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

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

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

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

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

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

{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

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

{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 $\operatorname{N}_m(\boldsymbol{\mu}(x_i), \mathbf{\Psi}^{-1})$, 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 $\mathbf{\Psi}^{-1}$.

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

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

{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 $f_j \in \mathcal{F}$, $j = 1 \dots, m$, defined by

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

with $\mathbf{w}_{i\cdot} := (w_{i1}, \dots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}(0, \boldsymbol{\Psi})$. 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 $i = 1, \dots, n$,

$y_i = \arg\max_{k=1}^m y^*_{ik} \in \{1,\dots,m\}$, where, for $j = 1,\dots,m$,

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

{eq:iprobitmod}

The parameters of the I-probit model are denoted by $\theta = \{\alpha_1,\dots,\alpha_m,\eta,\boldsymbol\Psi\}$. 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 $\boldsymbol\epsilon \sim \text{MN}_{n,m}(\mathbf{0},\mathbf{I}_n,\boldsymbol\Psi^{-1})$;
- $\mathbf{w} \in \mathbb{R}^{n\times m}$ 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 $\mathbf{w} \sim \text{MN}_{n,m}(\mathbf{0},\mathbf{I}_n,\boldsymbol\Psi)$;
- $\mathbf{f},\mathbf{f}_0 \in \mathbb{R}^{n\times m}$ denote the matrices containing (i,j) entries $f_j(x_i)$ and $f_0(x_i,j)$ respectively, so that $\mathbf{f} = \mathbf{f}_0 + \mathbf{H}_\eta\mathbf{w} \sim \text{MN}_{n,m}(\mathbf{1}_n\mathbf{f}_0^\top,\mathbf{H}_\eta^2,\boldsymbol\Psi)$;
- $\boldsymbol\alpha = (\alpha+\alpha_1,\dots,\alpha+\alpha_m)^\top \in \mathbb{R}^m$ 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 $\mathbf{y}^*|\mathbf{w} \sim \text{MN}_{n,m}(\boldsymbol\mu = \mathbf{1}_n\boldsymbol\alpha^\top + \mathbf{H}_\eta\mathbf{w},\mathbf{I}_n,\boldsymbol\Psi^{-1})$ and $\operatorname{vec}\mathbf{y}^* \sim \text{N}_{nm}\big(\operatorname{vec}(\mathbf{1}_n\boldsymbol\alpha^\top),\boldsymbol\Psi\otimes\mathbf{H}_\eta^2 + \boldsymbol\Psi^{-1}\otimes\mathbf{I}_n\big)$ (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 $\text{N}_m\big(\boldsymbol\mu(x_i), \boldsymbol\Psi^{-1}\big)$ of the underlying latent variables $\mathbf{y}^*_{i\cdot}$ would yield the same class probabilities as the multivariate normal distribution $\text{N}_m\big(a\mathbf{1}_m + c\boldsymbol\mu(x_i), c^2\boldsymbol\Psi^{-1}\big)$, 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 $\alpha_1, \dots, \alpha_m$ (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 $= m+1$).

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

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. $y^*_{im} = 0, \forall i = 1,\dots,n$ has the effect of shrinking $\mathbf{\Psi}$ to $(m-1)\times(m-1)$ in size, and thus one more restriction needs to be made (typically, the first element $\mathbf{\Psi}_{11}$ 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 $i = 1,\dots,n$,

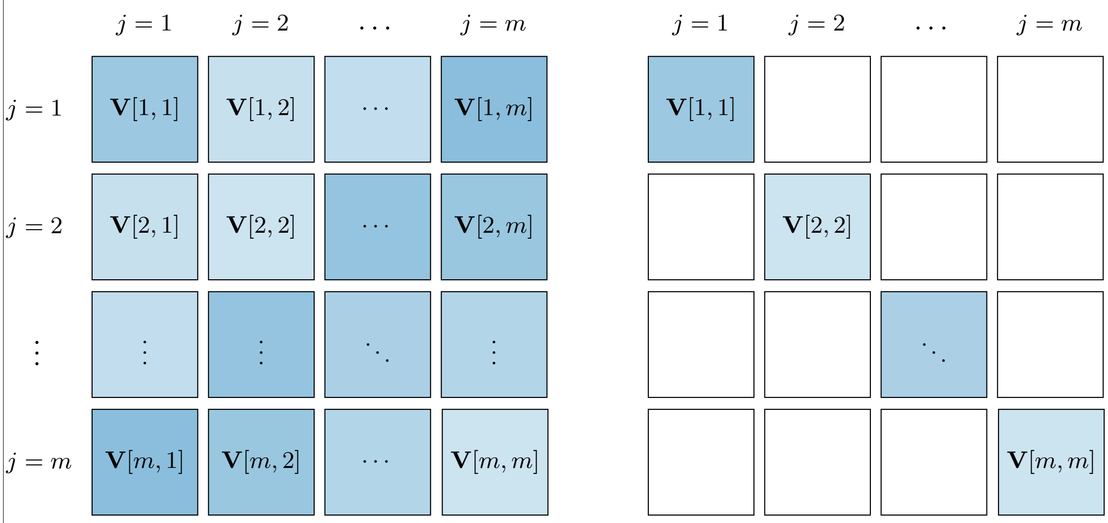
$$\begin{aligned} p_{i1} &= \mathrm{P}(y^*_{i1} > y^*_{i2} = 0) \\ &= \mathrm{P}\big(\alpha_1 + f_1(x_i) + \epsilon_{i1} > 0 \,|\, \epsilon_{i1} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)\big) \\ &= \Phi\big(\alpha_1 + f_1(x_i)\big) \qquad (5.5) \end{aligned}$$

{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 $\mathbf{\Psi}$ 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 $\mathbf{\Psi}$. Specifically, the off-diagonal elements $\mathbf{\Psi}_{jk}$ capture the correlation between alternatives j and k. Allowing all $m(m+1)/2$ covariance elements of $\mathbf{\Psi}$ 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, $\mathbf{V}_{y^*} = \mathbf{\Psi} \otimes \mathbf{H}_\eta^2 + \mathbf{\Psi}^{-1} \otimes \mathbf{I}_n$, and of the I-prior $\mathbf{V}_f = \mathbf{\Psi} \otimes \mathbf{H}_\eta^2$.



