \leftarrow Z$, $\tilde{\alpha} \leftarrow \alpha$, go to step 2.

To enact this procedure, the key difficulty is step 3. That is, we need to be able to make progress on the surrogate problem $\tilde{L}$. It is to this problem we now turn our attention.

1.3 Progress on the surrogate problem $\tilde{L}$

Here we consider the problem of optimizing

$$\tilde{L}_{M,k}(Z, \alpha) = \sum_{c,g} \left(X_{c,g} \left(\sum_k Z_{ck} \alpha_{gk} \right) - \frac{1}{2} M_{cg} \left(\sum_k Z_{ck} \alpha_{gk} \right)^2 \right)$$

Note we have dropped the κs that appeared in the previous section, since it is constant with respect to our objects of interest.

This problem can be optimized via coordinate ascent, alternating between Z and α. For example, let us consider only the case that we fix α and try to optimize Z. Note that with α fixed the problem is now separable over the cs. In particular, dropping constants, we see that for each c separately we need to optimize a problem of the form

$$f_c(z_c) = \sum_g \left(X_{c,g} \left(\sum_k Z_{ck} \alpha_{gk} \right) - \frac{1}{2} M_{cg} \left(\sum_k Z_{ck} \alpha_{gk} \right)^2 \right)$$

Take derivatives:

$$\frac{\partial}{\partial z_{ck}} f_c(z_c) = \sum_g X_{c,g} \alpha_{gk} - M_{cg} \alpha_{gk} \left(\sum_{k'} Z_{ck'} \alpha_{gk'} \right)$$

Setting equal to zero, we see that the optimal α_g will be achieved by taking

$$\Gamma_{k,k'} = \sum_g M_{cg} \alpha_{gk} \alpha_{gk'}$$
$$z_c^* = \Gamma^{-1} \alpha^T X_c$$

We can do the same kind of update for α.

1.4 Initialization

A reasonable initial condition for M is given by $M_{cg} = \lim_{\epsilon \to 0} \tanh(\epsilon)/(2\epsilon) = .5$. This leads to the surrogate problem

$$\tilde{L}_{M,k}(Z,\alpha) = \sum_{c,g} \left(X_{c,g} \left(\sum_k Z_{ck}\alpha_{gk} \right) - \frac{1}{4} \left(\sum_k Z_{ck}\alpha_{gk} \right)^2 \right)$$

It is easy to see that this problem is solved by taking Z, α as the first left and right singular vectors of $8X$. This gives a good initialization.

1.5 Regularization

If the matrix isn't roughly square, regularization can be helpful. The most trivial regularization is simply an $\mathscr{L}^2$ penalty. The inner objective becomes something like

$$f_c(z_c) = -\frac{\lambda}{2} \left\| z_c \right\|^2 + \sum_g \left(x_{c,g} \left(\sum_k z_{ck}\alpha_{gk} \right) - \frac{1}{2} M_{cg} \left(\sum_k z_{ck}\alpha_{gk} \right)^2 \right)$$

Which yields updates like

$$\begin{aligned} \Gamma_{k,k'} &= \sum_g M_{cg}\alpha_{gk}\alpha_{gk'} \\ z_c^* &= (\Gamma + \lambda I)^{-1}\alpha^T X_c \end{aligned}$$

You can also approximate an $\mathscr{L}^1$ penalty with $\log\cosh$. This suggests we compute $\zeta_{ck} = \tanh(z_{ck})/z_{ck}$ and consider the objective

$$f_c(z_c) = -\lambda \sum_k \zeta_{ck} z_{ck}^2 + \sum_g \left(x_{c,g} \left(\sum_k z_{ck}\alpha_{gk} \right) - \frac{1}{2} M_{cg} \left(\sum_k z_{ck}\alpha_{gk} \right)^2 \right)$$

Which yields the update

$$\begin{aligned} \Gamma_{k,k'} &= \sum_g M_{cg}\alpha_{gk}\alpha_{gk'} \\ z_c^* &= (\Gamma + \lambda \mathrm{diag}(\zeta_c))^{-1}\alpha^T X_c \end{aligned}$$

When z_c is large, this will make the penalization less significant.

2 Bayesian perspective

2.1 Model

Consider the model $p(Z, \alpha, Y, X)$ defined by

- $Z_c \sim \mathcal{N}(0, I)$
- $\alpha_g \sim \mathcal{N}(0, I)$
- $X_{cg}|Z, \alpha \sim \text{Bernoulli}\left(\frac{\exp(\sum_k Z_{ck}\alpha_{ck})}{1+\exp(\sum_k Z_{ck}\alpha_{ck})}\right)$
- $Y_{cg}|Z, \alpha \sim \text{PolyaGamma}(1, \sum_k Z_{ck}\alpha_{ck})$

That is, dropping normalizers,

$$\log p(z, \alpha, y, x) = -\frac{\sum_c \|z_c\|^2 + \sum_g \|\alpha_g\|^2}{2} + \sum_{cg}\left(\Gamma_{cg} - \log 2\cosh\frac{1}{2}\Gamma_{cg} - \frac{1}{2}\Gamma_{cg}^2 y_{cg}\right) + \cdots$$

where $\Gamma_{c,g} \triangleq \sum_k z_{ck}\alpha_{ck}$.

