To-do list

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

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

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

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

As with the normal I-prior model, an estimate of the posterior regression function with optimised hyperparameters is sought. The log likelihood function $L(\cdot)$ for θ using all n observations $\{(y_1, x_1), \dots, (y_n, x_n)\}$ is obtained by integrating out the I-prior from the categorical likelihood, as follows:

$$
\begin{aligned}
L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w}, \theta) \, p(\mathbf{w}|\theta) \, \mathrm{d}\mathbf{w} \\
&= \log \int \prod_{i=1}^n \prod_{j=1}^m \Big(g^{-1} \big(\alpha_k + \overbrace{f_k(x_i)}^{\sum_{i'=1}^n h_\eta(x_i, x_{i'}) w_{i'j}} \big)_{k=1}^m \Big)^{[y_i = j]} \cdot \phi(\mathbf{w}|\mathbf{0}, \boldsymbol{\Psi} \otimes \mathbf{I}_n) \, \mathrm{d}\mathbf{w} \qquad (5.2)
\end{aligned}
$$

{eq:intractablelikelihood}

where we have denoted the probit relationship from (**??**) using the function $g^{-1} : \mathbb{R}^m \to [0, 1]$. Unlike in the continuous response models, the integral does not present itself in closed form due to the conditional categorical PMF of the y_i's, which they themselves involve integrals of multivariate normal densities. For binary response models, g^{-1} is simply the probit function, but for multinomial responses, this can be quite challenging to evaluate—more on this in Section X.

Furthermore, the posterior distribution of the regression function, which requires the density of $\mathbf{w}|\mathbf{y}$, depends on the marginalisation provided by (5.2). The challenge of estimation is then to first overcome this intractability by means of a suitable approximation of the integral. We present three possible avenues to achieve this aim, namely the Laplace approximation, Markov chain Monte Carlo (MCMC) methods, and variational Bayes.

5.4.1 Laplace approximation

To compute the posterior density $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) =: e^{Q(\mathbf{w})}$ with normalising constant equal to the marginal density of $\mathbf{y}$, $p(\mathbf{y}) = \int e^{Q(\mathbf{w})} \, \mathrm{d}\mathbf{w}$, we have established that this is intractable. Laplace's method (Kass and Raftery, 1995, §4.1.1, pp. 777–778) entails expanding a Taylor series for Q about its posterior mode $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, which gives the relationship

$$
\begin{aligned}
Q(\mathbf{w}) &= Q(\hat{\mathbf{w}}) + \cancelto{0}{(\mathbf{w} - \hat{\mathbf{w}})^\top \nabla Q(\hat{\mathbf{w}})} - \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w} - \hat{\mathbf{w}}) + \cdots \\
&\approx Q(\hat{\mathbf{w}}) + -\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w} - \hat{\mathbf{w}}),
\end{aligned}
$$

because, assuming that Q has a unique maxima, ∇Q evaluated at its mode is zero. This is recognised as the logarithm of an unnormalised Gaussian density, implying $\mathbf{w}|\mathbf{y} \sim \mathrm{N}_n(\hat{\mathbf{w}}, \boldsymbol{\Omega}^{-1})$. Here, $\boldsymbol{\Omega} = -\nabla^2 Q(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}$ is the negative Hessian of Q evaluated at the posterior mode, and is typically obtained as a byproduct of the maximisation routine of Q using gradient or quasi-gradient based methods.

The marginal distribution is then approximated by

$$
\begin{aligned}
p(\mathbf{y}) &\approx \int \exp \overbrace{Q(\mathbf{w})}^{Q(\hat{\mathbf{w}}) - \frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w} - \hat{\mathbf{w}})} \, \mathrm{d}\mathbf{w} \\
&= (2\pi)^{n/2} |\boldsymbol{\Omega}|^{-1/2} e^{Q(\hat{\mathbf{w}})} \int (2\pi)^{-n/2} |\boldsymbol{\Omega}|^{1/2} \exp\left(-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^\top \boldsymbol{\Omega}(\mathbf{w} - \hat{\mathbf{w}}) \right) \mathrm{d}\mathbf{w} \\
&= (2\pi)^{n/2} |\boldsymbol{\Omega}|^{-1/2} p(\mathbf{y}|\hat{\mathbf{w}}) p(\hat{\mathbf{w}}).
\end{aligned}
$$

The log marginal density of course depends on the parameters θ, which becomes the objective function to maximise in a likelihood maximising approach. Note that, should a fully Bayesian approach be undertaken, i.e. priors prescribed on the model parameters using $\theta \sim p(\theta)$, then this approach is viewed as a maximum a posteriori approach.

In any case, each evaluation of the objective function $L(\theta) = \log p(\mathbf{y}|\boldsymbol{\theta})$ involves finding the posterior modes $\hat{\mathbf{w}}$. This is a slow and difficult undertaking, especially for large sample sizes n—even assuming computation of the class probabilities g^{-1} is efficient—because the dimension of this integral is exactly the sample size. Furthermore, obtaining standard errors for the parameters are cumbersome, and it is likely that a computationally burdensome bootstrapping approach is needed. Lastly, as a comment, Laplace's method only approximates the true marginal likelihood well if the true function is small far away from the mode.

5.4.2 Variational inference

We turn to variational inference as a method of estimation. Variational methods are widely discussed in the machine learning literature, but there have been efforts to popularise it in statistics (Blei et al., 2017). In a fully Bayesian setting, one obtains an approximation to the intractable posterior distribution of interest, which is then used for inferential purposes in lieu of the actual posterior distribution.

In addition to the I-probit model, suppose that prior distributions are assigned on the hyperparameters of the model, $\theta \sim p(\theta)$. By appending the latent variables $\{\mathbf{y}^*, \mathbf{w}\}$ to the hyperparameters θ, we seek an approximation

$$p(\mathbf{y}^*, \mathbf{w}, \theta | \mathbf{y}) \approx \tilde{q}(\mathbf{y}^*, \mathbf{w}, \theta),$$

where $\tilde{q}$ satisfies $\tilde{q} = \arg\min_q \operatorname{KL}(q \| p)$, subject to certain constraints. The constraint considered by us in this thesis is that q satisfies a *mean-field* factorisation

$$q(\mathbf{y}^*, \mathbf{w}, \theta) = q(\mathbf{y}^*) q(\mathbf{w}) q(\theta).$$

Under this scheme, the posterior for $\mathbf{y}^*$ is found to be a *conically truncated multivariate normal* distribution, and for $\mathbf{w}$, a multivariate normal distribution. The posterior density $q(\theta)$ is often of a recognisable form, and usually one of the exponential family densities (normal, Wishart or gamma). This is useful, because point estimates of the hyperparameters can be taken to be either the mean or mode of these well-known distributions. In cases where $q(\theta)$ does not conform to an exponential family type density, then inference can still be done by sampling methods.