$$\begin{array}{c|cccc} & j=1 & j=2 & \dots & j=m \\ \hline j=1 & \mathbf{V}[1,1] & \mathbf{V}[1,2] & \cdots & \mathbf{V}[1,m] \\ j=2 & \mathbf{V}[2,1] & \mathbf{V}[2,2] & \cdots & \mathbf{V}[2,m] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ j=m & \mathbf{V}[m,1] & \mathbf{V}[m,2] & \cdots & \mathbf{V}[m,m] \end{array} \qquad \begin{array}{cccc} j=1 & j=2 & \dots & j=m \\ \hline \mathbf{V}[1,1] & & & \\ & \mathbf{V}[2,2] & & \\ & & \ddots & \\ & & & \mathbf{V}[m,m] \end{array}$$

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 $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_m)$, 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 $j = 1, \dots, m$, where $\sigma_j^2 = \psi_j^{-1}$. As a continuation of line (5.3), we can show the class probabilities p_{ij} to be

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

{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 $\{(y_1, x_1), \dots, (y_n, x_n)\}$ is obtained by performing the following integration:

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

{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 $p(\mathbf{y}|\mathbf{y}^*, \theta) = \prod_{i=1}^n \prod_{j=1}^m \big[y_{ij}^* = \max \mathbf{y}_{i\cdot}^* \big]^{[y_i = j]}$, with $0^0 := 1$. Note

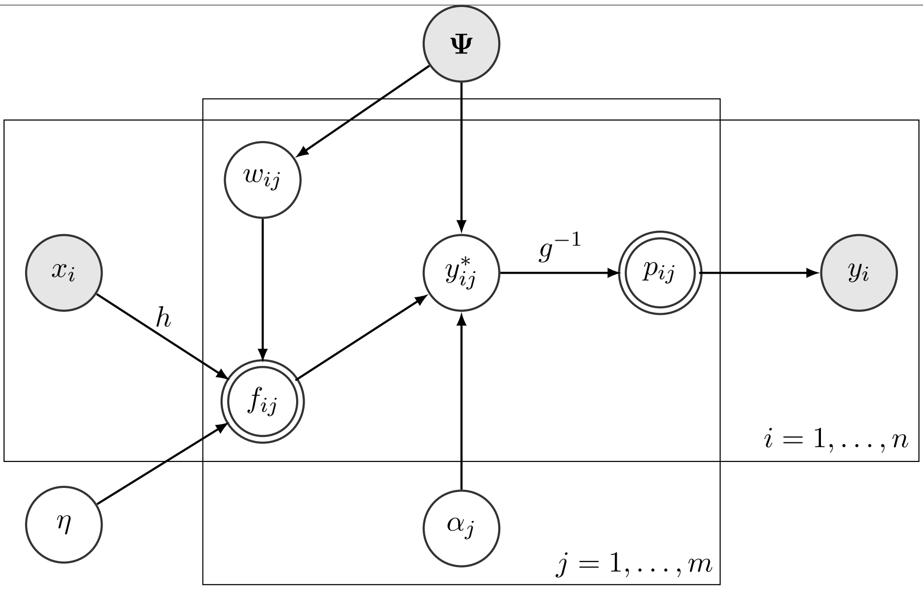


Figure 5.2: A directed acyclic graph (DAG) of the I-probit model. Observed/fixed nodes are shaded, while double-lined nodes represents calculable quantities.

fig:iprobit dag

that, given the corresponding latent propensities $\mathbf{y}^*_{i\cdot} = (y^*_{i1}, \dots, y^*_{im})^\top$, 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

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

{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

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

{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 $\mathbf{w}|\mathbf{y} \sim \text{N}_n(\hat{\mathbf{w}}, \boldsymbol{\Omega}^{-1})$. 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

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

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

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

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 $\operatorname{KL}(q\|p)$ 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

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

{eq:iprobitQEstep}

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

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

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.

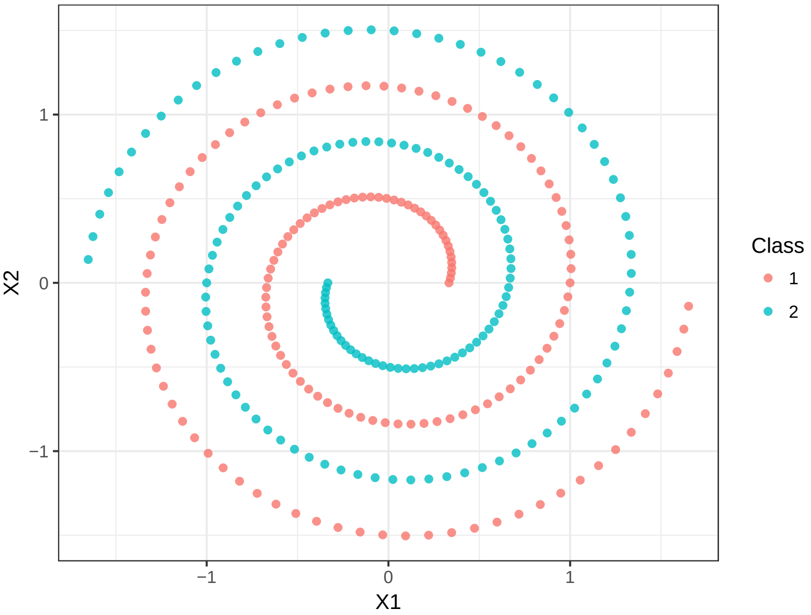


fig:example iprobit

Figure 5.3: A plot of simulated spiral data set.

