Group Sparse Variational Bayes

1 Problem formulation

1.1 Setting

Consider the model,

$$y = \langle x, \beta \rangle + \epsilon \tag{1}$$

{eq:model}

where $y \in \mathbb{R}$ is the response, $x = (x_1, \ldots, x_p)^\top \in \mathbb{R}^p$ a feature vector of explanatory variables, $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ the coefficient vector, and ϵ a noise term.

Under a *group-sparse* setting, it is assumed that features can be grouped, and that few groups have non-zero coefficient values (Giraud, 2021). Formally, define the groups as $G_k = \{G_{k,1}, \ldots, G_{k,m_k}\}$ for $k = 1, \ldots, M$ as disjoint sets of indices such that $\bigcup_{k=1}^{M} G_k = \{1, \ldots, p\}$ and let $G_k^c = \{1, \ldots, p\} \setminus G_k$. Further, denote $x_{G_k} = \{x_j : j \in G_k\}$ and $\beta_{G_k} = \{\beta_j : j \in G_k\}$.

Under the group structure (1) can be written as:

$$y = \left(\sum_{k=1}^{M} \langle x_{G_k}, \beta_{G_k} \rangle \right) + \epsilon \tag{2}$$

Furthermore we are going to assume the error term $\epsilon \overset{\text{iid.}}{\sim} N(0, \tau^2)$, under which the log-likelihood is given by,

$$\begin{aligned} \ell(\mathcal{D}; \beta) = & -\frac{1}{2} \sum_{i=1}^{n} \left[\log(2\pi\tau^2) + \frac{1}{\tau^2}(y_i - \langle x_i, \beta \rangle)^2 \right] \\ = & -\frac{n}{2} \log(2\pi\tau^2) - \frac{1}{2\tau^2} \|y - X\beta\|^2 \end{aligned} \tag{3}$$

{eq:log-likeli

where $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^n, y_i \in \mathbb{R}, x_i \in \mathbb{R}^p, y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$.

1.2 Prior

We consider a group spike-and-slab (GSpSL) prior for the model parameters β, which has a hierarchical representation,

$$\begin{aligned} \beta_{G_k} | z_k &\overset{\text{ind}}{\sim} z_k \Psi(\beta_{G_k}; \lambda) + (1 - z_k)\delta_0(\beta_{G_k}) \\ z_k | \theta_k &\overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_k) \\ \theta_k &\overset{\text{iid}}{\sim} \text{Beta}(a_0, b_0) \end{aligned} \tag{4}$$

for $k = 1, \ldots, M$, where δ_0 is the multivariate Dirac mass on zero with dimension $m_k = \dim(\beta_k)$, and $\Psi(\beta_{G_k})$ is the multivariate double exponential distribution with density

$$\psi(\beta_{G_k}; \lambda) = C_k \lambda^{m_k} \exp\left(-\lambda \|\beta_{G_k}\|\right) \tag{5}$$

{eq:density_mv

where $C_k = \left[2^{m_k} \pi^{(m_k - 1)/2} \Gamma\left((m_k + 1)/2\right)\right]^{-1}$ and $\|\cdot\|$ is the ℓ_2-norm. For a visual representation, the density (5) for the two-dimensional multivariate double exponential is shown in Figure 1.

It follows that the prior distribution for β is given by,

$$\Pi(\beta | z) = \bigotimes_{k=1}^{M} \left[z_k \Psi(\beta_{G_k}; \lambda) + (1 - z_k)\delta_0(\beta_{G_k}) \right] \tag{6}$$

{eq:prior}

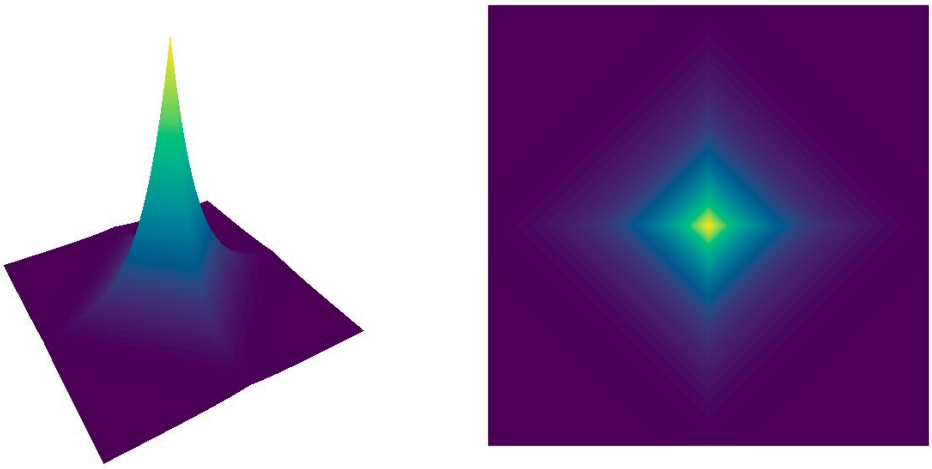


Figure 1: Two dimensional double exponential density with $\lambda = 1$. fig:double_exp

where $z = (z_1, \ldots, z_M)$ and $\otimes$ is the product measure.

Regarding τ^2, there are many popular choices that a practitioner may wish to use, for example:

- *Locally uniform*, wherein $\tau^2 \sim U(0, 1/\varepsilon)$ for a small positive ε.
- *Inverse-Gamma*, wherein $\tau^2 \sim \Gamma^{-1}(a, b)$, where Γ^{-1} denotes an inverse-Gamma distribution with shape a and scale b, with a common choice for each being $a = b = \varepsilon$, where ε is a small positive constant around 0.001.

In the meantime, we place do not place a prior on τ^2 and assume it is known.

1.3 Posterior

The posterior density is given by,

$$d\Pi(\beta|\mathcal{D}) = \Pi_D^{-1} e^{\ell(\mathcal{D};\beta)} d\Pi(\beta) \tag{7}$$ {eq:posterior}

where $\Pi_{\mathcal{D}} = \int_{\mathbb{R}^p} e^{\ell(\mathcal{D};\beta)} d\Pi(\beta)$ is a normalization constant and $\ell(\mathcal{D};\beta)$ is the log-likelihood function. (This isn't exactly correct, we still need to integrate our the z terms and re-write the prior).

1.4 Variational Family

In turn, our aim is to approximate (7) by a member of a tractable family of distributions, referred to as the variational family. We have chosen the family,

$$\mathcal{Q} = \left\{ Q = \bigotimes_{k=1}^{M} \left[\gamma_k \; N\left(\mu_{G_k}, \mathrm{diag}(\sigma_{G_k})\right) + (1-\gamma_k)\delta_0 \right] \right\} \tag{8}$$

where $N(\mu, \Sigma)$ denotes the multivariate Normal distribution with mean parameter μ and covariance Σ. In turn the variational posterior is given by solving,

$$\tilde{\Pi} = \underset{Q \in \mathcal{Q}}{\mathrm{argmin}} \; D_{\mathrm{KL}}\left(Q \| \Pi(\cdot|\mathcal{D})\right) \tag{9}$$

{eq:optim}

and is used in subsequent analysis as a proxy for the true posterior.

2 Co-ordinate ascent algorithm

Throughout our derivation we exploit the group independence structure within the prior and variational distribution, allowing the Radon-Nikodym derivative of Q with respect to the prior Π to be expressed as,

$$\frac{dQ}{d\Pi}(\beta) = \prod_{k=1}^{M} \frac{dQ_k}{d\Pi_k}(\beta_{G_k}) \tag{10}$$

Consider, the optimization problem (9) and recall the definition of the KL divergence

(A.1), it follows that

$$D_{\mathrm{KL}}(Q\|\Pi(\cdot|\mathcal{D})) = \mathbb{E}_Q\left[\log\frac{dQ}{d\Pi(\cdot|\mathcal{D})}\right] = \mathbb{E}_Q\left[\log\frac{\Pi_{\mathcal{D}}dQ}{e^{\ell(\mathcal{D};\beta)}d\Pi}\right]$$
$$= \mathbb{E}_Q\left[-\ell(\mathcal{D};\beta) + \log\frac{dQ}{d\Pi}\right] + \log\Pi_{\mathcal{D}} \quad (11)$$ {eq:opt}

As optimization of the objective (9) is invariant to constant terms, to simplify upcoming expressions we write them as C (the value of which may change line by line).

