To-do list

1. Exponential family for y not really necessary, it just follows nicely from the latent variable motivation. 2
2. Section 4.3.3 9
3. Write down the mean and variance for lambda 9
4. Proof? 11
5. Computational considerations? I think go through the entire chapter without discussing computational issues, and discuss them here. Less cluttered this way. 11
6. How is this calculated? Simulation usually, but also quadrature methods not too bad if m not too large. Stata sheet useful? Talk about if iid errors. . . . 12
7. can use Hamiltonian Monte Carlo? 12
8. Chapter 3 13
9. Lemma X 29
10. Section 4.3.3 33
11. expression XXX 35

Contents

Haziq Jamil
Department of Statistics
London School of Economics and Political Science
PhD thesis: ‘Regression modelling using Fisher information covariance kernels (I-priors)’

Chapter 5

I-priors for categorical responses

In a regression setting, consider polytomous response variables $y_1, \dots, y_n$, where each y_i takes on exactly one of the values $\{1, \dots, m\}$ from a set of m possible choices. Modelling categorical response variables is of profound interest in statistics, econometrics and machine learning, with applications aplenty. In the social sciences, categorical variables often arise from survey responses, and one may be interested in studying correlations between explanatory variables and the categorical response of interest. Economists are often interested in discrete choice models to explain and predict choices between several alternatives, such as consumer choice of goods or modes of transport. In this age of big data, machine learning algorithms are used for classification of observations based on what is usually a large set of variables or features.

As an extension to the I-prior methodology, we propose a flexible modelling framework suitable for regression of categorical response variables.

In the spirit of generalised linear models (McCullagh and Nelder, 1989), we relate the class probabilities of the observations to the I-prior regression model via a link function. Perhaps though, it is more intuitive to view it as machine learners do: Since the regression function is ranged on the entire real line, it is necessary to “squash” it through some sigmoid function to conform it to the interval $[0, 1]$ suitable for probability measures. As in GLMs, the y_i's are assumed to follow anexponential family distribution, and in this case, the categorical distribution. We denote this by

$$y_i \sim \text{Cat}(p_{i1}, \dots, p_{im}),$$

1. Exponential family for y not really necessary, it just follows nicely from the latent variable motivation.

with the class probabilities satisfying $p_{ij} \geq 0, \forall j = 1, \dots, m$ and $\sum_{j=1}^m p_{ij} = 1$. The probability mass function (PMF) of y_i is given by

$$p(y_i) = p_{i1}^{[y_i=1]} \cdots p_{im}^{[y_i=m]} \tag{5.1}$$

{eq:catdist}

where the notation $[\cdot]$ refers to the Iverson bracket[1]. The dependence of the class probabilities on the covariates is specified through the relationship

$$g(p_{ij}) = \big(\alpha_j + f_j(x_i)\big)_{j=1}^m$$

where $g : [0,1] \to \mathbb{R}^m$ is some specified link function. As we will see later, the normality assumption of the errors naturally implies a *probit* link function, i.e., g is the inverse cumulative distribution function (CDF) of a standard normal distribution (or more precisely, a function that *involves* the standard normal CDF). Normality is also a required assumption for I-priors to be specified on the regression functions. We call this method of probit regression using I-priors the *I-probit* regression model.

Note that the probabilities are modelled per class $j \in \{1, \dots, m\}$ by individual regression curves f_j, and in the most general setting, m sets of intercepts α_j and kernel hyperparameters η_j must be estimated. The dependence of these m curves are specified through covariances $\sigma_{jk} := \text{Cov}[\epsilon_{ij}, \epsilon_{ik}]$, for each $j, k \in \{1, \dots, m\}$ and $j \neq k$. While it may be of interest to estimate these covariances, this paper considers cases where the regression functions are class independent, i.e. $\sigma_{jk} = 0, \forall j \neq k$. This violates the independence of irrelevant alternatives (IIA) assumption (see Section 5.3 for details) crucial in choice models, but not so much necessary for classification when the alternatives are distinctively different.

The many advantages of the I-prior methodology of Jamil and Bergsma, 2017 transfer over quite well to the I-probit model for classification and inference. In particular, by choosing appropriate RKHSs for the regression functions, we are able to fit a multitude of binary and multinomial models, including multilevel or random-effects models, linear and non-linear classification models, and even spatio-temporal models. Examples of these models applied to real-world data is shown in Section **??**. Working in a Bayesian setting together with variational inference allows us to estimate the model much faster than traditional MCMC sampling methods, yet provides us with the conveniences that come with posterior estimates. For example, inferences around log-odds is usually cumbersome

[1] $[A]$ returns 1 if the proposition A is true, and 0 otherwise. The Iverson bracket is a generalisation of the Kronecker delta.

for probit models, but a credibility interval can easily be obtained by resampling methods from the relevant posteriors, which are normally exponential family distributions in the I-probit model.

5.1 A naïve model

5.2 A latent variable motivation: the I-probit model

5.3 Identifiability and IIA

sec:iia

5.4 Estimation

5.5 A variational algorithm

sec:iprobitvar

We present a variational inference algorithm to estimate the I-probit latent variables $\mathbf{y}^*$ and $\mathbf{w}$, together with the parameters $\theta = \{\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_m), \eta, \boldsymbol{\Psi}\}$. Begin by assuming some prior distribution on the parameters $p(\theta) = p(\boldsymbol{\alpha})p(\eta)p(\boldsymbol{\Psi})$. Although one may devote more attention to the prior specification of these parameters, for our purposes it suffices that they are independent component-wise, and the PDFs belong to the exponential family of distributions with known hyperparameters. The exponential family requirement greatly eases the complexity of deriving the variational algorithm later on[2].

Recall that $\mathbf{y}^*|\mathbf{w} \sim \text{MN}_{n,m}(\mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$ and $\mathbf{w} \sim \text{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$. The required posterior distribution is then $p(\mathbf{y}^*, \mathbf{w}, \theta|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{w}, \theta)p(\mathbf{w}|\theta)p(\theta)$. This is approximated by a mean-field distribution of the form $q(\mathbf{y}^*, \mathbf{w}, \theta) \equiv q(\mathbf{y}^*)q(\mathbf{w})q(\theta)$, and also $q(\theta) = q(\boldsymbol{\alpha})q(\eta)q(\boldsymbol{\Psi})$. Denote by $\tilde q$ the distributions which minimise the Kullbeck-Leibler divergence (maximise the variational lower bound). By appealing to Bishop (2006, equation 10.9, p. 466), we find that for each $\xi \in \{\mathbf{y}^*, \mathbf{w}, \theta\} =: \mathcal{Z}$, $\tilde q$ satisfies

$$\log \tilde q(\xi) = \text{E}_{-\xi}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)] + \text{const.} \tag{5.2}$$

{eq:qtilde}

where expectation of the log joint density of $(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)$ is taken with respect to all of the unknowns $\mathcal{Z}$ except the one currently in consideration, under their respective q densities. Estimates of the latent variables and parameters are then obtained by taking the mean of their respective approximate posterior distribution.

[2] Of interest, one may even opt to assign improper priors on θ and the algorithm would still work. This is akin to obtaining empirical Bayes estimate of the θ if seen from an EM algorithm standpoint.

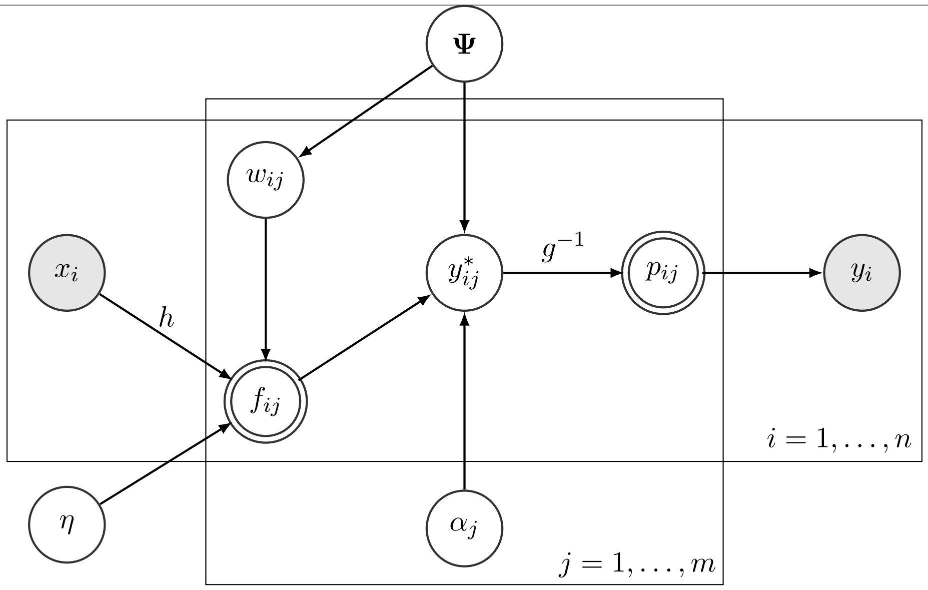

Figure 5.1: A DAG of the I-probit model. Observed nodes are shaded, while double-lined nodes represents calculable quantities.

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (5.2) to recognise it as a known log-density function, which is the case when exponential family distributions are considered. That is, suppose that each complete conditional $p(\xi|\mathcal{Z}_{-\xi}, \mathbf{y})$ follows an exponential family distribution,

$$p(\xi|\mathcal{Z}_{-\xi}, \mathbf{y}) = B(\xi) \exp \big(\langle \zeta_\xi(\mathcal{Z}_{-\xi}, \mathbf{y}), \xi \rangle - A(\zeta_\xi) \big).$$

Then, from (5.2),

$$\begin{aligned}
\tilde q(\xi) &\propto \exp \big(\operatorname{E}_{-\xi}[\log p(\xi|\mathcal{Z}_{-\xi}, \mathbf{y})] \big) \\
&= \exp \Big(\log B(\xi) + \operatorname{E}\langle \zeta_\xi(\mathcal{Z}_{-\xi}, \mathbf{y}), \xi \rangle - \operatorname{E}[A(\zeta_\xi)] \Big) \\
&\propto B(\xi) \exp \operatorname{E}\langle \zeta_\xi(\mathcal{Z}_{-\xi}, \mathbf{y}), \xi \rangle
\end{aligned}$$

is also in the same exponential family. In situations where there is no closed form expression for $\tilde q$, then one resorts to sampling methods such as a Metropolis random walk to estimate quantities of interest. This stochastic step within a deterministic algorithm has been explored before in the context of a Monte Carlo EM algorithm—see Meng and Van Dyk (1997, §4, pp. 537–538) and references therein.

We now present the mean-field variational distributions $\tilde{q}$. On notation: we will typically refer to posterior means of the parameters $\mathbf{y}^*$, $\mathbf{w}$, θ and so on by the use of a tilde. For instance, we write $\tilde{\mathbf{w}}$ to mean $\mathrm{E}_{\mathbf{w}\sim\tilde{q}}[\mathbf{w}]$, the expected value of $\mathbf{w}$ under the pdf $\tilde{q}(\mathbf{w})$. The distributions are simply stated, but a full derivation is given in the appendix.

5.5.1 Latent propensities $\mathbf{y}^*$

The fact that the rows of $\mathbf{y}^*$ are independent can be exploited. Write $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{im}^*)^\top$. Then $\mathbf{y}_i^*|\theta, x_i \sim \mathrm{N}_m(\boldsymbol{\alpha} + \mathbf{f}(x_i), \boldsymbol{\Psi}^{-1})$, and we have the induced factorisation of the distribution $q(\mathbf{y}^*) = \prod_{i=1}^n q(\mathbf{y}_i^*)$, where each $q(\mathbf{y}_i^*)$ is the density of a *conically truncated multivariate normal disribution.* That is, for each $i = 1, \dots, n$ and noting the observed values $y_i = j \in \{1, \dots, m\}$, the $\mathbf{y}_i^*$'s are distributed according to

$$\mathbf{y}_i^* \overset{\text{iid}}{\sim} \begin{cases} \mathrm{N}_m(\tilde{\boldsymbol{\alpha}} + \tilde{\mathbf{f}}(x_i), \tilde{\boldsymbol{\Sigma}}) & \text{if } y_{ij}^* > y_{ik}^*, \forall k \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

{eq:ystardist}

The required expectations $\mathrm{E}\,\mathbf{y}_i^* = \mathrm{E}(y_{i1}^*, \dots, y_{im}^*)^\top$ are tricky to compute. One strategy might be Monte Carlo integration: using samples from $\mathrm{N}_m(\tilde{\alpha} + \tilde{\mathbf{f}}(x_i), \tilde{\boldsymbol{\Psi}}^{-1})$, zero out those that do not satisfy the condition $y_{ij}^* > y_{ik}^*, \forall k \neq j$, then take the sample average. If the independent I-probit model is considered, where $\boldsymbol{\Psi} = \mathrm{diag}(\sigma_1^{-2}, \dots, \sigma_m^{-2})$, the expected value can be considered component-wise, and each component of this expectation is given by

$$\tilde{y}_{ik}^* = \begin{cases} \tilde{\alpha}_k + \tilde{f}_{ik} - \tilde{\sigma}_k C_i^{-1} \displaystyle\int \phi_{ik}(z) \prod_{l \neq k,j} \Phi_{il}(z) \phi(z) \, \mathrm{d}z & \text{if } k \neq y_i \\ \tilde{\alpha}_{y_i} + \tilde{f}_{iy_i} - \tilde{\sigma}_{y_i} \sum_{k \neq y_i} \big(\tilde{y}_{ik}^* - \tilde{f}_{ik} \big) & \text{if } k = y_i \end{cases} \tag{5.4}$$

{eq:ystarupdate}

with

$$\begin{aligned} \phi_{ik}(Z) &= \phi\left(\frac{\tilde{\sigma}_{y_i}}{\tilde{\sigma}_k} Z + \frac{\tilde{\alpha}_{y_i} + \tilde{f}_{iy_i} - \tilde{\alpha}_k - \tilde{f}_{ik}}{\tilde{\sigma}_k} \right) \\ \Phi_{ik}(Z) &= \Phi\left(\frac{\tilde{\sigma}_{y_i}}{\tilde{\sigma}_k} Z + \frac{\tilde{\alpha}_{y_i} + \tilde{f}_{iy_i} - \tilde{\alpha}_k - \tilde{f}_{ik}}{\tilde{\sigma}_k} \right) \\ C_i &= \int \prod_{l \neq j} \Phi_{il}(z) \phi(z) \, \mathrm{d}z \end{aligned}$$

and $Z \sim \mathrm{N}(0,1)$ with pdf and cdf $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. The integrals that appear above are functions of a unidimensional Gaussian pdf, and these can be computed rather efficiently using quadrature methods.

