1 Model

Stochastic Block Model (SBM)

- Choose the community proportions $\pi \sim \text{Dir}(\gamma)$, where $\pi \in \mathbb{R}^C$ with C latent communities
- For each representative u
  1. Choose a community membership assignment $M_u \overset{\text{iid}}{\sim} \text{Cat}(\pi)$
- For each pair of communities $k, l \in \{1, \ldots, C\}$, draw coexpression rate $P_{kl} \overset{\text{iid}}{\sim} \text{Gamma}(\lambda_0, \lambda_1)$
- For each pair of representatives $u, v \in \{1, \ldots, U\}$, draw $R_{uv} \mid P, M_u = k, M_v = l \sim \text{Poisson}(P_{kl})$

Ideal Point Model (IPM)

- For each document d
  1. Choose a discrimination $a_d \sim \mathcal{N}(\eta_a, \sigma_d^2)$
  2. Choose a difficulty $b_d \sim \mathcal{N}(\eta_b, \sigma_d^2)$
- For each representative u
  1. Choose a position $x_u \mid \nu \sim \mathcal{N}(\nu, \sigma_x^2)$
- Draw representative u's vote on document d as $V_{ud} \mid x_u, a_d, b_d \sim \text{Bern}(\sigma(a_d \cdot (x_u - b_d)))$

Latent Dirichlet Allocation (LDA)

- Draw a topic $\varphi_k \overset{\text{iid}}{\sim} \text{Dir}(\beta), \varphi_k \in \mathbb{R}^V$ as a distribution over words, for each $k \in \{1, \ldots, K\}$
- For each document, draw the topic proportions $\theta_d \overset{\text{iid}}{\sim} \text{Dir}(\alpha)$, where $\theta_d \in \mathbb{R}^K$
- For each document $d \in \{1, \ldots, D\}$ and each word $n \in \{1, \ldots, N_d\}$ in the document
  1. Choose a topic $z_{dn} \mid \theta_d \overset{\text{ind}}{\sim} \text{Mult}(\theta_d)$
  2. Choose a word $W_{dn} \mid z_{dn} = k, \varphi_k \overset{\text{ind}}{\sim} \text{Mult}(\varphi_k)$

1.1 Frankenstein Model

The ideal point model (IPM) is useful to us as a baseline model for the roll call voting data (V_{ud}) for a couple of reasons. For one, using it alone we can attempt to predict missing votes, a problem of interest in political science. Another problem of more qualitative interest is analyzing and interpreting the factors a_d, b_d specific to a document and those x_u specific to the representative. All are assumed to reside in some latent space $\mathbb{R}^S$ and so depending on how we set up the model, we might be able to interpret quantities like x_u as u's *political stance* or *ideological position* or $x_u - b_d$ as representative u's propensity for the bill/document d. There are a number of problems we cannot address in IPM. A major problem is predicting on heldout documents (the 'cold start'), which is a potentially useful performance measure. Similarly if we have relatively junior representatives, they may not have had enough votes for the inferred x_u to represent something (1) meaningful / interpretable or (2) reliable. We want to incorporate more information to inform the choices of a_d, b_d and x_u.

Ideal Point Allocator (IPA)

- Run the generative processes for SBM and LDA as described above. Then,
- For each document d
  1. Calculate the empirical topic proportions $\overline{z}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} z_d$ (a $K \times 1$ vector)
  2. Generate $S \times K$ matrices η_a, η_b with iid normal entries
  3. Choose a discrimination $a_d \sim \mathcal{N}(\eta_a' \overline{z}_d, \sigma_d^2)$
  4. Choose a difficulty $b_d \sim \mathcal{N}(\eta_b' \overline{z}_d, \sigma_d^2)$
- For each representative u
  1. Generate the community means $\nu_k \sim \mathcal{N}(\tau, \sigma_x^2)$
  2. Choose a position $x_u \mid M_u = k, \nu \sim \mathcal{N}(\nu_k, \sigma_x^2)$
- Draw representative u's vote on document d as $V_{ud} \mid x_u, a_d, b_d \sim \text{Bern}(\sigma(a_d \cdot (x_u - b_d)))$

