Problem formulation

We are interested in modelling,

$$\mathbb{E}\left[Y|X,\beta\right] = f(X\beta) \tag{1}$$

{eq:prob_formu

where for n observations and p features, $Y = (Y_1, \dots, Y_n)^\top$ is a random vector in $\mathbb{R}^n$ whose realizations are denoted by $y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$ is the design matrix with $x_i = (x_{i1}, \dots, x_{ip})^\top \in \mathbb{R}^p$ being the feature vector for the ith sample, $\beta = (\beta_1, \dots, \beta_p)^\top \in \mathbb{R}^p$ is the model coefficient vector and $f : \mathbb{R} \to \mathbb{R}$ is a link function applied element-wise to $X\beta$.

Notation

Throughout we define the groups $G_k = \{G_{k,1}, \dots, G_{k,m_k}\}$ for $k = 1, \dots, M$, to be disjoint sets of indices of size m_k such that $\bigcup_{k=1}^M G_k = \{1, \dots, p\}$ and let $G_k^c = \{1, \dots, p\} \setminus G_k$. Further, denote $X_{G_k} = (x_{1,G_k}, \dots, x_{n,G_K})^\top \in \mathbb{R}^{n \times m_k}$ where $x_{i,G_k} = \{x_{ij} : j \in G_k\}$, $X_{G_k^c} = (x_{1,G_k^c}, \dots, x_{n,G_K^c})^\top \in \mathbb{R}^{n \times (p - m_k)}$ where $x_{i,G_k^c} = \{x_{ij} : j \in G_k^c\}$, $\beta_{G_k} = \{\beta_j : j \in G_k\}$ and $\beta_{G_k^c} = \{\beta_j : j \in G_k^c\}$.

Contributions

- TODO

Prior and Posterior

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

$$\begin{aligned} \beta_{G_k}|z_k &\overset{\text{ind}}{\sim} z_k\Psi(\beta_{G_k};\lambda) + (1-z_k)\delta_0(\beta_{G_k}) \\ z_k|\theta_k &\overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_k) \\ \theta_k &\overset{\text{iid}}{\sim} \text{Beta}(a_0, b_0) \end{aligned} \tag{2}$$

for $k = 1, \ldots, M$, where $\delta_0(\beta_{G_k})$ is the multivariate Dirac mass on zero with dimension $m_k = \dim(\beta_{G_k})$, and $\Psi(\beta_{G_k})$ is the multivariate double exponential distribution with density

$$\psi(\beta_{G_k};\lambda) = C_k \lambda^{m_k} \exp\left(-\lambda \|\beta_{G_k}\|\right) \tag{3}$$

{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.

Variational Families

We introduce two variational families. The first is a fully factorized mean-field variational family,

$$\mathcal{Q} = \left\{ Q(\mu,\sigma,\gamma) = \bigotimes_{k=1}^{M} \left[\gamma_k \ N\left(\mu_{G_k}, \operatorname{diag}(\sigma^2_{G_k})\right) + (1-\gamma_k)\delta_0\right] \right\} \tag{4}$$

where $\mu \in \mathbb{R}^p$ with $\mu_{G_k} = \{\mu_j : j \in G_k\}$, $\sigma^2 \in \mathbb{R}^p_+$ with $\sigma^2_{G_k} = \{\sigma^2_j : j \in G_k\}$, $\gamma = (\gamma_1, \ldots, \gamma_M)^\top \in [0,1]^M$, and $N(\mu, \Sigma)$ denotes the multivariate Normal distribution with mean parameter μ and covariance Σ. The second, is variational family with unrestricted covariance within groups,

$$\mathcal{Q}' = \left\{ Q'(\mu,\Sigma,\gamma) = \bigotimes_{k=1}^{M} \left[\gamma_k \ N\left(\mu_{G_k}, \Sigma_{G_k}\right) + (1-\gamma_k)\delta_0\right] \right\} \tag{5}$$

where $\Sigma \in \mathbb{R}^{p\times p}$ is a covariance matrix for which $\Sigma_{ij} = 0$, for $i \in G_k, j \in G_l, k \neq l$ (i.e. there is independence between groups) and $\Sigma_{G_k} = (\Sigma_{ij})_{i,j\in G_k} \in \mathbb{R}^{m_k \times m_k}$ denotes the covariance matrix of the kth group.