5.5.2 I-prior random effects w

Given that both $\operatorname{vec}\mathbf{y}^*|\operatorname{vec}\mathbf{w}$ and $\operatorname{vec}\mathbf{w}$ are normally distributed, we find that the conditional posterior distribution $p(\mathbf{w}|\mathcal{Z}_{-\mathbf{w}},\mathbf{y})$ is also normal, and therefore the approximate posterior density $\tilde q$ for $\operatorname{vec}\mathbf{w} \in \mathbb{R}^{nm}$ is also normal with mean and precision given by

$$\operatorname{vec}\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w(\tilde{\boldsymbol{\Psi}} \otimes \tilde{\mathbf{H}}_\eta)\operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = (\tilde{\boldsymbol{\Psi}} \otimes \tilde{\mathbf{H}}_\eta^2) + (\tilde{\boldsymbol{\Psi}}^{-1} \otimes \mathbf{I}_n). \tag{5.5}$$

{eq:variposttw}

We note the similarity between (5.5) above and the posterior distribution for the I-prior random effects in a normal model (4.7) seen in the previous chapter. Naïvely computing the inverse $\tilde{\mathbf{V}}_w^{-1}$ presents a computational challenge, as this takes $O(n^3m^3)$ time. By exploiting the Kronecker product structure in $\tilde{\mathbf{V}}_w^{-1}$, we are able to efficiently compute the required inverse in roughly $O(n^3m)$ time—see the appendix for details. Equivalently, we can express the distribution for $\mathbf{w} \sim \tilde q$ as a matrix normal distribution

$$\mathrm{MN}_{nm}\left(\overbrace{\tilde{\mathbf{H}}_\eta^{-1}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top)\tilde{\boldsymbol{\Psi}}^2}^{\tilde{\mathbf{w}}}, \tilde{\mathbf{H}}_\eta^{-2}, \tilde{\boldsymbol{\Psi}}\right). \tag{5.6}$$

{eq:variposttw2}

If the independent I-probit model is assumed, i.e. $\tilde{\boldsymbol{\Psi}} = \operatorname{diag}(\tilde\sigma_1^{-2},\dots,\tilde\sigma_m^{-2})$, then the posterior covariance matrix $\tilde{\mathbf{V}}_w$ has a simpler structure. This means that the random matrix $\mathbf{w}$ will have columns which are independent of each other. By writing $\mathbf{w}_j = (w_{1j},\dots,w_{nj})^\top \in \mathbb{R}^n$, $j=1,\dots,m$, to denote the column vectors of $\mathbf{w}$ and with a slight abuse of notation, we have that

$$\mathrm{N}_{nm}(\operatorname{vec}\mathbf{w}|\operatorname{vec}\tilde{\mathbf{w}},\tilde{\mathbf{V}}_w) = \prod_{j=1}^m \mathrm{N}_n(\mathbf{w}_j|\tilde{\mathbf{w}}_j,\tilde{\mathbf{V}}_{w_j}),$$

where

$$\tilde{\mathbf{w}}_j = \sigma_j^{-2}\tilde{\mathbf{V}}_{w_j}\tilde{\mathbf{H}}_\eta(\tilde{\mathbf{y}}_j^* - \tilde\alpha_j\mathbf{1}_n) \ \text{ and } \ \tilde{\mathbf{V}}_{w_j} = \left(\sigma_j^{-2}\tilde{\mathbf{H}}_\eta^2 + \sigma_j^2\mathbf{I}_n\right)^{-1}.$$

5.5.3 RKHS parameters η

The posterior density $\tilde q$ involving the RKHS parameters is of the form

$$\log \tilde q(\eta) = -\frac{1}{2} \operatorname{tr} \operatorname{E}_{-\eta} \Big[(\mathbf{y}^* - \mathbf{1}_n \boldsymbol\alpha^\top - \mathbf{H}_\eta \mathbf{w}) \boldsymbol\Psi (\mathbf{y}^* - \mathbf{1}_n \boldsymbol\alpha^\top - \mathbf{H}_\eta \mathbf{w})^\top \Big] + \log p(\eta) + \text{const.},$$

where $p(\eta)$ is an appropriate prior distribution for η. Generally, samples $\eta^{(1)}, \dots, \eta^{(T)}$ from $\tilde q(\eta)$ may be obtained using a Metropolis algorithm, and quantities such as $\tilde{\mathbf{H}}_\eta = \operatorname{E}_q[\mathbf{H}_\eta]$ may be approximated using $\frac{1}{T}\sum_{t=1}^T \mathbf{H}_{\eta^{(t)}}$.

However, when only RKHS scale parameters are involved, then the distribution $\tilde q$ can be found in closed-form, much like in the exponential family EM algorithm described in Section 4.3.3. Under the same setting as in that subsection, assume that only $\eta = \{\lambda_1, \dots, \lambda_p\}$ need be estimated, and for each $k = 1, \dots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k \mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}_\eta^2 = \lambda_k^2 \mathbf{R}_k^2 + \lambda_k \mathbf{U}_k + \mathbf{S}_k^2$. Additionally, we impose a further mean-field restriction on $q(\eta)$, and that is $q(\eta) = \prod_{k=1}^p p(\lambda_k)$. Then, by using independent and identical normal priors for λ_k, say $\lambda_k \sim \operatorname{N}(0, v_\lambda)$, each $\tilde q(\lambda_k)$ density is normal with mean and variance

Write down the mean and variance for lambda

5.5.4 Error precision $\boldsymbol\Psi$

A small reparameterisation of the I-prior random effects is necessary to achieve conjugacy for the $\boldsymbol\Psi$ parameter. Let $\mathbf{u} \in \mathbb{R}^{n \times m}$ be a matrix defined by $\boldsymbol\Psi^{-1}\mathbf{w}$. Then $\mathbf{u} \sim \operatorname{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol\Psi^{-1})$ a priori. From (5.6), the optimal variational distribution for $\mathbf{u}$ would be $\operatorname{MN}_{n,m}(\tilde{\mathbf{w}} \tilde{\boldsymbol\Psi}^{-1}, \tilde{\mathbf{H}}_\eta^2, \tilde{\boldsymbol\Psi}^{-1})$. With a Wishart prior on the precision matrix $\boldsymbol\Psi \sim \operatorname{Wis}_m(\mathbf{G}, g)$, where $g \geq m$, the optimal variational density for $\boldsymbol\Psi$ is found to satisfy

$$\log \tilde q(\boldsymbol\Psi) = \text{const.} - \frac{1}{2} \sum_{i=1}^n \operatorname{tr}\big((\mathbf{G}_1 + \mathbf{G}_2 + \mathbf{G}) \boldsymbol\Psi \big) + \frac{g - m - 1}{2} \log |\boldsymbol\Psi|$$

which is recognised as the log density of a Wishart distribution with scale matrix $\mathbf{G}_1 + \mathbf{G}_2 + \mathbf{G}$ and g degrees of freedom, where

$$\begin{aligned} \mathbf{G}_1 &= \mathrm{E}_{\mathcal{Z}\backslash\{\boldsymbol{\Psi}\}\sim q}\left[\sum_{i=1}^n (\mathbf{y}_i^* - \boldsymbol{\alpha} - \mathbf{f}(x_i))(\mathbf{y}_i^* - \boldsymbol{\alpha} - \mathbf{f}(x_i))^\top\right] \\ \mathbf{G}_2 &= \sum_{i=1}^n \mathrm{E}_{\mathbf{u}\sim q}\left[\mathbf{u}_i\mathbf{u}_i^\top\right]. \end{aligned} \tag{5.7}$$

The challenge here is that it involves the second posterior moment of the conically truncated multivariate normal distribution for $\mathbf{y}^*$, which may be obtained by sampling or numerical integration as described earlier.

If the independent I-probit model is considered, then $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \dots, \sigma_m^2)$, class independence holds so we can use independent inverse gamma distributions as a prior for $\boldsymbol{\Sigma}$, i.e. $p(\boldsymbol{\Sigma}) = \prod_{j=1}^m p(\sigma_j^2)$, where each $p(\sigma_j) \equiv \Gamma^{-1}(r,s)$. The posterior for $\boldsymbol{\Sigma}$ will also be of a similar factorised form , namely $\tilde q(\boldsymbol{\Sigma}) = \prod_{j=1}^m \tilde q(\sigma_j^2)$, where $\tilde q(\sigma_j^2)$ is the PDF of an inverse gamma distribution with shape and scale parameters $\tilde r = 2n + r - 1$ and $\tilde s = \frac{1}{2}\|\tilde{\mathbf{y}}_j^* - \tilde\alpha_j - \tilde{\mathbf{f}}_j\|^2 + \frac{1}{2}\|\tilde{\mathbf{u}}_j\|^2 + s$ respectively.

Finally, the posterior distribution for the intercepts follow a normal distribution should the prior specified on the intercepts also be a normal distribution, e.g. $\boldsymbol{\alpha} \sim \mathrm{N}_m(\mathbf{0}, \mathbf{A})$. The posterior mean and variance for the intercepts are given by

$$\tilde{\boldsymbol{\alpha}} = \tilde{\mathbf{V}}_\alpha \tilde{\boldsymbol{\Sigma}}^{-1}\big(\tilde{\mathbf{y}}_i^* - \tilde{\mathbf{f}}(x_i)\big) \ \text{ and } \ \tilde{\mathbf{V}}_\alpha = \big(n\tilde{\boldsymbol{\Sigma}}^{-1} + \mathbf{A}^{-1}\big)^{-1}.$$

Note that the evaluation of each of the component of the posterior depends on some of the components itself, and so this circular dependence is dealt with by using some arbitrary starting values and after which an iterative updating scheme of the components ensues. The updating scheme is performed until a maximum number of iterations is reached, or ideally until some of convergence criterion is met. In variational inference, the *variational lower bound* is typically used to asses convergence. The lower bound is given by

$$\begin{aligned} \mathcal{L} &= \int q(\mathbf{y}^*, \mathbf{w}, \theta) \log\left[\frac{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)}{q(\mathbf{y}^*, \mathbf{w}, \theta)}\right] \mathrm{d}\mathbf{y}^*\mathrm{d}\mathbf{w}\mathrm{d}\theta \\ &= \mathrm{E}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)] - \mathrm{E}[\log q(\mathbf{y}^*, \mathbf{w}, \theta)]. \end{aligned}$$

These are calculable once the posterior distributions $\tilde{q}$ are known—the first term is the expectation of the logarithm of the joint density, whereas the second term factorises into the entropy of its individual components. Similar to the EM algorithm, this quantity isexpected to increase with every iteration.

4. Proof?

5.6 Post-estimation

5.7 Computational consideration

Computational considerations? I think go through the entire chapter without discussing computational issues, and discuss them here. Less cluttered this way.

This is where talk about computational complexity. Of course, $O(n^3)$ (at least) for binary, otherwise $O(mn^3)$ in general, although can be $O((m-1)n^3)$. Worst case is $O(m^3n^3)$, but manage to reduce this. Storage is $O(n^2)$. Prediction is $O(mn^2)$.

Next, issue with probit link function (so to speak) and that is to compute the truncated normal probabilities. Accuracy? But in the variational algorithm, also require moments involving this truncated normal distribution. For independent probit, this is easier. Otherwise, concerned about accuracy as well.

5.8 Examples

5.9 Discussion

I-prior extended to non-normal data. Naive works good, but can be better. Simply transform the normal model through a squashing function. All the nice things about I-prior can be applied here too. Probit model variety of binary and multinomial regression models.

Laplace slow, unreliable modes. MCMC also slow. Variational has similarity to EM, but advantageous: easier to calculate posterior s.d., ability to do inference on transformed parameters.

As with the normal model, storage and time requirements slow. again, look to machine learning. improvements in variational algorithm.

Extend to include class-specific covariates.

improvement in calculating the normal integral? Need to see timing where takes longest

In terms of similarity to other works, the generalised additive models (GAMs) of Hastie and Tibshirani, 1986 comes close. The setup of GAMs is near identical to the I-probit model, although estimation is done differently. GAMs do not assume smooth functions from any RKHS, but instead estimates the f's using a local scoring method or a local likelihood method. Kernel methods for classification are extremely popular in computer science and machine learning; examples include support vector machines (Schölkopf and Smola, 2002) and Gaussian process classification (Rasmussen and Williams, 2006), with the latter being more closely related to the I-probit method. I-priors differ from Gaussian process priors in the specification of the covariance kernel. Gaussian process classification typically uses the logistic link function (or squashing function, to use machine learning nomenclature), and estimation is done most commonly using the Laplace approximation, but other methods such as expectation propagation (Minka, 2001) and MCMC (Neal, 1999) have been explored as well. Variational inference for Gaussian process probit models have been studied by Girolami and Rogers, 2006, with their work providing a close reference to the variational algorithm employed by us.

5.10 Miscellanea

5.10.1 A brief introduction to variational inference

5.10.2 Similarity between EM algorithm and variational Bayes

5.10.3 A note on computing the multivariate normal integral

misc:mnint

How is this calculated? Simulation usually, but also quadrature methods not too bad if m not too large. Stata sheet useful? Talk about if iid errors.

Much research has been devoted into developing efficient computational methods for computing these integral, and MCMC methods seem to be the tool of choice in Bayesian analysis(R. McCulloch and Rossi, 1994; Nobile, 1998; R. E. McCulloch et al., 2000). Things get more tractable if $\mathbf{\Sigma}$ is assumed to be diagonal (which corresponds to abandoning the independence of irrelevant alternatives assumption) and much more

7. can use Hamiltonian Monte Carlo?

so if we assume that $\mathbf{\Sigma} = \mathbf{I}_m$. The latter yields the *normalised I-probit model*, and a discussion of the merits of this model is given later.

Appendix

5.11 Some distributions and their properties

This is a reference relating to the multivariate normal, matrix normal, truncated univariate and multivariate normal, Wishart, and gamma distributions which are collated from various sources for convenience. Of interest are probability density functions, first and second moments, and entropy (as defined in Chapter 3).

5.11.1 Multivariate normal distribution

Let $X \in \mathbb{R}^d$ be distributed according to a multivariate normal (Gaussian) distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^d$ (a square, symmetric, positive-definite matrix). We say that $X \sim \mathrm{N}_d(\mu, \Sigma)$. Then,

- **Pdf**. $p(X|\mu,\Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2} \exp\big(-\frac{1}{2}(X-\mu)^\top\Sigma^{-1}(X-\mu)\big)$.
- **Moments**. $\mathrm{E}\,X = \mu$, $\mathrm{E}[XX^\top] = \Sigma + \mu\mu^\top$.
- **Entropy**. $H(p) = \frac{1}{2}\log|2\pi e\Sigma| = \frac{d}{2}(1+\log 2\pi) + \frac{1}{2}\log|\Sigma|$.