2.1 Updates of μ_{G_k} and σ_{G_k}

In order to update μ_{G_k} and σ_{G_k} we must assume that the group takes a non-zero value, i.e. $z_k = 1$. Hence,

$$\begin{aligned}
&\mathbb{E}_{Q|z_K=1}\left[-\ell(\mathcal{D};\beta) + \log\frac{dQ}{d\Pi}(\beta)\right] \\
&= \mathbb{E}_{Q|z_K=1}\left[-\ell(\mathcal{D};\beta) + \log\prod_{k=1}^{M}\frac{dQ_k}{d\Pi_k}(\beta_{G_k})\right] \\
&= \mathbb{E}_{Q|z_K=1}\left[-\ell(\mathcal{D};\beta) + \log\frac{dQ_K}{d\Pi_K}(\beta_{G_K}) + \log\prod_{k\neq K}\frac{dQ_k}{d\Pi_k}(\beta_{G_k})\right] \\
&= \mathbb{E}_{Q|z_K=1}\left[\frac{1}{2\tau^2}\|y - X\beta\|^2 + \log\frac{dQ_K}{d\Pi_K}(\beta_{G_K})\right] + C \\
&= \mathbb{E}_{Q|z_K=1}\left[\frac{1}{2\tau^2}\left\{\|X\beta\|^2 - 2\langle y, X\beta\rangle\right\} + \log\frac{dQ_K}{d\Pi_K}(\beta_{G_K})\right] + C \\
&= \mathbb{E}_{Q|z_K=1}\left[\frac{1}{2\tau^2}\left\{\mathrm{tr}\left(X^\top X\beta\beta^\top\right) - 2\sum_{k=1}^{M}\langle y, X_{G_k}\beta_{G_k}\rangle\right\} + \log\frac{dQ_K}{d\Pi_K}(\beta_{G_K})\right] + C \\
&= \mathbb{E}_{Q|z_K=1}\left[\frac{1}{2\tau^2}\,\mathrm{tr}\left(X^\top X\beta\beta^\top\right) - \frac{1}{\tau^2}\langle y, X_{G_K}\beta_{G_K}\rangle + \log\frac{dQ_K}{d\Pi_K}(\beta_{G_K})\right] + C \quad (12)
\end{aligned}$$ {eq:mu_sigma_1

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix and C is a constant term whose value does not depend on μ_{G_K} or σ_{G_K} (and may change line by line).

Consider the matrix $X^\top X\beta\beta^\top \in \mathbb{R}^{p\times p}$, using the fact that $\left(X^\top X\beta\beta^\top\right)_{ii} = \sum_{j=1}^{p}(X^\top X)_{ji}\beta_j\beta_i$ for $i, j = 1, \ldots, p$, we have

$$
\begin{aligned}
\mathbb{E}_{Q|z_K=1}\left[\operatorname{tr}\left(X^\top X\beta\beta^\top\right)\right] &= \mathbb{E}_{Q|z_K=1}\left[\sum_{i=1}^{p}\sum_{j=1}^{p}(X^\top X)_{ji}\beta_j\beta_i\right] \\
&= \sum_{i=1}^{p}\sum_{j=1}^{p}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right] \\
&= \sum_{i\in G_K}\left(\sum_{j=1}^{p}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]\right) + \sum_{i\in G_K^c}\left(\sum_{j=1}^{p}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]\right) \\
&= \sum_{i\in G_K}\left(\sum_{j\in G_K}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right] + \sum_{j\in G_K^c}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]\right) \\
&+ \sum_{i\in G_K^c}\left(\sum_{j\in G_K}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right] + \sum_{j\in G_K^c}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]\right) \\
&= \sum_{i\in G_K}\left(\sum_{j\in G_K}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right] + 2\sum_{j\in G_K^c}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]\right) + C
\end{aligned}
$$

Consider, $\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]$ and note $\operatorname{cov}(\beta_j, \beta_i) = \mathbb{E}\left[\beta_j\beta_i\right] - \mathbb{E}\left[\beta_j\right]\mathbb{E}\left[\beta_i\right]$, the previous display results in three cases

$$
\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right] = \begin{cases} \sigma_j^2 + \mu_j^2 & i, j \in G_K, i = j \\ \mu_j\mu_i & i, j \in G_K, i \neq j \\ \gamma_J\mu_j\mu_i & i \in G_K, j \in G_J, J \neq K \end{cases} \tag{13}
$$

The second term in (12) is straightforward and is given as,

$$
\mathbb{E}_{Q|z_K=1}\left[\langle y, X_{G_K}\beta_{G_K}\rangle\right] = \langle y, X_{G_K}\mathbb{E}_{Q|z_K=1}\left[\beta_{G_K}\right]\rangle = \langle y, X_{G_K}\mu_{G_K}\rangle \tag{14}
$$

{eq:mu_sigma_t

Finally, the third term in (12) is given as,

$$
\begin{aligned}
&\mathbb{E}_{Q|z_K=1}\left[\log \frac{dQ_K}{d\Pi_K}(\beta_{G_K})\right] \\
&\quad = \mathbb{E}_{Q|z_K=1}\left[\log \frac{\prod_{j\in G_K}\left(2\pi\sigma_j^2\right)^{-1/2}\exp\left\{-(2\sigma_j^2)^{-1}(\beta_j-\mu_j)^2\right\}}{C_K\lambda^{m_K}\exp\left(-\lambda\|\beta_{G_K}\|\right)}\right] \\
&\quad = \mathbb{E}_{Q|z_K=1}\left[\lambda\|\beta_{G_K}\| - \sum_{j\in G_K}\left(\log\sigma_j + \frac{1}{2\sigma_j^2}(\beta_j-\mu_j)^2\right)\right] + C \\
&\quad = \lambda\mathbb{E}_{Q|z_K=1}\left[\|\beta_{G_K}\|\right] - \sum_{j\in G_K}\log\sigma_j - \frac{m_K}{2} + C
\end{aligned} \tag{15}
$$

{eq:mu_simga_t

Evaluating the remaining expectation in (15) is non-trivial, in turn we derive an upper bound using Jensen's inequality,

$$
\mathbb{E}_{Q|z_K=1}\left[\|\beta_{G_K}\|\right] \le \left(\sum_{j\in G_K}\mathbb{E}_{Q|z_K=1}\left[\beta_j^2\right]\right)^{1/2} = \left(\sum_{j\in G_K}\sigma_j^2+\mu_j^2\right)^{1/2} \tag{16}
$$

{eq:mu_sigma_u

Putting these components together gives

$$
\begin{aligned}
&\mathbb{E}_{Q|z_K=1}\left[-\ell(\mathcal{D};\beta) + \log\frac{dQ}{d\Pi}(\beta)\right] \\
&\quad \le \frac{1}{2\tau^2}\sum_{i\in G_K}\left((X^\top X)_{ii}(\sigma_i^2+\mu_i^2) + \sum_{j\in G_K, j\neq i}(X^\top X)_{ji}\mu_j\mu_i\right) \\
&\quad + \frac{1}{\tau^2}\sum_{i\in G_K}\left(\sum_{j\in G_K^c}(X^\top X)_{ji}\gamma_J\mu_j\mu_i\right) - \frac{1}{\tau^2}\langle y, X_{G_K}\mu_{G_K}\rangle \\
&\quad - \sum_{i\in G_K}\log\sigma_i + \lambda\left(\sum_{i\in G_K}\sigma_i^2+\mu_i^2\right)^{1/2} + C
\end{aligned} \tag{17}
$$

{eq:mu_sigma_m

2.1.1 Group-wise update for μ_{G_K}

Re-writing the RHS of (17) in terms of μ_{G_K}, we have

$$\frac{1}{2\tau^2}\mu_{G_K}^\top X_{G_K}^\top X_{G_K}\mu_{G_K} + \frac{1}{\tau^2}\sum_{J\neq K}\gamma_J \mu_{G_K}^\top X_{G_K}^\top X_{G_J}\mu_{G_J} - \frac{1}{\tau^2}\langle y, X_{G_K}\mu_{G_K}\rangle + \lambda\left(\sigma_{G_K}^\top\sigma_{G_K} + \mu_{G_K}^\top\mu_{G_K}\right)^{1/2} + C \tag{18}$$

{eq:mu_gk}

Insight into the update of μ_{G_K}

Using the fact that $\left(\sum_{j\in G_K}\sigma_j^2 + \mu_j^2\right)^{1/2} \leq 1 + \sum_{j\in G_K}\sigma_j^2 + \mu_j^2$, which follows from $x \leq 1 + x^2$, we can upper bound (18).

$$\frac{1}{2\tau^2}\|X_{G_K}\mu_{G_K}\|^2 + \frac{1}{\tau^2}\sum_{J\neq K}\gamma_J \mu_{G_K}^\top X_{G_K}^\top X_{G_J}\mu_{G_J} - \frac{1}{\tau^2}\langle y, X_{G_K}\mu_{G_K}\rangle + \lambda\|\mu_{G_K}\|^2 + C \tag{19}$$

{eq:mu_gk_upper}

And in turn (19) is minimized by

$$\hat{\mu}_{G_K} = \Xi^{-1}X_{G_K}^\top y - \Xi^{-1}\sum_{J\neq K}\gamma_J X_{G_K}^\top X_{G_J}\mu_{G_J} \tag{20}$$

{eq:mu_gk_min}

where $\Xi = X_{G_K}^\top X_{G_K} + 2\lambda\tau^2 I_{m_K}$.