It can be shown that, for some variational density q, the marginal log-likelihood is an upper-bound for the quantity $\mathcal{L}$

$$\log p(\mathbf{y}) \geq \operatorname{E}_q \log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta) - \operatorname{E}_q \log \tilde{q}(\mathbf{y}^*, \mathbf{w}, \theta) =: \mathcal{L},$$

a quantity often referred to as the *evidence lower bound* (ELBO). It turns out that minimising $\operatorname{KL}(q \| p)$ is equivalent to maximising the ELBO, a quantity that is more practical to work with than the KL divergence. That is, if $\tilde{q}$ approximates the true posterior well, then the ELBO is a suitable proxy for the maximised marginal log-likelihood.

The algorithm to obtain $\tilde{q}$ which maximises the ELBO is known as the *coordinate ascent variational inference* (CAVI) algorithm. The algorithm itself typically condenses to that of a simple, sequential updating scheme, akin to the expectation-maximisation (EM) algorithm for exponential families we saw in Chapter 4, which is very fast to implement compared to the other methods described in the previous subsections. A full derivation of the variational algorithm used by us will be described in Section 5.5.

5.4.3 Markov chain Monte Carlo methods

As an alternative to the deterministic Bayesian approach of variational inference, it is possible to use Markov chain Monte Carlo sampling methods as an approach to stochastically approximate the intractable posterior distribution.

Albert and Chib (1993) showed that the latent variable approach to probit models can be analysed using exact Bayesian methods, due to the underlying normality structure. Paired with corresponding conjugate prior choices, sampling from the posterior is very simple using a Gibbs sampling approach. On the other hand, this data augmentation scheme enlarges the variable space to $n + q$ dimensions, where q is the number of parameters to estimate, which is inefficient and computationally challenging especially when n is large. It is no longer possible to marginalise the normal latent variables from the model, as this is intractable, as discussed previously.

Hamiltonian Monte Carlo is another possibility, since it does not require conjugacy. For binary models, this is a feasible approach because the class probabilities are a function of the normal CDF, which means that it is doable using off-the-shelf software such as Stan. However, with multinomial responses, the problem of computing class probabilities, which involve integration of an at most m-dimensional normal density, must be addressed separately.

5.4.4 Comparison of estimation methods

Compare: Laplace, variational and HMC.

The three estimation methods described aim to overcome the intractable integral by means of either a deterministic approximation (Laplace and variational inference) or a stochastic approximation (MCMC). In the Laplace and variational method, the posterior distribution of $\mathbf{w}$ ends up being approximated by a Gaussian distribution, although the mean and variance is different in each method. In essence, once $\mathbf{w}|\mathbf{y}$ is approximately normal, then estimation of the parameters θ using a direct optimisation approach or an EM-type approach is straightforward. On the other hand, MCMC approximates the density $p(\mathbf{w}|\mathbf{y})$ using samples generated via Gibbs sampling or HMC, and these samples would asymptotically be representative of draws from the true posterior.

Consider the data set... Plot the data. Explain priors for HMC and variational. Compare.

5.5 A variational algorithm

sec:iprobit var

We present a variational inference algorithm to estimate the I-probit latent variables $\mathbf{y}^*$ and $\mathbf{w}$, together with the parameters $\theta = \{\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top, \eta, \boldsymbol{\Psi}\}$. Begin by assuming some prior distribution on the parameters $p(\theta) = p(\boldsymbol{\alpha})p(\eta)p(\boldsymbol{\Psi})$. The following flat, uninformative priors are suggested:

- **Kernel parameters** η. This may include parameters such as the Hurst index, lengthscale and offset parameters, in addition to the RKHS scale parameters $\lambda_1, \dots, \lambda_p$, and each with their own support. For the scale parameters, assign each λ_k the vague prior

$$\lambda_k \overset{\text{iid}}{\sim} \mathrm{N}(0, v_\lambda = 0.001^{-1}), \ k = 1, \dots, p.$$

As $v_k^{-1} \to 0$, the prior becomes $p(\lambda_k) \propto \text{const.}$, an improper prior. The default choice for the rest of the kernel parameters is an improper prior $p(\eta) \propto \text{const.}$

- **Error precision** $\boldsymbol{\Psi}$. For the full I-probit model,

$$\boldsymbol{\Psi} \sim \mathrm{Wis}(\mathbf{G}, g)$$

with known scale matrix $\mathbf{G} = \mathrm{diag}(0.001, \dots, 0.001)$ and degrees of freedom $g = m + 1.001$. This implies a vague gamma prior for the precisions, and a near uniform prior for the off-diagonal elements (Alvarez et al., 2014). For the independent I-probit model such that $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \dots, \psi_m)$,

$$\psi_j \overset{\text{iid}}{\sim} \Gamma(s, r), \ j = 1, \dots, m,$$

with vague shape $s = 0.001$ and rate $r = 0.001$ parameters. Note that as $s, r \to 0$ then $p(\psi_j) \propto \psi_j^{-1}$, an improper Jeffreys' prior for scale parameters.

- **Intercepts** $\alpha_1, \dots, \alpha_m$. Also assign a vague normal prior for each intercept

$$\alpha_j \overset{\text{iid}}{\sim} \text{N}(0, v_\alpha = 0.001^{-1}).$$

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 that they are conjugate priors for the complete conditional density $p(\theta|\mathbf{y}, \mathbf{y}^*, \mathbf{w})$.

The posterior density of $\mathcal{Z} = \{\mathbf{y}^*, \mathbf{w}, \theta\}$ is approximated by a mean-field variational density q, i.e.

$$p(\mathbf{y}^*, \mathbf{w}, \theta|\mathbf{y}) = \tilde{q}(\mathbf{y}^*)q(\mathbf{w})q(\theta).$$

Additionally, we assume independence among the components of θ so that $q(\theta) = \prod_k q(\theta_k)$. We now present the mean-field variational distributions for each of unknowns in $\mathcal{Z}$. 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 $\text{E}_{\mathbf{w}\sim\tilde{q}}[\mathbf{w}]$, 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 $\mathbf{y}_{i\cdot}^* \in \mathbb{R}^m$, $i = 1, \dots, n$ of $\mathbf{y}^* \in \mathbb{R}^{n \times m}$ are independent can be exploited, which yields an induced factorisation $q(\mathbf{y}^*) = \prod_{i=1}^n q(\mathbf{y}_i^*)$. Define the set $\mathcal{C}_j = \{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}$. Then $q(\mathbf{y}_{i\cdot}^*)$ is the density of a multivariate normal distribution with mean $\tilde{\boldsymbol{\mu}}_{i\cdot} = \tilde{\boldsymbol{\alpha}} + \tilde{\mathbf{w}}^\top \tilde{\mathbf{h}}_\eta(x_i)$, and variance $\tilde{\boldsymbol{\Psi}}^{-1}$ subject to the truncation of its components to the set $\mathcal{C}_{y_i}$. That is, for each $i = 1, \dots, n$ and noting the observed value $y_i \in \{1, \dots, m\}$, the $\mathbf{y}_i^*$'s are distributed according to

$$\mathbf{y}_{i\cdot}^* \overset{\text{iid}}{\sim} \begin{cases} \text{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \tilde{\boldsymbol{\Psi}}^{-1}) & \text{if } y_{iy_i}^* > y_{ik}^*, \forall k \neq y_i \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

{eq:ystardist}

We denote this by $\mathbf{y}_{i\cdot}^* \overset{\text{iid}}{\sim} {}^{\text{t}}\text{N}(\tilde{\boldsymbol{\mu}}_{i\cdot}, \tilde{\boldsymbol{\Psi}}^{-1}, \mathcal{C}_{y_i})$, and the important properties of this distribution are explored in the appendix.