2.2 Variational formulation

Given observations of X, we can get an estimate for the likelihood of the data and the posterior on Z, α using variational methods. Specifically, consider the variational family $q(Z, \alpha, Y)$ defined by

- $\alpha_g \sim \mathcal{N}(\hat{\mu}_{\alpha,g}, \hat{\Sigma}_{\alpha,g})$
- $Z_c \sim \mathcal{N}(\hat{\mu}_{Z,c}, \hat{\Sigma}_{Z,c})$
- $Y_{cg} \sim \text{PolyaGamma}(1, \hat{\xi}_{cg})$

and the corresponding ELBO

$$\mathcal{L} = \mathbb{E}_q[\log p(x|Z, \alpha, Y)] + \mathbb{E}_q\left[\log \frac{p(Z, \alpha, Y)}{q(Z, \alpha, Y)}\right]$$

This can be computed as

$$\begin{aligned}
\mathcal{L} = &\sum_{cg}(x_{cg}\mathbb{E}_q[\Gamma_{cg}] - \log 2 - \omega_{cg}) \\
&+ \sum_{cg}\left(\frac{1}{2}(\xi_{cg}^2 - \mathbb{E}_q[\Gamma_{cg}^2]\mathbb{E}_q[Y_{cg}]) - \log\cosh\frac{1}{2}\hat{\xi}_{cg} + \omega_{cg}\right) \\
&- \sum_c \texttt{KLN}(\hat{\mu}_{Z,c}, \hat{\Sigma}_{Z,c}||0, I) - \sum_g \texttt{KLN}(\hat{\mu}_{\alpha,g}, \hat{\Sigma}_{\alpha,g}||0, I)
\end{aligned}$$

where `KLN` indicates the Gaussian KL divergence and

$$\begin{aligned}
\mathbb{E}_q[\Gamma_{cg}] &= \hat{\mu}_{Z,c}^T\hat{\mu}_{\alpha,c} \\
\mathbb{E}_q[\Gamma_{cg}^2] &= \text{tr}\left((\hat{\mu}_{Z,c}\hat{\mu}_{Z,c}^T + \hat{\Sigma}_{Z,c})(\hat{\mu}_{\alpha,c}\hat{\mu}_{\alpha,c}^T + \hat{\Sigma}_{\alpha,c})\right) \\
\mathbb{E}_q[Y_{cg}] &= \tanh(\hat{\xi}_{cg}/2)/(2\hat{\xi}_{cg}) \\
\omega_{cg} &\triangleq \mathbb{E}_q\left[\log\cosh\frac{1}{2}\Gamma_{cg}\right]
\end{aligned}$$

Note that the intractable ω terms cancel in $\mathcal{L}$ and everything else we have in closed form.

2.3 Updates

This ELBO can be optimized via coordinate ascent. In particular, taking gradients and setting them equal to zero, we get three different kinds of updates, each of which won't make our objective worse and might make them better:

$$\begin{aligned}
&\hat{\xi}_{cg} \leftarrow \sqrt{\mathbb{E}[\Gamma_{cg}^2]} \\
&\hat{\Sigma}_{Z,c}^{-1} \leftarrow I + \sum_g \mathbb{E}_q[Y_{cg}]\mathbb{E}_q[\alpha_g \alpha_g^T] \qquad && \hat{\mu}_{Z,c} \leftarrow \hat{\Sigma}_{Z,c}\left(\sum_g \mathbb{E}_q[\alpha_g] x_{cg}\right) \\
&\hat{\Sigma}_{\alpha,g}^{-1} \leftarrow I + \sum_c \mathbb{E}_q[Y_{cg}]\mathbb{E}_q[Z_c Z_c^T] \qquad && \hat{\mu}_{\alpha,g} \leftarrow \hat{\Sigma}_{\alpha,g}\left(\sum_c \mathbb{E}_q[Z_c] x_{cg}\right)
\end{aligned}$$

2.4 ELBO computation

Fix our variational posterior parameters for Z, α. After performing the ξ update, we calculate that:

$$\begin{aligned}
\mathbb{E}_q[\Gamma_{cg}] &= \hat{\mu}_{Z,c}^T \hat{\mu}_{\alpha,c} \\
\mathbb{E}_q[\Gamma_{cg}^2] &= \operatorname{tr}\left((\hat{\mu}_{Z,c}\hat{\mu}_{Z,c}^T + \hat{\Sigma}_{Z,c})(\hat{\mu}_{\alpha,c}\hat{\mu}_{\alpha,c}^T + \hat{\Sigma}_{\alpha,c})\right) \\
\mathbb{E}_q[Y_{cg}] &= \tanh\left(\sqrt{\mathbb{E}_q[\Gamma_{cg}^2]}/2\right) / \left(2\sqrt{\mathbb{E}_q[\Gamma_{cg}^2]}\right) \\
\mathcal{L} &= \sum_{cg}\left(x_{cg}\mu_{Z,c}^T \mu_{\alpha,g} - \log 2 - \omega_{cg}\right) \\
&\quad + \sum_{cg}\left(\frac{1}{2}\mathbb{E}_q[\Gamma_{cg}^2](1 - \mathbb{E}_q[Y_{cg}]) - \log\cosh\frac{1}{2}\hat{\xi}_{cg} + \omega_{cg}\right) \\
&\quad - \frac{1}{2}\sum_c\left(\operatorname{tr}(\hat{\Sigma}_{Z,c}) + \hat{\mu}_{Z,c}\hat{\mu}_{Z,c}^T - \log|\hat{\Sigma}_{Z,c}| - |I|\right) \\
&\quad - \frac{1}{2}\sum_g\left(\operatorname{tr}(\hat{\Sigma}_{\alpha,g}) + \hat{\mu}_{\alpha,g}\hat{\mu}_{\alpha,g}^T - \log|\hat{\Sigma}_{\alpha,g}| - |I|\right)
\end{aligned}$$

2.5 Initialization

The optimization perspective may give a reasonable initial conditions; let's denote them z^0, α^0. It turns out a sensible starting condition is actually $\Sigma_Z = \Sigma_\alpha = 0$. This yields an infinitely poor ELBO, but one round of updates quickly sorts that out.