Let $P := (X_{G_K}^\top X_{G_K} + 2\lambda\tau^2 I_{m_K})^{-1}X_{G_K}^\top$ and the prediction of y from μ_{G_J} as $\hat{y}_{G_J} = X_{G_J}\mu_{G_J}$, then we can re-express (20) as

$$\hat{\mu}_{G_K} = P(y - \sum_{J\neq K}\gamma_J\hat{y}_{G_J}) \tag{21}$$

In other words the minimizer $\hat{\mu}_{G_K}$ seeks a vector that explains the remaining signal in y given the signal explained by $\sum_{J\neq K}\gamma_J\hat{y}_{G_J}$, for example, consider the extreme case, $y - \sum_{J\neq K}\gamma_J\hat{y}_{G_J} = 0_n$ (the n-dimensional zero vector), then the resulting minimizer $\hat{\mu}_{G_K} = 0_{m_K}$.

It turns out, updating using (20) doesn't work that well in practise, in fact, using optimization routines to minimize (18) leads to better results in less time.

We've seen that using a looser upper bound does not work well, Q: how loose it too loose, there is a point at which we move from obtaining a good estimate to one where we obtain a bad one - and this depends on the bound. When does this transition happen?

2.1.2 Group-wise update for σ_{G_K}

Re-writing the RHS of (17) in terms of σ_{G_K}, we have

$$\sum_{i \in G_K} \left(\frac{1}{2\tau^2} (X^\top X)_{ii} \sigma_i^2 - \log \sigma_i \right) + \lambda \left(\sum_{i \in G_K} \sigma_i^2 + \mu_i^2 \right)^{1/2} + C \tag{22}$$

{eq:sig_gk}

2.2 Updates for γ_K

Similarly for γ_K we evaluate the expectation with respect to Q, however without conditioning on the group being non-zero.

$$\begin{aligned}
&\mathbb{E}_Q \left[-\ell(\mathcal{D}; \beta) + \log \frac{dQ}{d\Pi}(\beta) \right] \\
&= \mathbb{E}_Q \left[-\ell(\mathcal{D}; \beta) + \mathbb{I}_{\{z_K = 1\}} \log \frac{\gamma_K dN_K}{\bar{w} d\Psi_K}(\beta_{G_K}) + \mathbb{I}_{\{z_K = 0\}} \log \frac{1 - \gamma_K}{1 - \bar{w}} \right] + C \\
&= \frac{1}{2\tau^2} \sum_{i \in G_K} \left(\sum_{j \in G_K} (X^\top X)_{ji} \mathbb{E}_Q \left[\beta_j \beta_i \right] + 2 \sum_{j \in G_K^c} (X^\top X)_{ji} \mathbb{E}_Q \left[\beta_j \beta_i \right] \right) \\
&- \frac{1}{\tau^2} \langle y, X_{G_K} \mathbb{E}_Q \left[\beta_{G_K} \right] \rangle - \frac{\gamma_K}{2} \sum_{j \in G_K} \log \left(2\pi\sigma_j^2 \right) - \gamma_K \log(C_K) \\
&- \gamma_K m_K \log(\lambda) + \mathbb{E}_Q \left[\mathbb{I}_{\{z_K = 1\}} \left(\lambda \| \beta_{G_K} \| - \sum_{j \in G_K} \frac{1}{2\sigma_j^2} (\beta_j - \mu_j)^2 \right) \right] \\
&+ \gamma_K \log \frac{\gamma_K}{\bar{w}} + (1 - \gamma_K) \log \frac{1 - \gamma_K}{1 - \bar{w}} + C
\end{aligned}$$

Noting $\mathbb{E}_Q\left[\beta_{G_K}\right] = \gamma_K \mu_{G_K}$ and

$$\mathbb{E}_Q\left[\beta_j \beta_i\right] = \begin{cases} \gamma_K(\sigma_j^2 + \mu_j^2) & i, j \in G_K, i = j \\ \gamma_K \mu_j \mu_i & i, j \in G_K, i \neq j \\ \gamma_K \gamma_J \mu_j \mu_i & i \in G_K, j \in G_J, J \neq K \end{cases} \tag{23}$$

and

$$\mathbb{E}_Q\left[\mathbb{I}_{\{z_K=1\}} \|\beta_{G_K}\|\right] = \gamma_K \mathbb{E}_{N_K}\left[\|\beta_{G_K}\|\right] \leq \gamma_K \left(\sum_{j \in G_K} \sigma_j^2 + \mu_j^2\right)^{1/2} \tag{24}$$

Substituting these expressions into the previous display gives,

$$\begin{aligned} &\mathbb{E}_Q\left[-\ell(\mathcal{D}; \beta) + \log \frac{dQ}{d\Pi}(\beta)\right] \\ &\leq \frac{\gamma_K}{2\tau^2} \sum_{i \in G_K} \left((X^\top X)_{ii}(\sigma_i^2 + \mu_i^2) + \sum_{j \in G_K, j \neq i} (X^\top X)_{ji} \mu_j \mu_i\right) \\ &+ \frac{\gamma_K}{\tau^2} \sum_{i \in G_K} \left(\sum_{j \in G_K^c} (X^\top X)_{ji} \gamma_J \mu_j \mu_i\right) - \frac{\gamma_K}{\tau^2} \langle y, X_{G_K} \mu_{G_K} \rangle \\ &- \frac{\gamma_K}{2} \sum_{j \in G_K} \log\left(2\pi\sigma_j^2\right) - \gamma_K \log(C_K) - \gamma_K m_K \log(\lambda) + \lambda \gamma_K \left(\sum_{j \in G_K} \sigma_j^2 + \mu_j^2\right)^{1/2} \\ &- \frac{\gamma_K m_K}{2} + \gamma_K \log \frac{\gamma_K}{\bar{w}} + (1 - \gamma_K) \log \frac{1 - \gamma_K}{1 - \bar{w}} + C \end{aligned}$$

Differentiating the RHS of the previous display with respect to γ_K, setting to zero and re-arranging gives the update equation for γ_K, formally,

$$\begin{aligned} \log \frac{\gamma_K}{1 - \gamma_K} &= \log \frac{\bar{w}}{1 - \bar{w}} + \frac{m_K}{2} + \frac{1}{\tau^2} \langle y, X_{G_K} \mu_{G_K} \rangle \\ &+ \frac{1}{2} \sum_{j \in G_K} \log\left(2\pi\sigma_j^2\right) + \log(C_K) + m_K \log(\lambda) - \left\{ \lambda \left(\sum_{j \in G_K} \sigma_j^2 + \mu_j^2\right)^{1/2} \right. \\ &\left. + \frac{1}{2\tau^2} \sum_{i \in G_K} \left((X^\top X)_{ii} \sigma_i^2 + \sum_{j \in G_K} (X^\top X)_{ji} \mu_j \mu_i + 2 \sum_{j \in G_K^c} (X^\top X)_{ji} \gamma_J \mu_j \mu_i\right) \right\} \end{aligned} \tag{25}$$

{eq:update_gam

2.3 Evidence Lower Bound

The evidence lower bound, acts as a lower bound for the model evidence $\Pi_{\mathcal{D}}$, and follows from the definition of the KL divergence,

$$0 \leq D_{\mathrm{KL}}(Q \| \Pi(\cdot|\mathcal{D})) = \mathbb{E}_Q \left[\Pi_{\mathcal{D}} - \ell(\mathcal{D}; \beta) - \log \frac{d\Pi}{dQ}(\beta) \right]$$
$$\implies \mathbb{E}_Q \left[\ell(\mathcal{D}; \beta) + \log \frac{d\Pi}{dQ}(\beta) \right] \leq \Pi_{\mathcal{D}}$$

Formally, the ELBO is defined as,

$$\mathcal{L}(\mathcal{D}) = \mathbb{E}_Q \left[\ell(\mathcal{D}; \beta) + \log \frac{d\Pi}{dQ}(\beta) \right] \tag{26}$$

In effect, maximizing the ELBO is equivalent to minimizing $D_{\mathrm{KL}}(Q \| \Pi(\cdot|\mathcal{D}))$, i.e. solving (9). Often the ELBO is used to assess the convergence of co-ordinate ascent algorithms, but can also act as a goodness of fit measure.

For our model the ELBO is given as

$$\begin{aligned} \mathcal{L}(\mathcal{D}) = &-\frac{n}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\left(\|y\|^2 + \sum_{i=1}^{p}\sum_{j=1}^{p} (X^\top X)_{ji} \mathbb{E}_Q\left[\beta_j \beta_i\right] \right) \\ &+ \sum_{K=1}^{M} \left(\frac{1}{\tau^2}\gamma_K \langle y, X_{G_K} \mu_{G_K} \rangle + \frac{\gamma_K}{2} \sum_{j \in G_K} \left(\log(2\pi\sigma_j^2) \right) + \gamma_K \log(C_K) + \frac{\gamma_K m_K}{2} \right. \\ &\left. + \gamma_K m_K \log(\lambda) - \mathbb{E}_Q\left[\mathbb{I}_{z_K = 1} \lambda \|\beta_{G_K}\| \right] - \gamma_K \log\frac{\gamma_K}{\bar{w}} - (1-\gamma_K)\log\frac{1-\gamma_K}{1-\bar{w}} \right) \end{aligned} \tag{27}$$

{eq:elbo}

where the expectation $\mathbb{E}_Q\left[\beta_i \beta_j\right]$ is defined earlier. Given $\mathbb{E}_Q\left[\mathbb{I}_{\{z_K=1\}} \lambda \|\beta_{G_K}\|\right]$ does not have a closed form expression, Monte-Carlo integration to evaluate it.

