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 an exponential family distribution, and in this case, the categorical distribution. We denote this by

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

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

with the class probabilities satisfying $p_{ij} \geq 0, \forall j = 1, \ldots, m$ 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 $j \in \{1, \ldots, m\}$ 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 $j, k \in \{1, \ldots, m\}$ 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

The I-prior methodology can be used naïvely to fit a categorical regression model. Suppose, as before, we observe data $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ where each $x_i \in \mathcal{X}$, for $i = 1, \ldots, n$. Here, the responses are categorical $y_i \in \{1, \ldots, m\}$, and additionally, write $y_i = (y_{i1}, \ldots, y_{im}) =: \mathcal{M}$ where the class responses $y_{ij} = 1$ if individual i's response category is $y_i = j$, and 0 otherwise. In other words, there is exactly one '1' at the j'th position in the vector $y_i = (y_{i1}, \ldots, y_{im})$, zeroes everywhere else. For $j = 1, \ldots, m$, we model

$$\begin{aligned} y_{ij} &= \alpha + \alpha_j + f_j(x_i) + \epsilon_{ij} \\ (\epsilon_{i1}, \ldots, \epsilon_{im})^\top &\overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}). \end{aligned} \tag{5.2}$$

{eq:naivecl assmod}

The idea here being that we attempt to model the class responses y_{ij} using class-specific regression functions f_j, and the class responses are assumed to be independent among individuals, but may or may not be correlated among classes for each individual. The class correlations are manifest themselves in the variance of the errors $\boldsymbol{\Psi}^{-1}$, which is an $m \times m$ matrix.

Denote the regression function f in (5.2) on the set $\mathcal{X} \times \mathcal{M}$ as $f(x_i, j) = \alpha_j + f_j(x_i)$. This regression function can be seen as an ANOVA decomposition of the spaces $\mathcal{F}_\mathcal{M}$ and $\mathcal{F}_\mathcal{X}$ of functions over $\mathcal{M}$ and $\mathcal{X}$ respectively. That is, $\mathcal{F} = \mathcal{F}_\mathcal{M} \oplus (\mathcal{F}_\mathcal{M} \otimes \mathcal{F}_\mathcal{X})$ is a decomposition into the main effects of 'class', and an interaction effect of the covariates for each class. Let $\mathcal{F}_\mathcal{M}$ and $\mathcal{F}_\mathcal{X}$ be RKHSs respectively with kernels $a : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ and $b : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Then, the ANOVA RKKS $\mathcal{F}$ possesses the reproducing kernel $h : (\mathcal{X} \times \mathcal{M})^2 \to \mathbb{R}$ as defined by

$$b_\eta\big((x, j), (x', j')\big) = a(j, j') + a(j, j')h_\eta(x, x'). \tag{5.3}$$

{eq:anovacl ass}

The kernel h_η may be any of the kernels described in this thesis, ranging from the linear kernel, to the fBm kernel, or even an ANOVA kernel. Choices for $a : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ include

1. **The Pearson kernel** (as defined in Definition 2.34). With $J \sim \mathrm{P}$, a probability measure over $\mathcal{M}$,
$$a(j, j') = \frac{\delta_{jj'}}{\mathrm{P}(J = j)} - 1.$$

2. **The identity kernel**. With δ denoting the Kronecker delta function,

$$a(j, j') = \delta_{jj'}.$$

The purpose of either of these kernels is to contribute to the class intercepts α_j, and to associate a regression function in each class. We have a slight preference for the identity kernel, which lends itself as being easy to handle computationally. The only difference between the two is the inverse probability weighting per class that is applied in the Pearson kernel, but not in the identity kernel.

As a remark, the functions in $\mathcal{F}_{\mathcal{M}}$ and $\mathcal{F}_{\mathcal{X}}$ need necessarily be zero-mean functions (as per the functional ANOVA definition in Definition 2.37). What this means is that $\sum_{j=1}^m \alpha_j = 0$, $\sum_{j=1}^m f_j(x_i) = 0$, and $\sum_{i=1}^n f_j(x_i) = 0$. In particular,

$$\begin{aligned}\sum_{j=1}^m y_{ij} &= \sum_{j=1}^m (\alpha + \alpha_j + f_j(x_i)) \\ &= m\alpha + \cancelto{0}{\sum_{j=1}^m \alpha_j} + \cancelto{0}{\sum_{j=1}^m f_j(x_i)}\end{aligned}$$

and since $\sum_{j=1}^m y_{ij} = 1$, we have that $\alpha = 1/m$ and can thus be fixed to resolve identification. The Pearson RKHS will contain zero mean functions, but the RKHS of constant functions induced by the identity kernel may not. If this is the case, then it should be ensured that $\sum_{j=1}^m \alpha_j = 0$ in other ways; perhaps during the estimation process.

With $f \in \mathcal{F}$ the RKKS with kernel h_η, it is straightforward to assign an I-prior on f. It is in fact

$$\begin{aligned} f(x_i, j) &= \sum_{j'=1}^m \sum_{i'=1}^n a(j, j')\big(1 + h_\eta(x_i, x_{i'})\big) w_{i'j'} \\ (w_{i'1}, \ldots, w_{i'm})^\top &\sim \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}) \end{aligned} \tag{5.4}$$

{eq:naiveclassiprior}

assuming a zero prior mean $f_0(x, j) = 0$. It is much convenient to work in vector and matrix form, so let us introduce some notation. Let $\mathbf{w}$ (c.f. $\mathbf{y}$, $\mathbf{f}$ and $\boldsymbol{\epsilon}$) be an $n \times m$ matrix whose (i, j) entries contain w_{ij} (c.f. y_{ij}, $f(x_i, j)$, and ϵ_{ij}). The row-wise entries of $\mathbf{w}$ are independent of each other (independence assumption of the n observations), while any two of their columns have covariance as specified in $\boldsymbol{\Psi}$. This means that $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$ which implies $\operatorname{vec} \mathbf{w} \sim \mathrm{N}_{nm}(\mathbf{0}, \boldsymbol{\Psi} \otimes \mathbf{I}_n)$, and similarly,

