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

chapter5

In a regression setting such as (1.1), consider polytomous response variables $y_1, \dots, y_n$, where each y_i takes on exactly one of the values from the set of m possible choices $\mathcal{M} = \{1, \dots, m\}$. 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.

The normality assumption (1.2) is not entirely appropriate anymore. 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 class probabilities of the observations to a normal 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 ranges.

Expanding on this idea further, assume that the y_i's follow a categorical distribution, denoted by

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

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

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

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

$$g(p_{i1}, \ldots, p_{im}) = \big(\alpha_1 + f_1(x_i), \ldots, \alpha_m + f_m(x_i)\big)$$

where $g : [0,1]^m \to \mathbb{R}^m$ is some specified link function. As we will see later, a normal regression model as in (1.1) subject to (1.2) naturally implies a *probit* link function. With an I-prior assumed on the f_j's, we call this method of probit regression using I-priors the *I-probit* regression model.

Due to the nature of the model, unfortunately, the posterior distribution of the regression functions cannot be found in closed form. In particular, marginalising the I-prior from the joint likelihood involves a high-dimensional intractable integral. We explore a fully Bayesian approach to estimate I-probit models using *variational inference*. The main idea is to replace the difficult posterior distribution with an approximation that is tractable. 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 Bayesian machinery. For example, inferences around log-odds is usually cumbersome for probit models, but a credibility interval can easily be obtained by resampling methods from the relevant posteriors, which are typically made up of densities which are familiar and readily available in software.

By choosing appropriate RKHSs/RKKSs 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 5.7. We find that the many advantages of the normal I-prior methodology transfer over quite well to the I-probit model for binary and multinomial regression.

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

5.1 A latent variable motivation: the I-probit model

It is convenient, as we did in Section X naive classification, to again think of the responses $y_i \in \{1, \dots, m\} = \mathcal{M}$ as comprising of a binary vector $\mathbf{y}_{i\cdot} = (y_{i1}, \dots, y_{im})^\top$, with a single '1' at the position corresponding to the value that y_i takes. That is,

$$y_{ik} = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{if } y_i \neq k. \end{cases}$$

With $y_i \overset{\text{iid}}{\sim} \text{Cat}(p_{i1}, \dots, p_{im})$ for $i = 1, \dots, n$, each y_{ij} is distributed as Bernoulli with probability p_{ij}, $j = 1, \dots, m$ according to the above formulation. Now, assume that, for each $y_{i1}, \dots, y_{im}$, there exists corresponding *continuous, underlying, latent variables* $y^*_{i1}, \dots, y^*_{im}$ such that

$$y_i = \begin{cases} 1 & \text{if } y^*_{i1} \geq y^*_{i2}, y^*_{i3}, \dots, y^*_{im} \\ 2 & \text{if } y^*_{i2} \geq y^*_{i1}, y^*_{i3}, \dots, y^*_{im} \\ \vdots & \\ m & \text{if } y^*_{im} \geq y^*_{i2}, y^*_{i3}, \dots, y^*_{i\,m-1}. \end{cases} \tag{5.1}$$

{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. In this sense, the y^*_{ij}'s represent individual i's *latent propensities* for choosing alternative j.

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

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

{eq:multinomial-latent}

We can see some semblance of this model with the one in **??**, and ultimately the aim is to assign I-priors to the regression function of these latent variables, which we shall describe shortly. For now, write $\boldsymbol{\mu}(x_i) \in \mathbb{R}^m$ whose j'th component is $\alpha + \alpha_j + f_j(x_i)$, and realise that each $\mathbf{y}^*_{i\cdot} = (y^*_{i1}, \dots, y^*_{im})^\top$ has the distribution $\text{N}_m(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1})$, conditional on the data x_i, the intercepts $\alpha, \alpha_1, \dots, \alpha_m$, the evaluations of the functions at x_i for each class $f_1(x_i), \dots, f_m(x_i)$, and the error covariance matrix $\boldsymbol{\Psi}^{-1}$.

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

$$\begin{aligned} p_{ij} &= \mathrm{P}(y_i = j) \\ &= \mathrm{P}\left(\{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}\right) \\ &= \int_{\{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}} \cdots \int \phi(y_{i1}^*, \ldots, y_{im}^* | \boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}y_{i1}^* \cdots \mathrm{d}y_{im}^*, \end{aligned} \tag{5.3}$$

{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 $\mathbf{y}_{i\cdot}^*$ belongs to the set $\mathcal{C}_j := \{y_{ij}^* > y_{ik}^* \,|\, \forall k \neq j\}$, 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 for 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 5.6.1 for a note regarding this matter.

Now, we'll see how to specify an I-prior on the regression problem (5.2). In the naïve I-prior model, we wrote $f(x_i, j) = \alpha_j + f_j(x_i)$, and called for f to belong to an ANOVA RKKS with kernel defined in **??**. Instead of doing the same, we take a different approach. Treat the α_j's in (5.2) 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 $f_j \in \mathcal{F}$, $j = 1 \ldots, m$, defined by

$$f_j(x_i) = f_0(x_i, j) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_{ik}$$

with $\mathbf{w}_{i\cdot} := (w_{i1}, \ldots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}(0, \boldsymbol{\Psi})$. This is similar to the naïve I-prior specification **??**, except that the intercept have been treated as parameters rather than accounting for them using an RKHS of functions (Pearson RKHS or identity kernel RKHS). Importantly, the overall regression relationship still satisfies the ANOVA functional decomposition, because the α_j's sum to zero. We find that this approach bodes well down the line computationally.

We call the multinomial probit regression model of (5.1) subject to (5.2) and I-priors on $f_j \in \mathcal{F}$, the *I-probit model*. For completeness, this is stated again: for $i = 1, \ldots, n$,

$y_i = \arg\max_{k=1}^m y_{ik}^* \in \{1, \ldots, m\}$, where, for $j = 1, \ldots, m$,

$$
\begin{aligned}
y_{ij}^* &= \alpha + \alpha_j + \overbrace{f_0(x_i, j) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_{ik}}^{f_j(x_i)} + \epsilon_{ij} \\
\boldsymbol{\epsilon}_{i\cdot} &:= (\epsilon_{i1}, \ldots, \epsilon_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\
\mathbf{w}_{i\cdot} &:= (w_{i1}, \ldots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}).
\end{aligned}
\tag{5.4}
$$

{eq:iprobit mod}

The parameters of the I-probit model are denoted by $\theta = \{\alpha_1, \ldots, \alpha_m, \eta, \boldsymbol{\Psi}\}$. To establish notation, let

- $\boldsymbol{\epsilon} \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries ϵ_{ij}, whose rows are $\boldsymbol{\epsilon}_{i\cdot}$, columns are $\boldsymbol{\epsilon}_{\cdot j}$, and is distributed $\boldsymbol{\epsilon} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$;
- $\mathbf{w} \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries w_{ij}, whose rows are $\mathbf{w}_{i\cdot}$, columns are $\mathbf{w}_{\cdot j}$, and is distributed $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$;
- $\mathbf{f}, \mathbf{f}_0 \in \mathbb{R}^{n \times m}$ denote the matrices containing (i, j) entries $f_j(x_i)$ and $f_0(x_i, j)$ respectively, so that $\mathbf{f} = \mathbf{f}_0 + \mathbf{H}_\eta \mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{1}_n \mathbf{f}_0^\top, \mathbf{H}_\eta^2, \boldsymbol{\Psi})$;
- $\boldsymbol{\alpha} = (\alpha + \alpha_1, \ldots, \alpha + \alpha_m)^\top \in \mathbb{R}^m$ be the vector of intercepts;
- $\boldsymbol{\mu} = \mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{f}$, whose (i, j) entries are $\mu_j(x_i) = \alpha + \alpha_j + f_j(x_i)$; and
- $\mathbf{y}^* \in \mathbb{R}^{n \times m}$ denote the matrix containing (i, j) entries y_{ij}^*, that is, $\mathbf{y}^* = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, so $\mathbf{y}^* | \mathbf{w} \sim \mathrm{MN}_{n,m}(\boldsymbol{\mu} = \mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$ and $\operatorname{vec} \mathbf{y}^* \sim \mathrm{N}_{nm}\left(\operatorname{vec}(\mathbf{1}_n \boldsymbol{\alpha}^\top), \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n\right)$ (note that the marginal distribution of $\mathbf{y}^*$ cannot be expressed as a matrix normal, except when $\boldsymbol{\Psi} = \mathbf{I}_m$).

Before proceeding with estimating the I-probit model (5.4), we lay out several standing assumptions:

ass:A4 A4 **Centred responses**. Set $\alpha = 0$.

ass:A5 A5 **Zero prior mean**. Assume a zero prior mean $f_0(x) = 0$ for all $x \in \mathcal{X}$.

ass:A6 A6 **Fixed error precision**. Assume $\boldsymbol{\Psi}$ is fixed.

Assumption A4 is a requirement for identifiability, while A5 is motivated by a similar argument to assumption A2 in the normal I-prior model. While estimation of $\boldsymbol{\Psi}$ would add flexibility to the model, several computational issues were not able to be resolved within the time limitations of completing this project (see Section 5.6.3).

5.2 Identifiability and IIA

sec:iia

The parameters in the standard linear multinomial probit model is well known to be unidentified (Michael P. Keane, 1992; Train, 2009), and we find this to be the case in the I-probit model as well. Unrestricted probit models are not identified for two reasons. Firstly, an addition of a non-zero constant $a \in \mathbb{R}$ to the latent variables y_{ij}^*'s in (5.1) will not change which latent variable is maximal, and therefore leaves the model unchanged. It is for this reason assumptions A4 and A5 are imposed. Secondly, all latent variables can be scaled by some positive constant $c \in \mathbb{R}_{>0}$ without changing which latent variable is largest. This means that m-variate normal distribution $\mathrm{N}_m\left(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}\right)$ of the underlying latent variables $\mathbf{y}_{i\cdot}^*$ would yield the same class probabilities as the multivariate normal distribution $\mathrm{N}_m\left(a\mathbf{1}_m + c\boldsymbol{\mu}(x_i), c^2\boldsymbol{\Psi}^{-1}\right)$, according to (5.3). Therefore, the multinomial probit model is not identified as there exists more than one set of parameters for which the categorical likelihood $\prod_{i,j} p_{ij}$ is the same.

Identification for the probit model is resolved by setting one restriction on the intercepts $\alpha_1, \ldots, \alpha_m$ (location) and $m+1$ restrictions on the precision matrix $\boldsymbol{\Psi}$ (scale). Restrictions on the intercepts include $\sum_{j=1}^m \alpha_j = 0$ or setting one of the intercepts to zero. In this work, we apply the former restriction to the I-probit model, as this is analogous to the requirement of zero-mean functions in the functional ANOVA decomposition. If A6 holds, then location identification is all that is needed to achieve identification. However, if $\boldsymbol{\Psi}$ is a free parameter to be estimated, only $m(m-1)/2 - 1$ parameters are identified. Many possible specifications of the restriction on $\boldsymbol{\Psi}$ is possible, depending on the number of alternatives m and the intended effect of $\boldsymbol{\Psi}$, for example:

- **Case $m = 2$** (minimum number of restrictions = 3).

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & \\ 0 & 0 \end{pmatrix}, \text{ or } \boldsymbol{\Psi} = \begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}$$

- **Case $m = 3$** (minimum number of restrictions = 4).

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & & \\ \psi_{12} & \psi_{22} & \\ 0 & 0 & 0 \end{pmatrix}, \text{ or } \boldsymbol{\Psi} = \begin{pmatrix} 1 & & \\ 0 & \psi_{22} & \\ 0 & 0 & \psi_{33} \end{pmatrix}$$

- **Case $m \geq 4$** (minimum number of restrictions $= m + 1$).

$$\boldsymbol{\Psi} = \begin{pmatrix} 1 & & & & \\ \psi_{12} & \psi_{22} & & & \\ \vdots & \vdots & \ddots & & \\ \psi_{1,m-1} & \psi_{2,m-1} & \cdots & \psi_{m-1,m-1} & \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix}, \text{ or } \boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & & & \\ & \psi_{22} & & \\ & & \ddots & \\ & & & \psi_{mm} \end{pmatrix}$$

Remark 5.1. Identification is most commonly achieved by fixing the latent propensities of one of the classes to zero and fixing one element the covariance matrix (Dansie, 1985; Bunch, 1991). Fixing the last class, say, to zero, i.e. $y^*_{im} = 0, \forall i = 1, \ldots, n$ has the effect of shrinking $\boldsymbol{\Psi}$ to $(m-1) \times (m-1)$ in size, and thus one more restriction needs to be made (typically, the first element $\boldsymbol{\Psi}_{11}$ is set to one). This speaks to the fact that the absolute values of the latent propensities themselves do not matter, but their relative differences do—see Section X. We also remark that for the binary case ($m = 2$), setting the latent propensities for the second class to zero and fixing the remaining variance parameter to one yields, for $i = 1, \ldots, n$,

$$\begin{aligned} p_{i1} &= \mathrm{P}(y^*_{i1} > y^*_{i2} = 0) \\ &= \mathrm{P}\left(\alpha_1 + f_1(x_i) + \epsilon_{i1} > 0 \,|\, \epsilon_{i1} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)\right) \\ &= \Phi\big(\alpha_1 + f_1(x_i)\big) \end{aligned} \tag{5.5}$$

{eq:iprobitbin}

and $p_{i2} = 1 - \Phi\big(\alpha_1 + f_1(x_i)\big)$, the familiar binary probit model. Note that in the binary case only one set of latent propensities need to be estimated, so we can drop the subscript '1' in the above equations. In fact, for m classes, only $m - 1$ sets of regression functions need to be estimated (since one of them needs to be fixed), but in the multinomial presentation of this thesis we define regression functions for each class.