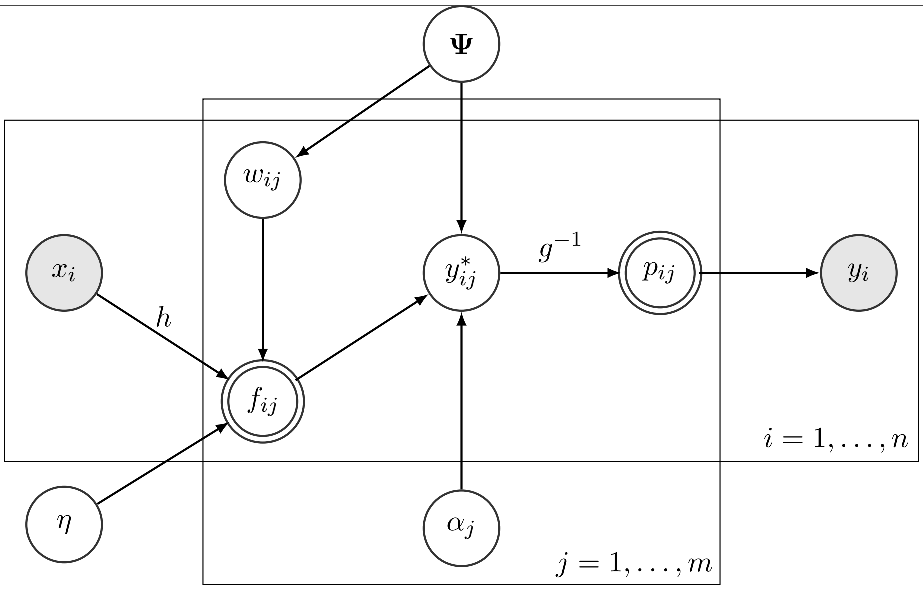

Figure 5.1: A DAG of the I-probit model. Observed nodes are shaded, while double-lined nodes represents calculable quantities.

The required expectations $\operatorname{E} \mathbf{y}^*_{i\cdot} = \operatorname{E}(y^*_{i1}, \dots, y^*_{im})^\top$ are tricky to compute. One strategy might be Monte Carlo integration: using samples from $\operatorname{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \tilde{\boldsymbol{\Psi}}^{-1})$, disregard those that do not satisfy the condition $y^*_{iy_i} > y^*_{ik}, \forall k \neq j$, and then take the sample average. This works reasonably well so long as the truncation region does not fall into the extreme tails of the multivariate normal. Alternatively, a fast, Gibbs based approach to estimating the mean or any other quantity $\operatorname{E} r(\mathbf{y}^*_{i\cdot})$ can be implemented, and this is detailed in the appendix.

If the independent I-probit model is considered, where the covariance matrix has the independent structure $\tilde{\boldsymbol{\Psi}} = \operatorname{diag}(\tilde\sigma_1^{-2}, \dots, \tilde\sigma_m^{-2})$, then the expected value can be considered component-wise, and each component of this expectation is given by

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

{eq:ystarupdate}

with

$$\phi_{ik}(Z) = \phi\left(\frac{\tilde\sigma_{y_i}}{\tilde\sigma_k}Z + \frac{\tilde\mu_{iy_i} - \tilde\alpha_k - \tilde f_{ik}}{\tilde\sigma_k}\right)$$
$$\Phi_{ik}(Z) = \Phi\left(\frac{\tilde\sigma_{y_i}}{\tilde\sigma_k}Z + \frac{\tilde\mu_{iy_i} - \tilde\alpha_k - \tilde f_{ik}}{\tilde\sigma_k}\right)$$
$$C_i = \int \prod_{l \neq j} \Phi_{il}(z)\phi(z)\,\mathrm{d}z$$

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 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 $\tilde{\mathbf{V}}_w$, we are able to efficiently compute the required inverse in roughly $O(n^3m)$ time—see the appendix for details.

If the independent I-probit model is assumed, i.e. $\tilde{\boldsymbol\Psi} = \operatorname{diag}(\tilde\psi_1,\dots,\tilde\psi_m)$, then the posterior covariance matrix $\tilde{\mathbf{V}}_w$ has a simpler structure: random matrix $\mathbf{w}$ will have columns which are independent of each other. By writing $\mathbf{w}_{\cdot j} = (w_{1j},\dots,w_{nj})^\top \in \mathbb{R}^n$, $j = 1,\dots,m$, to denote the column vectors of $\mathbf{w}$ and with a slight abuse of notation, we have that

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

where

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

The consequence of this is that the posterior regression functions are class independent, the exact intended effect by specifying a diagonal precision matrix $\boldsymbol{\Psi}$.

5.5.3 Kernel parameters η

The posterior density q involving the kernel parameters is of the form

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

where $p(\eta)$ is an appropriate prior density for η. Generally, samples $\eta^{(1)}, \dots, \eta^{(T)}$ from $\tilde q(\eta)$ may be obtained using a Metropolis algorithm, so that quantities such as $\tilde{\mathbf{H}}_\eta = \mathrm{E}_{\eta\sim q} \mathbf{H}_\eta$ and the like may be approximated using $\frac{1}{T}\sum_{t=1}^T \mathbf{H}_{\eta^{(t)}}$. Details of the Metropolis sampler is available in the appendix.

When only RKHS scale parameters are involved, then the distribution 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 $q(\eta)$, i.e., $q(\eta) = \prod_{k=1}^p p(\lambda_k)$. Then, by using independent and identical normal priors on the λ_k's, such as the one listed at the beginning of this section, we find that $q(\lambda_k)$ is the density of a normal distribution with mean $d_k c_k^{-1}$ and variance c_k^{-1}, where

$$c_k = \operatorname{tr}\left(\tilde{\boldsymbol{\Psi}} \, \mathrm{E}[\mathbf{w}^\top \mathbf{R}_k^2 \mathbf{w}] \right) + v_\lambda^{-2}$$

and

$$d_k = \operatorname{tr}\Big(\tilde{\boldsymbol{\Psi}} (\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top)^\top \mathbf{R}_k \tilde{\mathbf{w}} - \frac{1}{2} \tilde{\boldsymbol{\Psi}} \, \mathrm{E}[\mathbf{w}^\top \mathbf{U}_k \mathbf{w}] \Big).$$

For a method of evaluating quantities such as $\operatorname{tr}(\mathbf{C} \, \mathrm{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}])$ for suitably sized matrices $\mathbf{C}$ and $\mathbf{D}$, refer to the appendix.

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

$\mathbf{u} \sim \text{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \mathbf{\Psi}^{-1})$ a priori. From (5.5), the posterior for vec $\mathbf{u}$ is normal with mean vec $\tilde{\mathbf{u}} = \text{vec}(\tilde{\mathbf{w}}\tilde{\mathbf{\Psi}}^{-1})$ and variance

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

In essence, this reparameterisation simply introduces an additional step in the CAVI algorithm.