Lemma 5.1 (Properties of multivariate normal)**.** *Assume that* $X \sim \mathrm{N}_d(\mu,\Sigma)$ *and* $Y \sim \mathrm{N}_d(\nu,\Psi)$, *where*

$$X = \begin{pmatrix} X_a \\ X_b \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \quad \textit{and} \;\; \Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ab}^\top & \Sigma_b \end{pmatrix}.$$

Then,

- ***Marginal distributions***.

$$X_a \sim \mathrm{N}_{\dim X_a}(\mu_a, \Sigma_a) \quad \textit{and} \quad X_b \sim \mathrm{N}_{\dim X_b}(\mu_b, \Sigma_b).$$

- ***Conditional distributions.***

  $$X_a|X_b \sim \mathrm{N}_{\dim X_a}(\tilde\mu_a, \tilde\Sigma_a) \quad \textit{and} \quad X_b \sim \mathrm{N}_{\dim X_b}(\tilde\mu_b, \tilde\Sigma_b),$$

  where

  $$\begin{aligned} \tilde\mu_a &= \mu_a + \Sigma_{ab}\Sigma_b^{-1}(X_b - \mu_b) & \qquad \tilde\mu_b &= \mu_b + \Sigma_{ab}^\top\Sigma_a^{-1}(X_a - \mu_a) \\ \tilde\Sigma_a &= \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ab}^\top & \qquad \tilde\Sigma_b &= \Sigma_b - \Sigma_{ab}^\top\Sigma_a^{-1}\Sigma_{ab} \end{aligned}$$

- ***Linear combinations.***

  $$AX + BY + C \sim \mathrm{N}_d(A\mu + B\nu + C, A\Sigma A^\top + B\Psi B^\top)$$

  where A and B are appropriately sized matrices, and $C \in \mathbb{R}^d$.

- ***Product of Gaussian densities.***

  $$p(X|\mu,\Sigma)p(Y|\nu,\Psi) \propto p(Z|m,S)$$

  where $p(Z)$ is a Gaussian density, $m = S(\Sigma^{-1}\mu + \Psi^{-1}\nu)$ and $S = (\Sigma^{-1} + \Psi^{-1})^{-1}$. The normalising constant is equal to the density of $\mu \sim \mathrm{N}(\nu, \Sigma + \Psi)$.

Proof. Omitted—see Petersen and Pedersen (2008, §8). □

Frequently, in Bayesian statistics especially, the following identities will be useful in deriving posterior distributions involving multivariate normals.

Lemma 5.2. *Let $x, b \in \mathbb{R}^d$ be a vector, $X, B \in \mathbb{R}^{n\times d}$ a matrix, and $A \in \mathbb{R}^{d\times d}$ a symmetric, invertible matrix. Then,*

$$\begin{aligned} -\frac{1}{2}x^\top Ax + b^\top x &= -\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) + \frac{1}{2}b^\top A^{-1}b \\ -\frac{1}{2}\operatorname{tr}(X^\top AX) + \operatorname{tr}(B^\top X) &= -\frac{1}{2}\operatorname{tr}\big((X - A^{-1}B)^\top A(X - A^{-1}B)\big) + \frac{1}{2}\operatorname{tr}(B^\top A^{-1}B). \end{aligned}$$

Proof. Omitted—see Petersen and Pedersen (2008, §8.1.6). □

5.11.2 Matrix normal distribution

The matrix normal distribution is an extension of the Gaussian distribution to matrices. Let $X \in \mathbb{R}^{n \times m}$ matrix, and let X follow a matrix normal distribution with mean $\mu \in \mathbb{R}^{n \times m}$ and row and column variances $\Sigma \in \mathbb{R}^{n \times n}$ and $\Psi \in \mathbb{R}^{m \times m}$ respectively, which we denote by $X \sim \text{MN}_{n,m}(\mu, \Sigma, \Psi)$. Then,

- **Pdf**. $p(X|\mu,\Sigma,\Psi) = (2\pi)^{-nm/2}|\Sigma|^{-m/2}|\Psi|^{-n/2}e^{-\frac{1}{2}\operatorname{tr}\left(\Psi^{-1}(X-\mu)^\top\Sigma^{-1}(X-\mu)\right)}$.
- **Moments**. $\operatorname{E} X = \mu$, $\operatorname{Var}(X_{i\cdot}) = \Psi$ for $i = 1,\dots,n$, and $\operatorname{Var}(X_{\cdot j}) = \Sigma$ for $j = 1,\dots,m$.
- **Entropy**. $H(p) = \frac{1}{2}\log|2\pi e(\Psi \otimes \Sigma)| = \frac{nm}{2}(1 + \log 2\pi) + \frac{1}{2}\log|\Sigma|^m|\Psi|^n$.

In the above, '$\otimes$' denotes the Kronecker matrix product defined by

$$\Psi \otimes \Sigma = \begin{pmatrix} \Psi_{11}\Sigma & \Psi_{12}\Sigma & \cdots & \Psi_{1m}\Sigma \\ \Psi_{21}\Sigma & \Psi_{22}\Sigma & \cdots & \Psi_{2m}\Sigma \\ \vdots & \vdots & \ddots & \vdots \\ \Psi_{m1}\Sigma & \Psi_{m2}\Sigma & \cdots & \Psi_{mm}\Sigma \end{pmatrix} \in \mathbb{R}^{nm \times nm}.$$

Of use will be these properties of the Kronecker product (Zhang and Ding, 2013).

- **Bilinearity and associativity**. For appropriately sized matrices A, B and C, and a scalar λ,

$$\begin{aligned} A \otimes (B + C) &= A \otimes B + A \otimes C \\ (A + B) \otimes C &= A \otimes C + B \otimes C \\ \lambda A \otimes B &= A \otimes \lambda B = \lambda(A \otimes B) \\ (A \otimes B) \otimes C &= A \otimes (B \otimes C) \end{aligned}$$

- **Non-commutative**. In general, $A \otimes B \neq B \otimes A$, but they are *permutation equivalent*, i.e. $A \otimes B \neq P(B \otimes A)Q$ for some permutation matrices P and Q.
- **The mixed product property**. $(A \otimes B)(C \otimes D) = AC \otimes BD$.
- **Inverse**. $A \otimes B$ is invertible if and only if A and B are both invertible, and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
- **Transpose**. $(A \otimes B)^\top = A^\top \otimes B^\top$.

- **Determinant**. If A is $n \times n$ and B is $m \times m$, then $|A \otimes B| = |A|^m |B|^n$. Note that the exponent of $|A|$ is the order of B and vice versa.
- **Trace**. Suppose A and B are square matrices. Then $\operatorname{tr}(A \otimes B) = \operatorname{tr} A \operatorname{tr} B$.
- **Rank**. $\operatorname{rank}(A \otimes B) = \operatorname{rank} A \operatorname{rank} B$.
- **Matrix equations**. $AXB = C \Leftrightarrow (B^\top \otimes A) \operatorname{vec} X = \operatorname{vec}(AXB) = \operatorname{vec} C$.

The vectorisation operation 'vec' stacks the columns of the matrices into one long vector, for instance,

$$\operatorname{vec} \Psi = (\Psi_{11}, \dots, \Psi_{m1}, \Psi_{12}, \dots, \Psi_{m2}, \dots, \Psi_{1m}, \dots, \Psi_{mm})^\top \in \mathbb{R}^{m \times m}.$$

Lemma 5.3 (Equivalence between matrix and multivariate normal)**.** $X \sim \mathrm{MN}_{n,m}(\mu, \Sigma, \Psi)$ *if and only if* $\operatorname{vec} X \sim \mathrm{N}_{nm}(\operatorname{vec} \mu, \Psi \otimes \Sigma)$.

Proof. In the exponent of the matrix normal pdf, we have

$$\begin{aligned} -\frac{1}{2} \operatorname{tr} \big(\Psi^{-1} (X-\mu)^\top \Sigma^{-1} (X - \mu) \big) \\ &= -\frac{1}{2} \operatorname{vec}(X-\mu)^\top \operatorname{vec}(\Sigma^{-1}(X-\mu)\Psi^{-1}) \\ &= -\frac{1}{2} \operatorname{vec}(X-\mu)^\top (\Psi^{-1} \otimes \Sigma^{-1}) \operatorname{vec}(X-\mu) \\ &= -\frac{1}{2} (\operatorname{vec} X - \operatorname{vec} \mu)^\top (\Psi \otimes \Sigma)^{-1} (\operatorname{vec} X - \operatorname{vec} \mu). \end{aligned}$$

Also, $|\Sigma|^{-m/2}|\Psi|^{-n/2} = |\Psi \otimes \Sigma|^{-1/2}$. This converts the matrix normal pdf to that of a multivariate normal pdf. □

Some useful properties of the matrix normal distribution are listed:

- **Expected values**.

$$\begin{aligned} \mathrm{E}(X-\mu)(X-\mu)^\top &= \operatorname{tr}(\Psi)\Sigma \in \mathbb{R}^{n \times n} \\ \mathrm{E}(X-\mu)^\top(X-\mu) &= \operatorname{tr}(\Sigma)\Psi \in \mathbb{R}^{m \times m} \\ \mathrm{E}\, XAX^\top &= \operatorname{tr}(A^\top \Psi)\Sigma + \mu A \mu^\top \\ \mathrm{E}\, X^\top B X &= \operatorname{tr}(\Sigma B^\top)\Psi + \mu^\top B \mu \\ \mathrm{E}\, XCX &= \Sigma C^\top \Psi + \mu C \mu \end{aligned}$$

- **Transpose**. $X^\top \sim \mathrm{MN}_{m,n}(\mu^\top, \Psi, \Sigma)$.
- **Linear transformation**. Let $A \in \mathbb{R}^{a \times n}$ be of full-rank $a \leq n$ and $B \in \mathbb{R}^{m \times b}$ be of full-rank $b \leq m$. Then $AXB \sim \mathrm{MN}_{a,b}(\mu^\top, A\Sigma A^\top, B^\top \Psi B)$.
- **Iid**. If $X_i \overset{\text{iid}}{\sim} \mathrm{N}_m(\mu, \Psi)$ for $i = 1, \dots, n$, and we arranged these vectors row-wise into the matrix $X = (X_1^\top, \dots, X_n^\top)^\top \in \mathbb{R}^{n \times m}$, then $X \sim \mathrm{MN}(1_n\mu^\top, I_n, \Psi)$.

5.11.3 Truncated univariate normal distribution

Let $X \sim \mathrm{N}(\mu, \sigma^2)$ with X lying in the interval (a, b). Then we say that X follows a truncated normal distribution, and we denote this by $X \sim {}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, a, b)$. Let $\alpha = (a - \mu)/\sigma$, $\beta = (b - \mu)/\sigma$, and $C = \Phi(\beta) - \Phi(\alpha)$. Then,

- **Pdf**. $p(X|\mu, \sigma, a, b) = C^{-1}(2\pi\sigma^2)^{-1/2}e^{-\frac{1}{2\sigma^2}(X-\mu)^2} = \sigma C^{-1}\phi(\frac{x-\mu}{\sigma})$.
- **Moments**.
$$\begin{aligned} \mathrm{E}\,X &= \mu + \sigma\frac{\phi(\alpha) - \phi(\beta)}{C} \\ \mathrm{E}\,X^2 &= \sigma^2 + \mu^2 + \sigma^2\frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{C} + 2\mu\sigma\frac{\phi(\alpha) - \phi(\beta)}{C} \\ \mathrm{Var}\,X &= \sigma^2\left[1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{C} - \left(\frac{\phi(\alpha) - \phi(\beta)}{C}\right)^2\right] \end{aligned}$$
- **Entropy**.
$$\begin{aligned} H(p) &= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{2C} \\ &= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{1}{2\sigma^2} \cdot \overbrace{\sigma^2\frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{C}}^{\mathrm{Var}\,X - \sigma^2 + (\mathrm{E}\,X - \mu)^2} \\ &= \frac{1}{2}\log 2\pi\sigma^2 + \log C + \frac{1}{2\sigma^2}\,\mathrm{E}[X - \mu]^2 \end{aligned}$$

  because $\mathrm{Var}\,X + (\mathrm{E}\,X - \mu)^2 = \mathrm{E}\,X^2 - \cancel{(\mathrm{E}\,X)^2} + \cancel{(\mathrm{E}\,X)^2} + \mu^2 - 2\mu\,\mathrm{E}\,X$.

For binary probit models, the distributions that come up are one-sided truncations at zero, i.e. ${}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, 0, +\infty)$ (upper tail/positive part) and ${}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, -\infty, 0)$ (lower tail/negative part), for which their moments are of interest. As an aside, if $\mu = 0$ then the truncation ${}^{\mathrm{t}}\mathrm{N}(0, \sigma^2, 0, +\infty)$ is called the *half-normal* distribution. For the positive one-sided truncation at zero, $C = \Phi(+\infty) - \Phi(-\mu/\sigma) = 1 - \Phi(-\mu/\sigma) = \Phi(\mu/\sigma)$, and for the negative one-sided truncation at zero, $C = \Phi(-\mu/\sigma) - \Phi(-\infty) = 1 - \Phi(\mu/\sigma)$.

One may simulate random draws from a truncated normal distribution by drawing from $\text{N}(\mu, \sigma^2)$ and discarding samples that fall outside (a, b). Alternatively, the inverse-transform method using

$$X = \mu + \sigma\Phi^{-1}\left(\Phi(\alpha) + UC\right)$$

with $U \sim \text{Unif}(0,1)$ will work too. Either of these methods will work reasonably well as long as the truncation region is not too far away from μ, but neither is particularly fast. Efficient algorithms have been explored which are along the lines of either accept/reject algorithms (Robert, 1995), Gibbs sampling (Damien and Walker, 2001), or pseudo-random number generation algorithms (Chopin, 2011). The latter algorithm is inspired by the Ziggurat algorithm (Marsaglia and Tsang, 2000) which is considered to be the fastest Gaussian random number generator.

5.11.4 Truncated multivariate normal distribution

Consider the restriction of $X \sim \text{N}_d(\mu, \Sigma)$ to a convex subset[3] $\mathcal{A} \subset \mathbb{R}^d$. Call this distribution the truncated multivariate normal distribution, and denote it $X \sim {}^{\text{t}}\text{N}_d(\mu, \Sigma, \mathcal{A})$. The pdf is $p(X|\mu, \Sigma, \mathcal{A}) = C^{-1}\phi(X|\mu, \Sigma)\,\mathbb{1}[X \in \mathcal{A}]$, where

$$C = \int_{\mathcal{A}} \phi(x|\mu, \Sigma)\,\text{d}x = \text{P}(X \in \mathcal{A}).$$

Generally speaking, there are no closed-form expressions for $\text{E}\,g(X)$ for any well-defined functions g on X. One strategy to obtain values such as $\text{E}\,X$ (mean), $\text{E}\,X^2$ (second moment) and $E \log p(X)$ (entropy) would be Monte Carlo integration. If $X^{(1)}, \dots, X^{(T)}$ are samples from $X \sim {}^{\text{t}}\text{N}_d(\mu, \Sigma, \mathcal{A})$, then $\widehat{\text{E}\,g(X)} = \frac{1}{T}\sum_{i=1}^T g(X^{(i)})$.