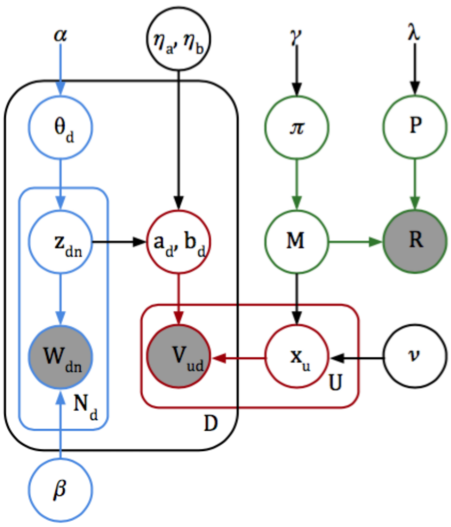

Figure 1: IPA graphical model

2 Variational Inference

2.1 For SBM.

After observing the symmetric matrix $R = (R_{uv})$, where R_{uv} is the number of caucuses that representatives u and v have in common, we see to find a distribution q over the latent community assignments $M = (M_u)$, the community coexpression rates $P = (P_{kl})$, and the community proportions $\pi = (\pi_k)$ which is close in relative entropy to the true posterior and lies in the factorized family $q(M)q(P)q(\pi)$. Each factor has free parameters described below and denoted with $\widehat{\text{hats}}$. The approximation q is equivalently scored by the ELBO objective $\mathcal{L}$, which we break down as:

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q\left[\log \frac{p(R, M, P, \pi)}{q(M, P, \pi)}\right] \\
&= \underbrace{\mathbb{E}_q\left[\log p(R \mid M, P) + \log \frac{p(P)}{q(P)}\right]}_{\mathcal{L}_{\text{data}}} + \underbrace{\mathbb{E}_q\left[-\log q(M)\right]}_{\mathcal{L}_{\text{ent}}} + \underbrace{\mathbb{E}_q\left[\log p(M \mid \pi)\right]}_{\mathcal{L}_{\text{local}}} + \underbrace{\mathbb{E}_q\left[\log \frac{p(\pi)}{q(\pi)}\right]}_{\mathcal{L}_{\text{global}}}
\end{aligned} \tag{1}$$

Variational Factors. To each u we associate variational parameters $\widehat{r}_u = (\widehat{r}_{uk})_{k=1}^{C}$, so

$$q(M) = \prod_{u=1}^{U} q(M_u \mid \widehat{r}_u) = \prod_{u=1}^{U} \prod_{k=1}^{C} \widehat{r}_{uk}^{\,\delta_k(M_u)}. \tag{2}$$

We define $q(\pi) \triangleq \text{Dir}(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_C)$ and $q(P) = \prod_{kl} q(P_{kl} \mid \widehat{\lambda}_{kl})$ where $q(P_{kl} \mid \widehat{\lambda}_{kl}) \triangleq \text{Gamma}(\widehat{\lambda}_{0kl}, \widehat{\lambda}_{1kl})$.

Computing the ELBO. Now we can write out the component terms of the ELBO more explicitly:

$$\begin{aligned}
\mathcal{L}_{\text{data}} &= \mathbb{E}_q\left[\log p(R \mid M, P) + \log \frac{p(P)}{q(P)}\right] = \sum_{kl} \mathbb{E}_q\left[\sum_{u,v} \delta_k(M_u)\delta_l(M_v) \log p(R_{uv} \mid P_{kl}) + \log \frac{p(P_{kl})}{q(P_{kl})}\right] \\
&= -\sum_{u,v} \log R_{uv}! + \sum_{k,l}\left(\lambda_0 \log \lambda_1 - \widehat{\lambda}_{0kl} \log \widehat{\lambda}_{1kl} - \log \frac{\Gamma(\lambda_0)}{\Gamma(\widehat{\lambda}_{0kl})}\right) + \sum_{k,l} \mathcal{L}_{kl}(R) \\
\mathcal{L}_{\text{ent}} &= \mathbb{E}_q\left[-\log q(M)\right] = -\sum_{u,k} \mathbb{E}_q\left[\delta_k(M_u) \log \widehat{r}_{uk}\right] = -\sum_{u,k} \widehat{r}_{uk} \log \widehat{r}_{uk} \\
\mathcal{L}_{\text{local}} &= \mathbb{E}_q\left[\log p(M \mid \pi)\right] = \sum_{u,k} \mathbb{E}_q\left[\delta_k(M_u) \log \pi_k\right] = \sum_{k} N_k \mathbb{E}_q\left[\log \pi_k\right] \\
\mathcal{L}_{\text{global}} &= \mathbb{E}_q\left[\log \frac{p(\pi)}{q(\pi)}\right] = \log \Gamma(C\gamma) - C \log \Gamma(\gamma) - \log \Gamma\left(\sum_k \widehat{\gamma}_k\right) + \sum_k \left\{\log \Gamma(\widehat{\gamma}_k) + (\gamma - \widehat{\gamma}_k)\mathbb{E}_q\left[\log \pi_k\right]\right\}
\end{aligned} \tag{3}$$