$\boldsymbol{\epsilon} \sim \mathrm{N}_{nm}(\mathbf{0}, \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)$. Denote by $\mathbf{H}_\eta$ the $n \times n$ kernel matrix with entries supplied by $1 + h_\eta$, and $\mathbf{A}$ the $m \times m$ matrix with entries supplied by a. From (5.4), we have that

$$\mathbf{f} = \mathbf{H}_\eta \mathbf{w} \mathbf{A} \in \mathbb{R}^{n \times m},$$

and thus $\operatorname{vec} \mathbf{f} \sim \mathrm{N}_{nm}(\mathbf{0}, \mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{H}_\eta^2)$. As $\mathbf{y} = \boldsymbol{\alpha} + \mathbf{f} + \boldsymbol{\epsilon}$, where $\boldsymbol{\alpha} \in \mathbb{R}^{n \times m}$ with (i, j) entries given by $\alpha + \alpha_j = \alpha_j + 1/m$, by linearity we have that

$$\operatorname{vec} \mathbf{y} \sim \mathrm{N}_{nm}\left(\operatorname{vec} \boldsymbol{\alpha}, (\mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \right) \tag{5.5}$$

and

$$\operatorname{vec} \mathbf{y} | \operatorname{vec} \mathbf{w} \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\boldsymbol{\alpha} + \mathbf{H}_\eta \mathbf{w} \mathbf{A}), (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \right). \tag{5.6}$$

which can then be estimated using the methods described in Chapter 4.

When using the identity kernel in conjunction with an assumption of iid errors ($\boldsymbol{\Psi} = \psi \mathbf{I}_n$), the above distributions simplify further. Specifically, the variance in the marginal distribution becomes

$$\begin{aligned} \operatorname{Var}(\operatorname{vec} \mathbf{y}) &= (\psi \mathbf{I}_m \otimes \mathbf{H}_\eta^2) + (\psi^{-1} \mathbf{I}_m \otimes \mathbf{I}_n) \\ &= (\mathbf{I}_m \otimes \psi \mathbf{H}_\eta^2) + (\mathbf{I}_m \otimes \psi^{-1} \mathbf{I}_n) \\ &= \mathbf{I}_m \otimes \overbrace{(\psi \mathbf{H}_\eta^2 + \psi^{-1} \mathbf{I}_n)}^{\mathbf{V}_y}. \end{aligned}$$

which implies independence and identical variances $\mathbf{V}_y$ for the vectors $(y_{1j}, \dots, y_{nj})^\top$ for each class $j = 1, \dots, m$. Evidently, this stems from the implied independence structure of the prior on f too, since now $\operatorname{Var}(\operatorname{vec} \mathbf{f}) = \operatorname{diag}(\psi \mathbf{H}_\eta^2, \dots, \psi \mathbf{H}_\eta^2)$, which could be interpreted as having independent and identical I-priors on the regression functions for each class $\mathbf{f}_j = \big(f(x_1, j), \dots, f(x_n, j)\big)^\top$.

There are several downfalls to using the model described above. Unlike in the case of continuous response variables, the normal I-prior model is highly inappropriate for categorical responses. For one, it violates the normality and homoscedasticity assumptions of the errors. For another, predicted values may be out of the range $[0, m]$ and thus poorly calibrated. Furthermore, it would be more suitable if the class probabilities—the probability of an observation belonging to a particular class—were also part of the model. In the next section, we propose an improvement to this naïve I-prior classification model by considering a probit-like transformation of the regression functions.

5.2 A latent variable motivation: the I-probit model

It is convenient, as we did in the previous subsection, to again think of the responses $y_i \in \{1, \dots, m\} = \mathcal{M}$ as comprising of a binary vector $(y_{i1}, \dots, y_{im})$, with a single '1' at the position corresponding to the value that y_i takes. In this formulation, each y_{ij} is distributed as Bernoulli with probability p_{ij}. Now, assume that, for each $y_i = (y_{i1}, \dots, y_{im})$, there exists corresponding *continuous, underlying, latent variables* $y^*_{i1}, \dots, y^*_{im}$ such that

$$y_i = \begin{cases} 1 & \text{if } y^*_{i1} \geq y^*_{i2}, y^*_{i3}, \dots, y^*_{im} \\ 2 & \text{if } y^*_{i2} \geq y^*_{i1}, y^*_{i3}, \dots, y^*_{im} \\ \vdots & \\ m & \text{if } y^*_{im} \geq y^*_{i2}, y^*_{i3}, \dots, y^*_{i\,m-1}. \end{cases} \tag{5.7}$$

{eq:latentmodel}

In other words, $y_{ij} = \arg\max_{k=1}^{m} y^*_{ik}$. Such a formulation is common in economic choice models, and is rationalised by a utility-maximisation argument: an agent faced with a choice from a set of alternatives will choose the one which benefits them most.

Instead of modelling the observed y_{ij}'s directly, we model instead the n latent variables in each class $j = 1, \dots, m$ according to the regression problem

$$\begin{aligned} y^*_{ij} &= \alpha_j + f_j(x_i) + \epsilon_{ij} \\ \boldsymbol{\epsilon}_i = (\epsilon_{i1}, \dots, \epsilon_{im})^\top &\overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}). \end{aligned} \tag{5.8}$$

{eq:multinomial-latent}

We can see some semblance of this model with the one in (5.4), and ultimately the aim is to assign I-priors to the regression function of these latent variables, and we will describe this shortly. For now, realise that each $\mathbf{y}^*_i := (y^*_{i1}, \dots, y^*_{im})^\top$ has the distribution $\mathrm{N}_m(\boldsymbol{\alpha} + \mathbf{f}(x_i), \boldsymbol{\Psi}^{-1})$, conditional on the data x_i, the intercepts $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_m)^\top$, the evaluations of the functions at x_i for each class $\mathbf{f}(x_i) = \big(f_1(x_i), \dots, f_m(x_i)\big)^\top$, and the error covariance matrix $\boldsymbol{\Psi}^{-1}$.