2.4 Updating τ

As of yet we have not placed a prior on τ^2 or considered updating τ^2 as part of our co-ordinate ascent algorithm. This is because, under any independent prior for τ^2

and independent factorization in the mean-field variational family, we can modify (without major alterations) our existing equations for μ_{G_K}, σ_{G_K}, γ_K and ELBO to include this modelling assumption.

We opt to use an inverse Gamma prior, predominately because it is a popular choice amongst practitioners (Browne and Draper, 2006). Formally, we let,

$$\tau^2 \sim \Gamma^{-1}(a, b) \tag{28}$$

where $\Gamma^{-1}(a, b)$ denotes the inverse Gamma distribution with density $\frac{b^a}{\Gamma(a)} x^{-a-1} e^{-b/x}$. Extending our prior (6) to include this modelling assumption for τ^2, we have,

$$\Pi'(\beta, \tau^2) = \Pi(\tau^2)\Pi(\beta) = \Gamma^{-1}(\tau^2; a, b)\Pi(\beta) \tag{29}$$

Similarly, extending our variational family with respect to τ^2, we have,

$$\mathcal{Q}' = \{\Gamma^{-1}(\tau^2; a', b') : a' > 0, b' > 0\} \times \mathcal{Q} \tag{30}$$

where we denote,

$$Q' = \Gamma^{-1}(\tau^2; a', b') \otimes Q(\beta) \in \mathcal{Q}' \tag{31}$$

Given these extension, our original optimization problem (11) is extended to be

$$\mathbb{E}_{Q'}\left[-\ell(\mathcal{D}; \beta) + \log \frac{dQ'}{d\Pi'}\right] = \mathbb{E}_{Q'}\left[-\ell(\mathcal{D}; \beta) + \log \frac{dQ}{d\Pi} + \log \frac{d\Gamma^{-1}(a', b')}{d\Gamma^{-1}(a, b)}(\tau^2)\right] \tag{32}$$

Consequently, when deriving the new update equations for μ_{G_K}, σ_{G_K} and γ_K, this new term $\mathbb{E}[\log \frac{d\Gamma^{-1}}{d\Gamma^{-1}}]$ is a constant and can be ignored. Therefore, we need only replace all occurrences of $1/\tau^2$ in our old update equations with $\mathbb{E}_{\Gamma^{-1}}[1/\tau^2] = a'/b'$. Regarding the ELBO we do the same, however we also include the term $\mathbb{E}_{\Gamma^{-1}(a',b')}[\log \frac{d\Gamma^{-1}(a',b')}{d\Gamma^{-1}(a,b)}]$ which is given by,

$$a' \log(b') - a \log(b) + \log \frac{\Gamma(a)}{\Gamma(a')} + (a - a')(\log(b') + \kappa(a')) + (b - b')\frac{a'}{b'} \tag{33}$$

noting $\mathbb{E}_{\Gamma^{-1}}[\log(\tau^2)] = \log(b') + \kappa(a')$ where $\kappa(\cdot)$ is the digamma function.

To update a' and b' we follow a similar procedure as before, writing (11) as a function of a' and b' and then finding the minimum,

$$\begin{aligned}
&\mathbb{E}_{Q'}\left[-\ell(\mathcal{D};\beta) + \log\frac{dQ'}{d\Pi'}\right] = \mathbb{E}_{Q'}\left[-\ell(\mathcal{D};\beta) + \log\frac{d\Gamma^{-1}(a',b')}{d\Gamma^{-1}(a,b)}(\tau^2)\right] + C \\
&= \mathbb{E}_{Q'}\left[\frac{n}{2}\log(\tau^2) + \frac{1}{2\tau^2}\|y - X\beta\|^2\right] + \mathbb{E}_{\Gamma^{-1}}\left[\log\frac{d\Gamma^{-1}(a',b')}{d\Gamma^{-1}(a,b)}(\tau^2)\right] + C \\
&= \mathbb{E}_{\Gamma^{-1}}\left[\frac{1}{2\tau^2}\right]\mathbb{E}_Q\left[\|y - X\beta\|^2\right] + \mathbb{E}_{\Gamma^{-1}}\left[\frac{n}{2}\log(\tau^2) + \log\frac{d\Gamma^{-1}(a',b')}{d\Gamma^{-1}(a,b)}(\tau^2)\right] + C \\
&= \frac{a'}{2b'}\left(\|y\|^2 - 2\langle y, X\mathbb{E}_Q[\beta]\rangle + \sum_{i=1}^{p}\sum_{j=1}^{p}(X^\top X)_{ji}\mathbb{E}_Q[\beta_j\beta_i]\right) + a'\log(b') \\
&- \log\Gamma(a') + \left(\frac{n}{2} + a - a'\right)(\log(b') + \kappa(a')) + (b - b')\frac{a'}{b'} + C
\end{aligned} \tag{34}$$

{eq:update_a_b

Note: finding the values of a' and b' that minimizes (34) needs to be done jointly. This is contrary to the outline of CAVI where parameters are optimized independently of one another. We've seen this before when optimizing μ_{G_k}. For instance, optimizing each element of μ_{G_k} independently of the rest does not lead to a stable algorithm.

Q: why is it the case that some parameters need to be optimized jointly and others do not? And when do they need to be optimized together or can be optimized separately.

H: maybe this has something to do with how the first moment is defined, i.e. the first moment of τ^2 is given by $b'/(a'-1)$ which depends on both a' and b'.

Q: does this hold for certain types of distributions? i.e. do exponential family distributions with multiple parameter means need to have those parameters optimized jointly?

2.5 Implementation details

Until convergence repeat the following steps:

1. For $k = 1, \ldots, M$

   (a) Update $\mu_{G_k} \leftarrow \underset{\mu_{G_k} \in \mathbb{R}^{m_k}}{\text{argmin}} \; f(\mu_{G_k}; \mu_{G_k^c}, \sigma, \gamma, \tau)$

   (b) Update $\sigma_{G_k} \leftarrow \underset{\sigma_{G_k} \in \mathbb{R}^{m_k}}{\text{argmin}} \; g(\mu_{G_k}; \mu_{G_k^c}, \sigma, \gamma, \tau)$

   (c) Update $\gamma_k \leftarrow \text{sigmoid } h(\mu_{G_k}; \mu_{G_k^c}, \sigma, \gamma, \tau)$

   (d) Update a', b' minimizers of (34)

The algorithm can be sensitive to initialization, in our implementation we used the group LASSO from the package `gglasso` to initialize μ. To initialize σ_k

3 Extension of Variational Family

We would like to capture the dependence within the groups. We therefore consider variational family

$$\mathcal{Q}_D = \left\{ Q_D = \bigotimes_{k=1}^{M} \left[\gamma_k \; N\left(\mu_{G_k}, \Sigma_k\right) + (1 - \gamma_k)\delta_0 \right] \right\} \tag{35}$$

where $\Sigma_k \in \mathbb{R}^{m_k \times m_k}$ is a positive semi-definite matrix. We proceed as before deriving the update equations for

3.1 Updates for μ_{G_k} and Σ_k

We follow a similar process as before, recall from (12), we have:

$$\mathbb{E}_{Q_D|z_K=1}\left[-\ell(\mathcal{D};\beta)+\log\frac{dQ_D}{d\Pi}(\beta)\right]$$
$$=\mathbb{E}_{Q_D|z_K=1}\left[\frac{1}{2\tau^2}\operatorname{tr}\left(X^\top X\beta\beta^\top\right)-\frac{1}{\tau^2}\langle y,X_{G_K}\beta_{G_K}\rangle+\log\frac{dQ_{D,K}}{d\Pi_K}(\beta_{G_K})\right]+C \quad (36)$$ {eq:QD_mu_sigm

Approaching each term separately using previous results we have,

$$\mathbb{E}_{Q_D|z_K=1}\left[\operatorname{tr}\left(X^\top X\beta\beta^\top\right)\right] \quad (37)$$
$$=\sum_{i\in G_K}\left(\sum_{j\in G_K}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]+2\sum_{j\in G_K^c}(X^\top X)_{ji}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]\right)+C$$

with

$$\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right]=\begin{cases}\Sigma_{i,j}+\mu_i\mu_j & i,j\in G_K\\ \gamma_J\mu_j\mu_i & i\in G_K, j\in G_J, J\neq K\end{cases} \quad (38)$$