where $N_k = \sum_u \widehat{r}_{uk}$, $S_{uk} = \sum_v \widehat{r}_{vk} R_{uv}$, $N_{kl} = \sum_{uv} \widehat{r}_{uk}\widehat{r}_{vl}$, $S_{kl} = \sum_{uv} \widehat{r}_{uk}\widehat{r}_{vl} R_{uv}$, and

$$\mathcal{L}_{kl}(R) = (S_{kl} + \lambda_0 - \widehat{\lambda}_{0kl})\mathbb{E}_q[\log P_{kl}] - (N_{kl} + \lambda_1 - \widehat{\lambda}_{1kl})\mathbb{E}_q[P_{kl}],$$

and the posterior expectations can also be computed explicitly as

$$\mathbb{E}_q[P_{kl}] = \frac{\widehat{\lambda}_{0kl}}{\widehat{\lambda}_{1kl}}, \ \mathbb{E}_q[\log P_{kl}] = \psi(\widehat{\lambda}_{0kl}) - \log \widehat{\lambda}_{1kl}, \ \mathbb{E}_q\left[\log \pi_k\right] = \psi\left(\widehat{\gamma}_k\right) - \psi\left(\sum_l \widehat{\gamma}_l\right)$$

CAVI Updates. The simplest approach to variational inference maximizes the ELBO $\mathcal{L}$ via coordinate-ascent, i.e. choosing the best value of a variational parameter with all others fixed. Iteratively applying these updates, the variational approximation q improves at every step toward some local optimum. Conditional conjugacy yields closed form updates for the global variational parameters.

- **Global Update to $q(\pi)$.** We have $\widehat{\gamma}_k = \gamma + N_k$.
- **Global Update to $q(P)$.** We have $\widehat{\lambda}_{0kl} = \lambda_0 + S_{kl}$ and $\widehat{\lambda}_{1kl} = \lambda_1 + N_{kl}$.
- **Local Update to $q(M)$.** First note

$$\frac{\partial \mathcal{L}_{kl}}{\partial \widehat{r}_{uk}} = \begin{cases} 2S_{uk}\mathbb{E}_q[\log P_{kk}] - 2N_k\mathbb{E}_q[P_{kk}], & k = l \\ S_{ul}\mathbb{E}_q[\log P_{kl}] - N_l\mathbb{E}_q[P_{kl}] & k \neq l \end{cases}$$

  We want to update $\widehat{r}_{uk}$ subject to the constraint that $\sum_k \widehat{r}_{uk} = 1$, so augment $\mathcal{L}$ with Lagrange multipliers $\widetilde{\mathcal{L}} = \mathcal{L} + \sum_u \kappa_u \left(\sum_k \widehat{r}_{uk} - 1\right)$.

$$\begin{aligned} 0 = \frac{\partial \mathcal{L}}{\partial \widehat{r}_{uk}} &= -\log \widehat{r}_{uk} - 1 + \mathbb{E}_q\left[\log \pi_k\right] + \frac{\partial \mathcal{L}_{kk}}{\partial \widehat{r}_{uk}} + \sum_{l \neq k} \frac{\partial \mathcal{L}_{kl}}{\partial \widehat{r}_{uk}} + \frac{\partial \mathcal{L}_{lk}}{\partial \widehat{r}_{uk}} \\ &= -\log \widehat{r}_{uk} - 1 + \mathbb{E}_q\left[\log \pi_k\right] + \sum_l S_{ul}\left(\mathbb{E}_q[\log P_{kl}] + \mathbb{E}_q[\log P_{lk}]\right) - N_l\left(\mathbb{E}_q[P_{kl}] + \mathbb{E}_q[P_{lk}]\right) \end{aligned}$$