Now, we turn to a discussion of the role of $\boldsymbol{\Psi}$ in the model. 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 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' choices

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.

To put it simply, the model is IIA if choice probabilities depend only on the choice in consideration, and not on any other alternatives. In the I-probit model, or rather, in probit models in general, choice dependency is controlled by the error precision matrix $\boldsymbol{\Psi}$. Specifically, the off-diagonal elements $\boldsymbol{\Psi}_{jk}$ capture the correlation between alternatives j and k. Allowing all $m(m+1)/2$ covariance elements of $\boldsymbol{\Psi}$ to be non-zero leads to the *full I-probit model*, and would not assume an IIA position. Figure 5.1 illustrates the covariance structure for the marginal distribution of the latent propensities, $\mathbf{V}_{y^*} = \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n$, and of the I-prior $\mathbf{V}_f = \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2$.

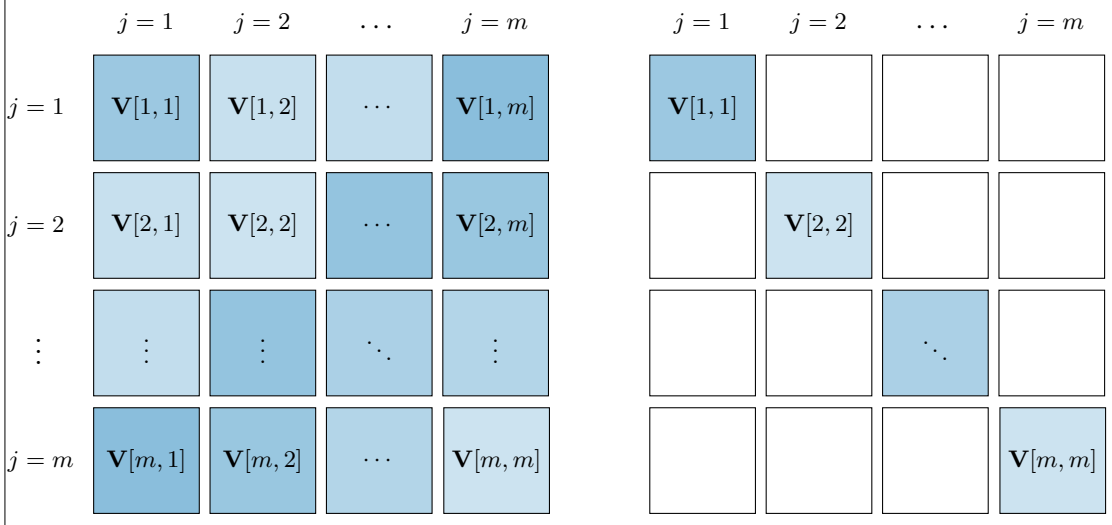

Figure 5.1: Illustration of the covariance structure of the full I-probit model (left) and the independent I-probit model (right). The full model has m^2 blocks of $n \times n$ symmetric matrices, and the blocks themselves are arranged symmetrically about the diagonal. The independent model, on the other hand, has a block diagonal structure, and its sparsity induces simpler computational methods for estimation.

fig:iprobcovstr

While it is an advantage to be able to model the correlations across choices (unlike in logistic models), there are applications where the IIA assumption would not adversely affect the analysis, such as classification tasks. Some analyses might also be indifferent

as to whether or not choice dependency exists. In these situations, it would be beneficial, algorithmically speaking, to reduce the I-probit model to a simpler version by assuming $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_m)$, which would trigger an IIA assumption in the I-probit model. We refer to this model as the *independent I-probit model*. The independence structure causes the distribution of the latent variables to be $y_{ij}^* \sim \mathrm{N}(\mu_k(x_i), \sigma_j^2)$ for $j = 1, \ldots, m$, where $\sigma_j^2 = \psi_j^{-1}$. As a continuation of line (5.3), we can show the class probabilities p_{ij} to be

$$\begin{aligned} p_{ij} &= \int_{\{y_{ij}^* > y_{ik}^* | \forall k \neq j\}} \cdots \int \prod_{k=1}^{m} \left\{ \phi(y_{ik}^* | \mu_k(x_i), \sigma_k^2) \, \mathrm{d}y_{ik}^* \right\} \\ &= \int \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{y_{ij}^* - \mu_k(x_i)}{\sigma_k} \right) \cdot \phi(y_{ij}^* | \mu_j(x_i), \sigma_j^2) \, \mathrm{d}y_{ij}^* \\ &= \mathrm{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\sigma_j}{\sigma_k} Z + \frac{\mu_j(x_i) - \mu_k(x_i)}{\sigma_k} \right) \right] \end{aligned} \tag{5.6}$$

{eq:pij2}

where $Z \sim \mathrm{N}(0, 1)$, $\Phi(\cdot)$ its cdf, and $\phi(\cdot | \mu, \sigma^2)$ is the pdf of $X \sim \mathrm{N}(\mu, \sigma^2)$. The equation (5.3) is thus simplified to a unidimensional integral involving the Gaussian pdf and cdf, which can be computed fairly efficiently using quadrature methods.

5.3 Estimation

The premise of the I-probit model is having regression functions capture the dependence of the covariates on a latent, continuous scale using I-priors, and then transforming these regression functions onto a probability scale. Therefore, as with the normal I-prior model, an estimate of the posterior regression function with optimised hyperparameters is sought. A schematic diagram depicting the I-probit model is shown in Figure 5.2.

The log likelihood function for θ using all n observations $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ is obtained by performing the following integration:

$$L(\theta | \mathbf{y}) = \log \iint p(\mathbf{y} | \mathbf{y}^*, \theta) p(\mathbf{y}^* | \mathbf{w}, \theta) p(\mathbf{w} | \theta) \, \mathrm{d}\mathbf{y}^* \, \mathrm{d}\mathbf{w}. \tag{5.7}$$

{eq:iprobit lik}

Here, $p(\mathbf{w}|\theta)$ is the pdf of $\mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$, $p(\mathbf{y}^*|\mathbf{w}, \theta)$ is the pdf of $\mathrm{MN}_{n,m}(\mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$, and $p(\mathbf{y}|\mathbf{y}^*, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{m} \left[y_{ij}^* = \max \mathbf{y}_{i\cdot}^* \right]^{[y_i = j]}$, with $0^0 := 1$. Note

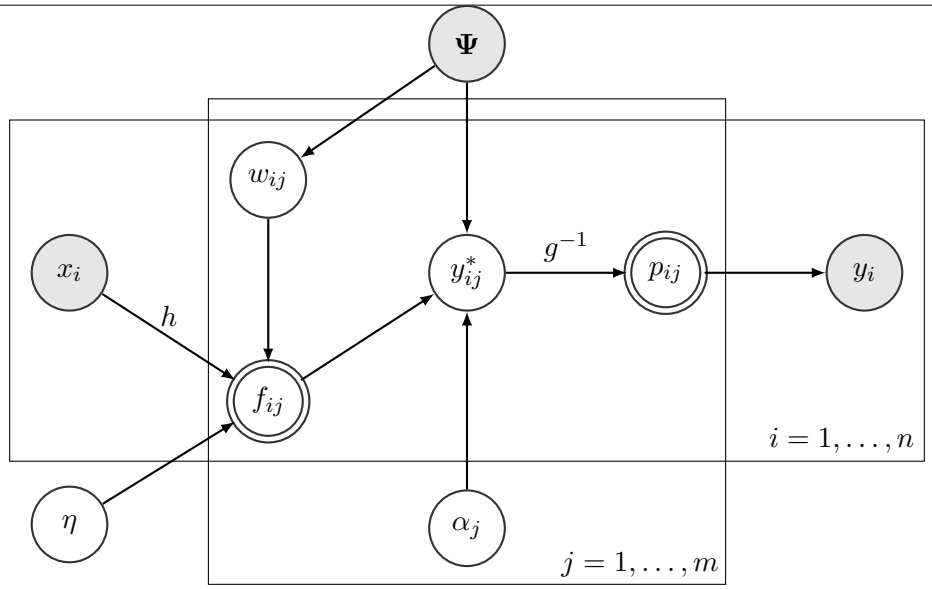

Figure 5.2: A directed acyclic graph (DAG) of the I-probit model. Observed/fixed nodes are shaded, while double-lined nodes represents calculable quantities.

fig:iprobit dag

that, given the corresponding latent propensities $\mathbf{y}^*_{i\cdot} = (y^*_{i1}, \ldots, y^*_{im})^\top$, the distribution $y_i|\mathbf{y}^*_{i\cdot}$ is tantamount to a degenerate categorical distribution, since after knowing which of the latent propensities is largest, knowledge of the outcome of the categorical response becomes a certainty.

The integral appearing in (5.7) is of order $2nm$, and so presents a massive computational challenge for classical numerical integration methods. This can be reduced by either integrating out the random effects $\mathbf{w}$ or the latent propensities $\mathbf{y}^*$ separately. Continuing on (5.7) gets us to either

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{y}^*, \theta) p(\mathbf{y}^*|\theta) \, \mathrm{d}\mathbf{y}^* \\ &= \log \int \left\{ \prod_{i=1}^{n} \prod_{j=1}^{m} \left[y^*_{ij} = \max \mathbf{y}^*_{i\cdot} \right]^{[y_i = j]} \right\} \phi(\mathbf{y}^*|\mathbf{1}_n \boldsymbol{\alpha}^\top, \boldsymbol{\Psi} \otimes \mathbf{H}^2_\eta + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \, \mathrm{d}\mathbf{y}^* \\ &= \log \int_{\bigcap_{i=1}^{n} \{y^*_{iy_i} > y^*_{ik} | \forall k \neq y_i\}} \phi(\mathbf{y}^*|\mathbf{1}_n \boldsymbol{\alpha}^\top, \boldsymbol{\Psi} \otimes \mathbf{H}^2_\eta + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \, \mathrm{d}\mathbf{y}^*, \end{aligned} \tag{5.8}$$

{eq:intractablelikelihood1}

by recognising that $\int p(\mathbf{y}^*|\mathbf{w},\theta)p(\mathbf{w}|\theta)\,\mathrm{d}\mathbf{w}$ has a closed-form expression since it is an integral involving two Gaussian densities, or

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w},\theta)\, p(\mathbf{w}|\theta)\,\mathrm{d}\mathbf{w} \\ &= \log \int \prod_{i=1}^{n}\prod_{j=1}^{m} \Big(g_j^{-1}(\overbrace{\boldsymbol{\alpha} + \mathbf{w}^\top \mathbf{h}_\eta(x_i)}^{\boldsymbol{\mu}(x_i)} \,|\, \boldsymbol{\Psi}) \Big)^{[y_i=j]} \cdot \mathrm{MN}_{n,m}(\mathbf{w}|\mathbf{0},\mathbf{I}_n,\boldsymbol{\Psi})\,\mathrm{d}\mathbf{w}, \end{aligned} \tag{5.9}$$

{eq:intractablelikelihood2}

where we have denoted the class probabilities p_{ij} from (5.3) using the function $g_j^{-1}(\cdot|\boldsymbol{\Psi}) : \mathbb{R}^m \to [0,1]$. Unfortunately, neither of these two simplifications are particularly helpful. In (5.8), the integral represents the probability of a mn-dimensional normal variate which is not straightforward to calculate, because its covariance matrix is dense. In (5.9), the integral has no apparent closed-form. Unavailability of an efficient, reliable way of calculating the log-likelihood hampers hope of obtaining parameter estimates via direct likelihood maximisation methods.

Furthermore, the posterior density of the regression function $\mathbf{f} = \mathbf{H}_\eta \mathbf{w}$, which requires the posterior density of $\mathbf{w}$ obtained via $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, has normalising constant equal to $L(\theta)$, which is intractable. The challenge of estimation is then to first overcome this intractability by means of a suitable approximation of the marginalising integral. We present three possible avenues to achieve this aim, namely the Laplace approximation, a variational EM algorithm, and Markov chain Monte Carlo (MCMC) methods.

5.3.1 Laplace approximation

The focus here is to obtain the posterior density $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) =: e^{R(\mathbf{w})}$ which has normalising constant equal to the marginal density of $\mathbf{y}$, $p(\mathbf{y}) = \int e^{R(\mathbf{w})}\,\mathrm{d}\mathbf{w}$, as per (5.9). Note that the dependence of the pdfs on θ is implicit, but is dropped for clarity. Laplace's method (Kass and Raftery, 1995, §4.1.1, pp. 777–778) entails expanding a Taylor series for R about its posterior mode $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, which gives the relationship

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

because, assuming that R has a unique maximum, ∇R 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, $\boldsymbol{\Omega} = -\nabla^2 R(\mathbf{w})|_{\mathbf{w}=\hat{\mathbf{w}}}$ is the negative Hessian of Q evaluated at the posterior mode, and is typically obtained as a byproduct of the maximisation routine of R using gradient or quasi-gradient based methods.

The marginal distribution is then approximated by

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

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.