Sampling from a truncated multivariate normal distribution is described by Robert (1995) and Damien and Walker (2001). In the latter, the authors explore a simple Gibbs-based approach that is easy to implement in practice. Assume that the one-dimensional slices of $\mathcal{A}$

$$\mathcal{A}_k(X_{-j}) = \{X_j | (X_1, \dots, X_{j-1}, X_j, X_{j+1}, \dots, X_d) \in \mathcal{A}\}$$

[3] A convex subset is a subset of a space that is closed under convex combinations. In Euclidean space, for every pair of points in a convex set, all the points that lie on the straight line segment which joins the pair of points are also in the set.

are readily available so that the bounds or anti-truncation region of X_j given the rest of the components X_{-j} are known to be (x_j^-, x_j^+). Using properties of the normal distribution, the full conditionals of X_j given X_{-j} is

$$\begin{gathered}
X_j \sim {}^{\text{t}}\text{N}(\tilde\mu_j, \tilde\sigma_j^2, x_j^-, x_j^+) \\
\tilde\mu_j = \mu_j + \Sigma_{j,-j}^\top \Sigma_{-j,-j}(x_{-j} - \mu_{-j}) \\
\tilde\sigma_j^2 = \Sigma_{11} - \Sigma_{j,-j}^\top \Sigma_{-j,-j} \Sigma_{j,-j}.
\end{gathered}$$

According to Robert (1995), if $\Psi = \Sigma^{-1}$, then

$$\Sigma_{-j,-j}^{-1} = \Psi_{-j,-j} - \Psi_{j,-j}\Psi_{-j,-j}^\top / \Psi_{jj}$$

which means that we need only compute one global inverse Σ^{-1}. Introduce a latent variable $Y \in \mathbb{R}$ such that the joint pdf of X and Y is

$$p(X_1, \dots, X_d, Y) \propto \exp(-Y/2)\, \mathbb{1}[y > (x-\mu)^\top \Sigma^{-1}(x-\mu)]\, \mathbb{1}[X \in \mathcal{A}].$$

Now, the Gibbs conditional densities for the X_k's are given by

$$p(X_j | X_{-j}, Y) \propto \mathbb{1}[X_j \in \mathcal{B}_j]$$

where

$$\mathcal{B}_j \in (x_j^-, x_j^+) \cap \{X_j | (X-\mu)^\top \Sigma^{-1}(X-\mu) < Y\}.$$

Thus, given values for X_{-j} and Y, the bounds for X_j involves solving a quadratic equation in X_j. The Gibbs conditional density for $Y|X$ is a shifted exponential distribution, which can be sampled using the inverse-transform method. Thus, both X and Y can be sampled directly from uniform variates.

For probit models, we are interested in the conical truncations $\mathcal{C}_j = \{X_j > X_k | k \neq j, \text{and } k = 1, \dots, m\}$ for which the j'th component of X is largest. These truncations form cones in d-dimensional space such that $\mathcal{C}_1 \cup \cdots \cup \mathcal{C}_d = \mathbb{R}^d$, and hence the name.

In the case where Σ is a diagonal matrix, the conically truncated multivariate normal distributions are easier to deal with due to the independence structure in the covariance matrix. In particular, most calculations of interest involve only a one dimensional integral of products of normal cdfs. We present some results that we have not previously seen before elsewhere.

thm:contruncn

Lemma 5.4. *Let $X \sim {}^{\mathrm{t}}\mathrm{N}_d(\mu, \Sigma, \mathcal{C}_j)$, with $\mu = (\mu_1, \dots, \mu_d)^\top$ and $\Sigma = \operatorname{diag}(\sigma_1^2, \dots, \sigma_d^2)$, and $\mathcal{C}_j = \{X_j > X_k | k \neq j, \text{and } k = 1, \dots, m\}$ a conical truncation of $\mathbb{R}^d$ such that the j'th component is largest. Then,*

(i) ***Pdf.*** *The pdf of X has the following functional form:*

$$p(X) = \frac{C^{-1}}{\sigma_1 \cdots \sigma_d (2\pi)^{d/2}} \exp\left[-\frac{1}{2} \sum_{i=1}^d \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \right]$$

where ϕ is the pdf of a standard normal distribution and

$$C = \mathrm{E}_Z \left[\prod_{\substack{i=1 \\ i \neq j}}^d \Phi\left(\frac{\sigma_j}{\sigma_i} Z + \frac{\mu_j - \mu_i}{\sigma_i} \right) \right]$$

where $Z \sim \mathrm{N}(0,1)$.

(ii) ***Moments.*** *The expectation $\mathrm{E}\, X = \big(\mathrm{E}\, X_1, \dots, \mathrm{E}\, X_d\big)^\top$ is given by*

$$\mathrm{E}\, X_i = \begin{cases} \mu_i - \sigma_i C^{-1} \mathrm{E}_Z \left[\phi_i \prod_{k \neq i,j} \Phi_k \right] & \textit{if } i \neq j \\ \mu_j - \sigma_j \sum_{i \neq j} \big(\mathrm{E}\, X_i - \mu_i\big) & \textit{if } i = j \end{cases}$$

and the second moments $\mathrm{E}[X - \mu]^2$ are given by

$$\mathrm{E}[X_i - \mu_i]^2 = \begin{cases} \sigma_i^2 + (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) + \sigma_i \sigma_j C^{-1} \mathrm{E}_Z \left[Z \phi_i \prod_{k \neq i,j} \Phi_k \right] & \textit{if } i \neq j \\ C^{-1} \sigma_j^2 \, \mathrm{E}_Z \left[Z^2 \prod_{k \neq j} \Phi_k \right] & \textit{if } i = j \end{cases}$$

where we had defined

$$\phi_i = \phi_i(Z) = \phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i} \right), \textit{ and}$$
$$\Phi_i = \Phi_i(Z) = \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i} \right).$$

(iii) ***Entropy.*** *The entropy is given by*

$$H(p) = \log C + \frac{d}{2} \log 2\pi + \frac{1}{2} \sum_{i=1}^d \log \sigma_i^2 + \frac{1}{2} \sum_{i=1}^d \frac{1}{\sigma_i^2} \mathrm{E}[x_i - \mu_i]^2.$$

Proof. See Section 5.12 for the proof. □

5.11.5 Wishart distribution

5.11.6 Gamma distribution

5.12 Proofs related to conically truncated multivariate normal distribution

apx:contrun proof

5.12.1 Proof of Lemma 5.4: Pdf

A derivation of the functional form for the pdf of $X \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given. Using the fact that $\int p(x) \mathrm{d}x = 1$, and that

$$\begin{aligned}
&\int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^d \mathrm{N}(\mu_i, \sigma_i^2) \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= \int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^d \left[\frac{1}{\sigma_i} \phi \left(\frac{x_i - \mu_i}{\sigma} \right) \right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= \int \mathbb{1}[x_i < x_j, \forall i \neq j] \frac{1}{\sigma_j} \phi \left(\frac{x_j - \mu_j}{\sigma_j} \right) \prod_{\substack{i=1 \\ i \neq j}}^d \left[\frac{1}{\sigma_i} \phi \left(\frac{x_i - \mu_i}{\sigma_i} \right) \right] \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= \int \prod_{\substack{i=1 \\ i \neq j}}^d \Phi \left(\frac{x_j - \mu_i}{\sigma_i} \right) \frac{1}{\sigma_j} \phi \left(\frac{x_j - \mu_j}{\sigma_j} \right) \mathrm{d}x_j \\
&= \int \prod_{\substack{i=1 \\ i \neq j}}^d \Phi \left(\frac{\sigma_j z_j + \mu_j - \mu_i}{\sigma_i} \right) \phi(z_j) \mathrm{d}z_j \\
&\qquad \text{(by using the standardisation } z_j = (x_j - \mu_j)/\sigma_j \text{)} \\
&= \mathrm{E} \left[\prod_{\substack{i=1 \\ i \neq j}}^d \Phi \left(\frac{\sigma_j}{\sigma_i} Z_j + \frac{\mu_j - \mu_i}{\sigma_i} \right) \right]
\end{aligned}$$

the proof follows directly.

5.12.2 Proof of Lemma 5.4: Moments

Recall that for $Y \sim {}^{\text{t}}\text{N}(\mu, \sigma^2, -\infty, b)$, for some function g of Y, we have that

$$\operatorname{E} g(Y) = \Phi(\beta)^{-1} \int g(y)\, \mathbb{1}[y < b] \phi(y|\mu, \sigma^2)\, \mathrm{d}y,$$

and in particular, we have

$$\operatorname{E}[Y - \mu] = -\sigma \frac{\phi(\beta)}{\Phi(\beta)} \tag{5.8}$$

{eq:tnXminMu}

$$\operatorname{E}[Y - \mu]^2 - \sigma^2 = -\sigma^2 \frac{\beta\phi(\beta)}{\Phi(\beta)} \tag{5.9}$$

{eq:tnXminMusq}

where $\beta = (b - \mu)/\sigma$. For the conically truncated multivariate normal distribution $X \sim {}^{\text{t}}\text{N}_d(\mu, \Sigma, \mathcal{A}_j)$, where $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$, the independence structure of Σ makes it possible to consider the expectations of each of the components separately by marginalising out the rest of the components. For simplicity, denote $p(x_k) = \phi(x_k|\mu_k, \sigma_k) = \sigma_k^{-1}\phi(\frac{x_k - \mu_k}{\sigma_k})$. For $i \neq j$, we have

$$\begin{aligned}
\operatorname{E} g(X_i) &= C^{-1} \int \cdots \int g(x_i)\, \mathbb{1}[x_k < x_j, \forall k \neq j] \prod_{k=1}^d p(x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \frac{\Phi((x_j - \mu_j)/\sigma_j)}{\Phi((x_j - \mu_j)/\sigma_j)} \iint g(x_i)\, \mathbb{1}[x_i < x_j] p(x_i) p(x_j) \prod_{\substack{k=1 \\ k \neq i,j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \mathrm{d}x_i\, \mathrm{d}x_j \\
&= C^{-1} \int \operatorname{E}_{X_i \sim {}^{\text{t}}\text{N}(\mu_i, \sigma_i^2, -\infty, x_j)}\, g(X_i) \prod_{\substack{k=1 \\ k \neq j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\, \mathrm{d}x_j
\end{aligned} \tag{5.10}$$

{eq:tnproofi}

where C is the normalising constant for X, while for the j'the component we have

$$\begin{aligned}
\operatorname{E} g(X_j) &= C^{-1} \int \cdots \int g(x_j)\, \mathbb{1}[x_k < x_j, \forall k \neq j] \prod_{k=1}^d p(x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \int g(x_j) \prod_{\substack{k=1 \\ k \neq j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\, \mathrm{d}x_d.
\end{aligned} \tag{5.11}$$

{eq:tnproofj}

Plugging in (5.8) for $g(X_i) = X_i - \mu_i$ in (5.10) we get

$$\begin{aligned}
\operatorname{E} X_i - \mu_i &= -C^{-1} \int \left(\sigma_i \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \Big/ \Phi\left(\frac{x_j - \mu_i}{\sigma_i}\right)\right) \prod_{\substack{k=1 \\ k\neq j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&= -\sigma_i C^{-1} \int \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k\neq i,j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&= -\sigma_i C^{-1} \int \phi\left(\frac{\sigma_j z + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k\neq j}}^d \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k}\right) \phi(z)\,\mathrm{d}z \\
&= -\sigma_i C^{-1} \operatorname{E}_Z \left[\phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k\neq j}}^d \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k}\right) \right]
\end{aligned}$$

where Z is the distribution of $\mathrm{N}(0,1)$, and we had used a change of variable $x_j = \sigma_j z + \mu_j$, so that $p(x_j) = \sigma_j^{-1}\phi(z)$ and $\mathrm{d}x_j = \sigma_j \mathrm{d}z$. For the j'th component, substitute $g(x_j) = x_j - \mu_j$ in (5.11) to get

$$\begin{aligned}
\operatorname{E} X_j - \mu_j &= C^{-1} \int (x_j - \mu_j) \prod_{\substack{k=1 \\ k\neq j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j)\,\mathrm{d}x_j \\
&= C^{-1}\sigma_j \int z \prod_{\substack{k=1 \\ k\neq j}}^d \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k}\right) \phi(z)\,\mathrm{d}z \\
&= \sigma_j \sum_{\substack{i=1 \\ i\neq j}}^d \sigma_i C^{-1} \operatorname{E} \left[\phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k\neq i,j}}^d \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k}\right) \right] \\
&= -\sigma_j \sum_{\substack{i=1 \\ i\neq j}}^d \big(\operatorname{E} X_i - \mu_i\big),
\end{aligned}$$

where we have made use of Lemma 5.5 in the second last step.