For a particular document d, the posterior distribution for its topic distribution θ_d and the latent topics $\{z_{dn}\}_{n=1}^{N_d}$ for the N_d words is given by

$$p(\theta_d, \{z_{dn}\}_{n=1}^{N_d} | \{w_{dn}\}_{n=1}^{N_d}, \alpha, \beta, a_d, b_d) = \frac{p(\theta_d|\alpha)\Big(\prod_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\Big)p(a_d, b_d|\{z_{dn}\}_{n=1}^{N_d}, \eta, \sigma^2)}{\int p(\theta_d|\alpha)\sum_{z_{\cdot d}}\Big(\prod_{n=1}^{N_d} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)\Big)p(a_d, b_d|\{z_{dn}\}_{n=1}^{N_d}, \eta, \sigma^2)\, d\theta_d}$$

In particular, the normalizer gives us the likelihood of the observed words $\{w_{dn}\}_{n=1}^{N_d}$ and the response a_d and b_d. This is computationally intractable, so we apply a variational method in which we assume a fully factorized posterior distribution for θ_d and z_{dn} of the form

$$q(\theta_d, \{z_{dn}\}_{n=1}^{N_d} | \gamma, \{\phi_n\}_{n=1}^{N_d}) = q(\theta_d|\gamma)\prod_{n=1}^{N} q(z_{dn}|\phi_n)$$

where ϕ_n is a vector on the symplex that gives the multinoulli probabilities of z_{dn}, and γ is the K dimensional Dirichlet parameter for θ_d. With respect to these two variational parameters, we seek to maximize the evidence lower bound (ELBO) given by

$$\begin{aligned} \mathcal{L}(\gamma, \phi; \alpha, \beta, \eta, \sigma) = E_q[\log p(\theta_d|\alpha)] &+ \sum_{n=1}^{N_d} E_q[\log p(z_{dn}|\theta)] + \sum_{n=1}^{N_d} E_q[\log p(w_{dn}|z_{dn}, \beta)] + \ldots \\ &+ E[\log p(a_d, b_d|\{z_{dn}\}_{n=1}^{N_d}, \eta, \sigma^2)] - E_q[\log q(\theta_d|\gamma)] - \sum_{i=1}^{N_d} E([\log q(z_{dn}|\phi_n)]) \end{aligned}$$

The updates for ϕ_j ($j \in \{1, \ldots, N_d\}$) and γ were derived in (Blei & McAuliffe, 2008) and given by

$$\phi_j^{new} \propto \exp\left\{E[\log \theta_d|\gamma] + E[\log p(w_{dn}|\beta)] + \Big(\frac{y}{N_d\sigma^2}\Big)\eta - \frac{[2(\eta^T\phi_{-j})\eta + (\eta \circ \eta)]}{2N_d^2\sigma^2}\right\}$$

where $\phi_{-j} := \sum_{n \neq j} \phi_n$; and $E(\log \theta_i | \gamma] = \Psi(\gamma_i) - \Psi(\sum \gamma_j)$, where Ψ is the digamma function. Now the updates for γ is:

$$\gamma^{new} \to \alpha + \sum_{n=1}^{N_d} \phi_n$$

These give the updates for a single document response pair. Having updated γ_d and ϕ_d for each document, the updates for β are given by examing the entire corpus (Blei et al, 2003):

$$\beta_{k,w}^{new} \propto \sum_{d=1}^{D} \sum_{n=1}^{N_d} 1\{(w_{dn} = w)\} \phi_{d,n}^k$$

2.2 For Ideal Point Model

We observe the votes matrix $V = (V_{ud})$ where V_{ud} is the vote of congressperson u on bill d. We have the ideal point for congressperson u, $x_u \in \mathbb{R}^s$, and the discrimination and difficulty for bill d, $a_d, b_d \in \mathbb{R}^s$. The variational distribution is the fully factorized family $\prod_{u=1}^{U} \prod_{d=1}^{D} q(x_u) q(a_D) q(b_d)$ where $q(x_u) \triangleq \text{Normal}(\hat{\tau}_u, \hat{\sigma}_\tau^2 I_S)$, $q(x_u) \triangleq \text{Normal}(\hat{\kappa}_{au}, \hat{\sigma}_{\kappa_a}^2 I_S)$, and $q(x_u) \triangleq \text{Normal}(\hat{\kappa}_{bu}, \hat{\sigma}_{\kappa_b}^2 I_S)$