5.3.2 Variational EM algorithm

We turn to variational methods as a means of approximating the posterior densities of interest and obtain parameter estimates. Variational methods are widely discussed in the machine learning literature, but there have been efforts to popularise it in statistics (Blei et al., 2017). Although variational inference is typically seen as a fully Bayesian method, whereby approximate posterior densities are sought for the latent variables and parameters, our goal is to apply variational inference to facilitate a pseudo maximum likelihood approach.

Consider employing an EM algorithm, similar to the one seen in the previous chapter, to estimate I-probit models. This time, treat both the latent propensities $\mathbf{y}^*$ and the I-prior random effects $\mathbf{w}$ as 'missing', so the complete data is $\{\mathbf{y}, \mathbf{y}^*, \mathbf{w}\}$. Now, due to the independence of the observations $i = 1, \dots, n$, the complete data log-likelihood is

$$
\begin{aligned}
L(\theta|\mathbf{y}, \mathbf{y}^*, \mathbf{w}) &= \log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta) \\
&= \sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{i\cdot}^*) + \log p(\mathbf{y}^*|\mathbf{w}) + \log p(\mathbf{w}) \\
&= \text{const.} + \cancel{\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})\right) \\
&\quad \cancel{-\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\mathbf{w}\right) \qquad (5.10)
\end{aligned}
$$

{eq:logjointprobit}

which looks like the complete data log-likelihood seen previously in (4.9) (Chapter 4, p. 14), except that here, together with $\mathbf{w}$, $\mathbf{y}_{i\cdot}^*$ is not observed.

For the E-step, it is of interest to determine the posterior density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}) = p(\mathbf{y}^*|\mathbf{w}, \mathbf{y})p(\mathbf{w}|\mathbf{y})$, which we have discerned from the discussion at the beginning of this section that this is hard to obtain, since it involves an intractable marginalising integral. We thus seek a suitable approximation

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

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

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

Under this scheme, the variational distribution for $\mathbf{y}^*$ is found to be a *conically truncated multivariate normal* distribution, and for $\mathbf{w}$, a multivariate normal distribution.

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

$$\log p(\mathbf{y}|\theta) \geq \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q} \log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta) - \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q} \log q(\mathbf{y}^*, \mathbf{w}) =: \mathcal{L}_\theta(q),$$

a quantity often referred to as the *evidence lower bound* (ELBO). It turns out that minimising $\mathrm{KL}(q\|p)$ is equivalent to maximising the ELBO, a quantity that is more practical to work with than the KL divergence, and certainly more tractable than the

log marginal density. Hence, if q approximates the true posterior well, then the ELBO is a suitable proxy for the marginal log-likelihood.

In practice, obtaining parameter estimates which maximise the ELBO and the approximate posterior distribution $q(\mathbf{y}^*, \mathbf{w})$ is achieved using a variational EM algorithm, an EM algorithm in which the conditional distributions are replaced with a variational approximation. The t'th E-step entails obtaining the density $q^{(t+1)}$ as a solution to $\arg\max_q \mathcal{L}_\theta(q)$, keeping θ fixed at the current estimate $\theta^{(t)}$. Let $\bar{\mathbf{y}}^* = \mathbf{y}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top$. The objective function to be maximised is computed as

$$\begin{aligned} Q(\theta) &= \mathrm{E}_{\mathbf{y}^*, \mathbf{w} \sim q^{(t+1)}} \left[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w} | \theta) \,|\, \theta^{(t)} \right] \\ &= \text{const.} - \frac{1}{2} \operatorname{tr} \Big(\boldsymbol{\Psi} \, \mathrm{E}[\mathbf{w}^\top \mathbf{H}_\eta^2 \mathbf{w}] + \boldsymbol{\Psi}^{-1} \, \mathrm{E}[\mathbf{w}^\top \mathbf{w}] \Big) \\ &\quad - \frac{1}{2} \operatorname{tr} \Big(\boldsymbol{\Psi} \big(\mathrm{E}[\mathbf{y}^{*\top} \mathbf{y}^*] + n \boldsymbol{\alpha} \boldsymbol{\alpha}^\top - 2 \boldsymbol{\alpha} \mathbf{1}_n \, \mathrm{E}[\mathbf{y}^*] - 2 \, \mathrm{E}[\mathbf{w}]^\top \mathbf{H}_\eta (\mathrm{E}[\mathbf{y}^*] - \mathbf{1}_n \boldsymbol{\alpha}^\top) \big) \Big), \end{aligned} \tag{5.11}$$

{eq:iprobit QEstep}

and this is maximised with respect to θ in the M-step to obtain $\theta^{(t+1)}$ The algorithm alternates between the E- and M-step until convergence of the ELBO. A full derivation of the variational EM algorithm used by us will be described in Section 5.4.

5.3.3 Markov chain Monte Carlo methods

Markov chain Monte Carlo (MCMC) methods is the tool of choice for a complete Bayesian analysis of multinomial probit models (McCulloch et al., 2000; Nobile, 1998; McCulloch et al., 2000). 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. That is, assuming a prior distribution on the parameters $\theta \sim p(\theta)$, the model with likelihood given by (5.7) obtains posterior samples $\{\mathbf{y}^{*(t)}, \mathbf{w}^{(t)}, \theta^{(t)}\}_{t=1}^T$ from their respective Gibbs conditional distributions. In particular, $\mathbf{y}^* | \mathbf{y}, \mathbf{w}, \theta$ is distributed according to a truncated multivariate normal, while $\mathbf{w} | \mathbf{y}, \mathbf{y}^*, \theta$ a multivariate normal. These conditional distributions are exactly of the same form as the ones obtained under a variational scheme. The difference is that in MCMC, sampling from posterior distributions is performed, whereas in a variational inference framework, a deterministic update of the variational distributions is performed. As such, a downside to this data augmentation scheme in an MCMC

framework is that it enlarges the variable space by an additional nm dimensions, which is memory inefficient for large n.

The models with likelihood (5.8) or (5.9) after integrating out $\mathbf{w}$ and $\mathbf{y}^*$ respectively, is less demanding for MCMC sampling than the model with likelihood (5.7). However, as mentioned already, (5.8) contains an integral involving a mn-variate normal distribution whose covariance matrix is dense, and as far as we are aware, the Kronecker product structure cannot be exploited for efficiency in sampling. This leaves (5.9), a non-conjugate model whose full conditional densities are not of recognisable form. Hamiltonian Monte Carlo is another possibility, since it does not require conjugacy. For binary models, this is a feasible approach because the class probabilities normal CDFs (see (5.5)), which means that it is doable using off-the-shelf software such as Stan. However, with multinomial responses, the arduous task of computing class probabilities, which involve integration of an at most m-dimensional normal density, must be addressed separately.

5.3.4 Comparison of estimation methods

In this subsection, we utilise a toy binary classification data set which has been simulated according to a spiral pattern, as in Figure 5.3. The predictor variables are X_1 and X_2, each of which are scaled similarly. Following (5.5), the binary I-probit model that is fitted is

$$\begin{aligned} y_i &\sim \text{Bern}(p_i) \\ \Phi^{-1}(p_i) &= \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k \\ w_1, \ldots, w_n &\overset{\text{iid}}{\sim} \text{N}(0, 1), \end{aligned}$$

where h_λ is the (scaled) kernel of the fBm RKHS.

We carry out the three estimation precodures described above (Laplace's method, variational EM, and Hamiltonian MC) to compare parameter estimates, (training) error rates, and runtime. The Laplace and variational EM methods were performed in the **iprobit** package, while Stan was used to code the Hamiltonian MC sampler. Prior choices for the fully Bayesian methods were: 1) a vague normal prior $\lambda \sim \text{N}_+(0, 100)$ for the RKHS scale parameter, and 2) a diffuse prior for the intercept $p(\alpha) \propto \text{const}$. Note that

restriction of λ to the positive orthant is required for identifiability. The results are presented in Table 5.1.

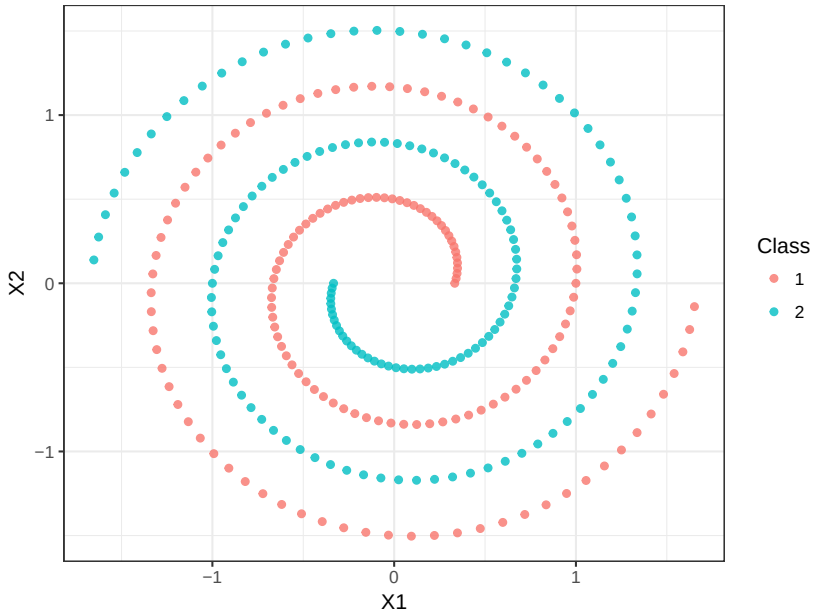

Figure 5.3: A plot of simulated spiral data set.

fig:exampleiprobit

The three methods pretty much concur on the estimation of the intercept, but not on the RKHS scale parameter. As a result, the log-density value at the optima is also different in all three methods. Notice the high posterior standard deviation for the scale parameter in the HMC method. The posterior density for λ was very positively skewed, and this contributed to the large posterior mean.

Table 5.1: Table comparing the estimated parameter values, (marginal) log-likelihood values, and also time taken for the three estimation methods.

tab:compreiprobit

	Laplace approximation	Variational EM	Hamiltonian MC
Intercept (α)	-0.02 (0.03)	0.00 (0.06)	0.00 (0.58)
Scale (λ)	0.85 (0.01)	5.67 (0.23)	29.3 (5.21)
Log density	-202.7	-140.7	-163.8
Error rate (%)	44.7	0.00	0.00
Brier score	0.20	0.02	0.01
Iterations	20	56	2000
Time taken (s)	>3600	5.32	>3600

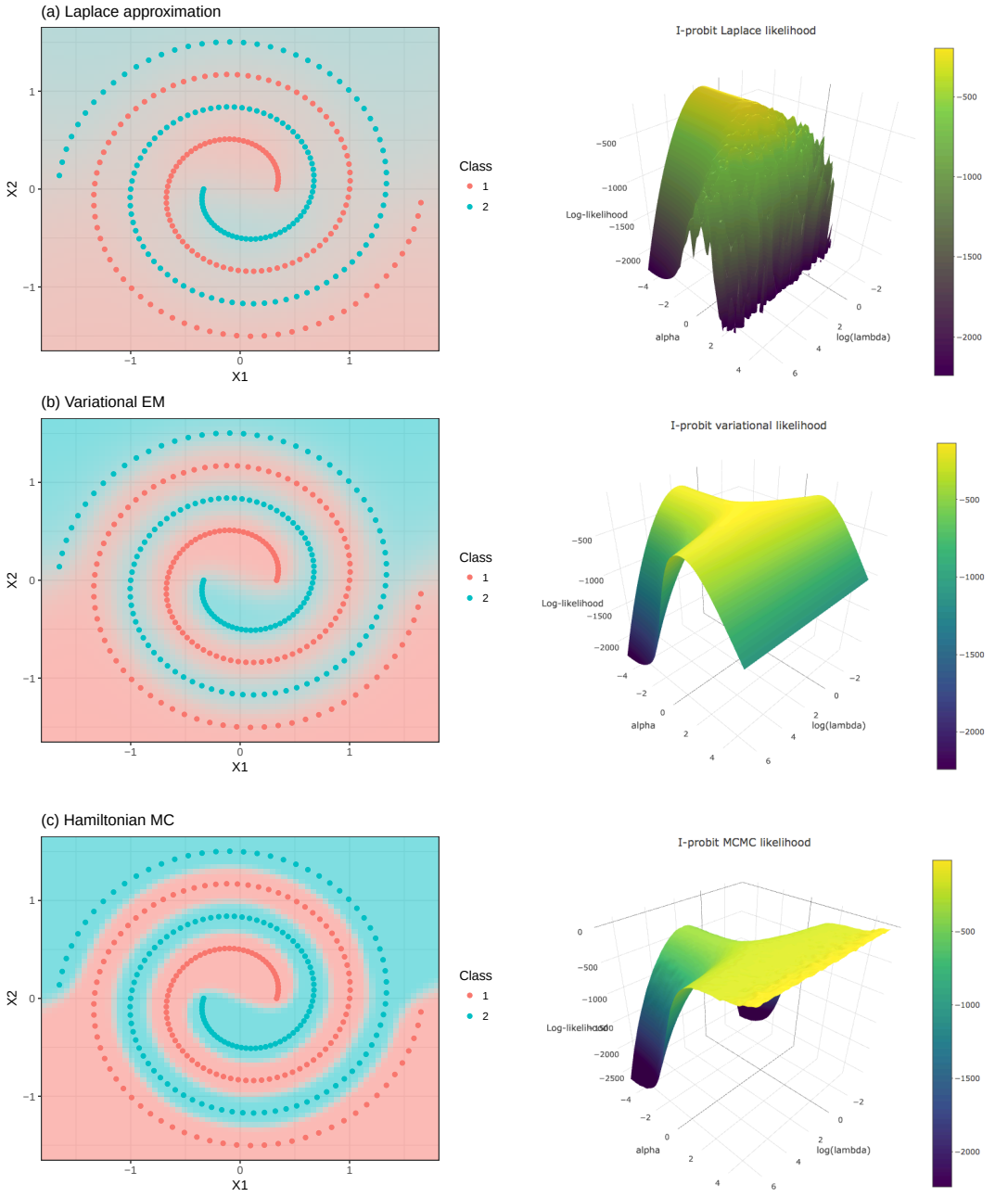

Figure 5.4: Plots showing predicted probabilities (shaded region) for belonging to class 1 or 2 indicated by colour and intensity, and likelihood surface plots for (a) Laplace's method, (b) variational EM, and (c) Hamiltonian MC.