With a Wishart prior on the precision matrix $\mathbf{\Psi} \sim \text{Wis}_m(\mathbf{G}, g)$, the mean-field variational density for $\mathbf{\Psi}$ is found to satisfy

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

which is recognised as the log density of a Wishart distribution with scale matrix $\tilde{\mathbf{G}} := \mathbf{G}_1 + \mathbf{G}_2 + \mathbf{G}$ and $\tilde{g} = 2n + g$ degrees of freedom, where

$$\begin{gathered} \mathbf{G}_1 = \text{E}\left[(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top\right] \\ \mathbf{G}_2 = \text{E}[\mathbf{u}^\top\mathbf{u}]. \end{gathered} \tag{5.6}$$

The challenge here is that this distribution involves the second posterior moment of the conically truncated multivariate normal distribution for $\mathbf{y}^*$, among other things. This is slightly awkward to calculate analytically, although sampling methods provide a reasonable way out.

Consider now the independent I-probit model for which $\mathbf{\Psi} = \text{diag}(\psi_1, \dots, \psi_m)$ with independent gamma priors on the ψ_j's. The posterior for $\mathbf{\Psi}$ is of a similar factorised form, namely $q(\mathbf{\Psi}) = \prod_{j=1}^m q(\psi_j)$, where each $q(\psi_j)$ is the pdf of a gamma distribution with shape and rate parameters $\tilde{s} = 2n+s-1$ and $\tilde{r} = \frac{1}{2}\,\text{E}\|\mathbf{y}^*_{\cdot j} - \alpha_j\mathbf{1}_n - \mathbf{H}_\eta\mathbf{w}_{\cdot j}\|^2 + \frac{1}{2}\,\text{E}\|\mathbf{u}_{\cdot j}\|^2 + r$ respectively.

As a remark, the fact that both parameterisations of the I-prior random effects $\mathbf{w}$ and $\mathbf{u}$ are used seems suspect. Because of the way $\mathbf{u}$ was defined, there is a linear dependence between the two sets of parameters.

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.

5. Proof?

5.6 Post-estimation

5.7 Computational consideration

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

Suppose that, in a fully Bayesian setting, we append the unknown model parameters to the latent variables to form $\mathbf{z} = \{\mathbf{y}^*, \mathbf{w}, \theta\}$. The crux of variational inference is this: find a suitably close distribution function $q(\mathbf{z})$ that approximates the true posterior $p(\mathbf{z}|\mathbf{y})$, where closeness here is defined in the Kullback-Leibler divergence sense,

$$\mathrm{KL}(q\|p) = \int \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{y})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z}.$$

One may then show that log marginal density (the log of the intractable integral) holds the following bound:

$$\begin{aligned}
\log p(\mathbf{y}) &= \log p(\mathbf{y},\mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\
&= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \mathcal{L}(q) + \mathrm{KL}(q\|p) \\
&\geq \mathcal{L}(q) \qquad (5.7)
\end{aligned}$$

{eq:varbound}

since the KL divergence is a non-negative quantity. The functional $\mathcal{L}(q)$ given by

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \mathrm{E}_{\mathbf{z}\sim q}[\log p(\mathbf{y},\mathbf{z})] + H(q), \qquad (5.8)
\end{aligned}$$

{eq:elbo1}

where H is the entropy functional, is known as the *evidence lower bound* (ELBO), which serves as the proxy objective function in the likelihood maximisation problem. Evidently, the closer q is to the true p, the better, and this is achieved by maximising $\mathcal{L}$, or equivalently, minimising the KL divergence[2] from p to q. Note that the bound (5.7) achieves equality if and only if $q \equiv p$, but of course the true form of the posterior is unknown to us. Maximising $\mathcal{L}(q)$ or minimising $\mathrm{KL}(q\|p)$ with respect to the density q is a problem of calculus of variations, which incidentally, is where variational inference takes its name.

Maximising $\mathcal{L}$ over all possible density functions q is not possible without considering certain constraints. Two such constraints are described. The first, is to make a distributional assumption regarding q, for which it is parameterised by ν. For instance, we might choose the closest normal distribution to the posterior $p(\mathbf{z}|\mathbf{y})$ in terms of KL divergence. In this case, the task is to find optimal mean and variance parameters of a normal distribution.

[2] The astute reader will realise that $\mathrm{KL}(q\|p)$ is impossible to compute, since one does not know the true distribution $p(\mathbf{z}|\mathbf{y})$. Efforts are concentrated on maximising the ELBO instead.

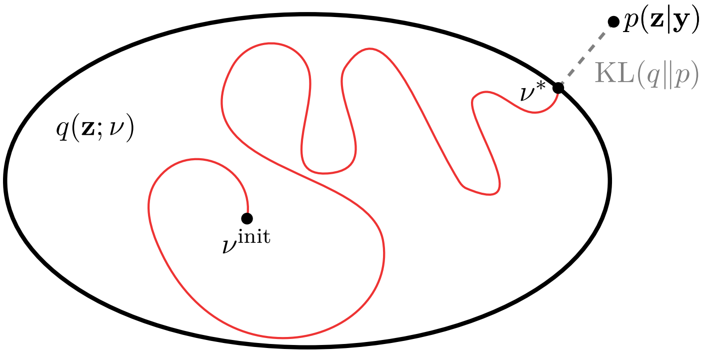

Figure 5.2: Schematic view of variational inference. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

The second type of constraint, and the one considered in this thesis, is simply an assumption that the approximate posterior q factorises into M disjoint factors. Supposing that the elements of $\mathbf{z}$ may indeed be partitioned into M disjoint groups $\mathbf{z} = (z^{(1)}, \dots, z^{(M)})$, then the structure

$$q(\mathbf{z}) = \prod_{k=1}^{M} q_k(z^{(k)})$$

for q is considered. This factorised form of variational inference is known in the statistical physics literature as the *mean-field theory* (Itzykson and Drouffe, 1991).

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

$$\log \tilde q(\xi) = \operatorname{E}_{-\xi}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)] + \text{const.} \tag{5.9}$$

{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.

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (5.9) 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 (5.9),

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

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.

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 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.

7. can use Hamiltonian Monte Carlo?

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

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

- ***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,*

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

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 $X \sim \mathrm{MN}_{n,m}(\mu, \Sigma, \Psi)$. 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

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

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.**

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

- **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 $a\le n$ and $B\in\mathbb{R}^{m\times b}$ be of full-rank $b\le m$. 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 $\mathrm{N}(\mu, \sigma^2)$ 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 $U \sim \mathrm{Unif}(0, 1)$ 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 $X \sim \mathrm{N}_d(\mu, \Sigma)$ to a convex subset[3] $\mathcal{A} \subset \mathbb{R}^d$. Call this distribution the truncated multivariate normal distribution, and denote it $X \sim {}^{\mathrm{t}}\mathrm{N}_d(\mu, \Sigma, \mathcal{A})$. The pdf is $p(X|\mu, \Sigma, \mathcal{A}) = C^{-1}\phi(X|\mu, \Sigma)\,\mathbb{1}[X \in \mathcal{A}]$, where

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

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

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 $\mathcal{A}$

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

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

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

[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.