For the second moments, plug in (5.9) for $g(X_i) = (X_i - \mu_i)^2 - \sigma_i^2$ in (5.10) to get

$$
\begin{aligned}
\mathrm{E}[X_i - \mu_i]^2 - \sigma_i^2 &= -\sigma_i^{\cancel{2}} C^{-1} \int \frac{\overbrace{x_j - \mu_i}^{x_j - \mu_i - \mu_j + \mu_j}}{\cancel{\sigma_i}} \cdot \frac{\phi\big((x_j - \mu_i)/\sigma_i\big)}{\Phi\big((x_j - \mu_i)/\sigma_i\big)} \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j) \,\mathrm{d}x_j \\
&= -\sigma_i C^{-1} \int (x_j - \mu_j) \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j) \,\mathrm{d}x_j \\
& + (\mu_j - \mu_i) \cdot \overbrace{-\sigma_i C^{-1} \int \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j) \,\mathrm{d}x_j}^{\mathrm{E}\, X_i - \mu_i} \\
&= (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) \\
& + \sigma_i C^{-1} \int \sigma_j z \phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j z + \mu_j - \mu_k}{\sigma_k}\right) \phi(z) \,\mathrm{d}z \\
&= (\mu_j - \mu_i)(\mathrm{E}\, X_i - \mu_i) \\
& + \sigma_i \sigma_j C^{-1}\, \mathrm{E}\left[Z \phi\left(\frac{\sigma_j Z + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z + \mu_j - \mu_k}{\sigma_k}\right) \right]
\end{aligned}
$$

And similarly, for the j'th component

$$
\begin{aligned}
\mathrm{E}[X_j - \mu_j]^2 &= C^{-1} \int (x_j - \mu_j)^2 \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j) \,\mathrm{d}x_j \\
&= C^{-1} \sigma_j^2 \int z^2 \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{z\sigma_j + \mu_j - \mu_k}{\sigma_k}\right) p(x_j) \,\mathrm{d}z \\
&= C^{-1} \sigma_j^2\, \mathrm{E}_Z \left[Z^2 \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi\left(\frac{Z\sigma_j + \mu_j - \mu_k}{\sigma_k}\right) \right].
\end{aligned}
$$

Lastly, we use this result in the derivation above.

lem:EZgZ

Lemma 5.5. *Let* $Z \sim \mathrm{N}(0,1)$*. Then for all* $m \in \{\mathbb{N} \,|\, m > 1\}$ *and* $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$*,*

$$\mathrm{E}\left[Z \prod_{\substack{k=1\\k\neq j}}^m \Phi(\sigma_k Z + \mu_k) \right] = \sum_{\substack{i=1\\i\neq j}}^m \mathrm{E}\left[\sigma_i \phi(\sigma_i Z + \mu_i) \prod_{\substack{k=1\\k\neq i,j}}^m \Phi(\sigma_k Z + \mu_k) \right]$$

for some $j \in \{1, \dots, m\}$*.*

Proof. Use the fact that for any differentiable function g, $\mathrm{E}[Zg(Z)] = \mathrm{E}[g'(Z)]$, and apply the result with the function $g_m : z \mapsto \prod_{k\neq j} \Phi(\sigma_k z + \mu_k)$. All that is left is to derive the derivative of g, and we use an inductive proof to do this. Introduce the following notation for convenience:

$$\begin{aligned} \phi_i &= \phi(\sigma_i z + \mu_i) \\ \Phi_i &= \Phi(\sigma_i z + \mu_i) \end{aligned}$$

The simplest case is when $m = 2$, which can be trivially shown to be true. Without loss of generality, let $j = 1$. Then

$$\begin{aligned} g_2(z) &= \Phi_2 \\ \Rightarrow \dot g_2(z) &= \sigma_2 \phi_2 = \sum_{\substack{i=1\\i\neq 1}}^2 \left[\sigma_i \phi_i \sum_{\substack{k=1\\k\neq 1,2}}^2 \Phi_k \right]. \end{aligned}$$

Now assume that the inductive hypothesis holds for some $m \in \{\mathbb{N} \,|\, m > 1\}$. That is, the derivative of $g_m(z) = \prod_{k\neq j} \Phi_k$,

$$\dot g_m(z) = \sum_{\substack{i=1\\i\neq j}}^m \left[\sigma_i \phi_i \prod_{\substack{k=1\\k\neq i,j}}^m \Phi_k \right],$$

is assumed to be true. Also assume that, without loss of generality, $j \neq m+1$. Then, the derivative of

$$g_{m+1}(z) = \prod_{\substack{k=1\\k\neq j}}^{m+1} \Phi_k = g_m(z)\Phi_{m+1}$$

is found to be

$$\begin{aligned}
\dot g_{m+1}(z) &= \sigma_{m+1}\phi_{m+1} g_m(z) + \dot g_m(z)\Phi_{m+1} \\
&= \sigma_{m+1}\phi_{m+1} \prod_{\substack{k=1 \\ k\neq j}}^{m} \Phi_k + \sum_{\substack{i=1 \\ i\neq j}}^{m} \Bigg[\sigma_i\phi_i \prod_{\substack{k=1 \\ k\neq i,j}}^{m} \Phi_k \Bigg] \Phi_{m+1} \\
&= \sigma_{m+1}\phi_{m+1} \prod_{\substack{k=1 \\ k\neq j,m+1}}^{m+1} \Phi_k + \sum_{\substack{i=1 \\ i\neq j}}^{m} \Bigg[\sigma_i\phi_i \prod_{\substack{k=1 \\ k\neq i,j}}^{m+1} \Phi_k \Bigg] \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{m+1} \Bigg[\sigma_i\phi_i \prod_{\substack{k=1 \\ k\neq i,j}}^{m+1} \Phi_k \Bigg],
\end{aligned}$$

as required for the inductive proof. Using linearity of expectations, the proof is complete. □

5.12.3 Proof of Lemma 5.4: Entropy

As a direct consequence of the definition of entropy,

$$\begin{aligned}
H(p) &= -\operatorname{E}\log p(X) \\
&= -\operatorname{E}\left[-\log C - \frac{d}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 - \frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2 \right] \\
&= \log C + \frac{d}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i^2}\operatorname{E}[x_i-\mu_i]^2.
\end{aligned}$$

5.13 Derivation of the CAVI algorithm

Let $\mathcal{Z} = \{\mathbf{y}^*, \mathbf{w}, \boldsymbol\alpha, \eta, \boldsymbol\Psi\}$. Approximate the posterior for $\mathcal{Z}$ by a mean-field variational distribution

$$\begin{aligned}
p(\mathbf{y}^*, \mathbf{w}, \alpha, \eta, \boldsymbol\Psi|\mathbf{y}) &\approx q(\mathbf{y}^*)q(\mathbf{w})q(\boldsymbol\alpha)q(\eta)q(\boldsymbol\Psi) \\
&= \prod_{i=1}^{n} q(\mathbf{y}_i^*)q(\mathbf{w})q(\boldsymbol\alpha)q(\eta)q(\boldsymbol\Psi).
\end{aligned}$$

The first line is by assumption, while the second line follows from an induced factorisation on the latent propensities, as we will see later. If needed, we also assume that $q(\eta)$ factorises into its constituents components. Recall that, for each $\xi \in \mathcal{Z}$, the optimal mean-field variational density $\tilde q$ for ξ satisfies

$$\log \tilde q(\xi) = \operatorname{E}_{-\xi}[\log p(\mathbf{y}, \mathcal{Z})] + \text{const.} \tag{5.2}$$

Write $\mathbf{f} = \mathbf{H}_\eta \mathbf{w} \in \mathbb{R}^{n \times m}$. The joint likelihood $p(\mathbf{y}, \mathcal{Z})$ is given by

$$\begin{aligned} p(\mathbf{y}, \mathcal{Z}) &= p(\mathbf{y}|\mathcal{Z})p(\mathcal{Z}) \\ &= p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi})p(\mathbf{w}|\boldsymbol{\Psi})p(\eta)p(\boldsymbol{\Psi})p(\boldsymbol{\alpha}). \end{aligned}$$

For reference, the relevant distributions are listed below.

- $\boldsymbol{p(\mathbf{y}|\mathbf{y}^*)}$. For each observation $i \in \{1, \dots, n\}$, given the corresponding latent propensities $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{im}^*)$, the distribution for y_i is a degenerate distribution which depends on the j'th component of $\mathbf{y}_i^*$ being largest, where the value observed for y_i was j. Since each of the y_i's are independent, everything is multiplicative.

$$p(\mathbf{y}|\mathbf{y}^*) = \prod_{i=1}^n \prod_{j=1}^m p_{ij} = \prod_{i=1}^n \prod_{j=1}^m \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*]^{\mathbb{1}[y_i = j]}.$$

- $\boldsymbol{p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi})}$. Given values for the parameters and I-prior random effects, the distribution of the latent propensities is matrix normal

$$\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi} \sim \text{MN}_{n,m}(\mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1}).$$

Write $\boldsymbol{\mu} = \mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}$. Its pdf is

$$\begin{aligned} p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi}) &= \exp\left[-\frac{nm}{2}\log 2\pi + \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}\left((\mathbf{y}^* - \boldsymbol{\mu})\boldsymbol{\Psi}(\mathbf{y}^* - \boldsymbol{\mu})^\top\right)\right] \\ &= \exp\left[-\frac{nm}{2}\log 2\pi + \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^n(\mathbf{y}_i^* - \boldsymbol{\mu}_i)^\top\boldsymbol{\Psi}(\mathbf{y}_i^* - \boldsymbol{\mu}_i)\right], \end{aligned}$$

where $\mathbf{y}_i^* \in \mathbb{R}^m$ and $\boldsymbol{\mu}_i \in \mathbb{R}^m$ are the rows of $\mathbf{y}^*$ and $\boldsymbol{\mu}$ respectively. The second line follows directly from the definition of the trace, but also emanates from the fact that $\mathbf{y}_i^*$ are independent multivariate normal with mean $\boldsymbol{\mu}_i$ and variance $\boldsymbol{\Psi}^{-1}$.

- $\boldsymbol{p(\mathbf{w}|\boldsymbol{\Psi})}$. The $\mathbf{w}$'s are normal random matrices $\mathbf{w} \sim \mathrm{N}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$ with pdf

$$\begin{aligned} p(\mathbf{w}|\boldsymbol{\Psi}) &= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}\left(\mathbf{w}\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\right)\right] \\ &= \exp\left[-\frac{nm}{2}\log 2\pi - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i=1}^n \mathbf{w}_i^\top\boldsymbol{\Psi}^{-1}\mathbf{w}_i\right]. \end{aligned}$$

- $\boldsymbol{p(\eta)}$. The most common scenario would be $\eta = \{\lambda_1, \dots, \lambda_p\}$ only. In this case, choose independent normal priors for each $\lambda_k \sim \mathrm{N}(m_k, v_k)$, $k = 1, \dots, p$, whose pdf is

$$p(\eta) = \prod_{k=1}^p \exp\left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\log v_k - \frac{1}{2v_k^2}(\lambda_k - m_k)^2\right].$$

  An improper prior $p(\eta) \propto$ const. can be used as well, and this is the same as letting $m_k \to 0$ and $v_k \to 0$. The resulting posterior will be proper. If η contains other parameters as well, such as the Hurst coefficient $\gamma \in (0,1)$, SE lengthscale $l > 0$ or polynomial offset $c > 0$, then appropriate priors should be used to match the support of the parameter. Choices include $p(\gamma) = \mathbb{1}\big(\gamma \in (0,1)\big)$ and $l, c \sim \Gamma(a,b)$.

- $\boldsymbol{p(\boldsymbol{\Psi})}$. For the precision matrix, a Wishart prior with scale matrix $\mathbf{G}^{-1}$ and g degrees of freedom, denoted $\boldsymbol{\Psi} \sim \mathrm{Wis}_m(\mathbf{G}^{-1}, g)$, is convenient. It has pdf

$$p(\boldsymbol{\Psi}) = \exp\left[\text{const.} + \frac{g-m-1}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\operatorname{tr}(\mathbf{G}\boldsymbol{\Psi})\right].$$

  For the independent I-probit model, $\boldsymbol{\Psi} = \mathrm{diag}($
  $psi_1, \dots, \psi_m)$, and we choose independent Gamma distributions for each precision $\sigma_j^{-2} \sim \Gamma(s_j, r_j)$, where s_j and r_j are the shape and rate parameters. Then,

$$p(\boldsymbol{\Psi}) = \prod_{j=1}^m \exp\left[\text{const.} + (s_j - 1)\log\psi_j - r_j\psi_j\right].$$

- $\boldsymbol{p(\boldsymbol{\alpha})}$. Choose independent normal priors for the intercept, $\alpha_j \sim \mathrm{N}(a_j, A_j)$ for $j = 1, \dots, m$. The pdf is

$$p(\boldsymbol{\alpha}) = \prod_{j=1}^m \exp\left[\log 2\pi - \log A_j - \frac{1}{2A_j}(\alpha_j - a_j)^2\right].$$

Remark 5.1. The priors on the parameters $\{\boldsymbol{\alpha}, \eta\}$ can be set to very vague or even improper priors, and the resulting posterior will still yield a proper distribution. Using improper priors eases the algebra slightly. For the precision matrix $\boldsymbol{\Psi}$, it is best to stick with the Wishart prior to avoid positive-definite issues, unless the independent I-probit model is used, in which case Jeffreys' prior for the precisions $p(\sigma_j^{-2}) \propto \sigma_j^2$ is a convenient choice.

5.13.1 Derivation of $\tilde{q}(\mathbf{y}^*)$

The rows of $\mathbf{y}^*$ are independent, and thus we can consider the variational density for each $\mathbf{y}_i^*$ separately. Consider the case where y_i takes one particular value $j \in \{1, \dots, m\}$. The mean-field density $q(\mathbf{y}_i^*)$ for each $i = 1, \dots, n$ is found to be

$$\begin{aligned}
\log \tilde{q}(\mathbf{y}_i^*) &= \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \operatorname{E}_{\mathcal{Z} \backslash \{\mathbf{y}^*\} \sim q} \left[-\frac{1}{2}(\mathbf{y}_i^* - \boldsymbol{\mu}_i)^\top \boldsymbol{\Psi} (\mathbf{y}_i^* - \boldsymbol{\mu}_i) \right] + \text{const.} \\
&= \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \left[-\frac{1}{2}(\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i)^\top \tilde{\boldsymbol{\Psi}} (\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i) \right] + \text{const.} \qquad (\star) \\
&\equiv \begin{cases} \phi(\mathbf{y}_i^* | \tilde{\boldsymbol{\mu}}_i, \tilde{\boldsymbol{\Psi}}) & \text{if } y_{ij}^* > y_{ik}^*, \forall k \neq j \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_i = \operatorname{E} \boldsymbol{\alpha} + (\operatorname{E} \mathbf{H}_\eta \operatorname{E} \mathbf{w})_i$, and expectations are taken under the optimal mean-field distribution $\tilde{q}$. The distribution $q(\mathbf{y}_i^*)$ is a truncated m-variate normal distribution such that the j'th component is always largest. Unfortunately, the expectation of this distribution cannot be found in closed-form, and must be approximated by techniques such as Monte Carlo integration. If, however, the independent I-probit model is used and $\tilde{\boldsymbol{\Psi}}$ is diagonal, then Lemma X provides a simplification.