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

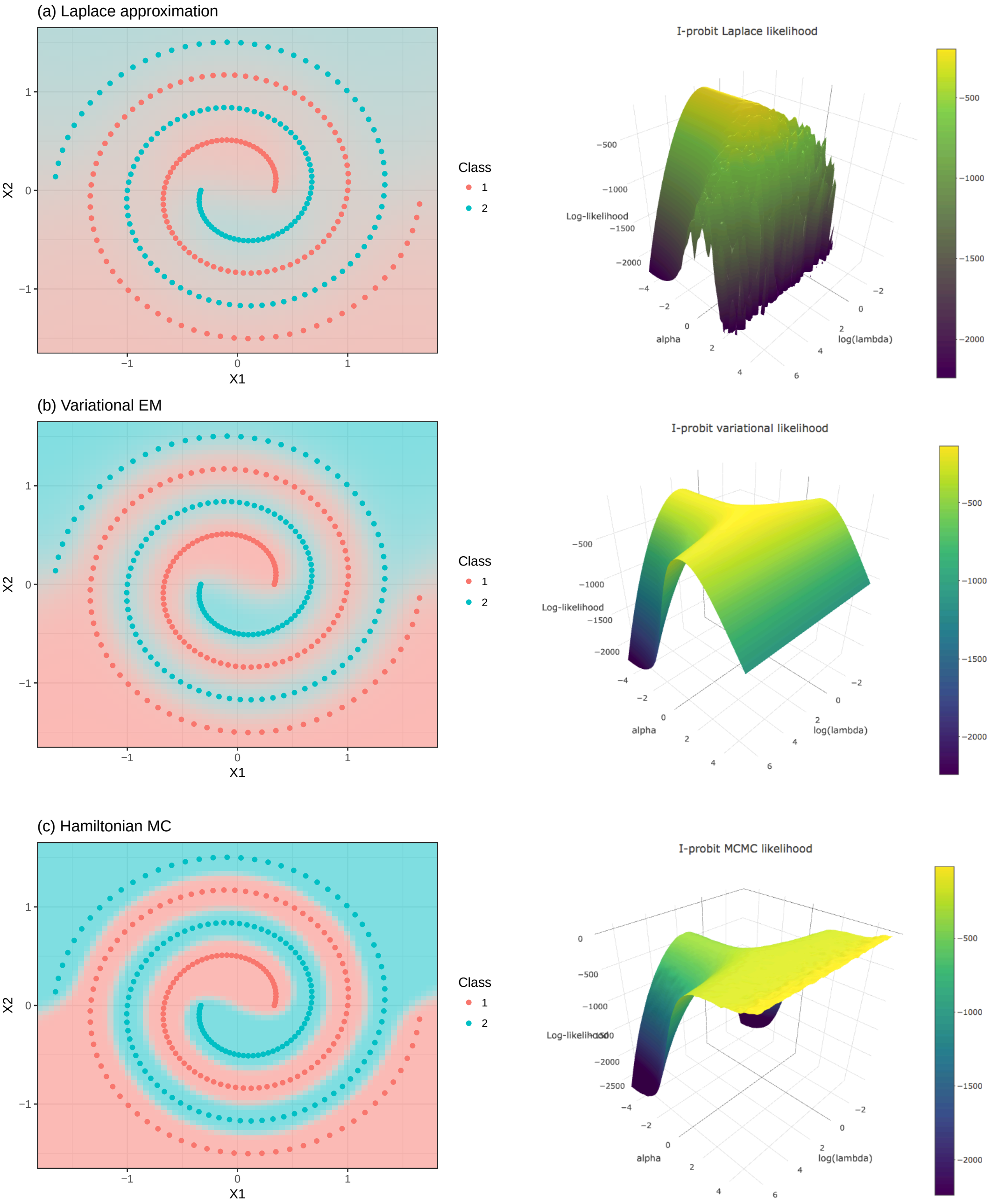


Figure 5.4: Plots showing predicted probabilities (shaded region) for belonging to class 1 or 2 indicated by colour and intensity, and likelihood surface plots for (a) Laplace's method, (b) variational EM, and (c) Hamiltonian MC.

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 $\{w_1, \dots, w_n, \alpha, \lambda\}$ 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.

5.4 The variational EM algorithm for I-probit models

sec:iprobit var

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 $\text{KL}\big[q(\mathbf{y}^*, \mathbf{w}) \| p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})\big]$, 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 $\text{KL}\big[q \| p\big]$ 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

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

{eq:logqystar}

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

{eq:logqw}

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 $\tilde{\mathbf{w}} = \text{E}_{\mathbf{w}\sim\tilde q}\,\mathbf{w}$, and variance $\boldsymbol{\Psi}^{-1}$, subject to a truncation of its components to the set $\mathcal{C}_{y_i}$. That is, for each $i = 1, \dots, n$ and noting the observed categorical response $y_i \in \{1, \dots, m\}$ 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 $\mathbf{y}_{i\cdot}^* \overset{\text{iid}}{\sim} {}^{\mathrm{t}}\mathrm{N}(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, and the important properties of this distribution are explored in the appendix.

The required expectation $\tilde{\mathbf{y}}^* := \mathrm{E}_{\mathbf{y}^* \sim \tilde{q}}\, \mathbf{y}_{i\cdot}^* = \mathrm{E}_{\mathbf{y}^* \sim \tilde{q}} (y_{i1}^*, \dots, y_{im}^*)^\top$ 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 $\boldsymbol{\Psi} = \mathrm{diag}(\sigma_1^{-2}, \dots, \sigma_m^{-2})$, then the first moment can be considered component-wise. Each component of this expectation is given by

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

{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 $\mathbf{w}$

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

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

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,

$$\operatorname{E}[r(\mathbf{w})|\mathbf{y}] = \operatorname{E}_{\mathbf{y}^*}\big[\operatorname{E}[r(\mathbf{w})|\mathbf{y}^*] \,|\, \mathbf{y}\big],$$

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.

sec:varupdeta

5.4.2 The M-step

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

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

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 $\operatorname{E}[\mathbf{y}^{*\top}\mathbf{y}^*]$ is absorbed into the constant. On closer inspection, the trace involving the second moments of $\mathbf{w}$ is found to be

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

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

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

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 $\tilde{\mathbf{w}}_{\cdot j} = \big(\operatorname{E}[w_{ij}]\big)_{i=1}^n$ is the expected value of the random effects for class j. Specifically, when the error precision is of the form $\boldsymbol{\Psi} = \operatorname{diag}(\psi_1, \dots, \psi_m)$, this trace reduces to