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

$$\Sigma^{-1}_{-j,-j} = \Psi_{-j,-j} - \Psi_{j,-j}\Psi^\top_{-j,-j}/\Psi_{jj}$$

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 $X \sim {}^{\text{t}}\text{N}_d(\mu,\Sigma,\mathcal{C}_j)$, with $\mu = (\mu_1,\dots,\mu_d)^\top$ and $\Sigma = \text{diag}(\sigma_1^2,\dots,\sigma_d^2)$, 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

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

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

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

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

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

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

where we had defined

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

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

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

Proof. See **??** for the proof. □

5.11.5 Wishart distribution

Let $X \in \mathbb{R}^{m \times m}$ be a symmetric, positive-definite matrix. The Wishart distribution with scale matrix Ψ and $s > m - 1$ degrees of freedom is denoted $X \sim \operatorname{Wis}_m(\Psi, s)$, and its pdf, moments and entropy are

- **Pdf**.

$$p(X) = \frac{|X|^{(m-s-1)/2} e^{-\operatorname{tr}(\Psi^{-1}X)/2}}{2^{sm/2}|\Psi|^{s/2}\Gamma_m(s/2)}.$$

- **Moments**. $\operatorname{E} X = s\Psi$, $\operatorname{Var} X_{ij} = s(\Psi_{ij}^2 + \Psi_{ii}\Psi_{jj})$.
- **Entropy**.

$$H(p) = \frac{m+1}{2}\log|\Psi| + \frac{1}{2}m(m+1)\log 2 + \log\Gamma_p\left(\frac{s}{2}\right) - \frac{s-m-1}{2}\psi_m\left(\frac{s}{2}\right) + \frac{sm}{2}.$$

In the above, $\Gamma_m(\cdot)$ and $\psi_m(\cdot)$ are the multivariate gamma and digamma functions, respectively, given by

$$\Gamma_m(a) = \int_{U>0} \exp(-\operatorname{tr} U)|U|^{a-(m+1)/2}\,\mathrm{d}U$$

for positive-definite, real, $m \times m$ matrices U, and

$$\psi_m(a) = \frac{\partial}{\partial a}\log\Gamma_p(a).$$

5.11.6 Gamma distribution

For $X \in \mathbb{R}^+$, let X be distributed according to the gamma distribution with shape s and rate r, denoted $X \sim \Gamma(s, r)$. Then,

- **Pdf**. $p(X) = \Gamma(s)^{-1}r^s X^{s-1}e^{-rX}$.
- **Moments**. $\operatorname{E} X = s/r$, $\operatorname{Var} X_{ij} = s/r^2$.
- **Entropy**. $H(p) = s - \log r + \log\Gamma(s) + (1-s)\psi(s)$.

In the above, $\Gamma(\cdot) = \Gamma_1(\cdot)$ and $\psi(\cdot) = \psi_1(\cdot)$ are the gamma and digamma functions.

5.12 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

$$\log \tilde{q}(\xi) = \operatorname{E}_{-\xi}[\log p(\mathbf{y}, \mathcal{Z})] + \text{const.} \tag{5.9}$$

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.

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

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

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

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

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

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 $\mathbf{w} \sim \text{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$ with pdf

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

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

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

An improper prior $p(\eta) \propto \text{const.}$ 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 $p(\gamma) = \mathbb{1}\left(\gamma \in (0,1)\right)$ 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 $\boldsymbol{\Psi} \sim \text{Wis}_m(\mathbf{G}^{-1}, g)$, 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, $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_m)$, 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,

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

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

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

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 $\mathbf{\Psi}$, 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.

5.12.1 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

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

where $\tilde{\boldsymbol\mu}_i = \mathrm{E}\,\boldsymbol\alpha + (\mathrm{E}\,\mathbf{H}_\eta\,\mathrm{E}\,\mathbf{w})_i$, 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 $\tilde{\mathbf{\Psi}}$ 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.

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

The terms involving $\mathbf{w}$ in (5.9) 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

$$\begin{gathered}
\operatorname{vec}\mathbf{y}^*|\boldsymbol{\alpha},\mathbf{w},\eta,\boldsymbol{\Psi} \sim \operatorname{N}_{nm}\big(\operatorname{vec}(\mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}), \boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_n\big) \\
\text{and} \\
\operatorname{vec}\mathbf{w}|\boldsymbol{\Psi} \sim \operatorname{N}_{nm}(\mathbf{0}, \boldsymbol{\Psi}\otimes\mathbf{I}_n)
\end{gathered}$$

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,

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

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

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

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

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

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.

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

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

$$\begin{aligned}
\mathbf{V}_w &= \big(\operatorname{diag}(\tilde\sigma_1^{-2}, \dots, \tilde\sigma_m^{-2}) \otimes \tilde{\mathbf{H}}_\eta^2 + (\operatorname{diag}(\tilde\sigma_1^{2}, \dots, \tilde\sigma_m^{2}) \otimes \mathbf{I}_n \big)^{-1} \\
&= \operatorname{diag}\Big(\big(\tilde\sigma_1^{-2}\tilde{\mathbf{H}}_\eta^2 + \tilde\sigma_1^2 \mathbf{I}_n\big)^{-1}, \cdots, \big(\tilde\sigma_m^{-2}\tilde{\mathbf{H}}_\eta^2 + \tilde\sigma_m^2 \mathbf{I}_n\big)^{-1} \Big) \\
&=: \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \dots, \tilde{\mathbf{V}}_{w_m}).
\end{aligned}$$

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

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

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

$$\begin{aligned}
\operatorname{tr}(\mathbf{C}\operatorname{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) &= \sum_{i,j=1}^m \mathbf{C}_{ij} \operatorname{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]_{ij} \\
&= \sum_{i,j=1}^m \mathbf{C}_{ij} \operatorname{E}[\mathbf{w}_{\cdot i}^\top\mathbf{D}\mathbf{w}_{\cdot j}]. \qquad (5.10)
\end{aligned}$$

{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 (5.10) leads us to

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

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

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

5.12.3 Derivation of $\tilde q(\eta)$

By looking at only the terms involving η in (5.9), we deduce that $\tilde q$ for η satisfies