Remark 5.2. In $(\star)$ above, we needn't consider the second order terms in the expectations because they do not involve $\mathbf{y}^*$ and can be absorbed into the constant. To see this,

$$\begin{aligned}
\operatorname{E}[(\mathbf{y}_i^* - \boldsymbol{\mu}_i)^\top \boldsymbol{\Psi} (\mathbf{y}_i^* - \boldsymbol{\mu}_i)] &= \operatorname{E}[\mathbf{y}_i^{*\top} \boldsymbol{\Psi} \mathbf{y}_i^* + \boldsymbol{\mu}_i^\top \boldsymbol{\Psi} \boldsymbol{\mu}_i - 2\boldsymbol{\mu}_i^\top \boldsymbol{\Psi} \mathbf{y}_i^*] \\
&= \mathbf{y}_i^{*\top} \boldsymbol{\Psi} \mathbf{y}_i^* - 2 \operatorname{E}[\boldsymbol{\mu}_i^\top] \operatorname{E}[\boldsymbol{\Psi}] \mathbf{y}_i^* + \text{const.} \\
&= \mathbf{y}_i^{*\top} \boldsymbol{\Psi} \mathbf{y}_i^* - 2\tilde{\boldsymbol{\mu}}_i^\top \tilde{\boldsymbol{\Psi}} \mathbf{y}_i^* + \text{const.} \\
&= (\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i)^\top \tilde{\boldsymbol{\Psi}} (\mathbf{y}_i^* - \tilde{\boldsymbol{\mu}}_i) + \text{const.}
\end{aligned}$$

We will see this occurring a lot later on and we shall take note of this fact.

5.13.2 Derivation of $\tilde{q}(\mathbf{w})$

The terms involving $\mathbf{w}$ in (5.2) are the $p(\mathbf{y}^*|\boldsymbol{\alpha},\mathbf{w},\eta,\boldsymbol{\Psi})$ and $p(\mathbf{w}|\boldsymbol{\Psi})$ terms, and the rest are absorbed into the constant. The easiest way to derive $\tilde{q}(\mathbf{w})$ is to vectorise $\mathbf{y}^*$ and $\mathbf{w}$. We know that

$$\operatorname{vec}\mathbf{y}^*|\boldsymbol{\alpha},\mathbf{w},\eta,\boldsymbol{\Psi} \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}), \boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_n\right)$$

and

$$\operatorname{vec}\mathbf{w}|\boldsymbol{\Psi} \sim \mathrm{N}_{nm}(\mathbf{0},\boldsymbol{\Psi}\otimes\mathbf{I}_n)$$

using properties of matrix normal distributions. We also use the fact that $\operatorname{vec}(\mathbf{H}_\eta\mathbf{w}) = (\mathbf{I}_m\otimes\mathbf{H}_\eta)\operatorname{vec}\mathbf{w}$. For simplicity, write $\bar{\mathbf{y}}^* = \operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)$, and $\mathbf{M} = (\mathbf{I}_m\otimes\mathbf{H}_\eta)$. Thus,

$$\begin{aligned}
\log\tilde{q}(\mathbf{w}) &= \mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[-\frac{1}{2}(\bar{\mathbf{y}}^* - \mathbf{M}\operatorname{vec}\mathbf{w})^\top(\boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_n)^{-1}(\bar{\mathbf{y}}^* - \mathbf{M}\operatorname{vec}\mathbf{w})\right] \\
&\quad + \mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[-\frac{1}{2}(\operatorname{vec}\mathbf{w})^\top(\boldsymbol{\Psi}\otimes\mathbf{I}_n)^{-1}\operatorname{vec}(\mathbf{w})\right] + \text{const.} \\
&= -\frac{1}{2}\mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[(\operatorname{vec}\mathbf{w})^\top\Big(\overbrace{\mathbf{M}^\top(\boldsymbol{\Psi}\otimes\mathbf{I}_n)\mathbf{M} + (\boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_n)}^{\mathbf{A}}\Big)\operatorname{vec}(\mathbf{w})\right] \\
&\quad + \mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[\overbrace{\bar{\mathbf{y}}^{*\top}(\boldsymbol{\Psi}\otimes\mathbf{I}_n)\mathbf{M}}^{\mathbf{a}^\top}\operatorname{vec}(\mathbf{w})\right] + \text{const.} \\
&= -\frac{1}{2}\mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[(\operatorname{vec}\mathbf{w} - \mathbf{A}^{-1}\mathbf{a})^\top\mathbf{A}(\operatorname{vec}\mathbf{w} - \mathbf{A}^{-1}\mathbf{a})\right] + \text{const.}
\end{aligned}$$

This is recognised as a multivariate normal of dimension nm with mean and precision given by $\operatorname{vec}\tilde{\mathbf{w}} = \mathrm{E}[\mathbf{A}^{-1}\mathbf{a}]$ and $\tilde{\mathbf{V}}_w^{-1} = \mathrm{E}[\mathbf{A}]$ respectively. With a little algebra, we find that

$$\begin{aligned}
\mathbf{V}_w^{-1} &= \mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}[\mathbf{A}] \\
&= \mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[(\mathbf{I}_m\otimes\mathbf{H}_\eta)^\top(\boldsymbol{\Psi}\otimes\mathbf{I}_n)(\mathbf{I}_m\otimes\mathbf{H}_\eta) + (\boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_n)\right] \\
&= \mathrm{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[(\boldsymbol{\Psi}\otimes\mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_n)\right] \\
&= (\tilde{\boldsymbol{\Psi}}\otimes\tilde{\mathbf{H}}_\eta^2) + (\tilde{\boldsymbol{\Psi}}^{-1}\otimes\mathbf{I}_n)
\end{aligned}$$

and making a first-order approximation $(\operatorname{E}\mathbf{A})^{-1} \approx \operatorname{E}[\mathbf{A}^{-1}]$[4],

$$\begin{aligned}
\operatorname{vec}\tilde{\mathbf{w}} &= \operatorname{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}[\mathbf{A}^{-1}\mathbf{a}] \\
&= \tilde{\mathbf{V}}_w \operatorname{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[(\mathbf{I}_m \otimes \mathbf{H}_\eta)(\mathbf{\Psi} \otimes \mathbf{I}_n)\operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right] \\
&= \tilde{\mathbf{V}}_w \operatorname{E}_{\mathcal{Z}\backslash\{\mathbf{w}\}\sim q}\left[(\mathbf{\Psi} \otimes \mathbf{H}_\eta)\operatorname{vec}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right] \\
&= \tilde{\mathbf{V}}_w(\tilde{\mathbf{\Psi}} \otimes \tilde{\mathbf{H}}_\eta)\operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top).
\end{aligned}$$

Ideally, we do not want to work with the $nm \times nm$ matrix $\mathbf{V}_w$, since its inverse is expensive to compute. We can exploit the Kronekcer product structure to compute the inverse efficiently. Perform an orthogonal eigendecomposition of $\mathbf{H}_\eta$ to obtain $\mathbf{H}_\eta = \mathbf{V}\mathbf{U}\mathbf{V}^\top$ and of $\mathbf{\Psi}$ to obtain $\mathbf{\Psi} = \mathbf{Q}\mathbf{P}\mathbf{Q}^\top$. This process takes $O(n^3 + m^3) \approx O(n^3)$ time if $m \ll n$. Then, manipulate $\mathbf{V}_w^{-1}$ as follows (for clarity, we drop the tildes from the notations):

$$\begin{aligned}
\mathbf{V}_w^{-1} &= (\mathbf{\Psi} \otimes \mathbf{H}_\eta^2) + (\mathbf{\Psi}^{-1} \otimes \mathbf{I}_n) \\
&= (\mathbf{Q}\mathbf{P}\mathbf{Q}^\top \otimes \mathbf{V}\mathbf{U}^2\mathbf{V}^\top) + (\mathbf{Q}\mathbf{P}^{-1}\mathbf{Q}^\top \otimes \mathbf{V}\mathbf{V}^\top) \\
&= (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P} \otimes \mathbf{U}^2)(\mathbf{Q}^\top \otimes \mathbf{V}^\top) + (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P}^{-1} \otimes \mathbf{I}_n)(\mathbf{Q}^\top \otimes \mathbf{V}^\top) \\
&= (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P} \otimes \mathbf{U}^2 + \mathbf{P}^{-1} \otimes \mathbf{I}_n)(\mathbf{Q}^\top \otimes \mathbf{V}^\top)
\end{aligned}$$

Its inverse is

$$\begin{aligned}
\mathbf{V}_w &= (\mathbf{Q}^\top \otimes \mathbf{V}^\top)^{-1}(\mathbf{P} \otimes \mathbf{U}^2 + \mathbf{P}^{-1} \otimes \mathbf{I}_n)^{-1}(\mathbf{Q} \otimes \mathbf{V})^{-1} \\
&= (\mathbf{Q} \otimes \mathbf{V})(\mathbf{P} \otimes \mathbf{U}^2 + \mathbf{P}^{-1} \otimes \mathbf{I}_n)^{-1}(\mathbf{Q}^\top \otimes \mathbf{V}^\top)
\end{aligned}$$

which is easy to compute since the middle term is an inverse of diagonal matrices.

In the case of the I-probit model, where $\mathbf{\Psi} = \operatorname{diag}(\sigma_1^{-2}, \dots, \sigma_m^{-2})$, then the covariance $\mathbf{V}_w$ takes a simpler form. Specifically, it has the block diagonal structure:

$$\begin{aligned}
\mathbf{V}_w &= \big(\operatorname{diag}(\tilde{\sigma}_1^{-2}, \dots, \tilde{\sigma}_m^{-2}) \otimes \tilde{\mathbf{H}}_\eta^2 + (\operatorname{diag}(\tilde{\sigma}_1^{2}, \dots, \tilde{\sigma}_m^{2}) \otimes \mathbf{I}_n\big)^{-1} \\
&= \operatorname{diag}\Big((\tilde{\sigma}_1^{-2}\tilde{\mathbf{H}}_\eta^2 + \tilde{\sigma}_1^2\mathbf{I}_n)^{-1}, \cdots, (\tilde{\sigma}_m^{-2}\tilde{\mathbf{H}}_\eta^2 + \tilde{\sigma}_m^2\mathbf{I}_n)^{-1}\Big) \\
&=: \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \dots, \tilde{\mathbf{V}}_{w_m}).
\end{aligned}$$

[4] Groves and Rothenberg (1969) show that $\operatorname{E}[\mathbf{A}^{-1}] = (\operatorname{E}\mathbf{A})^{-1} + \mathbf{B}$, where $\mathbf{B}$ is a positive-definite matrix.

The mean $\tilde{\mathbf{w}}$ in matrix form is

$$\begin{aligned}
\tilde{\mathbf{w}} &= \tilde{\mathbf{V}}_w(\operatorname{diag}(\tilde\sigma_1^{-2},\dots,\tilde\sigma_m^{-2}) \otimes \tilde{\mathbf{H}}_\eta)(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol\alpha}^\top) \\
&= \operatorname{diag}(\tilde{\mathbf{V}}_{w_1},\dots,\tilde{\mathbf{V}}_{w_m}) \operatorname{diag}(\tilde\sigma_1^{-2}\tilde{\mathbf{H}}_\eta,\dots,\tilde\sigma_m^{-2}\tilde{\mathbf{H}}_\eta)(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol\alpha}^\top) \\
&= \operatorname{diag}(\tilde\sigma_1^{-2}\tilde{\mathbf{V}}_{w_1}\tilde{\mathbf{H}}_\eta,\dots,\tilde\sigma_m^{-2}\tilde{\mathbf{V}}_{w_m}\tilde{\mathbf{H}}_\eta)(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol\alpha}^\top) \\
&= \begin{array}{c} \begin{array}{ccc} \tilde{\mathbf{w}}_{\cdot 1} & \cdots & \tilde{\mathbf{w}}_{\cdot m} \end{array} \\ \begin{pmatrix} \tilde\sigma_1^{-2}\tilde{\mathbf{V}}_{w_1}\tilde{\mathbf{H}}_\eta(\tilde{\mathbf{y}}^*_{\cdot 1} - \tilde\alpha_1\mathbf{1}_n) & \cdots & \tilde\sigma_m^{-2}\tilde{\mathbf{V}}_{w_m}\tilde{\mathbf{H}}_\eta(\tilde{\mathbf{y}}^*_{\cdot m} - \tilde\alpha_m\mathbf{1}_n) \end{pmatrix} \end{array}.
\end{aligned}$$

Therefore, we can consider the distribution of $\mathbf{w} = (\mathbf{w}_{\cdot 1},\dots,\mathbf{w}_{\cdot m})$ columnwise, and each are normally distributed with mean and variance

$$\tilde{\mathbf{w}}_{\cdot j} = \tilde\sigma_j^{-2}\tilde{\mathbf{V}}_{w_j}\tilde{\mathbf{H}}_\eta(\tilde{\mathbf{y}}^*_{\cdot j} - \tilde\alpha_j\mathbf{1}_n) \quad \text{and} \quad \tilde{\mathbf{V}}_{w_j} = \big(\tilde\sigma_j^{-2}\tilde{\mathbf{H}}_\eta^2 + \tilde\sigma_j^2\mathbf{I}_n\big)^{-1}.$$

A quantity that we will be requiring time and again will be $\operatorname{tr}(\mathbf{C}\operatorname{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}])$, where $\mathbf{C} \in \mathbb{R}^{m\times m}$ and $\mathbf{D} \in \mathbb{R}^{n\times n}$ are both square and symmetric matrices. Using the definition of the trace directly, we get

$$\begin{aligned}
\operatorname{tr}(\mathbf{C}\operatorname{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) &= \sum_{i,j=1}^m \mathbf{C}_{ij}\operatorname{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]_{ij} \\
&= \sum_{i,j=1}^m \mathbf{C}_{ij}\operatorname{E}[\mathbf{w}_{\cdot i}^\top\mathbf{D}\mathbf{w}_{\cdot j}]. \qquad (5.12)
\end{aligned}$$

{eq:trCEwDw }

The expectation of the univariate quantity $\mathbf{w}_{\cdot i}^\top\mathbf{D}\mathbf{w}_{\cdot j}$ is inspected below:

$$\begin{aligned}
\operatorname{E}[\mathbf{w}_{\cdot i}^\top\mathbf{D}\mathbf{w}_{\cdot j}] &= \operatorname{tr}(\mathbf{D}\operatorname{E}[\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot i}^\top]) \\
&= \operatorname{tr}\big(\mathbf{D}(\operatorname{Cov}(\mathbf{w}_{\cdot j},\mathbf{w}_{\cdot i}) + \operatorname{E}[\mathbf{w}_{\cdot j}]\operatorname{E}[\mathbf{w}_{\cdot i}]^\top)\big) \\
&= \operatorname{tr}\big(\mathbf{D}(\mathbf{V}_w[i,j] + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot i}^\top)\big).
\end{aligned}$$

where $\mathbf{V}_w[i,j] \in \mathbb{R}^{n\times n}$ refers to the (i,j)'th submatrix block of $\mathbf{V}_w$. Of course, in the independent the I-probit model, this is equal to

$$\mathbf{V}_w[i,j] = \delta_{ij}(\psi_j\mathbf{H}_\eta^2 + \psi_j^{-1}\mathbf{I}_n)^{-1}$$

where δ is the Kronecker delta. Continuing on (5.12) leads us to

$$\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top \mathbf{D}\mathbf{w}]) = \sum_{i,j=1}^m \mathbf{C}_{ij} \Big(\operatorname{tr}\big(\mathbf{D}(\delta_{ij}\mathbf{V}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot i}^\top)\big). \Big) \, .$$