Note that $\mathcal{Q} \subset \mathcal{Q}'$, therefore $\mathcal{Q}'$ should provide greater flexibility in approximating the posterior. Additionally $\mathcal{Q}'$ should capture the dependence between coefficients in the same group, i.e. between the elements of β_{G_k}.

Computing the variational posterior

The variational posterior is given by solving,

$$\tilde{\Pi} = \operatorname*{argmin}_{\mu,\sigma,\gamma} \mathbb{E}_Q \left[\log \frac{dQ}{d\Pi} - \ell(\mathcal{D};\beta) \right] \tag{6}$$

{eq:opt}

where ℓ is the log-likelihood for a given model. As this optimization problem is generally not convex, we approach it via co-ordinate ascent variational inference. Wherein, for each group $k = 1, \dots, M$, we update the parameters for the group keeping the remainder fixed, formally this is detailed in Algorithm 1.

alg:cavi_gsvb

Algorithm 1 General CAVI strategy for computing the variational posterior

Initialize μ, σ, γ
while not converged
  for $k = 1, \dots, M$
    $\mu_{G_k} \leftarrow \operatorname{argmin}_{\mu_{G_k} \in \mathbb{R}^{m_k}} \mathbb{E}_{Q|z_k=1} \left[\log(dQ/d\Pi) - \ell(\mathcal{D};\beta) \mid \mu_{G_k^c}, \sigma, \gamma_{-k} \right]$
    $\sigma_{G_k} \leftarrow \operatorname{argmin}_{\sigma_{G_k} \in \mathbb{R}_+^{m_k}} \mathbb{E}_{Q|z_k=1} \left[\log(dQ/d\Pi) - \ell(\mathcal{D};\beta) \mid \mu, \sigma_{G_k^c}, \gamma_{-k} \right]$
    $\gamma_k \leftarrow \operatorname{argmin}_{\gamma_k \in [0,1]} \mathbb{E}_Q \left[\log(dQ/d\Pi) - \ell(\mathcal{D};\beta) \mid \mu, \sigma, \gamma_{-k} \right]$
return μ, σ, γ.

Regardless of the form of log-likelihood, the Radon-Nikodym derivative between the variational family and the prior can be expressed as

$$\log \frac{dQ}{d\Pi}(\beta) = \sum_{k=1}^{M} \log \frac{dQ_k}{d\Pi_k}(\beta_{G_k}) = \sum_{k=1}^{M} \mathbb{I}_{z_k=1} \log \frac{\gamma_k dN_k}{\bar{w} d\Psi_k}(\beta_{G_k}) + \mathbb{I}_{z_k=0} \log \frac{1-\gamma_k}{1-\bar{w}} \frac{d\delta_0}{d\delta_0}(\beta_{G_k})$$

Under this factorization it follows that

$$\mathbb{E}_Q\left[\log\frac{dQ}{d\Pi}\right] = \sum_{k=1}^{M}\Bigg(\gamma_k\log\frac{\gamma_k}{\bar{w}} - \frac{\gamma_k}{2}\log(\det(2\pi\Sigma_k)) - \frac{\gamma_k m_k}{2} - \gamma_k\log(C_k)$$
$$- \gamma_k m_k\log(\lambda) + \mathbb{E}_Q\left[\mathbb{I}_{z_k=1}\lambda\|\beta_{G_k}\|\right] + (1-\gamma_k)\log\frac{1-\gamma_k}{1-\bar{w}}\Bigg) \tag{7}$$