The probability of belonging to class j for observation i, i.e. p_{ij}, is calculated as

$$\begin{aligned} p_{ij} &= \mathrm{P}(y_i = j) \\ &= \mathrm{P}\left(\{y^*_{ij} > y^*_{ik} \,|\, \forall k \neq j\}\right) \\ &= \int_{\{y^*_{ij} > y^*_{ik} \,|\, \forall k \neq j\}} \phi(\mathbf{y}^*_i | \boldsymbol{\alpha} + \mathbf{f}(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}\mathbf{y}^*, \end{aligned} \tag{5.9}$$

{eq:pij}

where $\phi(\cdot|\mu, \Sigma)$ is the density of the multivariate normal with mean μ and variance Σ. This is the probability that the normal random variable $\mathbf{y}_i^*$ belongs to the set $\{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}$, which are cones in $\mathbb{R}^m$. Since the union of these cones is the entire m-dimensional space of reals, the probabilities add up to one and hence they represent a proper probability mass function of the classes. While this does not have a closed-form expression and highlights one of the difficulties of working with probit models, the integral is by no means impossible to compute—see Section 5.9.1 for a note regarding this matter.

Note that the dimension of the integral (5.9) is $m-1$, since the j'th coordinates is fixed relative to the others. Alternatively, we could have specified the model in terms of *relative differences* of the latent variables. Choosing the first category as the reference category, define new random variables $z_{ij} = y_{ij}^* - y_{i1}^*$, for $j = 2, \ldots, m-1$. The model (5.7) is equivalently represented by

$$y_i = \begin{cases} 1 & \text{if } \max(z_{i2}, \ldots, z_{im}) < 0 \\ j & \text{if } \max(z_{i2}, \ldots, z_{im}) = z_{ij} \geq 0. \end{cases} \tag{5.10}$$

Write $\mathbf{z}_i = (z_{i2}, \ldots, z_{im})^\top \in \mathbb{R}^{m-1}$. Then $\mathbf{z}_i = \mathbf{Q}\mathbf{y}_i^*$, where $\mathbf{Q} \in \mathbb{R}^{(m-1)\times m}$ is the $(m-1)$ identity matrix pre-augmented with a column vector of minus ones. We have that $\mathbf{z}_i \overset{\text{iid}}{\sim} \mathrm{N}_{m-1}\left(\mathbf{Q}(\boldsymbol{\alpha} + \mathbf{f}(x_i)), \mathbf{Q}\boldsymbol{\Psi}^{-1}\mathbf{Q}^\top\right)$. Thus, the class probabilities for $j = 2, \ldots, m$ are

$$p_{ij} = \int\limits_{\{z_{ik}<0\,|\,\forall k\neq j\}} \mathbb{1}(z_{ij} \geq 0)\, \phi(\mathbf{z}_i)\, \mathrm{d}\mathbf{z}_i, \tag{5.11}$$

{eq:pij2}

with $\phi(\mathbf{z}_i)$ representing the $(m-1)$-variate normal density for $\mathbf{z}_i$. The class probability p_{i1} is simply

$$p_{i1} = \int\limits_{\{z_{ik}<0\}} \phi(\mathbf{z}_i)\, \mathrm{d}\mathbf{z}_i = 1 - \sum_{k\neq 1} p_{ik}.$$

From this representation of the model, with $m = 2$ (binary outcomes) we see that

$$p_{i1} = \Phi\left(\frac{z_{i2} - \mu}{\sigma}\right) \quad \text{and} \quad p_{i2} = 1 - \Phi\left(\frac{z_{i2} - \mu}{\sigma}\right),$$

where $\Phi(\cdot)$ is the CDF of the standard normal univariate distribution, and μ and σ are the mean and standard deviation of the random variable z_{i2}.

Now we'll see how to specify an I-prior on the regression problem (5.8). In the naïve I-prior model, we wrote $f(x_i, j) = \alpha_j + f_j(x_i)$, and specified for f to belong to an ANOVA RKKS with kernel defined in (5.3). Instead of doing the same, we take a different approach. Treat the α_j's in (5.8) as intercept parameters to estimate with the additional requirement that $\sum_{j=1}^m \alpha_j = 0$. Further, let $\mathcal{F}$ be a (centred) RKHS/RKKS of functions over $\mathcal{X}$ with reproducing kernel h_η. Now, consider putting an I-prior on the regression functions $f_j \in \mathcal{F}$, $j = 1 \dots, m$, defined by

$$f_j(x_i) = \sum_{k=1}^{n} h_\eta(x_i, x_k) w_{ik}$$

with $\mathbf{w}_i := (w_{i1}, \dots, w_{im})^\top \overset{\text{iid}}{\sim} \text{N}(0, \boldsymbol{\Psi})$. This is similar to the naïve I-prior specification (5.4), except that the intercept have been treated as parameters rather than accounting for them using an RKHS of constant functions. In particular, the overall regression relationship still satisfies the ANOVA functional decomposition. We find that this method bodes well down the line computationally.

We call the multinomial probit regression model of (5.7) subject to (5.8) and I-priors on $f_j \in \mathcal{F}$, the *I-probit model*. For completeness, this is stated again: for $i = 1, \dots, n$, $y_i = \arg\max_{k=1}^m y_{ik}^* \in \{1, \dots, m\}$, where, for $j = 1, \dots, m$,

$$\begin{aligned} y_{ij}^* &= \alpha_j + \overbrace{\sum_{k=1}^{n} h_\eta(x_i, x_k) w_{ik}}^{f_j(x_i)} + \epsilon_{ij} \\ \boldsymbol{\epsilon}_i &= (\epsilon_{i1}, \dots, \epsilon_{im})^\top \overset{\text{iid}}{\sim} \text{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\ \mathbf{w}_i &:= (w_{i1}, \dots, w_{im})^\top \overset{\text{iid}}{\sim} \text{N}(0, \boldsymbol{\Psi}). \end{aligned} \tag{5.12}$$

The parameters of the I-probit model are denoted by $\theta = \{\alpha_1, \dots, \alpha_m, \eta, \boldsymbol{\Psi}\}$. Let $\mathbf{y}^* \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries y_{ij}^*. Using the results in Chapter 4, the marginal distribution of the latent variables is

$$\text{vec}\, \mathbf{y}^* \sim \text{N}_{nm}\left(\boldsymbol{\alpha}, (\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right).$$

5.2.1 IIA

In decision theory, the independence axiom states that an agent's choice between a set of alternatives should not be affected by the introduction or elimination of a (new) choice option. The probit model is suitable for modelling multinomial data where the independence axiom, which is also known as the *independence of irrelevant alternatives* (IIA) assumption, is not desired. Such cases arise frequently in economics and social science, and the famous Red-Bus-Blue-Bus example is often used to illustrate IIA. Suppose commuters face the decision between taking cars and red busses. The addition of blue busses to commuters' choice should in theory be more likely chosen by those who prefer taking the bus over cars. That is, assuming commuters are indifferent about the colour of the bus, commuters who are predisposed to taking the red bus would see the blue bus as an identical alternative. Yet, if IIA is imposed, then the three choices are distinct, and the fact that red and blue busses are substitutable is ignored.