`fig:example iprobitfit`

A plot of the log-likelihood surface for three methods in Figure 5.4 reveals some insight. The variational likelihood has two ridges, with the maxima occurring around the intersection of these two ridges. The Laplace likelihood seems to indicate a similar shape—in both the Laplace and variational method, the posterior distribution of $\mathbf{w}$ is approximated by a Gaussian distribution, with different means and variances. However, parts of the Laplace likelihood are poorly approximated resulting in a loss of fidelity around the supposed maxima, which might have contributed to the set of values that were estimated. Laplace's method is known to yield poor approximations to probit model likelihoods (Kuss and Rasmussen, 2005). On the other hand, the log-likelihood calculated using a Hamiltonian MC sampler (treating parameters as fixed values) yields a slightly different graph: the log-likelihood increases as values of α become larger, resulting in the upwards inflection of the log-likelihood surface (as opposed to a downward inflection seen in the variational and Laplace likelihood).

In terms of predictive abilities, both the variational and Hamiltonian MC methods, even though the posteriors are differently estimated, have good predictive performance as indicated by their error rates and Brier scores[2]. Figure 5.4 shows that HMC is more confident of new data predictions compared to variational inference, as indicated by the intensity of the shaded regions (HMC is shaded stronger than variational EM). Laplace's method gave poor predictive performance.

Finally, on the computational side, variational inference was by far the fastest method to fit the model. Sampling using Hamiltonian MC was very slow, because the parameter space is in effect $O(n+2)$ (parameters are $\{w_1, \ldots, w_n, \alpha, \lambda\}$ under the model with likelihood (5.9), i.e. without the data augmentation scheme). As for Laplace, each Newton step involves obtaining posterior modes of the w_i's, and this contributed to the slowness of this method. The reality is that variational inference takes seconds to complete what either the Laplace or full MCMC methods would take hours to. The predictive performance, while not as good as HMC, is certainly an acceptable compromise in favour of speed.

[2]The Brier score is defined as $\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\hat{p}_{ij})$ with $y_{ij}=1$ if $y_i=j$ and zero otherwise, and $\hat{p}_{ij}$ is the fitted probability of $y_i=j$ occurring. It gives a better sense of "training/test error", compared to simple misclassification rates, by accounting for the forecasted probabilities of the events happening. The Brier score is a proper scoring rule, i.e., it is uniquely minimised by the true probabilities.

sec:iprobit var

5.4 The variational EM algorithm for I-probit models

We present an EM algorithm to estimate the I-probit latent variables $\mathbf{y}^*$ and $\mathbf{w}$, in which the E-step consists of a mean-field variational approximation of the conditional density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}, \theta) = q(\mathbf{y}^*)q(\mathbf{w})$. As per assumptions *A4*, *A5* and *A6*, the parameters of the I-probit model consists of $\theta = \{\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\top, \eta\}$.

The algorithm cycles through a variational inference E-step, in which the variational density $q(\mathbf{y}^*, \mathbf{w}) = \prod_{i=1}^n q(\mathbf{y}_{i\cdot}^*)q(\mathbf{w})$ is optimised with respect to the Kullbeck-Leibler divergence $\mathrm{KL}\left[q(\mathbf{y}^*, \mathbf{w}) \| p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})\right]$, and an M-step, in which the approximate expected joint density (5.11) is maximised with respect to the parameters θ. Convergence is assessed by monitoring the ELBO. Apart from the fact that the variational EM algorithm uses approximate conditional distributions and involves matrices $\mathbf{y}^*$ and $\mathbf{w}$, it is very similar to the EM described in Chapter 4, and as such, the efficient computational work derived there is applicable.

5.4.1 The variational E-step

Let $\tilde{q}(\mathbf{y}^*, \mathbf{w})$ be the pdf that minimises the Kullbeck-Leibler divergence $\mathrm{KL}\left[q\|p\right]$ subject to the mean-field constraint $q(\mathbf{y}^*, \mathbf{w}) = q(\mathbf{y}^*)q(\mathbf{w})$. By appealing to Bishop (2006, equation 10.9, p. 466), the optimal mean-field variational density $\tilde{q}$ for the latent variables $\mathbf{y}^*$ and $\mathbf{w}$ satisfy

$$\log \tilde{q}(\mathbf{y}^*) = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}\left[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})\right] + \text{const.} \tag{5.12}$$

$$\log \tilde{q}(\mathbf{w}) = \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}\left[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})\right] + \text{const.} \tag{5.13}$$

where $p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}) = p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w})$ is as per (5.7). We now present the variational densities $\tilde{q}(\mathbf{y}^*)$ and $\tilde{q}(\mathbf{w})$. For further details on the derivation of these densities, please refer to the appendix.

Variational distribution for the latent propensities $\mathbf{y}^*$

The fact that the rows $\mathbf{y}_{i\cdot}^* \in \mathbb{R}^m$, $i = 1, \ldots, n$ of $\mathbf{y}^* \in \mathbb{R}^{n\times m}$ are independent can be exploited, and this results in a further 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 $\tilde{\boldsymbol{\mu}}_{i\cdot} = \boldsymbol{\alpha} + \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i)$, where $\tilde{\mathbf{w}} = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}\,\mathbf{w}$, and variance $\boldsymbol{\Psi}^{-1}$,

subject to a truncation of its components to the set $\mathcal{C}_{y_i}$. That is, for each $i = 1, \ldots, n$ and noting the observed categorical response $y_i \in \{1, \ldots, m\}$ for the i'th observation, the $\mathbf{y}_i^*$'s are distributed according to

$$\mathbf{y}_{i\cdot}^* \overset{\text{iid}}{\sim} \begin{cases} \mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}) & \text{if } y_{iy_i}^* > y_{ik}^*, \forall k \neq y_i \\ 0 & \text{otherwise.} \end{cases} \tag{5.14}$$

{eq:ystardist}

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

The required expectation $\tilde{\mathbf{y}}^* := \mathrm{E}_{\mathbf{y}^* \sim \tilde{q}}\, \mathbf{y}_{i\cdot}^* = \mathrm{E}_{\mathbf{y}^* \sim \tilde{q}}(y_{i1}^*, \ldots, y_{im}^*)^\top$ in the M-step can be tricky to obtain. One strategy that can be considered is Monte Carlo integration: using samples from $\mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1})$, 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 (Robert, 1995) for sampling from a truncated multivariate normal can be implemented, and this is detailed in the appendix.

If the independent I-probit model is under consideration, whereby the covariance matrix has the independent structure $\boldsymbol{\Psi} = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_m^{-2})$, then the first moment can be considered component-wise. Each component of this expectation is given by

$$\tilde{y}_{ik}^* = \begin{cases} \tilde{\mu}_{ik} - \sigma_k C_i^{-1} \displaystyle\int \phi_{ik}(z) \prod_{l \neq k, y_i} \Phi_{il}(z) \phi(z)\, \mathrm{d}z & \text{if } k \neq y_i \\ \tilde{\mu}_{iy_i} - \sigma_{y_i} \sum_{k \neq y_i} \left(\tilde{y}_{ik}^* - \tilde{\mu}_{ik} \right) & \text{if } k = y_i \end{cases} \tag{5.15}$$

{eq:ystarupdate}

with

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

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

Variational distribution for the I-prior random effects w

Given that both $\operatorname{vec}\mathbf{y}^*|\operatorname{vec}\mathbf{w}$ and $\operatorname{vec}\mathbf{w}$ are normally distributed as per the model (5.4), we find that the full conditional distribution $p(\mathbf{w}|\mathbf{y}^*,\mathbf{y}) \propto p(\mathbf{y}^*,\mathbf{y},\mathbf{w}) \propto p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w})$ is also normal. The variational density q for $\operatorname{vec}\mathbf{w} \in \mathbb{R}^{nm}$ is found to be Gaussian with mean and precision given by

$$\operatorname{vec}\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w(\boldsymbol{\Psi}\otimes\mathbf{H}_\eta)\operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = \boldsymbol{\Psi}\otimes\mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1}\otimes\mathbf{I}_n = \mathbf{V}_{y^*}. \tag{5.16}$$

{eq:varipostw}

As a computational remark, computing the inverse $\tilde{\mathbf{V}}_w^{-1}$ presents a challenge, as this takes $O(n^3m^3)$ time if computed naïvely. 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 Section X for details. Storage requirement is $O(n^2m^2)$, as a result of the covariance matrix in (5.16).

If the independent I-probit model is assumed, i.e. $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1,\dots,\psi_m)$, then the posterior covariance matrix $\tilde{\mathbf{V}}_w$ has a simpler structure which implies column independence in the matrix $\mathbf{w}$. 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

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

We note the similarity between (5.16) above and the posterior distribution for the I-prior random effects in a normal model (4.7) seen in the previous chapter, with the difference being (5.16) uses the continuous latent propensities $\mathbf{y}^*$ instead of the the observations $\mathbf{y}$. The consequence of this is that the posterior regression functions are class independent, the exact intended effect by specifying a diagonal precision matrix $\boldsymbol{\Psi}$. Storage requirement is $O(n^2m)$, since we need $\mathbf{V}_{w_1},\dots,\mathbf{V}_{w_m}$.

Remark 5.2. The variational distribution $q(\mathbf{w})$ which approximates $p(\mathbf{w}|\mathbf{y})$ is in fact exactly $p(\mathbf{w}|\mathbf{y}^*)$, the conditional density of the I-prior random effects given the latent

propensities. By the law of total expectations,

$$\mathrm{E}[r(\mathbf{w})|\mathbf{y}] = \mathrm{E}_{\mathbf{y}^*}\left[\,\mathrm{E}[r(\mathbf{w})|\mathbf{y}^*]\,|\,\mathbf{y}\right],$$

where $r(\cdot)$ is some function of $\mathbf{w}$, and expectations are taken under the posterior distribution of $\mathbf{y}^*$. Hypothetically, if the true pdf $p(\mathbf{y}^*|\mathbf{y})$ were tractable, then the E-step can be computed using the true conditional distribution. Since it is not tractable, we resort to an approximation, and in the case of a variational approximation, (5.16) is obtained.

5.4.2 The M-step

sec:varupdeta

From (5.11), the function to be maximised in the M-step is

$$\begin{aligned} Q(\theta) = \text{const.} &- \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\right) \\ &- \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Psi}\Big(\,\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*] + n\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\tilde{\mathbf{y}}^{*\top}\mathbf{1}_n\boldsymbol{\alpha}^\top - 2\tilde{\mathbf{w}}^\top\mathbf{H}_\eta(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\Big)\right), \quad (5.11) \end{aligned}$$

where expectations are taken with respect to the variational distributions of $\mathbf{y}^*$ and $\mathbf{w}$. Note that since $\boldsymbol{\Psi}$ is treated as fixed, the term $\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*]$ is absorbed into the constant. On closer inspection, the trace involving the second moments of $\mathbf{w}$ is found to be

$$\operatorname{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\right) = \sum_{i,j=1}^{m}\left\{\psi_{ij}\operatorname{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}_{ij}) + \psi_{ij}^{-}\operatorname{tr}(\tilde{\mathbf{W}}_{ij})\right\}$$

by the results of equation derived in the appendix. In the above, we had defined ψ_{ij}^{-} to be the (i,j)'th element of $\boldsymbol{\Psi}^{-1}$, and

$$\tilde{\mathbf{W}}_{ij} = \mathrm{E}[\mathbf{w}_{\cdot i}\mathbf{w}_{\cdot j}^\top] = \mathbf{V}_w[i,j] + \tilde{\mathbf{w}}_{\cdot i}\tilde{\mathbf{w}}_{\cdot j}^\top,$$

where $\mathbf{V}_w[i,j] \in \mathbb{R}^{n\times n}$ refers to the (i,j)'th submatrix block of $\mathbf{V}_w$, and the n-vector $\tilde{\mathbf{w}}_{\cdot j} = \big(\mathrm{E}[w_{ij}]\big)_{i=1}^n$ is the expected value of the random effects for class j. Specifically, when the error precision is of the form $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1,\ldots,\psi_m)$, this trace reduces to

$$\begin{aligned} \operatorname{tr}\left(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\right) &= \sum_{j=1}^{m}\left\{\psi_j\operatorname{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}_{jj}) + \psi_j^{-1}\operatorname{tr}(\tilde{\mathbf{W}}_{jj})\right\} \\ &= \sum_{j=1}^{m}\operatorname{tr}\big(\overbrace{(\psi_j\mathbf{H}_\eta^2 + \psi_j^{-1}\mathbf{I}_n)}^{\boldsymbol{\Sigma}_{\theta,j}}\tilde{\mathbf{W}}_{jj}\big) \end{aligned}$$

The bulk of the computational effort required to evaluate $Q(\theta)$ stems from the trace involving the second moments of $\mathbf{w}$, and the fact that $\mathbf{H}_\eta^2$ needs to be reevaluated each time $\theta = \{\boldsymbol{\alpha}, \eta\}$ changes. As discussed previously, each E-step takes $O(n^3 m)$ time to compute the required first and second (approximate) posterior moments of $\mathbf{w}$. Once this is done, we can use the 'front-loading of the kernel matrices' trick described in Section 4.3.2, which effectively renders the evaluation of Q to be linear in θ (after an initial $O(n^2)$ procedure at the beginning).