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

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 $O(n^3m)$ 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 $\eta = \{\lambda_1, \dots, \lambda_p\}$ need be estimated, and for each $k = 1, \dots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k\mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}_\eta^2 = \lambda_k^2\mathbf{R}_k^2 + \lambda_k\mathbf{U}_k + \mathbf{S}_k^2$. 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

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

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 ${}^{\text{t}}\text{N}_m(\boldsymbol{\alpha}+\mathbf{w}^\top\mathbf{h}_\eta(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, with $\mathcal{C}_{y_i} = \{y_{iy_i}^* > y_{ik}^* | \forall k \neq y_i\}$. That is,

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

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 $\operatorname{vec}\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_{\cdot 1}, \dots, \tilde{\mathbf{w}}_{\cdot m})^\top$ to separate out the I-prior random effects per class, we have that $\mathbf{w}_{\cdot j}|\hat{\theta} \sim \text{N}_n(\tilde{\mathbf{w}}_{\cdot j}, \tilde{\mathbf{V}}_w[j,j])$, and $\text{Cov}(\mathbf{w}_{\cdot j}, \mathbf{w}_{\cdot k}) = \tilde{\mathbf{V}}_w[j,k]$, where the '$[\cdot,\cdot]$' indexes the $n \times n$ sub-block of the block alg:varemiprobit

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

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

while not converged do
    procedure Variational E-step
        for $i = 1, \dots, n$ do                                                    ▷ Update $\mathbf{y}^*$
            $\tilde q^{(t+1)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\text{t}}\text{N}_m\big(\tilde{\boldsymbol{\alpha}}^{(t)} + \tilde{\mathbf{w}}^{(t)\top}\mathbf{h}_{\eta^{(t)}}(x_i), \boldsymbol{\Psi}, \mathcal{C}_{y_i}\big)$
            $\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} \leftarrow \text{E}_{q^{(t+1)}}[\mathbf{y}^*_{i\cdot}]$
        end for

        $\tilde{\mathbf{V}}_w^{(t+1)} \leftarrow \big((\boldsymbol{\Psi} \otimes \mathbf{H}^2_{\eta^{(t)}}) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\big)^{-1}$      ▷ Update $\mathbf{w}$
        $\operatorname{vec} \tilde{\mathbf{w}}^{(t+1)} \leftarrow \tilde{\mathbf{V}}_w^{(t+1)}(\boldsymbol{\Psi} \otimes \mathbf{H}_{\eta^{(t)}}) \operatorname{vec}(\tilde{\mathbf{y}}^{*(t+1)} - \mathbf{1}_n\boldsymbol{\alpha}^{(t)\top})$
        $\tilde q^{(t+1)}(\mathbf{w}) \leftarrow \text{N}_{nm}\big(\operatorname{vec} \tilde{\mathbf{w}}^{(t+1)}, \tilde{\mathbf{V}}_w^{(t+1)}\big)$
    end procedure

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

        $\mathbf{a} \leftarrow \frac{1}{n}\sum_{i=1}^n \big(\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} - \tilde{\mathbf{w}}^{(t+1)\top}\tilde{\mathbf{h}}_{\eta^{(t+1)}}(x_i)\big)$     ▷ 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 $j = 1, \dots, m$ 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 $y^*_{\text{new},j}$ for each class are normally distributed with mean, variance, and covariances

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

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

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

The predicted class is inferred from the latent variables via

$$\hat y_{\text{new}} = \arg\max_k \hat\mu_k(x_{\text{new}}),$$

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

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

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

$$\hat p_{\text{new},j} = \mathrm{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^m \Phi\left(\frac{\hat\sigma_j(x_{\text{new}})}{\hat\sigma_k(x_{\text{new}})} Z + \frac{\hat\mu_j(x_{\text{new}}) - \hat\mu_k(x_{\text{new}})}{\hat\sigma^2_k(x_{\text{new}})} \right) \right],$$

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). Prediction of a single new data point takes $O(n^2m)$ time, because there are essentially m I-prior posterior regression functions, and each take $O(n^2)$ to evaluate. This is assuming negligible time to compute the class probabilities.

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 $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(T)}$ 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 $\{p_{xj}^{(1)}, \dots, p_{xj}^{(T)}\}_{j=1}^m$, for a given data point $x \in \mathcal{X}$ by evaluating

$$p_{xj}^{(t)} = \mathop{\int \cdots \int}_{\{y_j^* > y_k^* | \forall k \neq j\}} \phi\big(y_1^*, \dots, y_m^* | \hat{\boldsymbol{\mu}}^{(t)}(x), \hat{\mathbf{V}}(x)\big) \, \mathrm{d}y_1^* \cdots \mathrm{d}y_m^*,$$

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 $\hat\sigma_j^2(x)$ 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 $\big\{r(p_{xj}^{(1)}), \dots, r(p_{xj}^{(T)})\big\}$.

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 $\mathcal{S} = \{(y_1, x_1), \dots, (y_n, x_n)\}$.
2. For $t = 1, \dots, T$, do
   (a) Obtain $\mathcal{S}^{(t)} = \{(y_1^{(t)}, x_1^{(t)}), \dots, (y_n^{(t)}, x_n^{(t)})\}$ 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

$$\widehat{\operatorname{Var}}(\hat\theta_l) = \frac{1}{T} \sum_{t=1}^T (\hat\theta_l^{(t)} - \bar\theta_l)^2 \quad \text{where} \quad \bar\theta_l = \frac{1}{T} \sum_{t=1}^T \hat\theta_l^{(t)}.$$

The obvious downside to this is computational time.

tab:BF

tab:bf

Table 5.2: Guidelines for interpreting Bayes factors.