Computing the ELBO. We can write the ELBO as

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q \left[-\sum_{u=1}^{U} \log q(x_u) - \sum_{d=1}^{D} \log q(a_d) - \sum_{d=1}^{D} \log q(b_d) \right] + \mathbb{E}_q \left[\sum_{u=1}^{U} \log p(X_u) + \sum_{d=1}^{D} \log p(a_d) + \sum_{d=1}^{D} \log p(b_d) \right] \\
&\quad + \mathbb{E}_q[\log p(V|x, a, b)] \\
&= H(q) + \mathbb{E}_q \left[\sum_{u=1}^{U} \log p(X_u) + \sum_{d=1}^{D} \log p(a_d) + \sum_{d=1}^{D} \log p(b_d) \right] + \mathbb{E}_q[\log p(V|x, a, b)]
\end{aligned}$$

We can break this up into

$$\begin{aligned}
H(q) &= U \frac{S}{2} \log 2\pi e \hat{\sigma}_\tau^2 + D \frac{S}{2} \log 2\pi e \hat{\sigma}_{\kappa_a}^2 + D \frac{S}{2} \log 2\pi e \hat{\sigma}_{\kappa_b}^2 \\
\mathbb{E}_q \left[\sum_{u=1}^{U} \log p(X_u) \right] &= \sum_{u=1}^{U} \mathbb{E}_q \left[-\frac{S}{2} \log 2\pi \sigma_x^2 - \frac{1}{2\sigma_x^2} \|x_u - \nu\|_2^2 \right] \\
&= -U \frac{S}{2} - \frac{1}{2\sigma_x^2} \sum_{u=1}^{U} \hat{\sigma}_\tau^2 S + \|\hat{\tau}_u - \nu\|_2^2 \\
\mathbb{E}_q \left[\sum_{d=1}^{D} \log p(a_d) \right] &= \sum_{d=1}^{D} \mathbb{E}_q \left[-\frac{S}{2} \log 2\pi \sigma_a^2 - \frac{1}{2\sigma_a^2} \|a_d - \eta_a\|_2^2 \right] \\
&= -D \frac{S}{2} - \frac{1}{2\sigma_a^2} \sum_{d=1}^{D} \hat{\sigma}_{\kappa_a}^2 S + \|\hat{\kappa}_{ad} - \eta_a\|_2^2 \\
\mathbb{E}_q \left[\sum_{d=1}^{D} \log p(b_d) \right] &= \sum_{d=1}^{D} \mathbb{E}_q \left[-\frac{S}{2} \log 2\pi \sigma_b^2 - \frac{1}{2\sigma_b^2} \|b_d - \eta_b\|_2^2 \right] \\
&= -D \frac{S}{2} - \frac{1}{2\sigma_b^2} \sum_{d=1}^{D} \hat{\sigma}_{\kappa_b}^2 S + \|\hat{\kappa}_{bd} - \eta_b\|_2^2
\end{aligned}$$

We can deal with the last expectation by using the 2nd order delta method (Braun McAullife 2008) which is the approximation

$$\mathbb{E}[f(V)] \approx f(\mathbb{E}[V]) + \frac{1}{2}\text{trace}\left(\nabla^2 \mathbb{E}[V]\text{Cov}(V)\right)$$

Letting $u(i), d(i)$ be the users and documents for data point i, and applying this gives the approximation to the ELBO contribution from the likelihood

$$\begin{aligned}
\mathbb{E}_q[\log p(V|x,a,b)] &= \sum_{i=1}^{n} \mathbb{E}_q[V_i(a_{d(i)} \cdot (X_{u(i)} - b_{d(i)}))] + \mathbb{E}_q[\log(1 - \sigma(a_{d(i)} \cdot (X_{u(i)} - b_{d(i)})))] \\
&\approx \sum_{i=1}^{n} V_i(\hat{\kappa}_{ad(i)} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)}) - \log(1 + \exp(\hat{\kappa}_{ad(i)} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)}) \\
&\quad - \frac{1}{2}\sigma''(\kappa_{a}\hat{d(i)} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)})(\hat{\sigma}^2_{\kappa_a}\|\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)}\|_2^2 + (\hat{\sigma}^2_\tau + \hat{\sigma}^2_{\kappa_b})\|\hat{\kappa}_{ad(i)}\|^2)
\end{aligned}$$

Where $\sigma(\cdot)$ is the sigmoid function.