Since, $\mathbb{E}_Q\left[\mathbb{I}_{z_k=1}\lambda\|\beta_{G_k}\|\right]$ does not have a closed form, we upper bound this quantity by

$$\mathbb{E}_Q\left[\mathbb{I}_{z_k=1}\lambda\|\beta_{G_k}\|\right] = \gamma_k\mathbb{E}_{N_k}\left[\lambda\|\beta_{G_k}\|\right] \leq \gamma_k\lambda\left(\sum_{i\in G_k}\Sigma_{ii}+\mu_i^2\right)^{1/2} \tag{8}$$

Model classes

We consider three common classes of models: Gaussian, Binomial and Poisson. For each class the co-ordinate update equations are provided for both variational families.

Gaussian

Under the Gaussian linear model $Y_i \overset{\text{iid}}{\sim} N(x_i^\top\beta, \tau^2)$ where $\tau^2 > 0$ is an unknown variance and the canonical link function is $f(x) = x$. Hence the log-likelihood is given as,

$$\ell(\mathcal{D};\beta,\tau^2) = -\frac{n}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\|y - X\beta\|^2 \tag{9}$$ {eq:log-likeli

where $\mathcal{D} = \{(y_i, x_i)\}_{i=1}^n$.

To model this family, an Inverse-Gamma prior is placed on the nuisance parameter τ^2, formally, $\tau^2 \overset{\text{ind}}{\sim} \Gamma^{-1}(a,b)$, where $a, b > 0$. Further, we extend the variational family to take τ^2 into account by letting $\mathcal{Q}_\tau = \mathcal{Q}\times\{\Gamma^{-1}(a',b') : a' > 0, b' > 0\}$.

Evaluating an expression for the expected value of the negative log-likelihood,

$$\begin{aligned}\mathbb{E}_{Q_\tau}\left[-\ell(\mathcal{D};\beta,\tau^2)\right] = \frac{a'}{2b'}\left(\|y\|^2 + \sum_{i,j=1}^{p}(X^\top X)_{ij}\mathbb{E}_Q\left[\beta_i\beta_j\right]\right) - \sum_{k=1}^{M}\left(\frac{a'}{b'}\gamma_k\langle y, X_{G_k}\mu_{G_k}\rangle\right) \\ + \frac{n}{2}(\log(2\pi) + \log(b') - \kappa(a'))\end{aligned} \tag{10}$$

where

Binomial

Under the binomial family $Y_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)$ where $p \in (0,1)$ and the canonical link function is the popular logistic function wherein,

$$\mathbb{E}[Y_i|x_i, \beta] = p = \text{logistic}(x^\top \beta) = \frac{\exp(x^\top \beta)}{1 + \exp(x^\top \beta)} \tag{11}$$

Under this model the log-likelihood is,

$$\ell(\mathcal{D}, \beta) = \sum_{i=1}^{n} y_i \left(x_i^\top \beta\right) - \log\left(1 + \exp(x_i^\top \beta)\right) \tag{12}$$

Bound using Jensen's inequality

Bounding the likelihood using Jensen's inequality is straightforward,

$$\begin{aligned} \mathbb{E}_Q\left[-\ell(\mathcal{D}; \beta)\right] = & \sum_{i=1}^{n} \mathbb{E}_Q\left[\log\left(1 + \exp(x_i^\top \beta)\right) - y_i\left(x_i^\top \beta\right)\right] \\ \leq & \sum_{i=1}^{n} \log\left(1 + \mathbb{E}_Q\left[\exp(x_i^\top \beta)\right]\right) - y_i \sum_{k=1}^{M} \gamma_k \sum_{j \in G_k} x_{ij}\mu_j \end{aligned} \tag{13}$$

{eq:logistic_j

where

$$\mathbb{E}_Q\left[\exp(x_i^\top \beta)\right] = \prod_{k=1}^{M} \gamma_k \exp\left\{\sum_{j \in G_k} x_{ij}\mu_j + \frac{1}{2} x_{ij}^2 \sigma_j^2\right\} + (1 - \gamma_k)$$

Based on (13) it is straightforward to derive the update equations for μ_{G_k}, σ_{G_k} and γ_k.