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

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

$$\begin{aligned}
\log\tilde q(\lambda_k) &= \mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[-\frac12 \operatorname{tr}\left((\mathbf y^* - \boldsymbol\mu)\boldsymbol\Psi(\mathbf y^* - \boldsymbol\mu)^\top\right)\right] - \frac{1}{2v_k^2}(\lambda_k - m_k)^2 + \text{const.} \\
&= -\frac12 \operatorname{tr}\mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[\boldsymbol\Psi\mathbf w^\top\mathbf H_\eta^2\mathbf w - 2\boldsymbol\Psi(\mathbf y^* - \mathbf 1\boldsymbol\alpha^\top)^\top\mathbf H_\eta\mathbf w\right] \\
&\ - \frac{1}{2v_k^2}(\lambda_k - m_k)^2 + \text{const.} \\
&= -\frac12 \operatorname{tr}\mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[\boldsymbol\Psi\mathbf w^\top(\lambda_k^2\mathbf R_k^2 + \lambda_k\mathbf U_k)\mathbf w - 2\boldsymbol\Psi(\mathbf y^* - \mathbf 1\boldsymbol\alpha^\top)^\top(\lambda_k\mathbf R_k)\mathbf w\right] \\
&\ - \frac{1}{2v_k^2}(\lambda_k^2 - 2m_k\lambda_k) + \text{const.} \\
&= -\frac12 \operatorname{tr}\mathrm{E}_{\mathcal Z\backslash\{\eta\}\sim q}\left[\lambda_k^2\boldsymbol\Psi\mathbf w^\top\mathbf R_k^2\mathbf w - 2\lambda_k\Big(\boldsymbol\Psi(\mathbf y^* - \mathbf 1\boldsymbol\alpha^\top)^\top\mathbf R_k\mathbf w - \frac12\boldsymbol\Psi\mathbf w^\top\mathbf U_k\mathbf w\Big)\right] \\
&\ - \frac12\left(\frac{1}{v_k^2}\lambda_k^2 - 2\frac{m_k}{v_k^2}\lambda_k\right) + \text{const.} \\
&= -\frac12\Big[\lambda_k^2\overbrace{\big(\operatorname{tr}(\tilde{\boldsymbol\Psi}\,\mathrm E[\mathbf w^\top\mathbf R_k^2\mathbf w]) + v_k^{-2}\big)}^{c_k} \\
&\qquad\qquad - 2\lambda_k\overbrace{\Big(\operatorname{tr}\Big(\tilde{\boldsymbol\Psi}(\tilde{\mathbf y}^* - \mathbf 1_n\tilde{\boldsymbol\alpha}^\top)^\top\mathbf R_k\tilde{\mathbf w} - \frac12\tilde{\boldsymbol\Psi}\,\mathrm E[\mathbf w^\top\mathbf U_k\mathbf w]\Big) + m_k v_k^{-2}\Big)}^{d_k}\Big]
\end{aligned}$$

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 $\lambda_k \mapsto \operatorname{E}_{\lambda_k \sim q}[\lambda_k]$ in the expression XXX. However, in the calculation of $\tilde{\mathbf{H}}_\eta^2$, we must replace $\lambda_k^2 \mapsto \operatorname{E}_{\lambda_k \sim q}[\lambda_k]^2 + \operatorname{Var}_{\lambda_k \sim q}[\lambda_k]$ in all occurrences of square terms. This can be cumbersome, so if felt necessary, use the approximation $\lambda_k^2 \mapsto \operatorname{E}_{\lambda_k \sim q}[\lambda_k]^2$ 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 $\lambda_k \propto$ 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

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

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

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

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

5.12.4 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 $\operatorname{vec}\mathbf{u} = (\operatorname{vec}\mathbf{w})^\top(\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)$, the optimal mean-field distribution for $\mathbf{u}$ is normal with mean $\operatorname{vec}\tilde{\mathbf{u}} = \operatorname{vec}(\tilde{\mathbf{w}}\tilde{\boldsymbol{\Psi}}^{-1})$ 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

$$\tilde{\mathbf{V}}_u = \operatorname{diag}(\tilde{\psi}_1^{-2}\tilde{\mathbf{V}}_{w_1}, \dots, \tilde{\psi}_m^{-2}\tilde{\mathbf{V}}_{w_m}).$$

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

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

which we recognise to be a Wishart distribution with scale matrix $(\mathbf{G} + \mathbf{G}_1 + \mathbf{G}_2)^{-1}$ and $\tilde g 2n + g$ 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 $\tilde{\mathbf{\Psi}} = (2n + g - m)(\mathbf{G} + \mathbf{G}_1 + \mathbf{G}_2)^{-1}$. 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 $\mathbf{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_m)$. Then, the variational density for each ψ_j

is found to be

$$\begin{aligned}\log \tilde q(\psi_j) &= \operatorname{E}_{\mathcal{Z}\backslash\{\boldsymbol\Psi\}\sim q}\left[\frac{n}{2}\log(\psi_1\cdots\psi_m) - \frac{1}{2}\sum_{j=1}^m\sum_{i=1}^n \psi_j(\mathbf{y}_{ij}^* - \boldsymbol\mu_{ij})^2\right] \\ &\quad + \operatorname{E}_{\mathcal{Z}\backslash\{\boldsymbol\Psi\}\sim q}\left[\frac{n}{2}\log(\psi_1\cdots\psi_m) - \frac{1}{2}\sum_{j=1}^m\sum_{i=1}^n \psi_j\mathbf{u}_{ij}^2\right] \\ &\quad + \sum_{j=1}^m \big((s_j-1)\log\psi_j - r_j\psi_j\big) + \text{const.} \\ &= (s_j+n-1)\log\psi_j - \psi_j\big(\frac{1}{2}\operatorname{E}\|\mathbf{y}_{\boldsymbol\cdot j}^* - \boldsymbol\mu_{\boldsymbol\cdot j}\|^2 + \frac{1}{2}\operatorname{E}\|\mathbf{u}_{\boldsymbol\cdot j}\|^2 + r_j\big) + \text{const.}\end{aligned}$$

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

5.12.5 Derivation of $\tilde q(\boldsymbol\alpha)$

The terms involving α_j in (5.9) are

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

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 $\operatorname{E}[\mathbf{H}_\eta\mathbf{w}] = \tilde{\mathbf{H}}_\eta\tilde{\mathbf{w}} \in \mathbb{R}^{n\times m}$.

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.

5.13 Deriving the ELBO expression

The evidence lower bound (ELBO) expression involves the following calculation:

$$
\begin{aligned}
\mathcal{L} &= \int \cdots \int q(\mathbf{y}^*, \mathbf{w}, \theta) \log \frac{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)}{q(\mathbf{y}^*, \mathbf{w}, \theta)} \, \mathrm{d}\mathbf{y}^* \, \mathrm{d}\mathbf{w} \, \mathrm{d}\theta \\
&= \mathrm{E} \log \overbrace{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)}^{\text{joint likelihood}} + \overbrace{\big(- \mathrm{E} \log q(\mathbf{y}^*, \mathbf{w}, \theta)\big)}^{\text{entropy}} \\
&= \mathrm{E} \Bigg[\cancel{\sum_{i=1}^n \sum_{j=1}^m \log p(y_i | y_{ij}^*)} + \sum_{i=1}^n \log p(\mathbf{y}_{i\cdot}^* | \boldsymbol{\alpha}, \mathbf{w}, \boldsymbol{\Psi}, \eta) + \log p(\mathbf{w}|\boldsymbol{\Psi}) + \log p(\boldsymbol{\Psi}) \\
&\qquad\qquad + \log p(\eta) + \log p(\boldsymbol{\alpha}) \Bigg] \\
&\quad + \sum_{i=1}^n H\big[q(\mathbf{y}_{i\cdot}^*)\big] + H\big[q(\mathbf{w})\big] + H\big[q(\boldsymbol{\Psi})\big] + H\big[q(\eta)\big] + H\big[q(\boldsymbol{\alpha})\big].
\end{aligned}
$$

Remark 5.3. As discussed, given the latent propensities $\mathbf{y}^*$, the pdf of $\mathbf{y}$ is degenerate and hence can be disregarded.