The middle term is trivial and finally,

$$\mathbb{E}_{Q_D|z_K=1}\left[\log\frac{dQ_K}{d\Pi_K}(\beta_{G_K})\right]=-\frac{1}{2}\log\det\Sigma_K+\lambda E_{Q_D|z_K=1}\|\beta_{G_K}\|+C \quad (39)$$

Using (16) to upper bound the previous display gives the objective function for μ_{G_K} and Σ_K. Formally,

$$\frac{1}{2\tau^2}\mu_{G_K}^\top X_{G_K}^\top X_{G_K}\mu_{G_K}+\frac{1}{2\tau^2}\operatorname{tr}(X_{G_K}^\top X_{G_K}\Sigma_K)+\frac{1}{\tau^2}\sum_{J\neq K}\gamma_J\mu_{G_K}^\top X_{G_K}^\top X_{G_J}\mu_{G_J}$$
$$-\frac{1}{\tau^2}\langle y,X_{G_K}\mu_{G_K}\rangle-\frac{1}{2}\log\det\Sigma_K+\lambda\left(\sum_{i\in G_K}\sigma_i^2+\mu_i^2\right)^{1/2}+C \quad (40)$$ {eq:QD_mu_sigm

Restricting the number of free parameters

We use similar ideas to those of (Seeger, 1999; Opper and Archambeau, 2009, pg. 119) and show that only $2m_k$ free parameters are needed to describe the optima of

(40). Naively we would expect to need $(m_k + 1)m_k/2$ for the covariance Σ_K and m_k for μ_k.

To show this, let $\Psi_K = X_{G_K}^\top X_{G_K}$ and $\nu_K = \lambda(\sum_{i \in G_K} \sigma_i^2 + \mu_i^2)^{1/2}$, further write,

$$\widetilde{\pi}_K = \frac{\partial \nu_K}{\partial \mu_{G_K}} = \left(\frac{\partial \nu_K}{\partial \mu_{G_K,1}}, \ldots, \frac{\partial \nu_K}{\partial \mu_{G_K,m_K}} \right)^\top, \quad \widetilde{W}_K = \frac{\partial \nu_K}{\partial \Sigma_K} = \operatorname{diag} \frac{\partial \nu_K}{\partial \sigma^2_{G_K,i}} \tag{41}$$

Differentiating (40) with respect to μ_{G_k} gives

$$\frac{1}{\tau^2}\Psi_K \mu_{G_K} + \frac{1}{\tau^2} \sum_{J \neq K} \gamma_J X_{G_K}^\top X_{G_J} \mu_{G_J} - \frac{1}{\tau^2} X_{G_K}^\top y + \widetilde{\pi}_K \tag{42}$$

setting to zero and re-arranging gives

$$\widehat{\mu}_{G_K} = -\Psi_K^{-1} \left(\sum_{J \neq K} \left(\gamma_J X_{G_K}^\top X_{G_J} \mu_{G_J} \right) - X_{G_K}^\top y + \tau^2 \widetilde{\pi}_K \right) \tag{43}$$

{eq:mu_free}

Similarly differentiating (40) wrt. Σ_K gives,

$$\frac{1}{2\tau^2}\Psi_K - \frac{1}{2}\Sigma_K^{-1} + \widetilde{W}_K \tag{44}$$

setting to zero and re-arranging gives,

$$\widehat{\Sigma}_K = \left(\tau^{-2}\Psi_K + 2\widetilde{W}_K \right)^{-1} \tag{45}$$

{eq:sigma_free

We see that (43) and (45) depend on m_K free parameters each. This follows for (45) since $\widetilde{W}_K$ is a diagonal matrix. Therefore writing,

$$\mu_{G_K} = -\Psi_K^{-1} \left(\sum_{J \neq K} (\gamma_J X_{G_K}^\top X_{G_J} \mu_{G_J} - X_{G_K}^\top y + \tau^2 \pi_K \right)$$

where $\pi_K = (\pi_{K,1}, \ldots, \pi_{K,m_K})^\top$, and $\Sigma_K = (\tau^{-1}\Psi_K + \operatorname{diag}(w_K))^{-1}$ where $w_K = (w_{K,1}, \ldots, w_{K,m_K})^\top$ allows us to represent the optima in terms of $2m_K$ free parameters. Notably, although the number of free parameters has not increased from the

independent to the unconstrained covariance case, optimization of Σ_K requires the inversion of an $m_K \times m_K$ matrix, which would be time consuming for large m_K (inversion is $O(m_K^3)$).

To finish things up we need the derivatives with respect to the new free parameters, π_K and w_K. Denoting the objective in (36) as f we have,

$$\begin{aligned}\frac{\partial f}{\partial \pi_{K,i}} &= \operatorname{tr}\left(\left(\frac{\partial f}{\partial \mu_K}\right)^\top \frac{\partial \mu_K}{\partial \pi_{K,i}}\right) \\ &= -\tau^2 \frac{\partial f}{\partial \mu_K}^\top \Psi_{K,:i}^{-1} = -\tau^2 \Psi_{K,i:}^{-1} \frac{\partial f}{\partial \mu_K}\end{aligned}$$

and therefore

$$\frac{\partial f}{\partial \pi_K} = \tau^2 \Psi_K^{-1} \left(\pi_K - \widetilde{\pi}_K\right) \tag{46}$$ {eq:QD_pi_grad

Similarly,

$$\begin{aligned}\frac{\partial f}{\partial w_{K,i}} &= \operatorname{tr}\left(\left(\frac{\partial f}{\partial \Sigma_K}\right)^\top \frac{\partial \Sigma_K}{\partial w_{K,i}}\right) \\ &= \operatorname{tr}\left(-\left(\widetilde{W} - \frac{1}{2}W\right)\Sigma_K \frac{\partial \Sigma_K^{-1}}{\partial w_{K,i}} \Sigma_K\right) \\ &= \sum_{j=1}^{m_K}\left(\frac{1}{2}w_{K,j} - \widetilde{W}_{j,j}\right)\Sigma_{K,ij}^2\end{aligned}$$

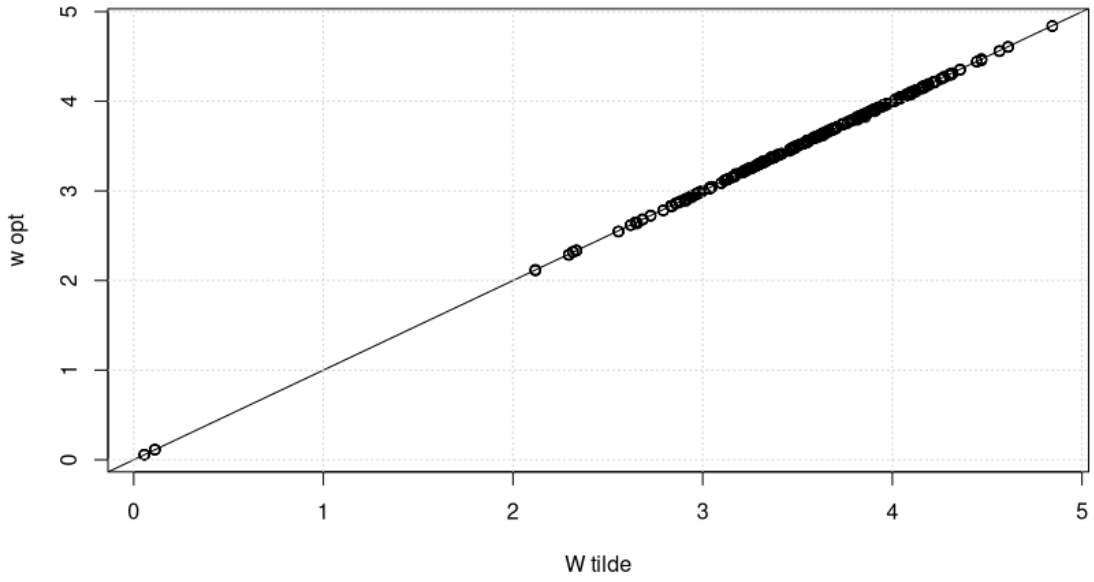
Hence

$$\frac{\partial f}{\partial w_K} = \frac{1}{2}\Sigma_K \circ \Sigma_K \left(w_K - \widetilde{w}_K\right) \tag{47}$$ {eq:QD_w_grad}

where $\widetilde{w}_K = \left(2\partial\nu/\partial\sigma^2_{G_K,1}, \ldots, 2\partial\nu/\partial\sigma^2_{G_K,m_K}\right)^\top$.