Bound based on Jaakkola and Jordan (1996)

Jaakkola and Jordan (1996) introduce a quadratic bound for the sigmoid function $\text{s}(x) = (1 + \exp(-x))^{-1}$, given as

$$\text{s}(x) \geq \text{s}(t) \exp\left\{\frac{x - t}{2} - \frac{a(t)}{2}(x^2 - t^2)\right\} \tag{14}$$

{eq:jj_bound}

where $a(t) = \frac{\mathrm{s}(t)-1/2}{t}$ and t is a variational parameter that must be optimized to ensure the bound is tight.

Using the fact that $\mathbb{P}(Y = y_i | X = x_i)$ can be written as $e^{y_i x_i^\top \beta} \mathrm{s}(-x_i^\top \beta)$, we have

$$
\begin{aligned}
-\ell(\mathcal{D};\, \beta) &= \sum_{i=1}^{n} -y_i x_i^\top \beta - \log \mathrm{s}(-x_i^\top \beta) \\
&\leq \sum_{i=1}^{n} -y_i x_i^\top \beta - \log \mathrm{s}(t_i) + \frac{x_i^\top \beta + t_i}{2} + \frac{a(t_i)}{2}((x_i^\top \beta)^2 - t_i^2) \\
&= -\langle y, X^\top \beta \rangle - \langle 1, \log s(t) \rangle + \frac{1}{2}\left(\langle 1, X^\top \beta + t \rangle + \beta^\top X^\top A_t X \beta - t^\top A_t t \right) \qquad (15)
\end{aligned}
$$

{eq:jj_likelih

where t_i is a variational parameter for each observation, $t = (t_1, \dots, t_n)^\top$, $A_t = \mathrm{diag}(a(t_i), \dots, a(t_n))$, $s(t) = (s(t_1), \dots, s(t_n))^\top$, and operators to vectors are performed element-wise, e.g. $\log s(t) = (\log s(t_i))_{i=1}^n$. Combining (15) with results seen in the linear regression setting, it is straightforward to obtain the necessary update equations.

Taking the expectation of (15) wrt. Q, gives,

$$
\begin{aligned}
\left(\sum_{k=1}^{M} \gamma_k \langle 1/2 - y, X_{G_k} \mu_{G_k} \rangle \right) + \frac{1}{2} \left(\sum_{i,j=1}^{p} (X^\top A_t X)_{ij} \mathbb{E}_Q[\beta_i \beta_j] \right) - \frac{t^\top A_t t}{2} \\
+ \langle 1, t/2 - \log s(t) \rangle
\end{aligned} \qquad (16)
$$

{eq:jj_expecta

Updates for μ_{G_k} and σ_{G_k} are given by finding the minimizers of

$$
\begin{aligned}
&\langle 1/2 - y, X_{G_k} \mu_{G_k} \rangle + \frac{1}{2} \left(\mu_{G_k}^\top X_{G_k}^\top A_t X_{G_k} \mu_{G_k} + \sum_{i \in G_k} (X^\top A_t X)_{ii} \sigma_i^2 \right) \\
&+ \sum_{j \neq k} \gamma_j \mu_{G_k}^\top X_{G_k}^\top A_t X_{G_j} \mu_{G_j} + \lambda \left(\sigma_{G_k}^\top \sigma_{G_k} + \mu_{G_k}^\top \mu_{G_k} \right)^{1/2} - \sum_{i \in G_k} \log \sigma_i
\end{aligned} \qquad (17)
$$