$2\log\text{BF}(M_1,M_0)$	$\text{BF}(M_1,M_0)$	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

Finally, a discussion on model comparison, which, in the variational inference literature, is achieved by comparing ELBO values of competing models (Beal and Ghahramani, 2003). The rationale is that the ELBO serves as a conservative estimate for the log marginal likelihood, which would allow model selection via (empirical) Bayes factors. This stems from the fact that (see note in section)

$$\log p(\mathbf{y}|\theta) = \mathcal{L}_q(\theta) + \text{KL}(q\|p) > \mathcal{L}_q(\theta),$$

since the Kullbeck-Leibler divergence from the true posterior density $p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})$ to the variational density $q(\mathbf{y}^*, \mathbf{w})$ is strictly positive (it is zero if and only if the two densities are equivalent), and is minimised under a variational inference scheme. Kass and Raftery (1995) suggest Section 5.5 as a way of interpreting observed Bayes factor values $\text{BF}(M_1, M_0)$ for comparing model M_1 against model M_0, where $\text{BF}(M_1, M_0)$ is approximated by

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

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 considerations

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 $\mathbf{\Psi}$ 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). As such, simulation techniques (Monte Carlo integration) are employed instead. The simplest strategy to overcome this is a frequency simulator (otherwise known as Monte Carlo integration): obtain random samples from $\text{N}_m\big(\boldsymbol\mu(x_i), \boldsymbol\Psi^{-1}\big)$, 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, in an extensive comparative study of various probability simulators, V. Hajivassiliou et al. (1996) concluded that the Geweke-Hajivassiliou-Keane (GHK) probability simulator (Geweke, 1989; V. A. Hajivassiliou and McFadden, 1998; Michael P Keane and Wolpin, 1994) is the most reliable under a multitude of scenarios. This is now described, and for clarity, we drop the subscript i denoting individuals.

Suppose that an observation $y = j$ has been made. Reformulate $\mathbf y^*$ in (5.1) by anchoring on the j'th latent variable y_j^* to obtain

$$\mathbf z := (\overbrace{y_1^* - y_j^*}^{z_1}, \dots, \overbrace{y_{j-1}^* - y_j^*}^{z_{j-1}}, \overbrace{y_{j+1}^* - y_j^*}^{z_j}, \dots, \overbrace{y_m^* - y_j^*}^{z_{m-1}},)^\top \in \mathbb R^{m-1}.$$

Note that we have indexed the vector $\mathbf z$ using $j' = k$ if $k < j$, and $j' = k-1$ if $k > j$ for $k = 1, \dots, m$, so that the index j' runs from 1 to $m-1$. Let $\mathbf Q \in \mathbb R^{(m-1)\times m}$ be a matrix formed by inserting a column of minus ones at the j'th position in an $(m-1)$ identity matrix. We can then write $\mathbf z = \mathbf Q\mathbf y^*$, and thus we have that $\mathbf z \sim \text{N}_{m-1}(\boldsymbol\nu_{(j)}, \boldsymbol\Omega_{(j)})$, where $\boldsymbol\nu_{(j)} = \mathbf Q\boldsymbol\mu(x_i)$ and $\boldsymbol\Omega_{(j)} = \mathbf Q\boldsymbol\Psi^{-1}\mathbf Q^\top$. These are indexed by '$(j)$' because the transformation is dependent on which latent variable the $\mathbf z$'s are anchored on.

Remark 5.5. Incidentally, the probit model in (5.1) is equivalently represented by

$$y_i = \begin{cases} 1 & \text{if } \max(y_{i2}^* - y_{i1}^*, \dots, y_{im}^* - y_{i1}^*) < 0 \\ j & \text{if } \max(y_{i2}^* - y_{i1}^*, \dots, y_{im}^* - y_{i1}^*) = y_{ij}^* - y_{i1}^* \ge 0, \end{cases} \tag{5.21}$$

{eq:latentmodel2}

which is obtained by anchoring on the first latent variable (referred to as the reference category), although the choice of reference category is arbitrary. This is similar to fixing the latent variables of the reference category to zero, and thus, as discussed previously in section, full identification is achieved by fixing one more element of the covariance matrix.

For the symmetric and positive definite covariance matrix $\boldsymbol{\Omega}_{(j)}$, obtain its Cholesky decomposition as $\boldsymbol{\Omega}_{(j)} = \mathbf{L}\mathbf{L}^\top$, where $\mathbf{L}$ is a lower triangular matrix. Then, $\mathbf{z} = \boldsymbol{\nu}_{(j)} + \mathbf{L}\boldsymbol{\zeta}$, where $\boldsymbol{\zeta} \sim \mathrm{N}_{m-1}(\mathbf{0}, \mathbf{I}_{m-1})$. That is,

$$
\begin{aligned}
\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_{m-1} \end{pmatrix}
&= \begin{pmatrix} \nu_{(j)1} \\ \nu_{(j)2} \\ \vdots \\ \nu_{(j)m-1} \end{pmatrix}
+ \begin{pmatrix} L_{11} & & & \\ L_{21} & L_{22} & & \\ \vdots & \vdots & \ddots & \\ L_{m-1,1} & L_{m-1,2} & \cdots & L_{m-1,m-1} \end{pmatrix}
\begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \vdots \\ \zeta_{m-1} \end{pmatrix} \\
&= \begin{pmatrix} \nu_{(j)1} + L_{11}\zeta_1 \\ \nu_{(j)2} + \sum_{k=1}^{2} L_{k2}\zeta_k \\ \vdots \\ \nu_{(j)m-1} + \sum_{k=1}^{m-1} L_{k,m-1}\zeta_k \end{pmatrix}.
\end{aligned}
$$

With this setup, the probability p_j of an observation belonging to class j, which is equivalent to the probability that each $z_{j'} < 0$, $j' = 1, \dots, m-1$, can be expressed as

$$
\begin{aligned}
p_j &= \mathrm{P}(z_1 < 0, \dots, z_{m-1} < 0) \\
&= \mathrm{P}(\zeta_1 < u_1, \dots, \zeta_{m-1} < u_{m-1}) \\
&= \mathrm{P}(\zeta_1 < u_1)\,\mathrm{P}(\zeta_2 < u_2 | \zeta_1 < u_1)\,\mathrm{P}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2) \cdots \\
&\qquad \cdots \mathrm{P}(\zeta_{m-1} < u_{m-1} | \zeta_1 < u_1, \dots, \zeta_{m-2} < u_{m-2}),
\end{aligned}
$$

where

$$
u_{j'} = u_{j'}(\zeta_1, \dots, \zeta_{j'-1}) =
\begin{cases}
-\nu_{(j)1}/L_{11} & \text{for } j' = 1 \\
-\big(\nu_{(j)j'} + \sum_{k=1}^{j'-1} L_{kj'}\zeta_k\big)/L_{j'j'} & \text{for } j' = 2, \dots, m-1
\end{cases}
$$