In the I-probit model, the choice dependency is controlled by the error precision matrix $\boldsymbol{\Psi}$. Specifically, the off-diagonal elements $\boldsymbol{\Psi}_{jk}$ capture the correlation between choices j and k. Allowing all $m(m+1)/2$ covariance elements of $\boldsymbol{\Psi}$ leads to the *full I-probit model*, and would not assume an IIA position.

While it is an advantage to be able to model the correlations across choices (unlike in logistic models), it would be a major simplification algorithmically to consider all covariances in $\boldsymbol{\Psi}$ to be zero. This would trigger the IIA assumption in the I-probit model. There are applications where the IIA assumption would not adversely affect the analysis, such as when all the choices are mutually exclusive and exhaustive. In these situations, it would be beneficial to reduce the I-probit model to a simpler version by assuming $\boldsymbol{\Psi} = \text{diag}(\psi_1, \ldots, \psi_m)$.

The independence assumption causes the distribution of the latent variables to be $y_{ij}^* \sim \text{N}(\alpha_j + f_j(x_i), \sigma_j^2)$ for $j = 1, \ldots, m$. As a continuation of line (5.9), we can show

the class probability p_{ij} to be

$$\begin{aligned} p_{ij} &= \int_{\{y_{ij}^*>y_{ik}^*|\forall k\neq j\}} \cdots \int \prod_{k=1}^{m} \left\{ p(y_{ik}^*|\alpha_j + f_k(x_i), \sigma_j^2) \mathrm{d}y_k^* \right\} \\ &= \int \prod_{\substack{k=1 \\ k\neq j}}^{m} \Phi\left(\frac{y_{ij}^* - \alpha_k - f_{ik}}{\sigma_k} \right) \cdot \frac{1}{\sigma_j} \phi\left(\frac{y_{ij}^* - \alpha_j - f_{ij}}{\sigma_j} \right) \mathrm{d}y_{ij}^* \\ &= \mathrm{E}_Z \left[\prod_{\substack{k=1 \\ k\neq j}}^{m} \Phi\left(\frac{\sigma_j}{\sigma_k} Z + \frac{\alpha_j + f_{ij} - \alpha_k - f_{ik}}{\sigma_k} \right) \right] \end{aligned}$$

where $Z \sim \mathrm{N}(0,1)$, and $\phi(\cdot)$ and $\Phi(\cdot)$ are its PDF and CDF respectively. The proof of this fact is included in the Appendix. With the exception of the binary case, these probabilities still do not have a closed-form expression (per se) and numerical methods are required to calculate them. In this simplified version of the I-probit model, the integral is unidimensional and involves the Gaussian PDF, and this can be efficiently obtained using quadrature methods.

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 $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ is obtained by integrating out the I-prior from the categorical likelihood, as follows:

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w}, \theta)\, p(\mathbf{w}|\theta)\, \mathrm{d}\mathbf{w} \\ &= \log \int \prod_{i=1}^{n} \prod_{j=1}^{m} \Big(g^{-1}\big(\alpha_k + \overbrace{f_k(x_i)}^{\sum_{i'=1}^{n} h_\eta(x_i, x_{i'}) w_{i'}}\big)_{k=1}^{m} \Big)^{[y_i=j]} \cdot \phi(\mathbf{w}|\mathbf{0}, \boldsymbol{\Psi} \otimes \mathbf{I}_n)\, \mathrm{d}\mathbf{w} \end{aligned} \tag{5.13}$$

eq:intractablelikelihood

where we have denoted the probit relationship from (5.9) 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 them-

selves involve integrals of normal densities. Furthermore, the posterior distribution of the regression function, which requires the density of $\mathbf{w}|\mathbf{y}$, depends on the marginalisation provided by (5.13). 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

One is interested in 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 $\mathbf{y}$, $p(\mathbf{y}) = \int e^{Q(\mathbf{w})}\,\mathrm{d}\mathbf{w}$, and we have established that the calculation of this marginal density 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 $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, and this 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.

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 fact, under an EM algorithm approach, using the approximate posterior density which is normally distributed is simply using the posterior mode in lieu of the actual posterior means.

2. Expand on this further.

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, because the dimension of this integral is exactly the sample size. Furthermore, standard errors for the parameters are cumbersome to calculate as well. 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 Markov chain Monte Carlo methods

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, just as we 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 in off-the-shelf software such as Stan. Things get out of hand with multinomial responses, because the intractability of computing class probabilities is not addressed.

In summary, the computational challenge here stems from two sources: 1) integrating out the random effects $\mathbf{w}$; and 2) evaluating class probabilities. Point 1) is addressed using a Gibbs sampling data augmentation scheme (latent variable approach), but this is not feasible with large n. Point 2) remains regardless whether Gibbs sampling or HMC is used.

5.4.3 Variational inference

We turn to variational inference as a method of estimation. Variational methods are widely discussed in the machine learning literature, and there have been efforts to popularise it in statistics (Blei et al., 2017). 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:

$$\begin{aligned}
\log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\
&= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \mathcal{L}(q) + \mathrm{KL}(q\|p) \\
&\geq \mathcal{L}(q) \qquad (5.14)
\end{aligned}$$

{eq:varboun d}

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

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} q(\mathbf{z})\,\mathrm{d}\mathbf{z} \\
&= \mathrm{E}_{\mathbf{z}\sim q}[\log p(\mathbf{y}, \mathbf{z})] + H(q), \qquad (5.15)
\end{aligned}$$

{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 (5.14) 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.

[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.

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.