If $\mathbf{C} = \operatorname{diag}(c_1, \dots, c_m)$, then

$$\begin{aligned} \operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top \mathbf{D}\mathbf{w}]) &= \sum_{j=1}^m c_j \Big(\operatorname{tr}\big(\mathbf{D}\tilde{\mathbf{V}}_{w_j}\big) + \tilde{\mathbf{w}}_{\cdot j}^\top \mathbf{D}\tilde{\mathbf{w}}_{\cdot j} \Big) \\ &= \sum_{j=1}^m c_j \operatorname{tr}\big(\mathbf{D}(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top)\big) \end{aligned}$$

5.13.3 Derivation of $\tilde{q}(\eta)$

By looking at only the terms involving η in (5.2), we deduce that $\tilde{q}$ for η satisfies

$$\begin{aligned} \log \tilde{q}(\eta) &= -\frac{1}{2} \operatorname{tr} \mathrm{E}_{\mathcal{Z}\backslash\{\eta\}\sim q} \Big[(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top \Big] + \log p(\eta) \\ &\quad + \text{const.} \\ &= -\frac{1}{2} \operatorname{tr} \mathrm{E}_{\mathcal{Z}\backslash\{\eta\}\sim q} \Big(\boldsymbol{\Psi}\mathbf{w}^\top \mathbf{H}_\eta^2 \mathbf{w} - 2\boldsymbol{\Psi}\mathbf{w}^\top \mathbf{H}_\eta(\mathbf{y}^* - \boldsymbol{\alpha}) \Big) + \log p(\eta) + \text{const.} \\ &= -\frac{1}{2} \operatorname{tr} \Big(\tilde{\boldsymbol{\Psi}}\,\mathrm{E}[\mathbf{w}^\top \mathbf{H}_\eta^2 \mathbf{w}] - 2\tilde{\boldsymbol{\Psi}}\tilde{\mathbf{w}}^\top \mathbf{H}_\eta(\tilde{\mathbf{y}}^* - \tilde{\boldsymbol{\alpha}}) \Big) + \log p(\eta) + \text{const.} \end{aligned}$$

with some appropriate prior $p(\eta)$. In general, this does not have a recognisable form in η, especially when it is not linearly dependent on the kernel matrix. This happens when considering parameters other than the scales of the RKHSs. Our interest would be to obtain $\tilde{\mathbf{H}}_\eta := \mathrm{E}_{\eta\sim q}\,\mathbf{H}_\eta$ and $\tilde{\mathbf{H}}_\eta^2 := \mathrm{E}_{\eta\sim q}\,\mathbf{H}_\eta^2$. We use a Metropolis random-walk algorithm to obtain these quantities, as detailed in the algorithm below.

Now consider the case where $\eta = \{\lambda_1, \dots, \lambda_p\}$ (RKHS scale parameters only), and the scenario described in the exponential family EM algorithm of Section 4.3.3 applies. In particular, for $k = 1, \dots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k \mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}_\eta^2 = \lambda_k^2 \mathbf{R}_k^2 + \lambda_k \mathbf{U}_k + \mathbf{S}_k^2$. Then, for $j = 1, \dots, m$, assuming each of

Algorithm 1 Metropolis random-walk to sample η

```
**inputs** $\tilde{\boldsymbol\alpha}$, $\tilde{\mathbf w}$, $\tilde{\boldsymbol\Psi}$, and $s$ Metropolis sampling s.d.
**initialise** $\eta^{(0)} \in \mathbb{R}^q$ and $t \leftarrow 0$
**for** $t = 1, \dots, T$ **do**
    Draw $\eta^* \sim \mathrm{N}_q(\eta^{(t)}, s^2)$
    Accept/reject proposal state, i.e.
        $$\eta^{(t+1)} \leftarrow \begin{cases} \eta^* & \text{if } u \sim \mathrm{Unif}(0,1) < \pi_{\text{acc}} \\ \eta^{(t)} & \text{otherwise} \end{cases}$$
    where
        $$\pi_{\text{acc}} = \min\Big(1, \exp\big(\log\tilde q(\eta^*) - \log\tilde q(\eta^{(t)})\big)\Big).$$
**end for**
$\tilde{\mathbf H}_\eta \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf H_{\eta^{(t)}}$ and $\tilde{\mathbf H}_\eta^2 \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf H_{\eta^{(t)}}^2$
```

the $q(\lambda_k)$ densities are independent of each other, we find that

$$\begin{aligned}
\log\tilde q(\lambda_k) &= \mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[-\frac{1}{2}\operatorname{tr}\big((\mathbf y^* - \boldsymbol\mu)\boldsymbol\Psi(\mathbf y^* - \boldsymbol\mu)^\top\big)\right] - \frac{1}{2v_k^2}(\lambda_k - m_k)^2 + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[\boldsymbol\Psi\mathbf w^\top\mathbf H_\eta^2\mathbf w - 2\boldsymbol\Psi(\mathbf y^* - \mathbf 1\boldsymbol\alpha^\top)^\top\mathbf H_\eta\mathbf w\right] \\
&\quad - \frac{1}{2v_k^2}(\lambda_k - m_k)^2 + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[\boldsymbol\Psi\mathbf w^\top(\lambda_k^2\mathbf R_k^2 + \lambda_k\mathbf U_k)\mathbf w - 2\boldsymbol\Psi(\mathbf y^* - \mathbf 1\boldsymbol\alpha^\top)^\top(\lambda_k\mathbf R_k)\mathbf w\right] \\
&\quad - \frac{1}{2v_k^2}(\lambda_k^2 - 2m_k\lambda_k) + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[\lambda_k^2\boldsymbol\Psi\mathbf w^\top\mathbf R_k^2\mathbf w - 2\lambda_k\Big(\boldsymbol\Psi(\mathbf y^* - \mathbf 1\boldsymbol\alpha^\top)^\top\mathbf R_k\mathbf w - \frac{1}{2}\boldsymbol\Psi\mathbf w^\top\mathbf U_k\mathbf w\Big)\right] \\
&\quad - \frac{1}{2}\left(\frac{1}{v_k^2}\lambda_k^2 - 2\frac{m_k}{v_k^2}\lambda_k\right) + \text{const.} \\
&= -\frac{1}{2}\Big[\lambda_k^2\overbrace{\big(\operatorname{tr}(\tilde{\boldsymbol\Psi}\,\mathrm{E}[\mathbf w^\top\mathbf R_k^2\mathbf w]) + v_k^{-2}\big)}^{c_k} \\
&\qquad\qquad - 2\lambda_k\overbrace{\Big(\operatorname{tr}\Big(\tilde{\boldsymbol\Psi}(\tilde{\mathbf y}^* - \mathbf 1\tilde{\boldsymbol\alpha}^\top)^\top\mathbf R_k\tilde{\mathbf w} - \frac{1}{2}\tilde{\boldsymbol\Psi}\,\mathrm{E}[\mathbf w^\top\mathbf U_k\mathbf w]\Big) + m_kv_k^{-2}\Big)}^{d_k}\Big]
\end{aligned}$$

By completing the squares, we recognise this is as the kernel of a univariate normal density. Specifically, $\lambda_k \sim \mathrm{N}(d_k/c_k, 1/c_k)$. The quantity $\tilde{\mathbf H}_\eta$ can be obtained by substituting $\lambda_k \mapsto \mathrm{E}_{\lambda_k \sim q}[\lambda_k]$ in the expression XXX. However, in the calculation of $\tilde{\mathbf{H}}_\eta^2$, we must replace $\lambda_k^2 \mapsto \mathrm{E}_{\lambda_k \sim q}[\lambda_k]^2 + \mathrm{Var}_{\lambda_k \sim q}[\lambda_k]$ in all occurrences of square terms. This can be cumbersome, so if felt necessary, use the approximation $\lambda_k^2 \mapsto \mathrm{E}_{\lambda_k \sim q}[\lambda_k]^2$ instead.

Example 5.1. Suppose $k = 1$, and we only have λ to estimate. Then, $\mathbf{H}_\eta = \lambda \mathbf{H}$, $\mathbf{R}_k = \mathbf{H}$, $\mathbf{R}_k^2 = \mathbf{H}^2$, and $\mathbf{U}_k = \mathbf{0}$. Suppose also we use an improper prior $\lambda_k \propto \text{const.}$, which is the same as having $v_k^2 \to 0$ and $m_k v_k^{-2} \to 0$. The mean field distribution for λ is then

$$\lambda \sim \mathrm{N}\left(\frac{\mathrm{tr}\left(\tilde{\boldsymbol{\Psi}}(\tilde{\mathbf{y}}^* - \mathbf{1}\tilde{\boldsymbol{\alpha}}^\top)^\top \mathbf{H}\tilde{\mathbf{w}} \right)}{\mathrm{tr}(\tilde{\boldsymbol{\Psi}}\, \mathrm{E}[\mathbf{w}^\top \mathbf{H}^2 \mathbf{w}])}, \frac{1}{\mathrm{tr}(\tilde{\boldsymbol{\Psi}}\, \mathrm{E}[\mathbf{w}^\top \mathbf{H}^2 \mathbf{w}])} \right)$$

Further, if $\tilde{\boldsymbol{\Psi}} = \tilde{\psi}\mathbf{I}_m$, then

$$\lambda \sim \mathrm{N}\left(\frac{\sum_{j=1}^m (\tilde{\mathbf{y}}_{\cdot j}^* - \tilde{\alpha}_j \mathbf{1})^\top \mathbf{H}\tilde{\mathbf{w}}_{\cdot j}}{\sum_{j=1}^m \mathrm{tr}\left(\mathbf{H}^2\, \mathrm{E}[\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot j}^\top] \right)}, \frac{1}{\sum_{j=1}^m \mathrm{tr}\left(\mathbf{H}^2\, \mathrm{E}[\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot j}^\top] \right)} \right)$$

which bears a resemblance to the exponential family EM algorithm solutions described in Chapter 4. Now, $\tilde{\mathbf{H}}_\eta = \mathrm{E}[\lambda \mathbf{H}] = \tilde{\lambda}\mathbf{H}$, and $\tilde{\mathbf{H}}_\eta^2 = \mathrm{E}[\lambda^2 \mathbf{H}^2] = (\mathrm{Var}\,\lambda + \tilde{\lambda}^2)\mathbf{H}^2$.

Derivation of $\tilde{q}(\boldsymbol{\Psi})$

Introduce the transformed random matrix $\mathbf{u} = \mathbf{w}\boldsymbol{\Psi}^{-1} \in \mathbb{R}^{n \times m}$. Since we have that $\mathrm{vec}\,\mathbf{u} = (\mathrm{vec}\,\mathbf{w})^\top(\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)$, the optimal mean-field distribution for $\mathbf{u}$ is normal with mean $\mathrm{vec}\,\tilde{\mathbf{u}} = \mathrm{vec}(\tilde{\mathbf{w}}\tilde{\boldsymbol{\Psi}}^{-1})$ and variance

$$\tilde{\mathbf{V}}_u = (\tilde{\boldsymbol{\Psi}}^{-1} \otimes \mathbf{I}_n)\tilde{\mathbf{V}}_w(\tilde{\boldsymbol{\Psi}}^{-1} \otimes \mathbf{I}_n).$$

In the case of the independent model, its mean is $\tilde{\mathbf{u}}_{\cdot j} = \tilde{\psi}_j^{-1}\tilde{\mathbf{u}}_{\cdot j}$ for $j = 1, \dots, m$ and its variance is

$$\tilde{\mathbf{V}}_u = \mathrm{diag}(\tilde{\psi}_1^{-2}\tilde{\mathbf{V}}_{w1}, \dots, \tilde{\psi}_m^{-2}\tilde{\mathbf{V}}_{wm}).$$

Now, to derive $\tilde q(\boldsymbol\Psi)$ for the full I-probit model, we inspect the equation

$$\begin{aligned}
\log \tilde q(\boldsymbol\Psi) &= \operatorname{E}_{\mathcal{Z}\backslash\{\boldsymbol\Psi\}\sim q}\left[\frac{n}{2}\log|\boldsymbol\Psi| - \frac{1}{2}\operatorname{tr}\big((\mathbf{y}^* - \boldsymbol\mu)^\top(\mathbf{y}^* - \boldsymbol\mu)\boldsymbol\Psi\big) + \frac{n}{2}\log|\boldsymbol\Psi| - \frac{1}{2}\operatorname{tr}\big(\mathbf{u}^\top\mathbf{u}\boldsymbol\Psi\big)\right] \\
& + \frac{g-m-1}{2}\log|\boldsymbol\Psi| - \frac{1}{2}\operatorname{tr}(\mathbf{G}\boldsymbol\Psi) + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\Big(\big(\mathbf{G} + \overbrace{\operatorname{E}[(\mathbf{y}^* - \boldsymbol\mu)^\top(\mathbf{y}^* - \boldsymbol\mu)]}^{\mathbf{G}_1} + \overbrace{\operatorname{E}[\mathbf{u}^\top\mathbf{u}]}^{\mathbf{G}_2}\big)\boldsymbol\Psi\Big) \\
& + \frac{2n+g-m-1}{2}\log|\boldsymbol\Psi| + \text{const.}
\end{aligned}$$

which we recognise to be a Wishart distribution with scale matrix $(\mathbf{G} + \mathbf{G}_1 + \mathbf{G}_2)^{-1}$ and $2n + g - m$ degrees of freedom. Note that using an improper prior, i.e. $\mathbf{G} = \mathbf{0}$ and $g = m$, will still yield a proper posterior distribution. The mean of this distribution is $\tilde{\boldsymbol\Psi} = (2n + g - m)(\mathbf{G} + \mathbf{G}_1 + \mathbf{G}_2)^{-1}$. The matrix $\mathbf{G}_1$ is given as

$$\begin{aligned}
\mathbf{G}_1 &= \operatorname{E}[(\mathbf{y}^* - \boldsymbol\mu)^\top(\mathbf{y}^* - \boldsymbol\mu)] \\
&= \operatorname{E}\left[\mathbf{y}^{*\top}\mathbf{y}^* + \boldsymbol\alpha\mathbf{1}_n^\top\mathbf{1}_n\boldsymbol\alpha^\top + \mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w} - 2\mathbf{y}^{*\top}\mathbf{1}_n\boldsymbol\alpha^\top - 2\mathbf{y}^{*\top}\mathbf{H}_\eta\mathbf{w} - 2\boldsymbol\alpha\mathbf{1}_n^\top\mathbf{H}_\eta\mathbf{w}\right] \\
&= \operatorname{E}\left[\mathbf{y}^{*\top}\mathbf{y}^*\right] + n\operatorname{E}[\boldsymbol\alpha\boldsymbol\alpha^\top] + \operatorname{E}[\mathbf{w}^\top\mathbf{H}_\eta\mathbf{w}] - 2(\tilde{\mathbf{y}}^{*\top}\mathbf{1}_n\tilde{\boldsymbol\alpha}^\top + \tilde{\mathbf{y}}^{*\top}\tilde{\mathbf{H}}_\eta\tilde{\mathbf{w}} + \tilde{\boldsymbol\alpha}\mathbf{1}_n^\top\tilde{\mathbf{H}}_\eta\tilde{\mathbf{w}})
\end{aligned}$$