The GHK algorithm entails making draws from truncated standard normal distributions (for instance, using an inverse transform method detailed in appendix):

- Draw $\tilde\zeta_1 \sim {}^{\text{t}}\mathrm{N}(0, 1, -\infty, u_1)$.
- Draw $\tilde\zeta_2 \sim {}^{\text{t}}\mathrm{N}(0, 1, -\infty, \tilde u_2)$, where $\tilde u_2 = u_2(\tilde\zeta_1)$.
- Draw $\tilde\zeta_3 \sim {}^{\text{t}}\mathrm{N}(0, 1, -\infty, \tilde u_3)$, where $\tilde u_2 = u_2(\tilde\zeta_1, \tilde\zeta_2)$.
- $\cdots$
- Draw $\tilde\zeta_{m-1} \sim {}^{\text{t}}\mathrm{N}(0, 1, -\infty, \tilde u_{m-2})$, where $\tilde u_{m-1} = u_m(\tilde\zeta_1, \dots, \tilde\zeta_{m-2})$.

These values are then used in the following manner:

- Use $\tilde{\zeta}_1$ to obtain a "draw" of $\text{P}(\zeta_2 < u_2 | \zeta_1 < \zeta_1)$,
$$\begin{aligned}\widetilde{\text{P}}(\zeta_2 < u_2 | \zeta_1 < \zeta_1) &= \text{P}(\zeta_2 < u_2 | \zeta_1 = \tilde{\zeta}_1) \\ &= \Phi\Big(- \big(\nu_{(j)2} + L_{12}\tilde{\zeta}_1\big)/L_{22}\Big)\end{aligned}$$
- Use $\tilde{\zeta}_1$ and $\tilde{\zeta}_2$ to obtain a "draw" of $\text{P}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2)$,
$$\begin{aligned}\widetilde{\text{P}}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2) &= \text{P}(\zeta_3 < u_3 | \zeta_1 = \tilde{\zeta}_1, \zeta_2 = \tilde{\zeta}_2) \\ &= \Phi\Big(- \big(\nu_{(j)3} + L_{13}\tilde{\zeta}_1 + L_{23}\tilde{\zeta}_2\big)/L_{33}\Big)\end{aligned}$$
- And so on.

Therefore, a simulated probability for p_j (denoted with a tilde) is obtained as

$$\tilde{p}_j = \Phi\left(-\nu_{(j)1}/L_{11}\right) \prod_{j'=2}^{m-1} \Phi\left(-\big(\nu_{(j)j'} + \textstyle\sum_{k=1}^{j'-1} L_{kj'}\zeta_k\big)/L_{j'j'}\right). \tag{5.22}$$

{eq:tildepj}

By performing the above scheme T number of times to obtain T such simulated probabilities $\{p_j^{(1)}, \dots, p_j^{(T)}\}$, the actual probability of interest p_j is then approximated by the sample mean of the draws,

$$\hat{p}_j = \frac{1}{T}\sum_{t=1}^{T} p_j^{(t)}.$$

If it so happens that one of the standard normal cdfs in (5.22) is extremely small, this can cause loss of significance due to floating-point errors (catastrophic cancellation). It is better to work on a log-probability scale, so the products in (5.22) turn into sums, and revert back by exponentiating.

Remark 5.6. The GHK algorithm provides reasonably fast and accurate calculations of class probabilities when $\mathbf{\Psi}$ is dense. 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 $\mathbf{\Psi}$ is diagonal. Note that if $\mathbf{\Psi}$ is diagonal, then the transformed $\mathbf{\Omega}_{(j)} = \mathbf{Q}\mathbf{\Psi}^{-1}\mathbf{Q}^\top$ is certainly not: the components of $\mathbf{z}$ are correlated because they are all anchored on the same random variable. Thus, direct evaluation of (5.6) using quadrature methods avoids the Cholesky step and random sampling employed by the GHK method.

sec:complxiprobit

5.6.2 Efficient Kronecker product inverse

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

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

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 variational EM algorithm down to a similar requirement as if $\boldsymbol{\Psi}$ were diagonal. Unfortunately, storage requirements remain at $O(n^2m^2)$ when $\boldsymbol{\Psi}$ is dense, because the entire $nm \times nm$ matrix $\mathbf{V}_w$ is needed to evaluate the posterior mean of $\operatorname{vec} \mathbf{w}$.

sec:difficultPsi

5.6.3 Estimation of $\boldsymbol{\Psi}$ in future work

Suppose that $\boldsymbol{\Psi} \in \mathbb{R}^{m \times m}$ is a free parameter to be estimated, bearing in mind that only $m(m-1)/2-1$ variance components are identified in the I-probit model (see Section 5.2). If so, the Q function from (5.11) conditional on the rest of the parameters can be written as

$$Q(\boldsymbol\Psi|\boldsymbol\alpha,\eta) = \text{const.} - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol\Psi\overbrace{\mathrm{E}[(\mathbf{y}^* - \boldsymbol\mu)^\top(\mathbf{y}^* - \boldsymbol\mu)]}^{\mathbf{G}_1} + \boldsymbol\Psi^{-1}\overbrace{\mathrm{E}[\mathbf{w}^\top\mathbf{w}]}^{\mathbf{G}_2}\Big)$$

with $\boldsymbol\mu = \mathbf{1}_n\boldsymbol\alpha^\top + \mathbf{H}_\eta\mathbf{w}$. This can be differentiated with respect to $\boldsymbol\Psi$ to obtain

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol\Psi}Q(\boldsymbol\Psi|\boldsymbol\alpha,\eta) &= -\frac{1}{2}\operatorname{tr}\left(\frac{\partial\boldsymbol\Psi}{\partial\boldsymbol\Psi}\mathbf{G}_1 + \frac{\partial\boldsymbol\Psi^{-1}}{\partial\boldsymbol\Psi}\mathbf{G}_2\right) \\ &= -\frac{1}{2}\operatorname{tr}\left(\mathbf{G}_1 - \boldsymbol\Psi^{-2}\mathbf{G}_2\right).\end{aligned}$$

Setting this to zero and solving for $\boldsymbol\Psi$ yields the M-step update equation for $\boldsymbol\Psi$. This can be solved numerically, though it must be ensured that the identifiability constraints and positive-definiteness are satisfied.