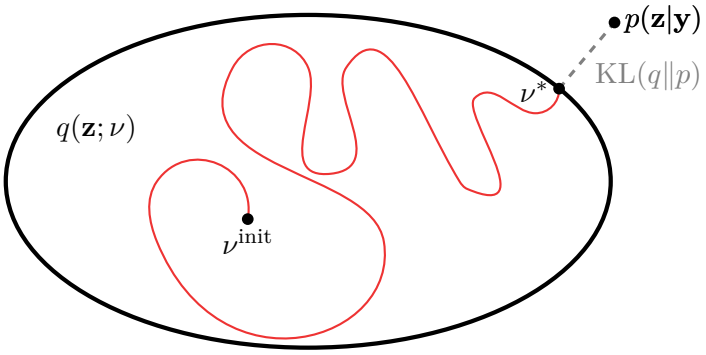

Figure 5.1: Schematic view of variational inference. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

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 $\mathbf{z} = (z^{(1)}, \ldots, z^{(M)})$, 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). By factorising appropriately, we can obtain approximated posteriors for the regression function and the parameters of the I-prior model. 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.

5.4.4 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 $\theta = \{\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_m), \eta, \boldsymbol{\Psi}\}$. Begin by assuming some prior distribution on the parameters $p(\theta) = p(\boldsymbol{\alpha})p(\eta)p(\boldsymbol{\Psi})$. Although one may devote more attention to the prior specification of these parameters, for our purposes it suffices that they are independent component-wise, and the PDFs belong to the exponential family of distributions with known hyperparameters. The exponential family requirement greatly eases the complexity of deriving the variational algorithm later on[3].

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

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

(2006, equation 10.9, p. 466), we find that for each $\xi \in \{\mathbf{y}^*, \mathbf{w}, \theta\} =: \mathcal{Z}$, $\tilde{q}$ satisfies

$$\log \tilde{q}(\xi) = \mathrm{E}_{-\xi}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \theta)] + \text{const.} \tag{5.16}$$

{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 (5.16) 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 (5.16),

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

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.

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

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

The fact that the rows of $\mathbf{y}^*$ are independent can be exploited. Write $\mathbf{y}_i^* = (y_{i1}^*, \ldots, y_{im}^*)^\top$. Then $\mathbf{y}_i^*|\theta, x_i \sim \mathrm{N}_m(\boldsymbol{\alpha} + \mathbf{f}(x_i), \boldsymbol{\Psi}^{-1})$, and we have the induced factorisation of the dis-

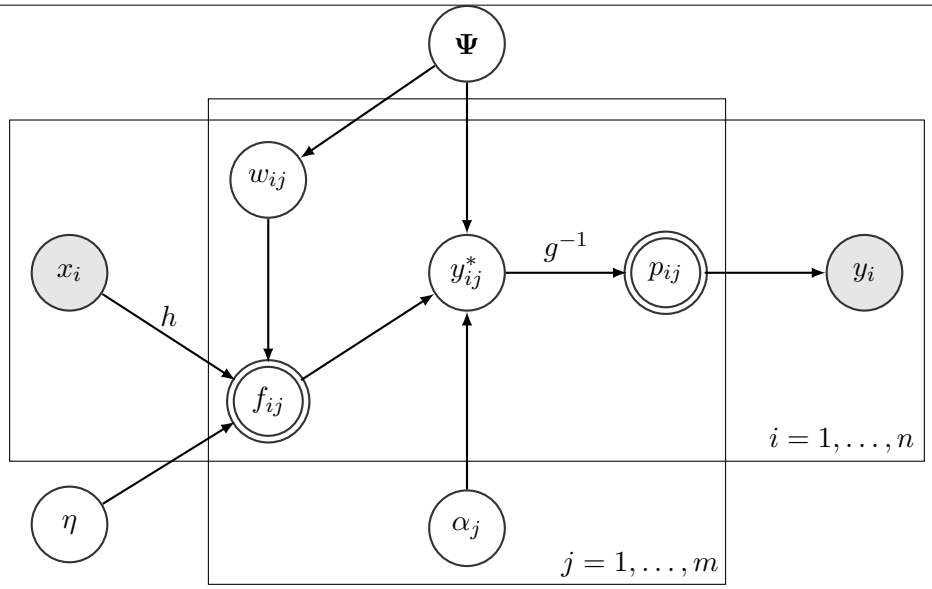

Figure 5.2: A DAG of the I-probit model. Observed nodes are shaded, while double-lined nodes represents calculable quantities.

tribution $q(\mathbf{y}^*) = \prod_{i=1}^n q(\mathbf{y}_i^*)$, where each $q(\mathbf{y}_i^*)$ is the density of a *conically truncated multivariate normal disribution*. That is, for each $i = 1, \ldots, n$ and noting the observed values $y_i = j \in \{1, \ldots, m\}$, the $\mathbf{y}_i^*$'s are distributed according to

$$\mathbf{y}_i^* \overset{\text{iid}}{\sim} \begin{cases} \mathrm{N}_m(\tilde{\boldsymbol{\alpha}} + \tilde{\mathbf{f}}(x_i), \tilde{\boldsymbol{\Sigma}}) & \text{if } y_{ij}^* > y_{ik}^*, \forall k \neq j \\ 0 & \text{otherwise.} \end{cases} \tag{5.17}$$

{eq:ystardist}

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

$$\tilde{y}_{ik}^* = \begin{cases} \tilde{\alpha}_k + \tilde{f}_{ik} - \tilde{\sigma}_k C_i^{-1} \displaystyle\int \phi_{ik}(z) \prod_{l \neq k,j} \Phi_{il}(z) \phi(z)\, \mathrm{d}z & \text{if } k \neq y_i \\ \tilde{\alpha}_{y_i} + \tilde{f}_{iy_i} - \tilde{\sigma}_{y_i} \sum_{k \neq y_i} \left(\tilde{y}_{ik}^* - \tilde{f}_{ik} \right) & \text{if } k = y_i \end{cases} \tag{5.18}$$

{eq:ystarupdate}

with

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

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

5.5.2 I-prior random effects w

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

$$\operatorname{vec}\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w(\tilde{\boldsymbol{\Psi}} \otimes \tilde{\mathbf{H}}_\eta)\operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = (\tilde{\boldsymbol{\Psi}} \otimes \tilde{\mathbf{H}}_\eta^2) + (\tilde{\boldsymbol{\Psi}}^{-1} \otimes \mathbf{I}_n). \tag{5.19}$$