CAVI Updates. There are no closed form updates for $\hat{\tau}_u$, $\hat{\kappa}_{ad}$, and $\hat{\kappa}_{bd}$, so we have to maximize the ELBO with respect to these parameters numerically. Let $V(u)$ be the set of votes for user u, and similarly let $V(d)$ be the set of votes on bill d. Also let
$rho_{ud} = \sigma(\kappa_{a}\hat{d(i)} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)})$. The gradients are

$$\begin{aligned}
\nabla_{\hat{\tau}_u}\mathcal{L} &= \frac{1}{\sigma_x^2}(\hat{\tau}_u - \nu) + \sum_{i \in V(u)} (V_i - \rho_{ud(i)})\hat{\kappa}_{ad(i)} \\
&\quad - \frac{1}{2}\sigma''(\hat{\kappa}_{ad(i)} \cdot (\hat{\tau}_u - \hat{\kappa}_{bd(i)})(\hat{\sigma}^2_{\kappa_a}\|\hat{\tau}_u - \hat{\kappa}_{bd(i)}\|_2^2 + (\hat{\sigma}^2_\tau + \hat{\sigma}^2_{\kappa_b})\|\hat{\kappa}_{ad(i)}\|^2)\hat{\kappa}_{ad(i)} \\
&\quad - \sigma'(\hat{\kappa}_{ad(i)} \cdot (\hat{\tau}_u - \hat{\kappa}_{bd(i)}))\hat{\sigma}^2_{ka}(\hat{\tau}_u - \hat{\kappa}_{bd(i)}) \\
\nabla_{\hat{\kappa}_{ad}}\mathcal{L} &= \frac{1}{\sigma_a^2}(\hat{\kappa}_{ad} - \eta_a) + \sum_{i \in V(d)} (V_i - \rho_{u(i)d})(\hat{\tau}_{u(i)} - \hat{\kappa}_{bd}) \\
&\quad - \frac{1}{2}\sigma''(\hat{\kappa}_{ad} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd})(\hat{\sigma}^2_{\kappa_a}\|\hat{\tau}_{u(i)} - \hat{\kappa}_{bd}\|_2^2 + (\hat{\sigma}^2_\tau + \hat{\sigma}^2_{\kappa_b})\|\hat{\kappa}_{ad}\|^2)(\hat{\tau}_{u(i)} - \hat{\kappa}_{bd}) \\
&\quad - \sigma'(\hat{\kappa}_{ad} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd})(\hat{\sigma}^2_\tau + \hat{\sigma}^2_{\kappa_b})\hat{\kappa}_{ad} \\
\nabla_{\hat{\kappa}_{bd}}\mathcal{L} &= \frac{1}{\sigma_b^2}(\hat{\kappa}_{bd} - \eta_b) - \sum_{i \in V(d)} (V_i - \rho_{u(i)d})\hat{\kappa}_{ad} \\
&\quad + \frac{1}{2}\sigma''(\hat{\kappa}_{ad} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd})(\hat{\sigma}^2_{\kappa_a}\|\hat{\tau}_{u(i)} - \hat{\kappa}_{bd}\|_2^2 + (\hat{\sigma}^2_\tau + \hat{\sigma}^2_{\kappa_b})\|\hat{\kappa}_{ad}\|^2)\hat{\kappa}_{ad} \\
&\quad + \sigma'(\hat{\kappa}_{ad} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd})\hat{\sigma}^2_{ka}(\hat{\tau}_{u(i)} - \hat{\kappa}_{bd})
\end{aligned}$$

For each parameter we solve this optimization problem using L-BFGS. Finally, there are closed form updates for the variational variance parameters by taking the derivative and setting to zero

$$\begin{aligned}
\hat{\sigma}^2_\tau &= \frac{US}{\frac{US}{\sigma_x^2} + \sum_{i=1}^n \sigma'(\kappa_{a}\hat{d(i)} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)}))(S\hat{\sigma^2}_{\kappa_a} + \|\hat{\kappa}_{ad(i)}\|_2^2)} \\
\hat{\sigma}^2_{\kappa_a} &= \frac{DS}{\frac{DS}{\sigma_a^2} + \sum_{i=1}^n \sigma'(\kappa_{a}\hat{d(i)} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)}))(S(\hat{\sigma^2}_\tau + \hat{\sigma}^2_{\kappa_b}) + \|\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)}\|_2^2)} \\
\hat{\sigma}^2_\tau &= \frac{DS}{\frac{DS}{\sigma_b^2} + \sum_{i=1}^n \sigma'(\kappa_{a}\hat{d(i)} \cdot (\hat{\tau}_{u(i)} - \hat{\kappa}_{bd(i)}))(S\hat{\sigma^2}_{\kappa_a} + \|\hat{\kappa}_{ad(i)}\|_2^2)}
\end{aligned}$$