As in the normal linear model, we employ a sequential update of the parameters (à la expectation conditional maximisation algorithm) by solving the first order conditions

$$\frac{\partial}{\partial \eta} Q(\eta|\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}\left(\frac{\partial \mathbf{H}_\eta^2}{\partial \eta}\tilde{\mathbf{W}}_{ij}\right) + \operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top \frac{\partial \mathbf{H}_\eta}{\partial \eta}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \boldsymbol{\alpha}^\top)\right) \tag{5.17}$$ {eq:vemeta}

$$\frac{\partial}{\partial \boldsymbol{\alpha}} Q(\boldsymbol{\alpha}|\eta) = 2n\boldsymbol{\Psi}\boldsymbol{\alpha} - 2\sum_{i=1}^{n} \boldsymbol{\Psi}\big(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i)\big) \tag{5.18}$$ {eq:vemalpha}

equated to zero, where $\mathbf{h}_\eta(x_i) \in \mathbb{R}^n$ is the i'th row of the kernel matrix $\mathbf{H}_\eta$. We now present the update equations for the parameters.

Update for kernel parameters η

When only ANOVA RKHS scale parameters are involved, then the conditional solution of η to (5.17) 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 $\eta = \{\lambda_1, \ldots, \lambda_p\}$ need be estimated, and for each $k = 1, \ldots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k \mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}_\eta^2 = \lambda_k^2 \mathbf{R}_k^2 + \lambda_k \mathbf{U}_k + \mathbf{S}_k^2$. As a follow-on from (5.17), the conditional solution for λ_k given the rest of the parameters is obtained by solving

$$\begin{aligned}
\frac{\partial}{\partial \lambda_k} Q(\lambda_k|\boldsymbol{\alpha}, \boldsymbol{\lambda}_{-k}) &= -\frac{1}{2}\sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}\left((2\lambda_k \mathbf{R}_k^2 + \mathbf{U}_k)\tilde{\mathbf{W}}_{ij}\right) + \operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top \mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) \\
&= -\lambda_k \sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}(\mathbf{R}_k^2 \tilde{\mathbf{W}}_{ij}) - \frac{1}{2}\sum_{i,j=1}^{m} \psi_{ij} \operatorname{tr}(\mathbf{U}_k \tilde{\mathbf{W}}_{ij}) \\
&\quad + \operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top \mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) \\
&= 0.
\end{aligned}$$

This yields the solution

$$\hat{\lambda}_k = \frac{\operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top \mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) - \frac{1}{2}\sum_{i,j=1}^m \psi_{ij}\operatorname{tr}(\mathbf{U}_k\tilde{\mathbf{W}}_{ij})}{\sum_{i,j=1}^m \psi_{ij}\operatorname{tr}(\mathbf{R}_k^2\tilde{\mathbf{W}}_{ij})}$$

In the case of the independent I-probit model, where $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_m)$, $\hat{\lambda}_k$ has the form

$$\hat{\lambda}_k = \frac{\sum_{j=1}^m \psi_j \left(\tilde{\mathbf{w}}_{\cdot j}^\top \mathbf{R}_k(\tilde{\mathbf{y}}^*_{\cdot j} - \alpha_j\mathbf{1}_n) - \frac{1}{2}\operatorname{tr}(\mathbf{U}_k\tilde{\mathbf{W}}_{jj})\right)}{\sum_{j=1}^m \psi_j \operatorname{tr}(\mathbf{R}_k^2\tilde{\mathbf{W}}_{jj})}.$$

Remark 5.3. There is no closed-form solution for η when the polynomial kernel is used, or when there are kernel parameters to optimise (e.g. Hurst coefficient or SE kernel lengthscale). In these situations, solutions for η are obtained using numerical methods (i.e. employ quasi-Newton methods such as L-BFGS algorithm for optimising $Q(\eta|\boldsymbol{\alpha})$).

Update for intercepts $\boldsymbol{\alpha}$

It is easy to see that the unique solution to (5.18) is

$$\hat{\boldsymbol{\alpha}} = \frac{1}{n}\boldsymbol{\Psi}^{-1}\left(\sum_{i=1}^n \boldsymbol{\Psi}\left(\mathbf{y}^*_{i\cdot} - \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i)\right)\right) = \frac{1}{n}\sum_{i=1}^n \left(\mathbf{y}^*_{i\cdot} - \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i)\right) \in \mathbb{R}^m.$$

Being free of $\boldsymbol{\Psi}$, the solution is the same whether the full or independent I-probit model is assumed. Furthermore, we must have that $\sum_{j=1}^m \alpha_j = 0$ for identifiability, so as an additional step to satisfy this condition, the solution $\boldsymbol{\alpha}$ is centred.

5.4.3 Summary

A summary of the variational EM algorithm is presented. Notice that the evaluation of each component of the posterior depends on knowing the posterior distribution of the other, i.e. $q(\mathbf{y}^*)$ depends on $q(\mathbf{w})$ and vice-versa. Similarly, each parameter update is obtained conditional upon the value of the rest of the parameters. These circular dependencies are dealt with by way of an iterative updating scheme: with arbitrary starting values for the distributions $q^{(0)}(\mathbf{y}^*)$ and $q^{(0)}(\mathbf{w})$, and for the parameters $\theta^{(0)}$, each are updated in turn according to the above derivations.

The updating sequence is repeated until no significant increase in the convergence criterion, the ELBO, is observed. The ELBO for the I-probit model is given by the

quantity

$$\mathcal{L}_q(\theta) = \frac{nm}{2} + \sum_{i=1}^{n} \log C_i(\theta) + \frac{1}{2}\log|\tilde{\mathbf{V}}_w| - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i,j=1}^{m} \psi_{ij}^{-} \operatorname{tr}(\tilde{\mathbf{W}}_{ij}), \tag{5.19}$$

where $C_i(\theta)$ is the normalising constant of the distribution ${}^{\mathrm{t}}\mathrm{N}_m(\boldsymbol{\alpha}+\mathbf{w}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, with $\mathcal{C}_{y_i} = \{y^*_{iy_i} > y^*_{ik} | \forall k \neq y_i\}$. That is,

$$C_i(\theta) = \int_{\{y^*_{iy_i} > y^*_{ik} | \forall k \neq y_i\}} \cdots \int \phi(y^*_{i1}, \ldots, y^*_{im} | \boldsymbol{\alpha} + \mathbf{w}^\top \mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}y^*_{i1} \cdots \mathrm{d}y^*_{im}.$$

Similar to the EM algorithm, each iteration of the algorithm increases the ELBO to a stationary point (Blei et al., 2017). Unlike the EM algorithm though, the variational EM algorithm does *not* guarantee an increase in the marginal log-likelihood at each step, nor does it guarantee convergence to the global maxima of the log-likelihood.

Further, the ELBO expression to be maximised is often not convex, which means the CAVI algorithm may terminate at local modes, for which they may be many. Note that the variational distribution with the higher ELBO value is the distribution that is closer, in terms of the KL divergence, to the true posterior distribution. In our experience, multiple random starts alleviates this issue for the I-probit model.

5.5 Post-estimation

Post-estimation procedures such as obtaining predictions for a new data point, the credibility interval for such predictions, and model comparison, are of interest. These are performed in a empirical Bayes manner using the variational posterior density of the regression function obtained from the output of the variational EM algorithm.

We first describe prediction of a new data point x_{new}. Step one is to determine the distribution of the posterior regression functions in each class, $\mathbf{f}(x_{\text{new}}) = \mathbf{w}^\top \mathbf{h}_\eta(x_{\text{new}})$, given values for the parameters θ of the I-probit model. To this end, we use the ELBO estimates for θ, i.e. $\hat{\theta} = \arg\max_\theta \mathcal{L}_q(\theta)$, as obtained from the variational EM algorithm. As we know, the variational distribution of $\operatorname{vec}\mathbf{w}$ is normally distributed with mean and variance according to (5.16). By writing $\operatorname{vec}\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_{\cdot 1}, \ldots, \tilde{\mathbf{w}}_{\cdot m})^\top$ to separate out the I-prior random effects per class, we have that $\mathbf{w}_{\cdot j}|\hat{\theta} \sim \mathrm{N}_n(\tilde{\mathbf{w}}_{\cdot j}, \tilde{\mathbf{V}}_w[j,j])$, and $\operatorname{Cov}(\mathbf{w}_{\cdot j}, \mathbf{w}_{\cdot k}) = \tilde{\mathbf{V}}_w[j,k]$, where the '$[\cdot,\cdot]$' indexes the $n \times n$ sub-block of the block

alg:varemip robit

Algorithm 1 Variational EM for the I-probit model (fixed $\boldsymbol{\Psi}$)

procedure INITIALISATION
  Initialise $\theta^{(0)} \leftarrow \{\boldsymbol{\alpha}^{(0)}, \eta^{(0)}\}$
  $\tilde{q}^{(0)}(\mathbf{w}) \leftarrow \mathrm{MN}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$
  $\tilde{q}^{(0)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\mathrm{t}}\mathrm{N}_m(\tilde{\boldsymbol{\alpha}}^{(0)}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$
  $t \leftarrow 0$
end procedure

while not converged **do**
  procedure VARIATIONAL E-STEP
    for $i = 1, \ldots, n$ **do** ▷ Update $\mathbf{y}^*$
      $\tilde{q}^{(t+1)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\mathrm{t}}\mathrm{N}_m\left(\tilde{\boldsymbol{\alpha}}^{(t)} + \tilde{\mathbf{w}}^{(t)\top}\mathbf{h}_{\eta^{(t)}}(x_i), \boldsymbol{\Psi}, \mathcal{C}_{y_i}\right)$
      $\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} \leftarrow \mathrm{E}_{q^{(t+1)}}[\mathbf{y}^*_{i\cdot}]$
    end for
    $\tilde{\mathbf{V}}_w^{(t+1)} \leftarrow \left((\boldsymbol{\Psi} \otimes \mathbf{H}^2_{\eta^{(t)}}) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right)^{-1}$ ▷ Update $\mathbf{w}$
    $\operatorname{vec} \tilde{\mathbf{w}}^{(t+1)} \leftarrow \tilde{\mathbf{V}}_w^{(t+1)}(\boldsymbol{\Psi} \otimes \mathbf{H}_{\eta^{(t)}}) \operatorname{vec}(\tilde{\mathbf{y}}^{*(t+1)} - \mathbf{1}_n \boldsymbol{\alpha}^{(t)\top})$
    $\tilde{q}^{(t+1)}(\mathbf{w}) \leftarrow \mathrm{N}_{nm}\left(\operatorname{vec} \tilde{\mathbf{w}}^{(t+1)}, \tilde{\mathbf{V}}_w^{(t+1)}\right)$
  end procedure

  procedure M-STEP
    if ANOVA kernel (closed-form updates) **then** ▷ Update η
      for $k = 1, \ldots, p$ **do**
        $T_{1k} \leftarrow \sum_{i,j=1}^m \psi_{ij} \operatorname{tr}(\mathbf{R}_k^2 \tilde{\mathbf{W}}_{ij})$
        $T_{2k} \leftarrow \operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top \mathbf{R}_k(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) - \frac{1}{2}\sum_{i,j=1}^m \psi_{ij} \operatorname{tr}(\mathbf{U}_k \tilde{\mathbf{W}}_{ij})$
        $\lambda_k^{(t+1)} \leftarrow T_{2k}/T_{1k}$
      end for
    else
      $\eta^{(t+1)} \leftarrow \arg\max_\eta Q(\eta|\boldsymbol{\alpha}^{(t)})$ by L-BFGS algorithm
    end if
    $\mathbf{a} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} - \tilde{\mathbf{w}}^{(t+1)\top}\tilde{\mathbf{h}}_{\eta^{(t+1)}}(x_i)\right)$ ▷ Update $\boldsymbol{\alpha}$
    $\boldsymbol{\alpha}^{(t+1)} \leftarrow \mathbf{a} - \frac{1}{m}\sum_{j=1}^m a_j$
  end procedure

  Calculate ELBO $\mathcal{L}^{(t+1)}$
  $t \leftarrow t + 1$
end while

matrix structured matrix $\mathbf{V}_w$. Thus, for each class $j = 1, \ldots, m$ and any $x \in \mathcal{X}$,

$$f_j(x)|\mathbf{y}, \hat{\theta} \sim \mathrm{N}\left(\mathbf{h}_{\hat{\eta}}(x)^\top \tilde{\mathbf{w}}_{\cdot j},\, \mathbf{h}_{\hat{\eta}}(x)^\top \tilde{\mathbf{V}}_w[j,j]\mathbf{h}_{\hat{\eta}}(x)\right),$$

and the covariance between the regression functions in two different classes is

$$\mathrm{Cov}\left[f_j(x), f_k(x)|\mathbf{y}, \hat{\theta}\right] = \mathbf{h}_{\hat{\eta}}(x)^\top \tilde{\mathbf{V}}_w[j,k]\tilde{\mathbf{h}}_{\hat{\eta}}(x).$$