8. Actually, I don't know how to solve this?

Specifically in the case where $\boldsymbol\Psi$ is a diagonal matrix $\operatorname{diag}(\psi_1,\dots,\psi_m)$, then

$$\begin{aligned}Q(\boldsymbol\Psi|\boldsymbol\alpha,\eta) = \text{const.} &- \frac{1}{2}\sum_{j=1}^m \psi_j \operatorname{tr}\mathrm{E}[(\mathbf{y}^*_{\cdot j} - \boldsymbol\mu_{\cdot j})(\mathbf{y}^*_{\cdot j} - \boldsymbol\mu_{\cdot j})^\top] \\ &- \frac{1}{2}\sum_{j=1}^m \psi_j^{-1}\operatorname{tr}\mathrm{E}[\mathbf{w}_{\cdot j}\mathbf{w}_{\cdot j}^\top]\end{aligned}$$

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

$$\hat\psi_j = \left(\frac{\mathrm{E}[\mathbf{w}_{\cdot j}^\top\mathbf{w}_{\cdot j}]}{\mathrm{E}[(\mathbf{y}^*_{\cdot j} - \boldsymbol\mu_{\cdot j})^\top(\mathbf{y}^*_{\cdot j} - \boldsymbol\mu_{\cdot j})]}\right)^{\frac{1}{2}},$$

independently of the rest of the other ψ_k's, $k \neq j$. As per the derivations in section, the numerator of this expression is equal to $\operatorname{tr}(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top) = \operatorname{tr}\tilde{\mathbf{W}}_{jj}$. The denominator on the other hand is

$$\mathrm{E}[\mathbf{y}^{*\top}_{\cdot j}\mathbf{y}^*_{\cdot j}] - n\alpha_j^2 - \operatorname{tr}(\mathbf{H}_\eta^2\tilde{\mathbf{W}}_{jj}) - 2\mathbf{y}^{*\top}_{\cdot j}\mathbf{H}_\eta\tilde{\mathbf{w}}_{\cdot j} - 2\alpha_j\sum_{i=1}^n\sum_{i'=1}^n(y^*_{ij} - h_\eta(x_i,x_{i'})\tilde w_{ij}).$$

In either the full or I-probit model, solving $\boldsymbol\Psi$ involves the second moments of a truncated normal distribution. In the case where $\boldsymbol\Psi$ is dense, this is obtained by Monte Carlo methods, where samples from a truncated multivariate normal distribution are obtained using Gibbs sampling. Although this strategy can be used when $\boldsymbol\Psi$ is diagonal, in Lemma 5.4, we show that the form for the second moments involve integration of standard normal cdfs and pdfs, much like in formula for the first moments.

5.7 Examples

sec:iprobit eg

We present analyses of real-data examples using the I-probit model for a variety of applicaitons, namely binary and multiclass classification, meta-analysis, and spatio-temporal modelling of point processes. Examples in this section have been analysed using the R package **iprobit** developed by us. All of these examples had assumed a fixed error precision $\boldsymbol{\Psi} = \mathbf{I}_m$.

5.7.1 Predicting cardiac arrhythmia

Statistical learning tools are being used in the field of medicine as a means to aid medical diagnosis of diseases. In this example, factors determining the presence or absence of heart diseses are studied. Traditionally, cardiologists inspect patients' cardiac activity (ECG data) in order to reach a diagnosis, which remains the "gold standard" method of obtaining diagnoses. The study by Guvenir et al. (1997) aimed to predict cardiac abnormalities by way of machine learning, and minimise the difference between the gold standard and computer-based classifications.

The data set[3] at hand contains a myriad of ECG readings and other patient attributes such as age, height, and weight. Altogether, there are $n = 451$ observations and $p = 279$ predictors. In order for a valid comparison to be made to other studies, we excluded nominal covariates, leaving us with $p = 194$ continuous predictors, which we then standardised. In the original data set, there are 13 distinct classes of cardiac arrhythmia—again, following the lead of other studies, we had combined all forms of cardiac diseases to form a single class, thus reducing the problem to a binary classification task (normal vs. arrhythmia).

The relationship between patient i's probability of having a form of cardiac arrhthmia p_i and the predictors $x_i \in \mathcal{X} \equiv \mathbb{R}^{194}$ is modelled as

$$\Phi(p_i) = \alpha + f(x_i).$$

[3] Data is made publicly available at https://archive.ics.uci.edu/ml/datasets/arrhythmia.

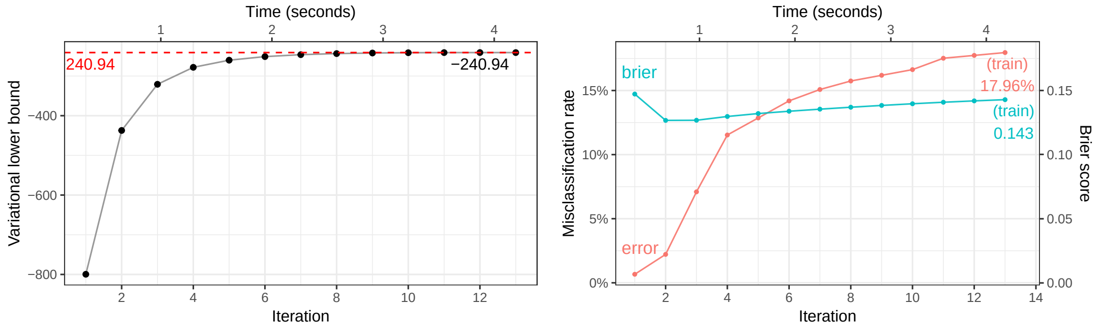


Figure 5.5: Plot of variational lower bound over time (left), and plot of training error rate and Brier scores over time (right).

fig:cardiac.mod.full.plot

Further, assuming $f \in \mathcal{F}$ a suitable RKHS with kernel h_λ, we may assign an I-prior on the (latent) regression function f. We consider three RKHSs: the canonical (linear) RKHS, the fBm-0.5 RKHS and the SE RKHS. The first of these three assumes an underlying linear relationship of the covariates and the probabilities, while the other two assumes a smooth relationship. As all covariates had been standardised, it is sufficient to assign a single scale parameter λ for the I-probit model.