SF: are we sure that after optimizing for $w_{K,i}$, the optimum ends up being the same as $2\tilde{W}_K$?

we've only managed to check this empirically. As it turns out $w_{K,i}$ is (approximately, ± 0.001 on average) the same as $2\tilde{W}_{K,ii}$



3.2 Update of γ_K

The update equation for γ_k is given by solving,

$$\log \frac{\gamma_K}{1-\gamma_K} = \log \frac{\bar{w}}{1-\bar{w}} + \frac{m_K}{2} + \frac{1}{\tau^2}\langle y, X_{G_K}\mu_{G_K}\rangle + \frac{1}{2}\log\det\left(2\pi\Sigma_K\right) + \log(C_K) + m_K\log(\lambda) - \left\{\lambda\left(\sum_{j\in G_K}\sigma_j^2 + \mu_j^2\right)^{1/2} + \frac{1}{2\tau^2}\sum_{i\in G_K}\left(\sum_{j\in G_K}\left[(X^\top X)_{ii}(\Sigma_{k,ij} + \mu_j\mu_i)\right] + 2\sum_{j\in G_K^c}(X^\top X)_{ji}\gamma_J\mu_j\mu_i\right)\right\} \tag{48}$$

{eq:QD_update_

initialization is important, particularly for μ, in some cases poor initialization can lead to optimization issues and a lack of convergence

3.3 Updates for a' and b'

Updates for a' and b' are essentially the same except the expectation in (34), which is now given by,

$$\mathbb{E}_Q\left[\beta_j\beta_i\right] = \begin{cases} \gamma_K(\Sigma_{i,j} + \mu_i\mu_j) & i,j \in G_K \\ \gamma_K\gamma_J\mu_j\mu_i & i \in G_K, j \in G_J, J \neq K \end{cases} \tag{49}$$

3.4 ELBO

Minor changes are also made to the evidence lower bound.

$$\begin{aligned} \mathcal{L}(\mathcal{D}) = &-\frac{n}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\left(\|y\|^2 + \sum_{i=1}^{p}\sum_{j=1}^{p}(X^\top X)_{ji}\mathbb{E}_Q\left[\beta_j\beta_i\right]\right) \\ &+ \sum_{K=1}^{M}\left(\frac{1}{\tau^2}\gamma_K\langle y, X_{G_K}\mu_{G_K}\rangle + \frac{\gamma_K}{2}\log(\det(2\pi\Sigma_K)) + \gamma_K\log(C_K) + \frac{\gamma_K m_K}{2}\right. \\ &\left. + \gamma_K m_K\log(\lambda) - \mathbb{E}_Q\left[\mathbb{I}_{z_K=1}\lambda\|\beta_{G_K}\|\right] - \gamma_K\log\frac{\gamma_K}{\bar{w}} - (1-\gamma_K)\log\frac{1-\gamma_K}{1-\bar{w}}\right) \end{aligned} \tag{50}$$

{eq:QD_elbo}

4 Extension to binary classification

Formally, we are going to be modelling a binary response $Y \in \{0, 1\}$ alongside a feature vector $x \in \mathbb{R}^p$, by using the popular logistic link function wherein,

$$\mathbb{P}(Y = 1|X = x) = \text{logistic}(x^\top \beta) = \frac{\exp(x^\top \beta)}{1 + \exp(x^\top \beta)} \tag{51}$$

where $\beta \in \mathbb{R}^p$ is the coefficient vector.

In a similar fashion as before we let the training data $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^n$ where $y_i \in \{0, 1\}, x_i \in \mathbb{R}^p, y = (y_1, \ldots, y_n)^\top$ and $X = (x_1, \ldots, x_n)^\top \in \mathbb{R}^{n \times p}$. We are going to assume that the group structure is known. We will be using the same notation introduced earlier for the groups.

Variational inference for logistic regression requires greater care than the linear counterpart, this is because there is no closed form for the expectation of the log-likelihood under the variational family. Several authors have proposed bounds or approximations to maintain tractability cite a few people e.g. JJ99, KM13 etc..

Recently Depraetere and Vandebroek (2017), compared several different approximations and bounds. Their results highlight that quadratic bounds do not perform well, Taylor series approximations are only accurate in specific settings. They conclude that non-quadratic bounds can perform well. Finally, the authors introduce an approximation based on quasi-Monte Carlo, TODO: define, and show that the method performs better, albeit slower, in their proposed simulations. Finally, the authors note that a potential avenue for future research would be to improve the variational approximation both in terms of accuracy and computation time. This could be accomplished through tighter bounds or faster approximations.

In light of this, we propose a non-quadratic bound based on the inequality

$$\log(1+\exp(x)) \leq \frac{1}{2}\left(x+\sqrt{2+x^2}\right) \tag{52}$$

A comparison presented in Figure 2 indicates the bound is tight, particularly around the origin.

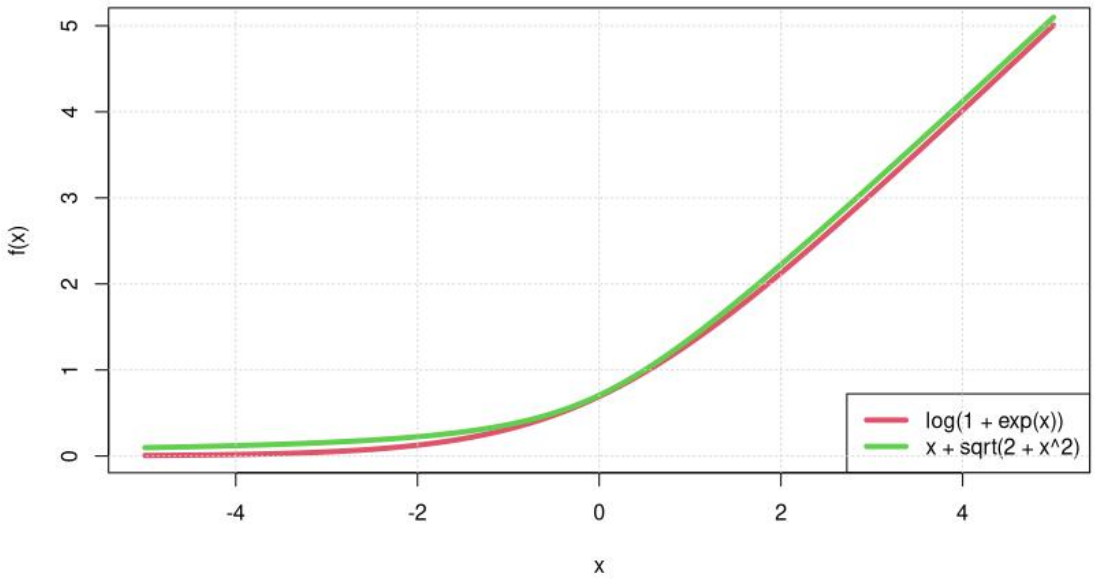


Figure 2: Comparison to bound

fig:bounds_com

4.1 Co-ordinate ascent update equations

We begin by writing the log-likelihood,

$$\begin{aligned}\ell(\mathcal{D}, \beta) &= \sum_{i=1}^{n} y_i \log\left(\frac{\exp(x_i^\top \beta)}{1+\exp(x_i^\top \beta)}\right) + (1-y_i)\log\left(\frac{1}{1+\exp(x_i^\top \beta)}\right) \\ &= \sum_{i=1}^{n} y_i \left(x_i^\top \beta\right) - \log\left(1+\exp(x_i^\top \beta)\right) \end{aligned} \tag{53}$$

And proceed as before, aiming to minimize

$$\mathbb{E}\left[-\ell(\mathcal{D};\beta) + \log \frac{dQ}{d\Pi}(\beta)\right] \tag{54}$$

{eq:l_opt_obje

As we've already derived results for the second term in the expectation, our focus falls to the expectation of the negative log-likelihood,

$$\sum_{i=1}^{n} \mathbb{E}\left[\log\left(1+\exp(x_i^\top \beta)\right) - y_i\left(x_i^\top \beta\right)\right] \tag{55}$$

{eq:l_objectiv

We note the first term is intractable and therefore propose bounding it using the fact that

$$\begin{aligned}\mathbb{E}\left[\log(1+\exp(x_i^\top \beta))\right] &\leq \mathbb{E}\left[\frac{1}{2}\left(x_i^\top \beta + \sqrt{2+(x_i^\top \beta)^2}\right)\right] \\ &\leq \frac{1}{2}\left(\mathbb{E}\left[x_i^\top \beta\right] + \sqrt{2+\mathbb{E}\left[(x_i^\top \beta)^2\right]}\right)\end{aligned} \tag{56}$$

where the first inequality follows form that fact that $\log(1+\exp(x)) \leq \frac{1}{2}\left(x+\sqrt{2+x^2}\right)$ and the second inequality from Jensens. Hence (55) is bound by

$$\sum_{i=1}^{n}\left(\frac{1}{2}-y_i\right)\mathbb{E}\left[x_i^\top \beta\right] + \frac{1}{2}\left(2+\sum_{j=1}^{p}\sum_{k=1}^{p} x_{ij}x_{ik}\mathbb{E}\left[\beta_j\beta_k\right]\right)^{1/2} \tag{57}$$

In the meantime, we will be approximating the posterior using the variational family