Then, in step two, using the results obtained in the previous chapter in Section 4.4, we have that the latent propensities $y^*_{\mathrm{new},j}$ for each class are normally distributed with mean, variance, and covariances

$$\begin{aligned}
\mathrm{E}[y^*_{\mathrm{new},j}|\mathbf{y}, \hat{\theta}] &= \hat{\alpha}_j + \mathrm{E}\left[f_j(x_{\mathrm{new}})|\mathbf{y}, \hat{\theta}\right] &&=: \hat{\mu}_j(x_{\mathrm{new}}) \\
\mathrm{Var}[y^*_{\mathrm{new},j}|\mathbf{y}, \hat{\theta}] &= \mathrm{Var}\left[f(x_{\mathrm{new}})|\mathbf{y}, \hat{\theta}\right] + \boldsymbol{\Psi}^{-1}_{jj} &&=: \hat{\sigma}^2_j(x_{\mathrm{new}}) \\
\mathrm{Cov}[y^*_{\mathrm{new},j}, y^*_{\mathrm{new},k}|\mathbf{y}, \hat{\theta}] &= \mathrm{Cov}\left[f_j(x), f_k(x)|\mathbf{y}, \hat{\theta}\right] + \boldsymbol{\Psi}^{-1}_{jk} &&=: \hat{\sigma}_{jk}(x_{\mathrm{new}}).
\end{aligned}$$

From here, step three would be to extract class information of data point x_{new}, which are contained in the normal distribution $\mathrm{N}_m\left(\hat{\boldsymbol{\mu}}_{\mathrm{new}}, \hat{\mathbf{V}}_{\mathrm{new}}\right)$, where

$$\hat{\boldsymbol{\mu}}_{\mathrm{new}} = \left(\mu_1(x_{\mathrm{new}}), \ldots, \mu_m(x_{\mathrm{new}})\right)^\top \quad \text{and} \quad \hat{\mathbf{V}}_{\mathrm{new},jk} = \begin{cases} \hat{\sigma}^2_j(x_{\mathrm{new}}) & \text{if } j = k \\ \hat{\sigma}_{jk}(x_{\mathrm{new}}) & \text{if } j \neq k. \end{cases}$$

The predicted class is inferred from the latent variables via

$$\hat{y}_{\mathrm{new}} = \arg\max_k \hat{\mu}_k(x_{\mathrm{new}}),$$

while the probabilities for each class are obtained via integration of a multivariate normal density, as per (5.3):

$$\hat{p}_{\mathrm{new},j} = \int_{\{y^*_j > y^*_k | \forall k \neq j\}} \cdots \int \phi(y^*_1, \ldots, y^*_m | \hat{\boldsymbol{\mu}}_{\mathrm{new}}, \hat{\mathbf{V}}_{\mathrm{new}})\, \mathrm{d}y^*_1 \cdots \mathrm{d}y^*_m. \tag{5.20}$$

For the independent I-probit model, class probabilities are obtained in a more compact manner via

$$\hat{p}_{\mathrm{new},j} = \mathrm{E}_Z\left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\hat{\sigma}_j(x_{\mathrm{new}})}{\hat{\sigma}_k(x_{\mathrm{new}})} Z + \frac{\hat{\mu}_j(x_{\mathrm{new}}) - \hat{\mu}_k(x_{\mathrm{new}})}{\hat{\sigma}^2_k(x_{\mathrm{new}})}\right)\right],$$

as per (5.6), since the m components of $\mathbf{f}(x_{\text{new}})$, and hence the $\mathbf{y}^*_{\text{new},j}$'s, are independent of each other ($\boldsymbol{\Psi}$ and $\hat{\mathbf{V}}_{\text{new}}$ are diagonal).

We are able to take advantage of the Bayesian machinery to obtain credibility intervals for probability estimates or any transformation of these probabilities (e.g. log odds or odds ratios). The procedure is as follows. First, obtain samples $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(T)}$ by drawing from its variational posterior distribution $\operatorname{vec} \mathbf{w}^{(i)} | \hat{\theta} \sim \mathrm{N}_{nm}(\operatorname{vec} \tilde{\mathbf{w}}, \mathbf{V}_w)$. Then, obtain samples of class probabilities $\{p^{(1)}_{xj}, \ldots, p^{(T)}_{xj}\}_{j=1}^m$, for a given data point $x \in \mathcal{X}$ by evaluating

$$p^{(t)}_{xj} = \int_{\{y^*_j > y^*_k | \forall k \neq j\}} \cdots \int \phi\big(y^*_1, \ldots, y^*_m | \hat{\boldsymbol{\mu}}^{(t)}(x), \hat{\mathbf{V}}(x)\big) \, \mathrm{d}y^*_1 \cdots \mathrm{d}y^*_m,$$

where $\hat{\boldsymbol{\mu}}^{(t)}(x) = \hat{\boldsymbol{\alpha}} + \mathbf{w}^{(t)\top} \mathbf{h}_{\hat{\eta}}(x)$, and $\hat{\mathbf{V}}(x)_{jk}$ equals $\hat{\sigma}^2_j(x)$ if $j = k$, and $\hat{\sigma}_{jk}(x)$ otherwise. To obtain a statistic of interest, say, a 95% credibility interval of a function $r(p_{xj})$ of the probabilities, simply take the empirical lower 2.5th and upper 97.5th percentile of the transformed sample $\big\{r(p^{(1)}_{xj}), \ldots, r(p^{(T)}_{xj})\big\}$.

Remark 5.4. Unfortunately, with the variational EM algorithm, standard errors for the parameters θ are not so easy to obtain. We could not ascertain as to the availability of an unbiased estimate of the asymptotic covariance matrix for θ under a variational framework. One strategy for obtaining standard errors is bootstrap (Chen et al., 2017):

1. Obtain $\hat{\theta} = \arg\max_\theta \mathcal{L}_q(\theta)$ using $\mathcal{S} = \{(y_1, x_1), \ldots, (y_n, x_n)\}$.
2. For $t = 1, \ldots, T$, do
   (a) Obtain $\mathcal{S}^{(t)} = \{(y^{(t)}_1, x^{(t)}_1), \ldots, (y^{(t)}_n, x^{(t)}_n)\}$ by sampling n points with replacement from $\mathcal{S}$.
   (b) Compute $\hat{\theta}^{(t)} = \arg\max_\theta \mathcal{L}_q(\theta)$ using the data $\mathcal{S}^{(t)}$.
3. For the l-th component of θ, compute its variance estimator using

$$\widehat{\operatorname{Var}}(\hat{\theta}_l) = \frac{1}{T} \sum_{t=1}^{T} (\hat{\theta}^{(t)}_l - \bar{\theta}_l)^2$$

where

$$\bar{\theta}_l = \frac{1}{T} \sum_{t=1}^{T} \hat{\theta}^{(t)}_l.$$

The obvious downside to this is computational time. In any case, inference surrounding kernel parameters or intercepts is not of interest.

Finally, on model selection. It is possible to perform model comparison by comparing the maximised ELBO quantity of several candidate models (Beal and Ghahramani, 2003), and the justification for this is that it supposedly gives a good approximation to the marginal log-likelihood (log model evidence), especially if the variational density is close in the KL divergence sense to the true posterior density. This would allow model selection using (empirical) Bayes factors as a model selection criterion. Kass and Raftery (1995) suggest the following interpretation of observed Bayes factor values for comparing model M_1 against model M_0.

Table 5.2: Guidelines for interpreting Bayes factors.

tab:bf

$2\log \mathrm{BF}(M_1, M_0)$	$\mathrm{BF}(M_1, M_0)$	Evidence against M_0
0–2	1–3	Not worth more than a bare mention
2–6	3–20	Positive
6–10	20–150	Strong
>10	>150	Very strong

where $\mathrm{BF}(M_1, M_0)$ is approximated by

$$\mathrm{BF}(M_1, M_0) \approx \frac{\mathcal{L}_q(\theta|M_1)}{\mathcal{L}_q(\theta|M_0)},$$

and $\mathcal{L}_q(\theta|M_k)$, $k = 0, 1$, is the ELBO for model M_k. It should be noted that while this works in practice, there is no theoretical basis for model comparison using the ELBO (Blei et al., 2017).

5.6 Computational consideration

Computational challenges for the I-probit model stems from two sources: 1) calculation of the class probabilities (5.3); and 2) storage and time requirements for the variational EM algorithm. Ways in which to overcome these challenges are discussed. In addition, we also discuss considerations to take into account if estimation of the error precision $\boldsymbol{\Psi}$ is desired, and thus path a road for future work.

5.6.1 Efficient computation of class probabilities

sec:mnint

The issue at hand here is that for $m > 4$, the evaluation of the class probabilities in (5.3) is computationally burdensome using classical methods such as quadrature methods Geweke et al. (1994).

The simplest strategy to overcome this is a frequency simulator (otherwise known as Monte Carlo integration): obtain random samples from $\mathrm{N}_{m-1}(\boldsymbol{\nu}_{i\cdot}, \boldsymbol{\Omega})$, and calculate how many of these samples fall within the required region. This method is fast and yields unbiased estimates of the class probabilities. However, accuracy of this method is questionable when the mean $\boldsymbol{\nu}_{i\cdot}$ of the multivariate normal is many standard deviations away from zero (the cutoff region as per **??**).

A more reliable method is the probability simulator of Geweke-Hajivassiliou-Keane (GHK) (Geweke, 1991; Hajivassiliou et al., 1996; Michael P Keane, 1994), which we describe now. For clarity, we drop the subscript i denoting individuals, and write $\mathbf{z} = (z_1, \ldots, z_m)$, remembering that $z_1 = 0$. Suppose that an observation $y = j$ has been made. Rewrite the model by anchoring on the j'th latent variable z_j as follows:

$$\tilde{\mathbf{z}} := (\overbrace{z_1 - z_j}^{\tilde{z}_1}, \ldots, \overbrace{z_{j-1} - z_j}^{\tilde{z}_{j-1}}, \overbrace{z_{j+1} - z_j}^{\tilde{z}_{j+1}}, \ldots, \overbrace{z_m - z_j}^{\tilde{z}_m},)^\top \in \mathbb{R}^{m-1}.$$

Let $\boldsymbol{\nu}_{(j)}$ and $\boldsymbol{\Omega}_{(j)}$ be the appropriately transformed mean vector and covariance matrix for $\tilde{\mathbf{z}}$. These are indexed by '(j)' because the transformation is dependent on which latent variable the $\mathbf{z}$'s are anchored on. Since this transformation is linear, $\tilde{\mathbf{z}} \sim \mathrm{N}_{m-1}(\boldsymbol{\nu}_{(j)}, \boldsymbol{\Omega}_{(j)})$. For the symmetric and positive definite matrix $\boldsymbol{\Psi}^{-1}$, obtain its Cholesky decomposition as $\boldsymbol{\Omega}_{(j)} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix. Then, $\tilde{\mathbf{z}} = \boldsymbol{\nu}_{(j)} + \mathbf{L}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \sim \mathrm{N}_{m-1}(\mathbf{0}, \mathbf{I}_{m-1})$. That is,

$$\begin{pmatrix} \tilde{z}_1 \\ \tilde{z}_2 \\ \vdots \\ \tilde{z}_m \end{pmatrix} = \begin{pmatrix} \nu_{(j)1} \\ \nu_{(j)2} \\ \vdots \\ \nu_{(j)m} \end{pmatrix} + \begin{pmatrix} L_{11} & & & \\ L_{21} & L_{22} & & \\ \vdots & \vdots & \ddots & \\ L_{m1} & L_{m2} & \cdots & L_{mm} \end{pmatrix} \begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_m \end{pmatrix} = \begin{pmatrix} \nu_{(j)1} + L_{11}\zeta_1 \\ \nu_{(j)2} + \sum_{k=1}^{2} L_{k2}\zeta_k \\ \vdots \\ \nu_{(j)m} + \sum_{k=1}^{m} L_{km}\zeta_k \end{pmatrix}.$$

With this setup, we can calculate p_j, the probability of class j, which is equivalent to the probability that each $\tilde{z}_k = z_k - z_j < 0$, as follows

$$\begin{aligned} p_j &= \mathrm{P}(\tilde{z}_1 < 0, \ldots, \tilde{z}_{j-1} < 0, \tilde{z}_{j+1} < 0, \ldots, \tilde{z}_m < 0) \\ &= \mathrm{P}(\zeta_1 < u_1, \ldots, \zeta_{j-1} < u_{j-1}, \zeta_{j+1} < u_{j+1}, \ldots, \zeta_m < u_m) \\ &= \mathrm{P}(\zeta_1 < u_1)\, \mathrm{P}(\zeta_2 < u_2 | \zeta_1 < u_1) \cdots \\ &\qquad \cdots \mathrm{P}(\zeta_m < u_m | \zeta_1 < u_1, \ldots, \zeta_{j-1} < u_{j-1}, \zeta_{j+1} < u_{j+1}, \ldots, \zeta_{m-1} < u_{m-1}), \end{aligned}$$

where $u_i = u_i(\zeta_1, \ldots, \zeta_{i-1}) = -(\nu_{(j)i} + \sum_{k=1}^{i-1} L_{ki}\zeta_k)/L_{ii}$. Thus, the integral involving a $(m-1)$-variate normal density **??** is turned into a product of $m-1$ univariate normal cdfs, which can be computed fairly efficiently in modern computer systems.