{eq:varipostw}

We note the similarity between (5.19) above and the posterior distribution for the I-prior random effects in a normal model (4.7) seen in the previous chapter. Naïvely computing the inverse $\tilde{\mathbf{V}}_w^{-1}$ presents a computational challenge, as this takes $O(n^3m^3)$ time. By exploiting the Kronecker product structure in $\tilde{\mathbf{V}}_w^{-1}$, we are able to efficiently compute the required inverse in roughly $O(n^3m)$ time—see the appendix for details. Equivalently, we can express the distribution for $\mathbf{w} \sim \tilde{q}$ as a matrix normal distribution

$$\mathrm{MN}_{nm}\left(\overbrace{\tilde{\mathbf{H}}_\eta^{-1}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top)\tilde{\boldsymbol{\Psi}}^2}^{\tilde{\mathbf{w}}}, \tilde{\mathbf{H}}_\eta^{-2}, \tilde{\boldsymbol{\Psi}}\right). \tag{5.20}$$

{eq:varipostw2}

If the independent I-probit model is assumed, i.e. $\tilde{\boldsymbol{\Psi}} = \operatorname{diag}(\tilde{\sigma}_1^{-2}, \ldots, \tilde{\sigma}_m^{-2})$, then the posterior covariance matrix $\tilde{\mathbf{V}}_w$ has a simpler structure. This means that the random matrix $\mathbf{w}$ will have columns which are independent of each other. By writing $\mathbf{w}_j = (w_{1j}, \ldots, w_{nj})^\top \in \mathbb{R}^n$, $j = 1, \ldots, m$, to denote the column vectors of $\mathbf{w}$ and with a slight

abuse of notation, we have that

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

where

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

5.5.3 RKHS parameters η

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

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

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

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

Write down the mean and variance for lambda

5.5.4 Error precision $\boldsymbol{\Psi}$

A small reparameterisation of the I-prior random effects is necessary to achieve conjugacy for the $\boldsymbol{\Psi}$ parameter. Let $\mathbf{u}\in\mathbb{R}^{n\times m}$ be a matrix defined by $\boldsymbol{\Psi}^{-1}\mathbf{w}$. Then $\mathbf{u}\sim\mathrm{MN}_{n,m}(\mathbf{0},\mathbf{I}_n,\boldsymbol{\Psi}^{-1})$ a priori. From (5.20), the optimal variational distribution for $\mathbf{u}$ would be $\mathrm{MN}_{n,m}(\tilde{\mathbf{w}}\tilde{\boldsymbol{\Psi}}^{-1},\tilde{\mathbf{H}}_\eta^2,\tilde{\boldsymbol{\Psi}}^{-1})$. With a Wishart prior on the precision matrix

$\boldsymbol{\Psi} \sim \text{Wis}_m(\mathbf{G}, g)$, where $g \geq m$, the optimal variational density for $\boldsymbol{\Psi}$ is found to satisfy

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

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

$$\begin{aligned} \mathbf{G}_1 &= \text{E}_{\mathcal{Z}\setminus\{\boldsymbol{\Psi}\}\sim q}\left[\sum_{i=1}^{n}(\mathbf{y}_i^* - \boldsymbol{\alpha} - \mathbf{f}(x_i))(\mathbf{y}_i^* - \boldsymbol{\alpha} - \mathbf{f}(x_i))^\top\right] \\ \mathbf{G}_2 &= \sum_{i=1}^{n} \text{E}_{\mathbf{u}\sim q}\left[\mathbf{u}_i\mathbf{u}_i^\top\right]. \end{aligned} \tag{5.21}$$

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

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

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

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

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 is expected to increase with every iteration.

6. Proof?

5.6 Post-estimation

5.7 Examples

5.8 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.

5.9 Miscellanea

5.9.1 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 $\boldsymbol{\Sigma}$ is assumed to be diagonal (which corresponds to abandoning the independence of irrelevant alternatives assumption) and much more so if we assume that $\boldsymbol{\Sigma} = \mathbf{I}_m$. The latter yields the *normalised I-probit model*, and a discussion of the merits of this model is given later.

8. can use Hamiltonian Monte Carlo?

5.9.2 Similarity of EM algorithm and variational Bayes

5.9.3 Conically truncated multivariate normal distributions

Crucial to the probit model, the properties of conically truncated multivariate normal distributions are worth investigating.

definition: conically-truncated-normal

Definition 5.1 (Conically-truncated multivariate normal distribution)**.** Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a d-dimensional random variable with pdf defined as

$$p(\mathbf{x}) = \begin{cases} \prod_{i=1}^{d} \mathrm{N}(\mu_i, \sigma_i) & \text{if } X_j > X_i, \forall i \neq j \\ 0 & \text{otherwise} \end{cases}$$

for some $j \in \{1, \ldots, d\}$. We denote the distribution of $\mathbf{X}$ by $\mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. The pdf of $\mathbf{X}$ has support on the set $\{\mathbb{R}^d \,|\, x_j > x_i, \forall i \neq j\}$ and the following functional form:

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

where ϕ is the pdf of a standard normal distribution and

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

where $Z \sim \mathrm{N}(0, 1)$. In the case where all variances are unity, the pdf of $\mathbf{X} \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \mathbf{I}_d)$ is

$$p(\mathbf{x}) = \left\{ (2\pi)^{d/2} \, \mathrm{E}_Z \left[\prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(Z + \mu_j - \mu_i \right) \right] \right\}^{-1} \exp\left[-\frac{1}{2} \sum_{i=1}^{d} (x_i - \mu_i)^2 \right].$$

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

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

the proof follows directly. □

lem:expectation-entropy-truncated-mvn

Lemma 5.1. *Let $X \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with pdf $p(\mathbf{x})$ as defined in Definition 5.1. Then*

(i) The expectation $\mathrm{E}[\mathbf{X}] = \big(\mathrm{E}[X_1], \ldots, \mathrm{E}[X_d]\big)$ is given by

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

(ii) The differential entropy $\mathcal{H}(p)$ is given by

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

where $C = \mathrm{E}\left[\prod_{i \neq j} \Phi_i\right]$*, and we had defined*

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

with $Z \sim \mathrm{N}(0, 1)$*, and* $\phi(\cdot)$ *and* $\Phi(\cdot)$ *the pdf and cdf of* Z *respectively.*

Appendix

Fact: $X \sim \mathrm{N}(a, A)$ and $Y \sim \mathrm{N}(b, B)$, then

$$p(x)p(y) \propto \mathrm{N}(c, C)$$

where $C = (A^{-1} + B^{-1})^{-1}$ and $c = C(A^{-1}a + B^{-1}b)$.