Remark 5.4. When using improper priors for the hyperparameters, i.e. $p(\boldsymbol{\Psi}, \eta, \boldsymbol{\alpha}) \propto$ const., then these terms can be disregarded.

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

$$
\begin{aligned}
&\sum_{i=1}^n \bigg(\mathrm{E} \log p(\mathbf{y}_{i\cdot}^* | \boldsymbol{\alpha}, \mathbf{w}, \boldsymbol{\Psi}, \eta) + H\big[q(\mathbf{y}_{i\cdot}^*)\big] \bigg) \\
&\qquad = -\frac{nm}{2} \log 2\pi + \frac{n}{2} \mathrm{E} \log |\boldsymbol{\Psi}| - \frac{1}{2} \mathrm{E} \sum_{i=1}^n (\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot})^\top \boldsymbol{\Psi} (\mathbf{y}_{i\cdot}^* - \boldsymbol{\mu}_{i\cdot}) \\
&\qquad\qquad + \frac{nm}{2} \log 2\pi - \frac{n}{2} \log |\tilde{\boldsymbol{\Psi}}| + \frac{1}{2} \mathrm{E} \sum_{i=1}^n (\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top \tilde{\boldsymbol{\Psi}} (\mathbf{y}_{i\cdot}^* - \tilde{\boldsymbol{\mu}}_{i\cdot}) + \log C_i \\
&\qquad = \text{const.} + \sum_{i=1}^n \log C_i
\end{aligned}
$$

where $\boldsymbol{\Omega}_i = \operatorname{Var} \mathbf{w}^\top \mathbf{h}_\eta(x_i)$, and C_i is the normalising constant for the distribution of multivariate truncated normal $\mathbf{y}_{i\cdot}$.

Notes:

1. $p(\mathbf{y}^*_{i\cdot})$ is the pdf of $\mathrm{N}(\boldsymbol{\mu}_{i\cdot}, \boldsymbol{\Psi}^{-1})$, and $q(\mathbf{y}^*_{i\cdot})$ is the pdf of ${}^{\mathrm{t}}\mathrm{N}(\tilde{\boldsymbol{\mu}}_{i\cdot}, \tilde{\boldsymbol{\Psi}}^{-1}, \mathcal{C}_{y_i})$, where $\boldsymbol{\mu}_{i\cdot} = \boldsymbol{\alpha} + \mathbf{w}^\top \mathbf{h}_\eta(x_i) \in \mathbb{R}^m$.
2. For $\boldsymbol{\Psi} \sim \mathrm{Wis}(\cdot,\cdot)$ with mean $\tilde{\boldsymbol{\Psi}}$, $\mathrm{E}\log|\boldsymbol{\Psi}| = \log\tilde{\boldsymbol{\Psi}} + \text{const.}$ (Bishop, 2006, §10.2).
3. It is simpler to use the approximation

$$\mathrm{E}(\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot})^\top \boldsymbol{\Psi} (\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot}) \approx \mathrm{E}(\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top \tilde{\boldsymbol{\Psi}} (\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot}). \tag{5.11}$$

{eq:elboyaprx}

rather than work out the actual quantity, which is

$$\mathrm{E}(\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot})^\top \boldsymbol{\Psi} (\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot}) = \mathrm{E}(\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top \tilde{\boldsymbol{\Psi}} (\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot}) + \mathrm{tr}(\tilde{\boldsymbol{\Psi}} \operatorname{Var} \boldsymbol{\mu}_{i\cdot}) \tag{5.12}$$

{eq:elboyact}

where $\operatorname{Var}\boldsymbol{\mu}_{i\cdot} = \operatorname{Var}\boldsymbol{\alpha} + \operatorname{Var}\mathbf{w}^\top\mathbf{h}_\eta(x_i)$, obtained by taking expectations with respect to everything except $\mathbf{y}^*_{i\cdot}$. The first term is a diagonal matrix of the posterior variances of the intercepts. The second term is where things get complicated Let $\boldsymbol{\Omega}_i = \operatorname{Var}\mathbf{w}^\top\mathbf{h}_\eta(x_i)$. Then $\boldsymbol{\Omega}_{i,kj} \approx \operatorname{Cov}(\mathbf{w}_{\cdot k}^\top \mathbf{h}_\eta(x_i), \mathbf{w}_{\cdot j}^\top \mathbf{h}_\eta(x_i)) = \mathbf{h}_\eta(x_i)^\top \tilde{\mathbf{V}}_w[k,j] \mathbf{h}_\eta(x_i)$. So

$$\mathrm{tr}(\tilde{\boldsymbol{\Psi}}\boldsymbol{\Omega}_i) \approx \sum_{k,j=1}^m \tilde{\boldsymbol{\Psi}}_{kj} \mathbf{h}_\eta(x_i)^\top \tilde{\mathbf{V}}_w[k,j] \mathbf{h}_\eta(x_i)$$

However, we know that $\operatorname{Var}XY = \mathrm{E}\,X^2Y^2 - (\mathrm{E}\,XY)^2 = \operatorname{Var}X\operatorname{Var}Y + \operatorname{Var}X(\mathrm{E}\,Y)^2 + \operatorname{Var}Y(\mathrm{E}\,X)^2$, so there is actually some covariance terms which need to be considered, and these are not so easily computed. In practice, we find that using (5.11) gives satisfactory results as far as determining convergence for the variational algorithm goes.

5.13.2 Terms involving distributions of w

$$\begin{aligned}
\mathrm{E}\log p(\mathbf{w}|\boldsymbol{\Psi}) + H\big[q(\mathbf{w})\big] &= \cancel{-\frac{nm}{2}\log 2\pi} - \frac{n}{2}\mathrm{E}\log|\boldsymbol{\Psi}| - \frac{1}{2}\mathrm{E}\,\mathrm{tr}\left(\mathbf{w}\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\right) \\
&\quad + \frac{nm}{2}(1 + \cancel{\log 2\pi}) + \frac{1}{2}\log|\tilde{\mathbf{V}}_w| \\
&= \text{const.} - \frac{n}{2}\log\tilde{\boldsymbol{\Psi}} - \frac{1}{2}\sum_{j=1}^m \mathrm{tr}\left(\tilde{\boldsymbol{\Psi}}^{-1}(\tilde{\mathbf{V}}_w[j,j] + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top)\right)
\end{aligned}$$

Notes:

1. $p(\mathbf{w})$ is the pdf of $\text{MN}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$, and $q(\mathbf{w})$ is the pdf of $\text{N}(\operatorname{vec} \tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$.
2. We used the first order approximation $\operatorname{E} \boldsymbol{\Psi}^{-1} \approx (\operatorname{E} \boldsymbol{\Psi})^{-1} = \tilde{\boldsymbol{\Psi}}^{-1}$.
3. $\tilde{\mathbf{V}}_w[j,j]$ are the $n \times n$ sub matrices along the diagonal of $\tilde{\mathbf{V}}_w$.