$\mathcal{Q}$ (the independent factorization). Wherein

$$\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_i\right] = \begin{cases} \sigma_j^2 + \mu_j^2 & i,j \in G_K, i = j \\ \mu_j\mu_i & i,j \in G_K, i \neq j \\ \gamma_J\mu_j\mu_i & i \in G_K, j \in G_J, J \neq K \\ \gamma_I\gamma_J\mu_j\mu_i & i \in G_I, j \in G_J, I \neq J \neq K \\ \gamma_J(\sigma_j^2 + \mu_j^2) & i,j \in G_J, i = j, J \neq K \\ \gamma_J\mu_j\mu_i & i,j \in G_K, i \neq j, J \neq K \end{cases} \tag{58}$$

4.1.1 Updates for μ_{G_K} and σ_{G_K}

Writing (54) as a function of μ_{G_K} and σ_{G_K} whilst keeping the remaining parameters fixed, gives

$$\begin{aligned} &\mathbb{E}_{Q|z_K=1}\Big[-\ell(\mathcal{D};\beta) + \log\frac{dQ}{d\Pi}(\beta)\Big] \\ &\quad\leq \sum_{i=1}^{n}\left[\left(\frac{1}{2} - y_i\right)x_{i,G_K}^\top\mu_{G_K} + \frac{1}{2}\left(2 + \sum_{j=1}^{p}\sum_{k=1}^{p} x_{ij}x_{ik}\mathbb{E}_{Q|z_K=1}\left[\beta_j\beta_k\right]\right)^{1/2}\right] \\ &\quad - \sum_{i\in G_K}\log\sigma_i + \lambda\left(\sum_{i\in G_K}\sigma_i^2 + \mu_i^2\right)^{1/2} + C \end{aligned} \tag{59}$$

{eq:l_mu_sig_u

where C is a constant term. Optimizing the surrogate functional in (59) with respect to either μ_{G_K} or σ_{G_K}, whilst keeping the other fixed, gives the update equations for μ_{G_K} and σ_{G_K}.

4.1.2 Updates for γ_K

Finally, for the update equation of γ_K, we are going to be using the factorization of the variational family to obtain a closed form update. Formally we use the fact that,

$$\mathbb{E}_Q\left[f(\beta)\right] = \gamma_K \mathbb{E}_{N_K \otimes Q_{\setminus K}}\left[f(\beta)\right] + (1-\gamma_K)\mathbb{E}_{\delta_{0,K}\otimes Q_{\setminus K}}\left[f(\beta)\right] \tag{60}$$

where $Q_{\setminus K} = \bigotimes_{j\neq K}\left[\gamma_j N(\mu_{G_K}, \mathrm{diag}(\sigma^2_{G_K})) + (1-\gamma_K\delta_0\right]$ for $j = 1,\ldots,M$, N_K is the Normal component of the Kth group, $\delta_{0,K}$ is the Dirac mass of the Kth component, and f is some function of β.

Writing (54) as a function of γ_K and fixing the remaining parameters we have,

$$\begin{aligned}
&\mathbb{E}_Q\Big[-\ell(\mathcal{D};\beta) + \log\frac{dQ}{d\Pi}(\beta)\Big] \\
&\leq \sum_{i=1}^{n}\Bigg[\left(\frac{1}{2}-y_i\right)\gamma_K x_{i,G_K}^\top \mu_{G_K} + \frac{\gamma_K}{2}\left(2+\sum_{j=1}^{p}\sum_{k=1}^{p} x_{ij}x_{ik}\mathbb{E}_{N_K\otimes Q_{\setminus K}}\left[\beta_j\beta_k\right]\right)^{1/2} \\
&+ \frac{1-\gamma_K}{2}\left(2+\sum_{j=1}^{p}\sum_{k=1}^{p} x_{ij}x_{ik}\mathbb{E}_{\delta_{0,K}\otimes Q_{\setminus K}}\left[\beta_j\beta_k\right]\right)^{1/2}\Bigg] - \frac{\gamma_K}{2}\sum_{j\in G_K}\log\left(2\pi\sigma_j^2\right) \qquad (61) \quad \texttt{eq:l_gamma_up} \\
&- \gamma_K\log(C_K) - \gamma_K m_K\log(\lambda) + \lambda\gamma_K\left(\sum_{j\in G_K}\sigma_j^2+\mu_j^2\right)^{1/2} - \frac{\gamma_K m_K}{2} \\
&+ \gamma_K\log\frac{\gamma_K}{\bar{w}} + (1-\gamma_K)\log\frac{1-\gamma_K}{1-\bar{w}} + C
\end{aligned}$$

Differentiating the RHS of the previous display wrt. γ_K, setting to 0 and re-arranging

gives the update equation for γ_K,

$$\log\frac{\gamma_K}{1-\gamma_K} = \log\frac{\bar{w}}{1-\bar{w}} + \frac{m_K}{2} + \frac{1}{2}\sum_{j\in G_K}\log\left(2\pi\sigma_j^2\right) + \log(C_K) - \lambda\left(\sum_{j\in G_K}\sigma_j^2+\mu_j^2\right)^{1/2}$$
$$+ m_K\log(\lambda) - \sum_{i=1}^{n}\left[\left(\frac{1}{2}-y_i\right)x_{i,G_K}^\top\mu_{G_K} + \frac{1}{2}\left(2+\sum_{j=1}^{p}\sum_{k=1}^{p}x_{ij}x_{ik}\mathbb{E}_{N_K\otimes Q_{\setminus K}}\left[\beta_j\beta_k\right]\right)^{1/2}\right.$$
$$\left. - \frac{1}{2}\left(2+\sum_{j\in G_K^c}^{p}\sum_{k\in G_K^c}^{p}x_{ij}x_{ik}\mathbb{E}_{Q_{\setminus K}}\left[\beta_j\beta_k\right]\right)^{1/2}\right] \tag{62}$$

5 Simulation study

5.1 Simulation design

Data is simulated for $i = 1, \ldots, n$ observations, each having a response $y_i \in \mathbb{R}$ and p continuous predictors $x_i \in \mathbb{R}^p$. The response is sampled independently from a Gaussian distribution with mean $\beta_0^\top x_i$ and variance σ^2, where the true coefficient vector $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^\top \in \mathbb{R}^p$ contains s non-zero groups each of size g. Non-zero elements of β_0 are sampled independently and uniformly from $[-3, -1] \cup [1, 3]$. Finally, the predictors are generated from one of # settings:

- **Setting 1**: $x_i \overset{\text{iid}}{\sim} N(0_p, I_p)$ where 0_p is the p-dimension zero vector and I_p the $p \times p$ identity matrix.
- **Setting 2**: $x_i \overset{\text{iid}}{\sim} N(0, \Sigma)$ where $\Sigma_{ij} = 0.6^{|i-j|}$ for $i, j = 1, \ldots, p$.
- **Setting 3**: $x_i \overset{\text{iid}}{\sim} N(0, \Sigma)$ where $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.6$ for $i \neq j$ and $i, j = 50k, \ldots, 50(k+1)$ for $k = 0, \ldots, p/50 - 1$ and $\Sigma_{ij} = 0$ otherwise.

5.2 Methods

Considered so far:

- **GSVB** (group sparse variational Bayes – ours): this is an implementation of
- **MCMC**: Gibb sampler for the group spike-and-slab prior
- **SSGL**: Spike-and-slab group LASSO, (similar to the spike-and-slab LASSO) where the multivariate Dirac mass is replaced with a multivariate double ex-

ponential distribution. (i.e. this is a continuous mixture), with one density acting as the spike and another the slab.

see Jonathan thesis eq. 4.28 on simulation designs for the group sparse setting

5.3 Results

Methods are comparable. Our method has the fastest runtime, however this may come down to the fact that our method (and the MCMC implementation) is written in C++, whereas GSVB is written in R.

Setting	Method	ℓ_2-error	ℓ_1-error	TPR	FDR	AUC	Runtime
Setting 1	GSVB_D	0.272 (0.18, 0.36)	0.846 (0.58, 1.14)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	1.4s (1.2s, 2.2s)
	GSVB	0.271 (0.18, 0.36)	0.851 (0.59, 1.15)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	1.6s (1.4s, 2.2s)
	MCMC	0.268 (0.18, 0.36)	0.840 (0.58, 1.13)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	3m 7s (3m 5s, 3m 13s)
	GSSL	0.273 (0.19, 0.36)	0.844 (0.59, 1.15)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	9.8s (9.4s, 12.6s)
Setting 2	GSVB_D	0.365 (0.25, 0.51)	1.151 (0.73, 1.62)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	1.7s (1.4s, 2.1s)
	GSVB	0.365 (0.25, 0.51)	1.165 (0.74, 1.65)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	1.8s (1.6s, 2.3s)
	MCMC	0.366 (0.24, 0.50)	1.141 (0.74, 1.63)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	3m 6s (3m 5s, 3m 8s)
	GSSL	0.365 (0.23, 0.51)	1.144 (0.74, 1.63)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	9.6s (9.4s, 11.3s)
Setting 3	GSVB_D	0.395 (0.28, 0.53)	1.263 (0.88, 1.78)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	2.3s (1.5s, 7.3s)
	GSVB	0.395 (0.29, 0.60)	1.278 (0.91, 1.97)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	2.7s (1.6s, 8.6s)
	MCMC	0.398 (0.29, 0.54)	1.264 (0.88, 1.76)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	3m 8s (3m 6s, 3m 13s)
	GSSL	0.396 (0.28, 0.54)	1.269 (0.89, 1.79)	1.000 (1.00, 1.00)	0.000 (0.00, 0.00)	1.000 (1.00, 1.00)	10.0s (9.6s, 12.7s)

Table 1: Companion of Group-sparse Bayesian variable selection methods

tab:bvs_compri

TODO, comparison of coverage

6 Application to real data

TODO

References

Browne2006 W. J. Browne and D. Draper. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514, 2006. ISSN 19360975. doi: 10.1214/06-BA117.

epraetere2017a N. Depraetere and M. Vandebroek. A comparison of variational approximations for fast inference in mixed logit models. *Computational Statistics*, 32(1):93–125, 2017. ISSN 16139658. doi: 10.1007/s00180-015-0638-y.