Updates for γ_k

$$
\begin{aligned}
\log \frac{\gamma_K}{1-\gamma_K} &= \log \frac{\bar{w}}{1-\bar{w}} + \frac{m_K}{2} + \langle y - 1/2, X_{G_K}\mu_{G_K}\rangle \\
&+ \frac{1}{2}\sum_{j\in G_K} \log\left(2\pi\sigma_j^2\right) + \log(C_K) + m_K\log(\lambda) - \Bigg\{\lambda\left(\sum_{i\in G_K}\sigma_i^2 + \mu_i^2\right)^{1/2} \\
&+ \frac{1}{2}\left(\mu_{G_k}^\top X_{G_k}^\top A_t X_{G_k}\mu_{G_k} + \sum_{i\in G_k}(X^\top A_t X)_{ii}\sigma_i^2\right) + \sum_{j\neq k}\gamma_j \mu_{G_k}^\top X_{G_k}^\top A_t X_{G_j}\mu_{G_j}\Bigg\}
\end{aligned}
\tag{18}
$$

ELBO given by combining the negation of (16) and $\mathbb{E}_Q[-\log dQ/d\Pi]$,

$$
\begin{aligned}
\mathcal{L}_Q(\mathcal{D}) = \frac{1}{2}&\left(t^\top A_t t - \sum_{i,j=1}^{p}(X^\top A_t X)_{ij}\mathbb{E}_Q[\beta_i\beta_j]\right) - \left(\sum_{k=1}^{M}\gamma_k\langle 1/2 - y, X_{G_k}\mu_{G_k}\rangle\right) \\
&- \langle 1, t/2 - \log s(t)\rangle + \sum_{k=1}^{M}\Bigg(\frac{\gamma_k}{2}\sum_{j\in G_k}\left(\log(2\pi\sigma_j^2)\right) + \gamma_k\log(C_k) + \frac{\gamma_k m_k}{2} \\
&+ \gamma_k m_k \log(\lambda) - \mathbb{E}_Q\left[\mathbb{I}_{z_k=1}\lambda\|\beta_{G_k}\|\right] - \gamma_k\log\frac{\gamma_k}{\bar{w}} - (1-\gamma_k)\log\frac{1-\gamma_k}{1-\bar{w}}\Bigg)
\end{aligned}
\tag{19}
$$

Updates for t_i are found by maximizing the ELBO, and are given by

$$
t_i = \left(\sum_{k=1}^{M}\gamma_k\left[(\mu_{G_k}^\top x_{i,G_k})^2 + \sum_{j\in G_k}\sigma_j^2 x_{i,j}^2\right]\right)^{1/2} \tag{20}
$$

Poisson

Under the Poisson family $Y_i \overset{\text{iid}}{\sim} \text{Poisson}(\lambda)$, where $\lambda > 0$. The canonical link function is $f(x) = \exp(x)$,

$$\ell(\mathcal{D}; \beta) = \sum_{i=1}^{n} y_i x_i^\top \beta - \exp(x_i^\top \beta) - \log(y!) \tag{21}$$

We are going to be using the full covariance matrix variational family. Taking the expectation under Q' gives

$$\begin{aligned} \mathbb{E}_{Q'}\left[\ell(\mathcal{D}; \beta)\right] &= \sum_{i=1}^{n} \mathbb{E}_{Q'}\left[y_i x_i^\top \beta - \exp(x_i^\top \beta) - \log(y!)\right] \\ &= \sum_{i=1}^{n} \left(\sum_{k=1}^{M} \gamma_k y_i x_{i,G_k}^\top \mu_{G_k} \right) + M_{Q'}(x_i) - \log(y!) \end{aligned} \tag{22}$$

where

$$M_{Q'}(x_i) = \prod_{k=1}^{M} \left(\gamma_k \exp\left\{ x_{i,G_k}^\top \mu_{G_k} + \frac{1}{2} x_{i,G_k}^\top \Sigma_{G_k} x_{i,G_k} \right\} + (1 - \gamma_k) \right) \tag{23}$$

References

Jakkola97 T. S. Jaakkola and M. I. Jordan. A variational approach to Bayesian logistic regression models and their extensions, 1996.