5.13.3 Terms involving distributions of η

If no closed-form expression for $q(\eta)$ is found, then the expression $\operatorname{E}[\log p(\eta) - q(\eta)]$ must be obtained by sampling methods. Otherwise, consider the case where $\eta = \{\lambda_1, \dots, \lambda_p\}$. Then, the contribution to the ELBO is

$$\begin{aligned}
&\operatorname{E} \log p(\lambda_1, \dots, \lambda_p) + H\big[q(\lambda_1, \dots, \lambda_p)\big] \\
&\qquad = -\frac{p}{2} \log 2\pi - \frac{1}{2} \log v_1 \cdots v_k - \frac{1}{2} \sum_{k=1}^{p} \frac{\operatorname{E}(\lambda_k - m_k)^2}{v_k} \\
&\qquad\quad + \frac{p}{2}(1 + \log 2\pi) + \frac{1}{2} \log \tilde{v}_1 \cdots \tilde{v}_p \\
&\qquad = \text{const.} + \frac{1}{2} \sum_{k=1}^{p} \log \tilde{v}_k - \frac{1}{2} \sum_{k=1}^{p} \frac{\tilde{v}_k + \tilde{\lambda}_k^2 - 2\tilde{\lambda}_k m_k}{v_k}
\end{aligned}$$

Notes:

1. The priors on the λ_k's are $\text{N}(m_k, v_k)$, and $q(\lambda_k)$ is the density of $\text{N}(\tilde{\lambda}_k, v_{\lambda_k})$.
2. When using improper priors $\lambda_k \propto \text{const.}$, then we need only consider the middle term involving the sums of $\log \tilde{v}_{\lambda_k}$.

5.13.4 Terms involving distributions of $\boldsymbol{\Psi}$

The terms involving $\boldsymbol{\Psi}$ are $\operatorname{E} \log p(\boldsymbol{\Psi}) + H\big[q(\boldsymbol{\Psi})\big]$. In the case of the full I-probit model, this becomes

$$\begin{aligned}
&\text{const.} + \frac{g - m - 1}{2} \operatorname{E} \log|\boldsymbol{\Psi}| - \frac{1}{2} \operatorname{E} \operatorname{tr}(\mathbf{G}\boldsymbol{\Psi}) + \log B - \frac{\tilde{g} - m - 1}{2} \operatorname{E} \log|\boldsymbol{\Psi}| + \frac{\tilde{g}m}{2} \\
&\qquad = \text{const.} + \log B + \frac{\tilde{g}m}{2} + \frac{g - \tilde{g}}{2} \log|\tilde{\boldsymbol{\Psi}}| - \frac{1}{2} \operatorname{tr}(\mathbf{G}\tilde{\boldsymbol{\Psi}})
\end{aligned}$$

where $B = \frac{\tilde g}{2}\log|\tilde G| + \frac{\tilde g m}{2}\log 2 + \log\Gamma_m(\tilde g/2)$. In the case of the independent I-probit model, we have

$$\begin{aligned}
\text{const.} + \sum_{j=1}^m \Big\{ (s_j - 1)\operatorname{E}\log\psi_j - r_j\operatorname{E}\psi_j \Big\} - \sum_{j=1}^m \log\tilde r_j \\
= \text{const.} + \sum_{j=1}^m \Big\{ (s_j - 1)\log\tilde\psi_j - r_j\tilde\psi_j \Big\} - \sum_{j=1}^m \log\tilde r_j
\end{aligned}$$

Notes:

1. The priors on the $\boldsymbol\Psi$ is $\operatorname{Wis}(\mathbf G, g)$, or if $\boldsymbol\Psi = \operatorname{diag}(\psi_1,\dots,\psi_m)$, then each $\psi_j \sim \Gamma(s_j, r_j)$. $q(\boldsymbol\Psi)$ is the density of $\operatorname{Wis}(\tilde{\mathbf G}, \tilde g)$ or in the case of the independent model, each $q(\psi_j)$ is the density of $\Gamma(s_j, \tilde r_j)$.
2. Use the first order Taylor expansion about $\operatorname{E}\psi_j$ to approximate $\operatorname{E}\log\psi_j \approx \log\operatorname{E}\psi_j = \log\tilde\psi_j$, as per Teh et al. (2007).

5.13.5 Terms involving distribution of $\boldsymbol\alpha$

For the intercepts, consider only

$$\begin{aligned}
\operatorname{E}\log p(\boldsymbol\alpha) + H\big[q(\boldsymbol\alpha)\big] &= \text{const.} - \frac{1}{2}\operatorname{E}\sum_{j=1}^m \frac{(\alpha_j - a_j)^2}{A_j} + \frac{1}{2}\log\tilde v_{\alpha_1}\cdots\tilde v_{\alpha_m} \\
&= \text{const.} + \frac{1}{2}\sum_{j=1}^m \log\tilde v_{\alpha_j} - \frac{1}{2}\sum_{j=1}^m \frac{v_{\alpha_j} + \tilde\alpha_j^2 - 2a_j\tilde\alpha_j}{A_j}
\end{aligned}$$

Notes:

1. $p(\boldsymbol\alpha)$ is $\prod_{j=1}^m \phi(\alpha_j|a_j, A_j)$, and $q(\boldsymbol\alpha)$ $\prod_{j=1}^m \phi(\alpha_j|\tilde\alpha_j, \tilde v_{\alpha_j})$.

5.13.6 ELBO summarised

In the example section of Chapter 5, we considered only 1) the independent I-probit model; 2) fixed $\boldsymbol\Sigma = \mathbf I_m$; 3) only RKHS scale parameters to estimate; and 4) and improper priors on the hyperparameters. In such situations, the ELBO expression is simply

$$\mathcal{L} = \text{const.} + \sum_{i=1}^{n} \log C_i - \frac{1}{2} \sum_{j=1}^{m} \operatorname{tr}\left(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j} \tilde{\mathbf{w}}_{\cdot j}^\top \right) + \frac{1}{2} \sum_{k=1}^{p} \log \tilde{v}_k.$$

As a final remark, often times the ELBO is treated as a proxy for the (penalised) marginal likelihood of the model, in which case it must be noted that the ELBO as we had derived is correct up to a constant. We find that keeping track of the constants is slightly tedious, and hence decided not to do so. When comparing ELBOs of two or more models, the comparison is still valid as only differences between the ELBOs matter, in which case the constants would cancel out.

Bibliography

albert1993bayesian Albert, James H and Siddhartha Chib (1993). “Bayesian analysis of binary and polychotomous response data”. In: *Journal of the American statistical Association* 88.422, pp. 669–679.

alvarez2014bayesian Alvarez, Ignacio, Jarad Niemi, and Matt Simpson (2014). “Bayesian inference for a covariance matrix”. In: *arXiv preprint arXiv:1408.4050*.

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

blei2017variational Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). “Variational inference: A review for statisticians”. In: *Journal of the American Statistical Association* just-accepted.

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.

itzykson1991statistical Itzykson, Claude and Jean Michel Drouffe (1991). *Statistical Field Theory: Volume 2, Strong Coupling, Monte Carlo Methods, Conformal Field Theory and Random Systems.* Cambridge University Press.

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

kass1995bayes Kass, Robert E and Adrian E Raftery (1995). "Bayes factors". In: *Journal of the american statistical association* 90.430, pp. 773–795.

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.

teh2007collapsed Teh, Yee W, David Newman, and Max Welling (2007). “A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation”. In: *Advances in neural information processing systems*, pp. 1353–1360.

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