5.9.4 Proof of Lemma

Proof. (i) Due to the independence structure in the pdf of $\mathbf{X}$, it is easy to consider the expectations of each of the components separately and marginalising out the

rest of the components. For $i \neq j$, we have

$$\begin{aligned}
\mathrm{E}[x_i] &= C^{-1} \int \cdots \int \mathbb{1}[x_k < x_j, \forall k \neq j] \cdot x_i \prod_{k=1}^{d} \frac{1}{\sigma_k} \phi\left(\frac{x_k - \mu_k}{\sigma_k}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \iint \mathbb{1}[x_i < x_j] \frac{x_i}{\sigma_i} \phi\left(\frac{x_i - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_i \mathrm{d}x_j \\
&= C^{-1} \iint \mathbb{1}[\sigma_i z_i + \mu_i < \sigma_j z_j + \mu_j](\sigma_i z_i + \mu_i) \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \mathrm{d}z_i \mathrm{d}z_j \\
&= \mu_i C^{-1} \iint \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \mathrm{d}z_i \mathrm{d}z_j \\
&\quad + \sigma_i C^{-1} \iint \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] z_i \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \mathrm{d}z_i \mathrm{d}z_j \\
&= \mu_i C^{-1} \overbrace{\int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \mathrm{d}z_j}^{C} \\
&\quad + \sigma_i C^{-1} \int \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] z_i \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \mathrm{d}z_i \mathrm{d}z_j
\end{aligned}$$

The integral involving z_i in the second part of the sum is recognised as the (unnormalised) expectation of the lower-tail of a univariate standard normal distribution truncated at $\tau_{ij} = (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i$. That is,

$$\mathrm{E}[Z_i | Z_i < \tau_{ij}] = \big[\Phi(\tau_{ij})\big]^{-1} \int \mathbb{1}[z_i < \tau_{ij}] z_i \phi(z_i) \mathrm{d}z_i = -\frac{\phi(\tau_{ij})}{\Phi(\tau_{ij})}$$

Plugging this expression back into the derivation of this expectation, we get

$$\begin{aligned}
\mathrm{E}[X_i] &= \mu_i - \sigma_i C^{-1} \int \phi\left(\frac{\sigma_j z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \mathrm{d}z_j \\
&= \mu_i - \sigma_i C^{-1} \, \mathrm{E}\left[\phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right)\right].
\end{aligned}$$

The expectation for the jth component is

$$\begin{aligned}
\mathrm{E}[X_j] &= C^{-1} \int \cdots \int \mathbb{1}[x_k < x_j, \forall k \neq j] \cdot x_j \prod_{k=1}^{d} \frac{1}{\sigma_k} \phi\left(\frac{x_k - \mu_k}{\sigma_k}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \int x_j \prod_{k \neq j} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \cdot \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_j \\
&= C^{-1} \int (\sigma_j z_j + \mu_j) \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot \phi(z_j) \mathrm{d}z_j \\
&= \mu_j C^{-1} \overbrace{\int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot \phi(z_j) \mathrm{d}z_j}^{C} \\
&\quad + \sigma_j C^{-1} \int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot z_j \phi(z_j) \mathrm{d}z_j \\
&= \mu_j + \sigma_j C^{-1} \,\mathrm{E}\left[Z_j \prod_{k \neq j} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right) \right] \\
&= \mu_j + \sigma_j \sum_{\substack{i=1 \\ i \neq j}}^{d} \sigma_i C^{-1} \,\mathrm{E}\left[\phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right) \right] \\
&= \mu_j - \sigma_j \sum_{i \neq j} \big(\mathrm{E}[X_i] - \mu_i \big)
\end{aligned}$$

where we have made use of Lemma 5.2 in the second last step of the above.

(ii) The differential entropy is given by

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

□

lem:EZgZ

Lemma 5.2. *Let $Z \sim \mathrm{N}(0,1)$. Then for all $m \in \{\mathbb{N} \,|\, m > 1\}$ and $(\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$,*

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

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

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

We adopt the following notation for convenience:

$$\phi_i = \phi(\sigma_i z + \mu_i)$$
$$\Phi_i = \Phi(\sigma_i z + \mu_i)$$

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

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

Now assume that the inductive hypothesis holds for some $m \in \{\mathbb{N} \,|\, m > 1\}$. That is, the derivative of

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

which is

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

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

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

is found to be

$$\begin{aligned} g'_{m+1}(z) &= \sigma_{m+1}\phi_{m+1}g_m(z) + g'_m(z)\Phi_{m+1} \\ &= \sigma_{m+1}\phi_{m+1}\prod_{\substack{k=1\\k\neq j}}^{m}\Phi_k + \sum_{\substack{i=1\\i\neq j}}^{m}\left[\sigma_i\phi_i\prod_{\substack{k=1\\k\neq i,j}}^{m}\Phi_k\right]\Phi_{m+1} \\ &= \sigma_{m+1}\phi_{m+1}\prod_{\substack{k=1\\k\neq j,m+1}}^{m+1}\Phi_k + \sum_{\substack{i=1\\i\neq j}}^{m}\left[\sigma_i\phi_i\prod_{\substack{k=1\\k\neq i,j}}^{m+1}\Phi_k\right] \\ &= \sum_{\substack{i=1\\i\neq j}}^{m+1}\left[\sigma_i\phi_i\prod_{\substack{k=1\\k\neq i,j}}^{m+1}\Phi_k\right] \\ &= g'_{m+1}(z). \end{aligned}$$

Thus, by induction and linearity of expectations, the proof is complete. □

5.10 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

$$\log \tilde{q}(\xi) = \mathrm{E}_{-\xi}[\log p(\mathbf{y}, \mathcal{Z})] + \text{const.} \tag{5.16}$$