This involves second order moments of a conically truncated multivariate normal distribution, which needs to be obtained via simulation. Meanwhile,

$$\begin{aligned}
\mathbf{G}_{2,ij} &= \operatorname{E}[\mathbf{u}^\top\mathbf{u}]_{ij} \\
&= \operatorname{E}[\mathbf{u}_{\cdot i}^\top\mathbf{u}_{\cdot j}] \\
&= \tilde{\mathbf{V}}_u[i,j] + \tilde{\mathbf{u}}_{\cdot i}^\top\tilde{\mathbf{u}}_{\cdot j}.
\end{aligned}$$

In the case of the I-probit model, we use a gamma prior on each of the precisions in the diagonal entries of $\boldsymbol\Psi = \operatorname{diag}(\psi_1, \dots, psi_m)$. Then, the variational density for each

ψ_j is found to be

$$
\begin{aligned}
\log\tilde q(\psi_j) &= \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\Psi}\}\sim q}\bigg[\cancel{\frac{n}{2}\log(\psi_1\cdots\psi_m)} - \frac{1}{2}\sum_{j=1}^m\sum_{i=1}^n \psi_j(\mathbf{y}^*_{ij}-\boldsymbol{\mu}_{ij})^2\bigg] \\
&\qquad + \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\Psi}\}\sim q}\bigg[\cancel{-\frac{n}{2}\log(\psi_1\cdots\psi_m)} - \frac{1}{2}\sum_{j=1}^m\sum_{i=1}^n \psi_j\mathbf{u}^2_{ij}\bigg] \\
&\qquad + \sum_{j=1}^m\big((s_j-1)\log\psi_j - r_j\psi_j\big) + \text{const.} \\
&= (s_j-1)\log\psi_j - \psi_j\big(\frac{1}{2}\,\mathrm{E}\|\mathbf{y}_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 + \frac{1}{2}\,\mathrm{E}\|\mathbf{u}_{\cdot j}\|^2 + r_j\big) + \text{const.}
\end{aligned}
$$

which is again a gamma distribution, and the shape and rate parameters can be read directly. The mean is given by $\tilde\psi_j = s_j(\frac{1}{2}\,\mathrm{E}\|\mathbf{y}_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 + \frac{1}{2}\,\mathrm{E}\|\mathbf{u}_{\cdot j}\|^2 + r_j)^{-1}$. Recall that each of the n components of $\mathbf{y}_{\cdot j}-\boldsymbol{\mu}_{\cdot j}$ are independent, and can be calculated using the methods described above. Also, we have $\mathbf{u}_{\cdot j} \sim \mathrm{N}_n(\tilde{\mathbf{u}}_{\cdot j}, \tilde{\mathbf{V}}_{u_j})$, and so $\mathrm{E}\|\mathbf{u}_{\cdot j}\|^2 = \operatorname{tr}(\tilde{\mathbf{V}}_{u_j} + \tilde{\mathbf{u}}_{\cdot j}\tilde{\mathbf{u}}_{\cdot j}^\top)$.

Derivation $\tilde q(\boldsymbol{\alpha})$

The terms involving α_j in (5.2) are

$$
\begin{aligned}
\log\tilde q(\alpha_j) &= \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\alpha}\}\sim q}\bigg[-\frac{1}{2}\sum_{k=1}^m\sum_{i=1}^n \psi_{ik}(y^*_{ik}-\alpha_j-f_{ik})^2\bigg] - \frac{A_j^{-1}}{2}(\alpha_j-a_j)^2 + \text{const.} \\
&= -\frac{1}{2}\,\mathrm{E}\left[\alpha_j^2\sum_{i=1}^n\psi_{ij} - 2\alpha_j\sum_{i=1}^n\psi_{ij}(y^*_{ij}-f_{ij})\right] - \frac{1}{2}\Big(A_j^{-1}\alpha_j^2 - 2A_j^{-1}a_j\alpha_j\Big) + \text{const.} \\
&= -\frac{\sum_{i=1}^n\tilde\psi_{ij} + A_j^{-1}}{2}\left(\alpha_j - \frac{\sum_{i=1}^n\tilde\psi_{ij}(\tilde y^*_{ij}-\tilde f_{ij}) + A_j^{-1}a_j}{\sum_{i=1}^n\tilde\psi_{ij} + A_j^{-1}}\right)^2 + \text{const.}
\end{aligned}
$$

which implies a normal distribution for α_j whose mean and variance can be read directly. Here, we used the notation $\tilde f_{ij}$ to mean the (i,j)'th element of $\mathrm{E}[\mathbf{H}_\eta\mathbf{w}] = \tilde{\mathbf{H}}_\eta\tilde{\mathbf{w}} \in \mathbb{R}^{n\times m}$.

As a remark, due to identifiability, only $m-1$ of these intercept are estimable. We can either put a constraint that one of the intercepts is fixed at zero, or the sum of the intercepts equals zero. The latter constraint is implemented in this thesis, and this is realised by estimating all the intercepts and then centring them.

5.14 Deriving the ELBO expression

A convergence criterion would be when there is no more significant increase in the lower bound $\mathcal{L}$, as defined by

$$\begin{aligned}
\mathcal{L} &= \int q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha) \log \left[\frac{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \lambda, \alpha)}{q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha)} \right] \mathrm{d}\mathbf{y}^* \mathrm{d}\mathbf{w} \mathrm{d}\lambda \mathrm{d}\alpha \\
&= \mathrm{E}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \lambda, \alpha)] - \mathrm{E}[\log q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha)] \\
&= \cancel{\mathrm{E}\left[\log \prod_{i=1}^n \prod_{j=1}^m p(y_i | y_{ij}^*) \right]} + \mathrm{E}\left[\log p(\mathbf{y}^* | \mathbf{f}) \right] + \mathrm{E}\left[\log p(\mathbf{w}) \right] + \cancel{\mathrm{E}\left[\log p(\lambda) \right]} + \cancel{\mathrm{E}\left[\log p(\alpha) \right]} \\
& \quad - \mathrm{E}\left[\log q(\mathbf{y}^*) \right] - \mathrm{E}\left[\log q(\mathbf{w}) \right] - \mathrm{E}\left[\log q(\lambda) \right] - \mathrm{E}\left[\log q(\alpha) \right]
\end{aligned}$$

Note that the categorical pmf $p(y_i | y_{ij}^*)$ becomes degenerate once the latent variables are known, so this term is cancelled out. With the exception of $q(\mathbf{y}^*)$, all of the distributions are Gaussian. The following results will be helpful.

Terms involving distributions of $\mathbf{y}^*$

$$\begin{aligned}
\mathrm{E}\left[\log p(\mathbf{y}^* | \mathbf{f}) \right] - \mathrm{E}\left[\log q(\mathbf{y}^*) \right] &= \sum_{i=1}^n \sum_{j=1}^m \mathrm{E}\left[\log p(y_{ij}^* | f_{ij}) \right] + \sum_{i=1}^n \mathcal{H}\big(q(y_i^*) \big) \\
&= \cancel{\sum_{i=1}^n \sum_{j=1}^m \left(-\frac{1}{2} \log 2\pi - \frac{1}{2} \mathrm{E}[y_{ij}^* - f_{ij}]^2 \right)} \\
& \quad + \cancel{\sum_{i=1}^n \sum_{j=1}^m \left(\frac{1}{2} \log 2\pi + \frac{1}{2} \mathrm{E}[y_{ij}^* - f_{ij}]^2 \right)} + \sum_{i=1}^n \log C_i
\end{aligned}$$

Terms involving distributions of w

$$
\begin{aligned}
\operatorname{E}[\log p(\mathbf{w})] - \operatorname{E}[\log q(\mathbf{w})] &= \sum_{j=1}^m \Big(\operatorname{E}[\log p(\mathbf{w}_j)] - \operatorname{E}[\log q(\mathbf{w}_j)] \Big) \\
&= \sum_{j=1}^m \left(-\frac{n}{2}\log 2\pi - \frac{1}{2}\operatorname{E}[\mathbf{w}_j^\top \mathbf{w}_j] + \mathcal{H}\big(q(\mathbf{w}_j)\big) \right) \\
&= \sum_{j=1}^m \left(\cancel{-\frac{n}{2}\log 2\pi} - \frac{1}{2}\operatorname{tr}\Big(\operatorname{E}[\mathbf{w}_j\mathbf{w}_j^\top]\Big) + \frac{n}{2}(1 + \cancel{\log 2\pi}) - \frac{1}{2}\log|\mathbf{A}_j| \right) \\
&= \frac{nm}{2} - \frac{1}{2}\sum_{j=1}^m \Big(\operatorname{tr}\widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j| \Big)
\end{aligned}
$$

Terms involving distribution of $q(\lambda)$

$$
\begin{aligned}
-\operatorname{E}[\log q(\lambda)] &= \sum_{j=1}^m \mathcal{H}\big(q(\lambda_j)\big) \\
&= \sum_{j=1}^m \left(\frac{1}{2}(1 + \log 2\pi) - \frac{1}{2}\log c_j \right) \\
&= \frac{m}{2}(1 + \log 2\pi) - \frac{1}{2}\sum_{j=1}^m \log c_j
\end{aligned}
$$

or if using single λ

$$
-\operatorname{E}[\log q(\lambda)] = \frac{1}{2}(1 + \log 2\pi) - \frac{1}{2}\log\sum_{j=1}^m c_j.
$$

Terms involving distribution of $q(\alpha)$

$$
\begin{aligned}
-\operatorname{E}[\log q(\alpha)] &= \sum_{j=1}^m \mathcal{H}\big(q(\alpha_j)\big) \\
&= \frac{m}{2}(1 + \log 2\pi - \log n)
\end{aligned}
$$

or if using single α

$$-\operatorname{E}[\log q(\alpha)] = \frac{1}{2}(1 + \log 2\pi - \log nm).$$

The lower bound

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^n \log C_i + \frac{nm}{2} - \frac{1}{2}\sum_{j=1}^m \Big(\operatorname{tr} \widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j| \Big) \\
& + \frac{m}{2}(1 + \log 2\pi) - \frac{1}{2}\sum_{j=1}^m \log c_j + \frac{m}{2}(1 + \log 2\pi - \log n) \\
&= \frac{m}{2}\big(n + 2(1+\log 2\pi) - \log n\big) - \frac{1}{2}\sum_{j=1}^m \Big(\operatorname{tr} \widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j| + \log c_j \Big) + \sum_{i=1}^n \log C_i
\end{aligned}$$

Of course, if using either single α or single λ, then the formula needs to be adjusted accordingly.

Bibliography

bishop2006pattern Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag.

chopin2011fast Chopin, Nicolas (2011). “Fast simulation of truncated Gaussian distributions”. In: *Statistics and Computing* 21.2, pp. 275–288.

damien2001sampling Damien, Paul and Stephen G Walker (2001). “Sampling truncated normal, beta, and gamma densities”. In: *Journal of Computational and Graphical Statistics* 10.2, pp. 206–215.

girolami2006variational Girolami, Mark and Simon Rogers (2006). “Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors”. In: *Neural Computation* 18.8, pp. 1790–1817.

groves1969note Groves, Theodore and Thomas Rothenberg (1969). “A note on the expected value of an inverse matrix”. In: *Biometrika* 56.3, pp. 690–691.

hastie1986 Hastie, Trevor and Robert Tibshirani (Aug. 1986). “Generalized Additive Models”. In: *Statist. Sci.* 1.3, pp. 297–310. DOI: 10.1214/ss/1177013604. URL: https://doi.org/10.1214/ss/1177013604.

jamil2017 Jamil, Haziq and Wicher Bergsma (2017). “iprior: An R Package for Regression Modelling using I-priors”. In: *Manuscript in submission.*

marsaglia2000ziggurat Marsaglia, George and Wai Wan Tsang (2000). “The ziggurat method for generating random variables”. In: *Journal of statistical software* 5.8, pp. 1–7.

mccullagh1989 McCullagh, P. and John A. Nelder (1989). *Generalized Linear Models.* 2nd. Chapman & Hall/CRC Press.

mcculloch2000bayesian McCulloch, Robert E, Nicholas G Polson, and Peter E Rossi (2000). “A Bayesian analysis of the multinomial probit model with fully identified parameters”. In: *Journal of econometrics* 99.1, pp. 173–193.

mcculloch1994exact McCulloch, Robert and Peter E Rossi (1994). “An exact likelihood analysis of the multinomial probit model”. In: *Journal of Econometrics* 64.1, pp. 207–240.

meng1997algorithm Meng, Xiao-Li and David Van Dyk (1997). “The EM Algorithm—an Old Folk-song Sung to a Fast New Tune”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 59.3, pp. 511–567.

minka2001expectation Minka, Thomas P (2001). “Expectation propagation for approximate Bayesian inference”. In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 362–369.

neal1999 Neal, Radford M. (1999). “Regression and Classification using Gaussian Process Priors”. In: *Bayesian Statistics*. Ed. by J M Bernardo, J O Berger, A P Dawid, and A F M Smith. Vol. 6. Oxford University Press. (with discussion), pp. 475–501.

nobile1998hybrid Nobile, Agostino (1998). “A hybrid Markov chain for the Bayesian analysis of the multinomial probit model”. In: *Statistics and Computing* 8.3, pp. 229–242.

petersen2008matrix Petersen, Kaare Brandt and Michael Syskind Pedersen (2008). “The matrix cookbook”. In: *Technical University of Denmark* 7.15, p. 510.

rasmussen2006gaussian Rasmussen, Carl Edward and Christopher K I Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

robert1995simulation Robert, Christian P (1995). “Simulation of truncated normal variables”. In: *Statistics and computing* 5.2, pp. 121–125.

scholkopf2002learning Schölkopf, Bernhard and Alexander J Smola (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

zhang2013kronecker Zhang, Huamin and Feng Ding (2013). “On the Kronecker products and their applications”. In: *Journal of Applied Mathematics* 2013.