As an aside, the GHK probability simulator, can be used to sample from a truncated multivariate normal distribution:

- Draw $\tilde{\zeta}_1 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, u_1)$.
- Draw $\tilde{\zeta}_2 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_2)$, where $\tilde{u}_2 = u_2(\tilde{\zeta}_1)$.
- $\cdots$
- Draw $\tilde{\zeta}_{j-1} \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_{j-1})$, where $\tilde{u}_{j-1} = u_{j-1}(\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{j-2})$.
- Draw $\tilde{\zeta}_{j+1} \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_{j+1})$, where $\tilde{u}_{j+1} = u_{j+1}(\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{j-1})$.
- $\cdots$
- Draw $\tilde{\zeta}_m \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_m)$, where $\tilde{u}_m = u_m(\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{j-1}, \tilde{\zeta}_{j+1}, \ldots, \tilde{\zeta}_{m-1})$.

Then, $\tilde{z} = \boldsymbol{\nu}_{(j)}\mathbf{L}\tilde{\zeta}$ will be distributed according to $\mathrm{N}_{m-1}(\boldsymbol{\nu}_{(j)}, \boldsymbol{\Omega}_{(j)})$. Any quantity of interest, e.g. $\mathrm{E}\, r(\tilde{z})$, can then be estimated by the sample mean. In the variational algorithm, we require quantities such as first and second moments and also the entropy of a truncated multivariate normal distribution. Alternative methods are also discussed in the appendix.

Finally, a point on independent probit models. As we alluded to earlier in the chapter, the class probabilities condense to a unidimensional integral involving products of normal cdfs (see (5.6)) if $\boldsymbol{\Psi}$ is diagonal. While this represents a massive simplification, care should be taken when dealing with the formula in (5.6). When at least one of the normal cdfs in the product is extremely small, this can cause loss of significance due to

floating-point errors. In the **iprobit** package, the product of normal cdfs is handled as a sum on the log scale to avoid this issue.

5.6.2 Computational complexity of the CAVI algorithm

sec:complxiprobit

As with the normal I-prior model, the time complexity of the variational inference algorithm for I-probit models is dominated by the step involving the posterior evaluation of the I-prior random effects $\mathbf{w}$, which essentially is the inversion of an $nm \times nm$ matrix. The matrix in question is

$$\mathbf{V}_w = \left[(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right]^{-1}. \qquad \text{(from 5.16)}$$

We can actually 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 $m \ll n$ or if done in parallel, and needs to be performed once per CAVI iteration. Then, manipulate $\mathbf{V}_w^{-1}$ as follows:

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

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. This brings time complexity of the CAVI down to a similar requirement as if $\boldsymbol{\Psi}$ was diagonal.

Storage requirements are still $O(n^2)$, and methods described in the previous chapter are applicable, particularly the discussion surrounding exponential family EM algorithm. Prediction of a new data point is $O(n^2 m)$, because there are essentially m 'separate' normal I-prior regressions, and each take $O(n^2)$ to evaluate.

sec:difficultPsi

5.6.3 Difficulties faced with estimating $\boldsymbol{\Psi}$

Suppose that, alongside the $\mathbf{y}^*$, $\mathbf{w}$, η and $\boldsymbol{\alpha}$ in the CAVI algorithm described in Section 5.4, $\boldsymbol{\Psi}$ is a free parameter to be estimated. If so, we find that the variational density q for $\boldsymbol{\Psi}$ satisfies

$$q(\boldsymbol{\Psi}) \propto \exp\left[-\frac{1}{2} \operatorname{tr}\left(\overbrace{\left(\mathrm{E}[(\mathbf{y}^* - \boldsymbol{\mu})^\top (\mathbf{y}^* - \boldsymbol{\mu})]\right)}^{\mathbf{G}_1} \boldsymbol{\Psi} + \overbrace{\mathrm{E}[\mathbf{w}^\top \mathbf{w}]}^{\mathbf{G}_2} \boldsymbol{\Psi}^{-1} \right) \right] \times p(\boldsymbol{\Psi})$$

where $p(\boldsymbol{\Psi})$ is a prior density chosen for $\boldsymbol{\Psi}$. Unfortunately, this does not resemble any known distribution, regardless of the prior choice for $\boldsymbol{\Psi}$. One can resort to sampling techniques to obtain quantities such as the mean or entropy, which are needed, but this has not been studied for this project due to time limitations. Even if this was possible, this requires, among other things, second moments of a truncated multivariate normal density $\mathbf{y}^*$, and also of $\mathbf{H}_\eta \mathbf{w} \in \mathbb{R}^{n \times m}$—of which both are a bit awkward to obtain.

What we realise, however, is that the *posterior mode* is relatively easy to obtain, especially with an improper prior $p(\boldsymbol{\Psi}) \propto \text{const}$. To see this, we look specifically at the case where $\boldsymbol{\Psi}$ is diagonal. On the log scale,

$$\log q(\psi_j) = \text{const.} - \frac{1}{2} \sum_{j=1}^{m} \psi_j \, \mathrm{E}\|\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j}\|^2 - \frac{1}{2} \sum_{j=1}^{m} \psi_j^{-1} \, \mathrm{E}\|\mathbf{w}_{\cdot j}\|^2$$

is maximised, for $j = 1, \ldots, m$, at

$$\hat{\psi}_j = \sqrt{\frac{\mathrm{E}\|\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j}\|^2}{\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2}}.$$

Perhaps, if the posterior mean is close to the mode, and not withstanding the involved calculations of the require second moments, then this quantity can be used instead in the CAVI algorithm. This ties with the idea of *variational Bayes EM algorithm*, which is an alternative to a fully Bayesian treatment of variational inference. This is discussed in **????**, but unfortunately, time constraints had made it impossible for us to examine this within the scope of this thesis.

sec:iprobit
eg

5.7 Examples

5.8 Conclusion

This work presents an extension of the normal I-prior methodology to fit categorical response models using probit link functions—a methodology we call the I-probit. The main motivation behind this work is to overcome the drawbacks of modelling probabilities using the normal I-prior model. We assumed latent variables that represent 'class propensities' exist, modelled these using a normal I-prior, and simply transformed them via a probit link function. In this way, all of the advantages of the I-prior methodology seen for the normal model are preserved for binary and multinomial regression as well.

The core of this work explores ways in which to overcome the intractable integral presented by the I-probit model in **??**. Techniques such as quadrature methods, Laplace approximation and MCMC tend to fail, or are unsatisfactorily slow to accomplish. The main reason for this is the dimension of this integral, which is nm, and thus for large sample sizes and/or number of classes, is unfeasible with such methods. We turned to variational inference in the face of an intractable posterior density that hampers an EM algorithm, and the result is a sequential updating scheme, similar in time and storage requirements to the EM algorithm.

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. However, Gaussian process classification typically uses the logistic sigmoid function, and estimation most commonly performed using 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.

Suggestions for future work include:

1. **Estimation of $\boldsymbol{\Psi}$**. A limitation we had to face in this work was to treat $\boldsymbol{\Psi}$ as fixed. This limitation was in part due to the non-conjugate nature of the variational density for $\boldsymbol{\Psi}$. We believe the variational Bayes EM algorithm, which estimates maximum a posteriori values for the parameters, could alleviate this issue. This would bring the estimation procedure on par with the frequentist objective of maximum likelihood via the EM algorithm, albeit with the use of approximate posterior densities (see **????** for further discussions).

2. **Inclusion of class-specific covariates**. Throughout the chapter, we assumed that covariates were unit-specific, rather than class-specific. One such example is modelling the choice of travel mode between two destinations (car, coach, train or aeroplane) as a function of travel time. Clearly, travel time depends on the mode of transport. This would require a careful rethink of the appropriate RKHS/RKKS to which the regression function belongs: the regression on the latent propensities could be extended as such:

$$y_{ij}^* = \alpha_j + f_j(x_i) + e(z_{ij})$$

   and $f_j \in \mathcal{F}_{\mathcal{X}}$, the RKHS with kernel $h : (\mathcal{X} \times \mathcal{M})^2 \to \mathbb{R}$ defined by $\delta_{jj'}h(x, x')$, and $e \in \mathcal{F}_{\mathcal{Z}}$, the RKHS of functions of the form $e : \{z_{ij} | i = 1, \ldots, n,\ j = 1, \ldots, m\} \to \mathbb{R}$. An I-prior would then be applied as usual, but the implications on the estimation would need to be considered as well.

3. **Improving computational efficiency**. The $O(n^3m)$ time requirement for estimating I-probit models hinder its use towards large-data applications. In a limited study, we did not obtain reliable improvements using low-rank approximations of the kernel matrix such as the Nyström method. The key to improving computational efficiency could lie in sparse variational methods, a suggestion that was made to improve normal I-prior models as well.

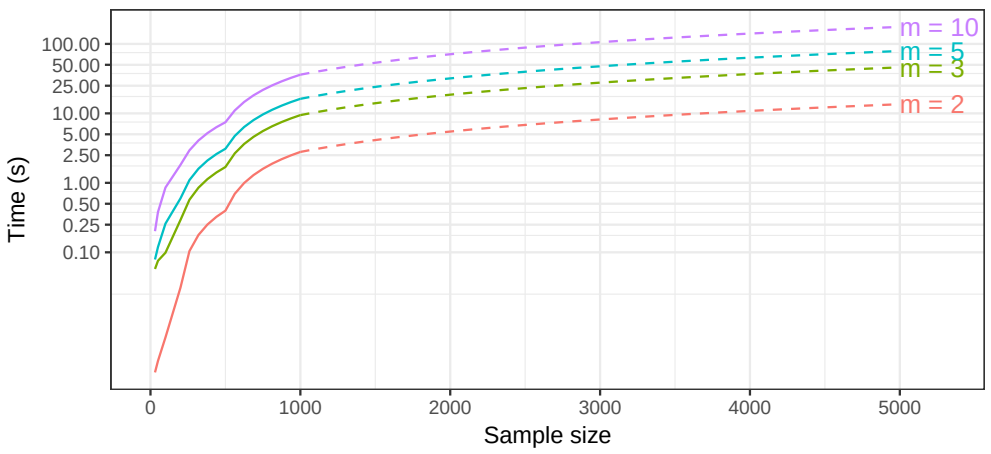

Figure 5.5: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes m. The solid line represents actual timings, while the dotted lines are linear extrapolations.

5.9 Miscellanea

Appendix

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

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

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.

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

- $\boldsymbol{p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \boldsymbol{\eta}, \boldsymbol{\Psi})}$. 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

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

  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 $\mathbf{w} \sim \mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$ 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 $\eta = \{\lambda_1, \ldots, \lambda_p\}$ 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

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

  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 $l, c \sim \Gamma(a, b)$.

- $\boldsymbol{p(\Psi)}$. Our analysis shows that regardless of prior choice of $\boldsymbol{\Psi}$, be it in the full or independent I-probit model, the posterior for $\boldsymbol{\Psi}$ will not be of a recognisable form. Without giving too much thought, assume an improper prior on $\boldsymbol{\Psi}$, i.e. $p(\boldsymbol{\Psi}) \propto$ const.

- $\boldsymbol{p(\alpha)}$. Choose independent normal priors for the intercept, $\alpha_j \sim \mathrm{N}(a_j, A_j)$ for $j = 1, \ldots, m$. 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 5.5. The priors on the parameters $\{\boldsymbol{\alpha}, \eta\}$ can be set to very vague or even improper priors, and the resulting posterior will still yield a proper distribution. Using improper priors eases the algebra slightly. For the precision matrix $\boldsymbol{\Psi}$, it is best to stick with the Wishart prior to avoid positive-definite issues, unless the independent I-probit model is used, in which case Jeffreys' prior for the precisions $p(\sigma_j^{-2}) \propto \sigma_j^2$ is a convenient choice.

5.10.1 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 $j \in \{1, \ldots, m\}$.

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

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

where $\tilde{\boldsymbol{\mu}}_i = \mathrm{E}\,\boldsymbol{\alpha} + (\mathrm{E}\,\mathbf{H}_\eta\,\mathrm{E}\,\mathbf{w})_i$, 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 **??** provides a simplification.

Remark 5.6. 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,

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

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

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

The terms involving $\mathbf{w}$ in **??** 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

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

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,

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

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 $(\mathrm{E}\,\mathbf{A})^{-1} \approx \mathrm{E}[\mathbf{A}^{-1}]$[3],

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

Ideally, we do not want to work with the $nm \times nm$ matrix $\mathbf{V}_w$, since its inverse is expensive to compute. Refer to Section 5.6.2 for details.

[3] 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. This approximation has been used also by Girolami and Rogers (2006) in their work.