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

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

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

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

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

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

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

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

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

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

pdf is

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

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

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

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

For the independent I-probit model, $\boldsymbol{\Psi} = \text{diag}(\sigma_1^{-2}, \ldots, \sigma_m^{-2})$, and we choose independent Gamma distributions for each precision $\sigma_j^{-2} \sim \Gamma(r_j, s_j)$, where r_j and s_j are the shape and scale parameters. Then,

$$p(\boldsymbol{\Psi}) = \prod_{j=1}^{m} \exp\left[\text{const.} + (r_j - 1)\log\sigma_j^{-2} - \frac{\sigma_j^{-2}}{s_j}\right].$$

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

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

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

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 $j \in \{1, \ldots, m\}$. The mean-field density $q(\mathbf{y}_i^*)$ for each $i = 1, \ldots, n$ is found to be

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

where $\tilde{\boldsymbol{\mu}}_i = \mathrm{E}\,\boldsymbol{\alpha} + (\mathrm{E}\,\mathbf{H}_\eta \,\mathrm{E}\,\mathbf{w})_i$, 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 $\tilde{\boldsymbol{\Psi}}$ is diagonal, then Lemma X provides a simplification.

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

The terms involving $\mathbf{w}$ in (5.16) 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

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

and

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

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

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

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

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

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

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

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 $\boldsymbol{\Psi}$ to obtain $\boldsymbol{\Psi} = \mathbf{Q}\mathbf{P}\mathbf{Q}^\top$. 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):

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

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.

$\tilde{q}(\lambda)$

For $j = 1, \ldots, m$,

$$\begin{aligned}
\log \tilde{q}(\lambda_j) &= \mathrm{E}_{\mathbf{y}^*, \mathbf{w}, \alpha}\left[-\frac{1}{2} \sum_{j=1}^{m} \|\mathbf{y}_j^* - \alpha_j \mathbf{1}_n - \lambda_j \mathbf{H}\mathbf{w}_j\|^2 \right] + \text{const.} \\
&= -\frac{1}{2} \sum_{j=1}^{m} \mathrm{E}_{\mathbf{y}^*, \mathbf{w}, \alpha}\left[\lambda_j^2 \, \mathbf{w}_j^\top \mathbf{H}^2 \mathbf{w}_j - 2\lambda_j (\mathbf{y}_j^* - \alpha_j \mathbf{1}_n)^\top \mathbf{H}\mathbf{w}_j \right] + \text{const.} \\
&= -\frac{1}{2} \sum_{j=1}^{m} \left(\lambda_j^2 \operatorname{tr}\left(\mathbf{H}^2 \, \mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top] \right) - 2\lambda_j (\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathbf{H} \, \mathrm{E}[\mathbf{w}_j] \right) + \text{const.}
\end{aligned}$$

By completing the squares, we recognise this is as the kernel of the product of independent univariate normal densities. Thus, each $\lambda_j \sim \mathrm{N}(d_j/c_j, 1/c_j)$, where

$$c_j = \operatorname{tr}\left(\mathbf{H}^2 \, \mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top] \right) \quad \text{and} \quad d_j = (\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathbf{H} \, \mathrm{E}[\mathbf{w}_j].$$

Supposing we use the same covariance kernel (and therefore scale parameter) for each regression class, the distribution for λ is easily seen as

$$\lambda \sim \mathrm{N}\left(\frac{\sum_{j=1}^m d_j}{\sum_{j=1}^m c_j}, \frac{1}{\sum_{j=1}^m c_j} \right).$$

$\tilde{q}(\alpha)$

For $j = 1, \ldots, m$, denote $\mathbf{H}_i$ as the row vector of the kernel matrix $\mathbf{H}$. Then,

$$\begin{aligned}
\log \tilde{q}(\alpha) &= \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\lambda}\left[-\frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{ij}^* - \alpha_j - \lambda_j \textstyle\sum_{k=1}^{n} h(x_i, x_k) w_{kj}\right)^2\right] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^{m}\mathrm{E}_{\mathbf{y}^*,\mathbf{w},\lambda}\left[n\alpha_j^2 - 2\alpha_j\sum_{i=1}^{n}(y_{ij}^* - \lambda_j \mathbf{H}_i \mathbf{w}_j)\right] + \text{const.} \\
&= -\frac{n}{2}\sum_{j=1}^{m}\left[\left(\alpha_j - \frac{1}{n}\sum_{i=1}^{n}(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\mathbf{w}_j)\right)^2\right] + \text{const.}
\end{aligned}$$

which is of course the kernel of the product of m univariate normal densities, each with mean and variance

$$\tilde{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\,\mathrm{E}[\mathbf{w}_j]\right) \quad \text{and} \quad v_{\alpha_j} = \frac{1}{n}.$$

Suppose that we use a single intercept parameter α. In this case, α is is also normally distributed with mean and variance

$$\tilde{\alpha} = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\,\mathrm{E}[\mathbf{w}_j]\right) \quad \text{and} \quad v_{\alpha} = \frac{1}{nm}.$$

5.10.1 Monitoring the lower bound

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

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

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

Definition 5.2 (Differential entropy)**.** The differential entropy $\mathcal{H}$ of a pdf $p(x)$ is given by

$$\mathcal{H}(p) = -\int p(x)\log p(x)\mathrm{d}x = -\,\mathrm{E}_p[\log p(x)].$$

thm:normentropy

Lemma 5.3. *Let $p(x)$ be the pdf of a random variable x. Then if*

(i) p is a univariate normal distribution with mean μ and variance σ^2,

$$\mathcal{H}(p) = \frac{1}{2}(1+\log 2\pi) + \frac{1}{2}\log\sigma^2$$

(ii) p is a d-dimensional normal distribution with mean μ and variance Σ,

$$\mathcal{H}(p) = \frac{d}{2}(1+\log 2\pi) + \frac{1}{2}\log|\Sigma|$$

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

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

Terms involving distributions of w

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

Terms involving distribution of $q(\lambda)$

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

or if using single λ

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

Terms involving distribution of $q(\alpha)$

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

or if using single α

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

The lower bound

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

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

Bibliography

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.

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.

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

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.

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.

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

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