Giraud2021 C. Giraud. *Introduction to High-Dimensional Statistics*. Chapman and Hall/CRC, Aug. 2021. doi: 10.1201/9781003158745. URL https://doi.org/10.1201/9781003158745.

Opper2009 M. Opper and C. Archambeau. The Variational Gaussian Approximation Revisited. *Neural Computation*, 21(3):786–792, 03 2009. ISSN 0899-7667. doi: 10.1162/neco.2008.08-07-592. URL https://doi.org/10.1162/neco.2008.08-07-592.

Seeger1999 M. Seeger. Bayesian methods for support vector machines and Gaussian processes. Master's thesis, University of Karlsruhe, 1999. URL http://people.mmci.uni-saarland.de/{~}mseeger/papers/bayes-svm.pdf.

A Definitions

Definition *Kullback-Leibler divergence.*

Let Q and P be probability measures on $\mathcal{X}$, such that Q is absolutely continuous

with respect to P, then the Kullback-Leibler divergence is defined as,

$$D_{\mathrm{KL}}(Q\|P) = \int_{\mathcal{X}} \log\left(\frac{dQ}{dP}\right) dQ \tag{A.1}$$ {eq:kl}

where dQ/dP is the Radon-Nikodym derivative of Q with respect to P.

B Co-ordinate ascent algorithm

B.1 Element-wise update equations

Updates for μ_{G_K}

$$\begin{aligned} f(\mu_i;\mu_{-i},\sigma,\gamma) := \; & \frac{1}{2\tau^2}\left((X^\top X)_{ii}\mu_i^2 + \sum_{j\in G_K, j\neq i}(X^\top X)_{ji}\mu_j\mu_i\right) \\ & + \frac{1}{\tau^2}\left(\left(\sum_{j\in G_K^c}(X^\top X)_{ji}\gamma_J\mu_j\mu_i\right) - \mu_i\langle y, X_{:i}\rangle\right) + \lambda\left(\sum_{j\in G_K}\sigma_j^2+\mu_j^2\right)^{1/2} + C \end{aligned} \tag{B.1}$$ {eq:mu_update}

The above expression in turn is minimized via optimization routines.

Using the fact that $\left(\sum_{j\in G_K}\sigma_j^2+\mu_j^2\right)^{1/2} \leq 1+\sum_{j\in G_K}\sigma_j^2+\mu_j^2$, we can obtain a looser upper bound on (17) and the resulting expression we need to minimize,

$$\begin{aligned} f_2(\mu_i;\mu_{-i},\sigma,\gamma) := \; & \frac{1}{2\tau^2}\left((X^\top X)_{ii}\mu_i^2 + \sum_{j\in G_K, j\neq i}(X^\top X)_{ji}\mu_j\mu_i\right) \\ & + \frac{1}{\tau^2}\left(\sum_{j\in G_K^c}\left[(X^\top X)_{ji}\gamma_J\mu_j\mu_i\right] - \mu_i\langle y, X_{:i}\rangle\right) + \lambda\mu_i^2 + C \end{aligned} \tag{B.2}$$

which in turn in minimized when

$$\mu_i = -\frac{\left(\sum_{j\in G_K^c}(X^\top X)_{ji}\gamma_J\mu_j\right) + \frac{1}{2}\left(\sum_{j\in G_K, j\neq i}(X^\top X)_{ji}\mu_j\right) - \langle y, X_{:i}\rangle}{(X^\top X)_{ii} + 2\tau^2\lambda} \tag{B.3}$$ {eq:mu_analyti

Interestingly, if we assume the columns of X are orthogonal, i.e. $(X^\top X)_{ij} = 0$ for $i \neq j$, and $\lambda = 0$, then (B.3) can be written as

$$\mu_i^{\text{ols}} = (X^\top X)_{ii}^{-1}(X_{:i})^\top y \tag{B.4}$$

which we recognize as the ordinary least squares estimator under an orthogonal design. Similarly, when $\lambda > 0$, the minimizer is given by

$$\mu_i^{\text{rr}} := \left((X^\top X)_{ii} + 2\tau^2\lambda\right)^{-1}(X_{:i})^\top y \tag{B.5}$$

which we recognize as the solution under the ridge penalty. It follows that the minimizer (B.3) is given by a ridge term under the assumption of an orthogonal design and some additional term, formally,

$$\mu_i = \mu_i^{\text{rr}} - \frac{\left(\sum_{j\in G_K^c}(X^\top X)_{ji}\gamma_J\mu_j\right) + \frac{1}{2}\left(\sum_{j\in G_K, j\neq i}(X^\top X)_{ji}\mu_j\right)}{(X^\top X)_{ii} + 2\tau^2\lambda} \tag{B.6}$$

Updates for σ_{G_K}

$$g(\sigma_i;\mu,\sigma_{-i},\gamma) := \frac{1}{2\tau^2}(X^\top X)_{ii}\sigma_i^2 - \log\sigma_i + \lambda\left(\sum_{j\in G_K}\sigma_j^2 + \mu_j^2\right)^{1/2} + C \tag{B.7}$$

As before, under the looser upper bound we have,

$$g_2(\sigma_i;\mu,\sigma_{-i},\gamma) := \frac{1}{2\tau^2}(X^\top X)_{ii}\sigma_i^2 - \log\sigma_i + \lambda\sigma_i^2 + C \tag{B.8}$$

which is minimized when,

$$\sigma_i = \left(\frac{(X^\top X)_{ii}}{\sigma^2} + 2\lambda\right)^{-1/2} \tag{B.9}$$

which we notice does not depend on the other parameters and in turn can be used to initialize σ_i.

C MCMC sampler

We construct a Gibbs sampler to sample from the posterior distribution. To begin, we note that the likelihood can be expressed as,

$$p(\mathcal{D}|\beta, z, \tau^2) = \prod \phi\left(y_i; \sum_{k=1}^{M} z_k \langle x_{G_k}, \beta_{G_k} \rangle, \tau^2\right) \tag{C.1}$$

In turn, we can re-write our prior as,

$$\begin{aligned} \beta_{G_k} &\overset{\text{ind}}{\sim} \Psi(\beta_{G_k}; \lambda) \\ z_k|\theta_k &\overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_k) \\ \theta_k &\overset{\text{iid}}{\sim} \text{Beta}(a_0, b_0) \end{aligned} \tag{C.2}$$

Finally, recall our prior on $\xi = \tau^2$, is

$$\xi \sim \Gamma^{-1}(\xi; a, b) \tag{C.3}$$

which has density $\frac{b^a}{\Gamma(a)} \left(\frac{1}{\xi}\right)^{a+1} \exp\left(-\frac{\beta}{x}\right)$.

To sample form the posterior we:

1. Initialize $\beta^{(i)}, z^{(i)}, \theta^{(i)}, \xi^{(i)}$
2. For $i = 1, \ldots, N$
   (a) For $k = 1, \ldots, M$
      i. Sample $\theta_k^{(i)} \overset{\text{iid.}}{\sim} \text{Beta}(a_0, b_0)$
   (b) For $k = 1, \ldots, M$
      i. Sample $z_k^{(i)} \overset{\text{ind.}}{\sim} \text{Bernoulli}(p_k)$ where

$$p_k = \frac{p(z_k = 1|\mathcal{D}, \beta, \theta, \xi)}{p(z_k = 1|\mathcal{D}, \beta, \theta, \xi) + p(z_k = 0|\mathcal{D}, \beta, \theta, \xi)} \tag{C.4}$$

(c) For $k = 1, \ldots, M$

i. Sample $\beta_{G_k}^{(i)} \sim p(\beta_{G_k} | \mathcal{D}, z, \beta^{(i-1)}, \xi)$

(d) Sample $\xi \overset{\text{iid.}}{\sim} \Gamma^{-1}(a + 0.5n, b + 0.5\|y - X(\beta \circ z)\|^2)$