In the case of the I-probit model, where $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \ldots, \psi_m)$, then the covariance matrix takes a simpler form. Specifically, it has the block diagonal structure:

$$\begin{aligned}
\tilde{\mathbf{V}}_w &= \mathrm{E}\left[\operatorname{diag}(\psi_1, \ldots, \psi_m) \otimes \mathbf{H}_\eta^2 + \operatorname{diag}(\psi_1, \ldots, \psi_m) \otimes \mathbf{I}_n\right]^{-1} \\
&= \operatorname{diag}\left(\mathrm{E}\left(\psi_1 \mathbf{H}_\eta^2 + \psi_1^{-1}\mathbf{I}_n\right)^{-1}, \cdots, \mathrm{E}\left(\psi_m \mathbf{H}_\eta^2 + \psi_m^{-1}\mathbf{I}_n\right)^{-1}\right) \\
&\approx \operatorname{diag}\left(\left(\tilde{\psi}_1 \tilde{\mathbf{H}}_\eta^2 + \tilde{\psi}_1^{-1}\mathbf{I}_n\right)^{-1}, \cdots, \left(\tilde{\psi}_m \tilde{\mathbf{H}}_\eta^2 + \tilde{\psi}_m^{-1}\mathbf{I}_n\right)^{-1}\right) \\
&=: \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \ldots, \tilde{\mathbf{V}}_{w_m}).
\end{aligned}$$

The mean $\operatorname{vec} \tilde{\mathbf{w}}$ is

$$\begin{aligned}
\operatorname{vec} \tilde{\mathbf{w}} &= \tilde{\mathbf{V}}_w (\operatorname{diag}(\tilde{\psi}_1, \ldots, \tilde{\psi}_m) \otimes \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top) \\
&= \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \ldots, \tilde{\mathbf{V}}_{w_m}) \operatorname{diag}(\tilde{\psi}_1 \tilde{\mathbf{H}}_\eta, \ldots, \tilde{\psi}_m \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top) \\
&= \operatorname{diag}(\tilde{\psi}_1 \tilde{\mathbf{V}}_{w_1} \tilde{\mathbf{H}}_\eta, \ldots, \tilde{\psi}_m \tilde{\mathbf{V}}_{w_m} \tilde{\mathbf{H}}_\eta)(\tilde{\mathbf{y}}^* - \mathbf{1}_n \tilde{\boldsymbol{\alpha}}^\top) \\
&\qquad\qquad \tilde{\mathbf{w}}_{\cdot 1} \qquad\quad \cdots \qquad\quad \tilde{\mathbf{w}}_{\cdot m} \\
&= \begin{pmatrix} \tilde{\psi}_1 \tilde{\mathbf{V}}_{w_1} \tilde{\mathbf{H}}_\eta (\tilde{\mathbf{y}}^*_{\cdot 1} - \tilde{\alpha}_1 \mathbf{1}_n) & \cdots & \tilde{\psi}_m \tilde{\mathbf{V}}_{w_m} \tilde{\mathbf{H}}_\eta (\tilde{\mathbf{y}}^*_{\cdot m} - \tilde{\alpha}_m \mathbf{1}_n) \end{pmatrix}^\top.
\end{aligned}$$

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

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

A quantity that we will be requiring time and again will be $\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}])$, 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

$$\begin{aligned}
\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}]) &= \sum_{i,j=1}^{m} \mathbf{C}_{ij}\, \mathrm{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}]_{ij} \\
&= \sum_{i,j=1}^{m} \mathbf{C}_{ij}\, \mathrm{E}[\mathbf{w}_{\cdot i}^\top \mathbf{D} \mathbf{w}_{\cdot j}]. \qquad (5.21)
\end{aligned}$$

{eq:trCEwDw}

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

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

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 (5.21) leads us to

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

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

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

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

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

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

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.

Algorithm 2 Metropolis random-walk to sample η

1: **inputs** $\tilde{\boldsymbol{\alpha}}$, $\tilde{\mathbf{w}}$, $\tilde{\boldsymbol{\Psi}}$, and s Metropolis sampling s.d.
2: **initialise** $\eta^{(0)} \in \mathbb{R}^q$ and $t \leftarrow 0$
3: **for** $t = 1, \ldots, T$ **do**
4: Draw $\eta^* \sim \mathrm{N}_q(\eta^{(t)}, s^2)$
5: 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

$$\pi_{\text{acc}} = \min\left(1, \exp\left(\log \tilde{q}(\eta^*) - \log \tilde{q}(\eta^{(t)})\right)\right).$$

6: **end for**
7: $\tilde{\mathbf{H}}_\eta \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf{H}_{\eta^{(t)}}$ and $\tilde{\mathbf{H}}_\eta^2 \leftarrow \frac{1}{T}\sum_{i=1}^T \mathbf{H}_{\eta^{(t)}}^2$

Now consider the case where $\eta = \{\lambda_1, \ldots, \lambda_p\}$ (RKHS scale parameters only), and the scenario described in the exponential family EM algorithm of Section 4.3.3 applies. In particular, for $k = 1, \ldots, p$, 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 $j = 1, \ldots, m$, assuming each of

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

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

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 $\lambda_k \mapsto \mathrm{E}_{\lambda_k\sim q}[\lambda_k]$ in the expression $\mathbf{H}_\eta = \lambda_k\mathbf{R}_k + \mathbf{S}_k$. However, in the calculation of $\tilde{\mathbf{H}}_\eta^2$ using $\lambda_k^2\mathbf{R}_k^2 + \lambda_k\mathbf{U}_k + \mathbf{S}_k^2$, we must replace occurrences of λ_k^2 with $\mathrm{E}_{\lambda_k\sim q}[\lambda_k]^2 + \mathrm{Var}_{\lambda_k\sim q}[\lambda_k]$. This can be cumbersome, so if felt necessary, use the approximation $\lambda_k^2 \mapsto \mathrm{E}_{\lambda_k\sim q}[\lambda_k]^2$ instead.

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

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

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

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

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

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

We find that $q(\boldsymbol{\Psi})$ satisfies

$$\begin{aligned}
\log q(\boldsymbol{\Psi}) &= \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\Psi}\}\sim q}\Big[-\frac{1}{2}\operatorname{tr}\left((\mathbf{y}^* - \boldsymbol{\mu})^\top (\mathbf{y}^* - \boldsymbol{\mu})\boldsymbol{\Psi}\right) - \frac{1}{2}\operatorname{tr}\left(\mathbf{w}^\top \mathbf{w}\boldsymbol{\Psi}^{-1}\right)\Big] \\
&\qquad + \log p(\boldsymbol{\Psi}) + \text{const.} \\
&= -\frac{1}{2}\operatorname{tr}\Big(\overbrace{\left(\mathrm{E}[(\mathbf{y}^* - \boldsymbol{\mu})^\top (\mathbf{y}^* - \boldsymbol{\mu})]\right)}^{\mathbf{G}_1}\boldsymbol{\Psi} + \overbrace{\mathrm{E}[\mathbf{w}^\top \mathbf{w}]}^{\mathbf{G}_2}\boldsymbol{\Psi}^{-1}\Big) \\
&\qquad + \log p(\boldsymbol{\Psi}) + \text{const.}
\end{aligned}$$

This seems to be the pdf of $\mathrm{Wis}(\mathbf{G} + \mathbf{G}_1, g)$ plus the pdf of a distribution which almost resembles an inverse Wishart pdf. Unfortunately, the properties such as its moments and entropy are unknown.

The matrix $\mathbf{G}_1$ is

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

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

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

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

for each ψ_j is found to be

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

which again, is a pdf of an unknown distribution. However, its posterior mode can be computed. Write $a = -\left(\frac{1}{2}\mathrm{E}\|\mathbf{y}^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 + r_j\right)$, $b = s_j - 1$, and $c = \left(\frac{1}{2}\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2\right)$. Then,

$$\frac{\partial}{\partial\psi_j}\log q(\psi_j) = \frac{\partial}{\partial\psi_j}\left(a\psi_j + b\log\psi_j - c\psi_j^{-1}\right) = a + b\psi_j^{-1} + c\psi_j^{-2}$$

equated to zero means solving a quadratic equation in ψ_j. Suppose that $p(\psi_j) \propto \text{const.}$, then $s_j = 1$ and $r_j = 0$ so $\tilde{\psi}_j$ can be solved directly to be

$$\hat{\psi}_j = \sqrt{\frac{\mathrm{E}\|\mathbf{y}^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2}{\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2}}.$$

If the posterior mean is close to its mode, then $\hat{\psi}_j$ is a good approximation for $\tilde{\psi}_j$.

To calculate $\mathrm{E}\|y^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 = \mathrm{E}\sum_{i=1}^{n}(\mathbf{y}^*_{ij}-\mu_{ij})^2$, one first needs $\mathrm{E}(y^*_{ij}-\alpha_j-\mathbf{w}_{\cdot j}^\top\mathbf{h}_\eta(x_i))^2$. This, in itself, presents a challenge to compute analytically, because it requires, among other things, the second moments $\mathrm{E}\,y^{*2}_{ij}$ and $\mathrm{E}[\mathbf{w}_{\cdot j}^\top\mathbf{h}_\eta(x_i)\mathbf{h}_\eta(x_i)^\top\mathbf{w}_{\cdot j}]$. Although not entirely accurate, it is simpler to use the approximation

$$\mathrm{E}\|\mathbf{y}^*_{\cdot j}-\boldsymbol{\mu}_{\cdot j}\|^2 \approx \|\tilde{\mathbf{y}}^*_{\cdot j}-\tilde{\boldsymbol{\mu}}_{\cdot j}\|^2.$$

(see note **??** on page **??**). Also, we have $\mathbf{w}_{\cdot j} \sim \mathrm{N}_n(\tilde{\mathbf{w}}_{\cdot j}, \tilde{\mathbf{V}}_{w_j})$, and so $\mathrm{E}\|\mathbf{w}_{\cdot j}\|^2 = \mathrm{tr}(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top)$.

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

Let $\mathbf{A} = \text{diag}(A_1, \ldots, A_m)$ and $\mathbf{a} = (a_1, \ldots, a_m)^\top$. The terms involving α_j in **??** are

$$\begin{aligned}
\log q(\boldsymbol{\alpha}) &= \mathrm{E}_{\mathcal{Z}\setminus\{\boldsymbol{\alpha}\}\sim q}\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_{i\cdot}^{*} - \boldsymbol{\alpha} - \mathbf{w}^\top \mathbf{h}_\eta(x_i)\right)^\top \boldsymbol{\Psi}\left(\mathbf{y}_{i\cdot}^{*} - \boldsymbol{\alpha} - \mathbf{w}^\top \mathbf{h}_\eta(x_i)\right)\right] \\
&\quad - \frac{1}{2}(\boldsymbol{\alpha} - \mathbf{a})^\top \mathbf{A}^{-1}(\boldsymbol{\alpha} - \mathbf{a}) + \text{const.} \\
&= -\frac{1}{2}\left[\boldsymbol{\alpha}^\top (\overbrace{n\boldsymbol{\Psi} + \mathbf{A}^{-1}}^{\tilde{\mathbf{A}}})\boldsymbol{\alpha} - 2\left(\overbrace{\sum_{i=1}^{n}\boldsymbol{\Psi}\left(\tilde{\mathbf{y}}_{i\cdot}^{*} - \tilde{\mathbf{w}}^\top \tilde{\mathbf{h}}_\eta(x_i)\right) + \mathbf{A}^{-1}\mathbf{a}}^{\tilde{\mathbf{a}}}\right)^\top \boldsymbol{\alpha}\right]
\end{aligned}$$

which implies a normal mean-field distribution for $\boldsymbol{\alpha}$ whose mean and variance are $\tilde{\boldsymbol{\alpha}} = \tilde{\mathbf{A}}^{-1}\tilde{\mathbf{a}}$ and $\tilde{\mathbf{A}}^{-1}$ respectively. If $\boldsymbol{\Psi}$ is diagonal, the components of $\boldsymbol{\alpha}$ would be independent.

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.

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.

beal2003 Beal, M. J. and Z. Ghahramani (2003). "The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures". In: *Bayesian Statistics 7*. Proceedings of the Seventh Valencia International Meeting. Ed. by José M. Bernardo, A. Philip Dawid, James O. Berger, Mike West, David Heckerman, M.J. Bayarri, and Adrian F.M. Smith. Oxford: Oxford University Press, pp. 453–464.

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.

bunch1991estimability Bunch, David S (1991). "Estimability in the multinomial probit model". In: *Transportation Research Part B: Methodological* 25.1, pp. 1–12.

chen2017use Chen, Yen-Chi, Y Samuel Wang, and Elena A Erosheva (2017). "On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example". In: *arXiv preprint arXiv:1711.11057*.

dansie1985parameter Dansie, BR (1985). "Parameter estimability in the multinomial probit model". In: *Transportation Research Part B: Methodological* 19.6, pp. 526–528.

geweke1991efficient Geweke, John (1991). *Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities*.

geweke1994alternative
Geweke, John, Michael Keane, and David Runkle (1994). "Alternative computational approaches to inference in the multinomial probit model". In: *The review of economics and statistics*, pp. 609–632.

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.

hajivassiliou1996simulation
Hajivassiliou, Vassilis, Daniel McFadden, and Paul Ruud (1996). "Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results". In: *Journal of econometrics* 72.1-2, pp. 85–134.

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.

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

keane1994computationally
Keane, Michael P (1994). "A computationally practical simulation estimator for panel data". In: *Econometrica: Journal of the Econometric Society*, pp. 95–116.

Keane1992
Keane, Michael P. (1992). "A Note on Identification in the Multinomial Probit Model". In: *Journal of Business & Economic Statistics* 10.2, pp. 193–200. ISSN: 0735-0015. DOI: 10.2307/1391677. URL: http://www.jstor.org/stable/1391677%5Cnhttp://www.jstor.org/stable/pdfplus/1391677.pdf?acceptTC=true.

kuss2005assessing
Kuss, Malte and Carl Edward Rasmussen (2005). "Assessing approximate inference for binary Gaussian process classification". In: *Journal of machine learning research* 6.Oct, pp. 1679–1704.

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.

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.

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.

train2009discrete Train, Kenneth E (2009). *Discrete choice methods with simulation*. Cambridge university press.