For reference, fitting an I-probit model on the full data set takes about 4 seconds only, with convergence reached in at most 15 iterations. Figure 5.5 plots the variational lower bound value over time and iterations for the cardiac arrhythmia data set. As expected, the lower bound value increases over time until a convergence criterion is reached.

To measure predictive ability, we fit the I-probit models on a random subset of the data and obtain the out-of-sample test error rates from the remaining held-out observations. We then compare the results against popular machine learning classifiers, namely: 1) linear and quadratic discriminant analysis (LDA/QDA); 2) k-nearest neighbours; 3) support vector machines (SVM) (Steinwart and Christmann, 2008); 4) Gaussian process classification (Rasmussen and Williams, 2006); 5) random forests (Breiman, 2001); 6) nearest shrunken centroids (NSC) (Tibshirani et al., 2002); and 7) L-1 penalised logistic regression. The experiment is set up as follows:

1. Form a training set by sub-sampling $s \in \{50, 100, 200\}$ observations.
2. The remaining unsampled data is used as the test set.

3. Fit model on training set, and obtain test error rates defined as

$$\text{test error rate} = \frac{1}{s}\sum_{i=1}^{n}[y_i^{\text{pred}} \neq y_i^{\text{test}}] \times 100\%.$$

4. Repeat steps 1-3 100 times to obtain the *average* test error rates and standard errors.

Results for the methods listed above were extracted from the in-depth study by Cannings and Samworth (2017), who also conducted an identical experiment using their random projection ensemble classification method (RP). The results are tabulated in Table 5.3.

Of the three I-probit models, the fBm model performed the best. That it performed better than the canonical linear I-probit model is unsurprising, since an underlying smooth function to model the latent variables is expected to generalise better than a rigid straight line function. The poor performance of the SE I-probit model may be due to the fact that the lengthscale parameter was not estimated (fixed at $l = 1$), but then again, we notice reliable performance of the fBm even with fixed Hurst index ($\gamma = 0.5$). It can be seen that the fBm I-probit model also outperform the more popular machine learning algorithms out there including k-nearest neighbours, support vector machines and Gaussian process classification. It came second only to random forests, an ensemble learning method, which depending on the number of random decisions trees generated simultaneously, might be slow. The time complexity of a random forest algorithm is $O(pqn\log(n))$, where p is the number of variables used for training, q is the number of random decision trees, and n is the number of observations.

Table 5.3: Mean out-of-sample misclassification rates and standard errors in parantheses for 100 runs of various training (s) and test ($451 - s$) sizes for the cardiac arrhythmia binary classification task.

tab:cardiac

	Misclassification rate (%)		
Method	$s = 50$	$s = 100$	$s = 200$
I-probit			
Linear	35.52 (0.44)	31.35 (0.33)	29.45 (0.38)
Smooth (fBm-0.5)	33.64 (0.66)	28.12 (0.34)	24.33 (0.24)
Smooth (SE-1.0)	48.26 (0.40)	48.32 (0.43)	47.11 (0.37)
Others			
RP-LDA	33.24 (0.42)	30.19 (0.35)	27.49 (0.30)
RP-QDA	30.47 (0.33)	28.28 (0.26)	26.31 (0.28)
RP-k-NN	33.49 (0.40)	30.18 (0.33)	27.09 (0.31)
Random forests	31.65 (0.39)	26.72 (0.29)	22.40 (0.31)
SVM (linear)	36.16 (0.47)	35.61 (0.39)	35.20 (0.35)
SVM (Gaussian)	48.39 (0.49)	47.24 (0.46)	46.85 (0.43)
GP (Gaussian)	37.28 (0.42)	33.80 (0.40)	29.31 (0.35)
NSC	34.98 (0.46)	33.00 (0.40)	31.08 (0.41)
L-1 logistic	34.92 (0.42)	30.48 (0.34)	26.12 (0.27)

5.7.2 Meta-analysis of smoking cessation

Conider the smoking cessation data set, as described in Skrondal and Rabe-Hesketh (2004). It contains observations from 27 separate smoking cessation studies in which participants are subjected to either a nicotine gum treatment or a placebo. The interest is to estimate the treatment effect size, and whether it is statistically significant, i.e. whether or not nicotine gum is an effective treatment to quit smoking. The studies are conducted at different times and due to various reasons such as funding and cultural effects, the results from all of the studies may not be in agreement. The number of effective participants plays a major role in determining the power of the statistical tests performed in individual studies. The question then becomes how do we meaningfully aggregate all the data to come up with one summary measure?

Several methods exist to analyse such data sets. One may consider a fixed-effects model, similar to a classical one-way ANOVA model to establish whether or not the effect size is significant. Because of the study-specific characteristics, it is natural to consider multilevel or random-effects models as a means to estimate the effect size. Regardless of method, the approach of analysing study-level treatment effects instead of patient-level data only is the paradigm for meta-analysis, and our I-prior model takes this approach as well.

A summary of the data is displayed by the box-plot in Figure 5.6. On the whole, there are a total of 5908 patients, and they are distributed roughly equally among the control and treatment groups (46.33% and 53.67% respectively, on average). From the box-plots, it is evident that there are more patients who quit smoking in the treatment group as compared to the placebo control group. There are various measures of treatment effect size, such as risk ratio or risk differences, but we shall concentrate on *odds ratios* as defined by

$$\text{odds ratio} = \frac{\text{odds of quitting smoking in } \textit{treatment} \text{ group}}{\text{odds of quitting smoking in } \textit{control} \text{ group}}.$$

The odds of quitting smoking in either group is defined as

$$\text{odds} = \frac{\mathrm{P}[\text{quit smoking}]}{1 - \mathrm{P}[\text{quit smoking}]},$$

and these probabilities, odds and ultimately the odds ratio can be estimated from sample proportions. This raw odds ratio for all study groups is calculated as $1.66 = e^{0.50}$. It is

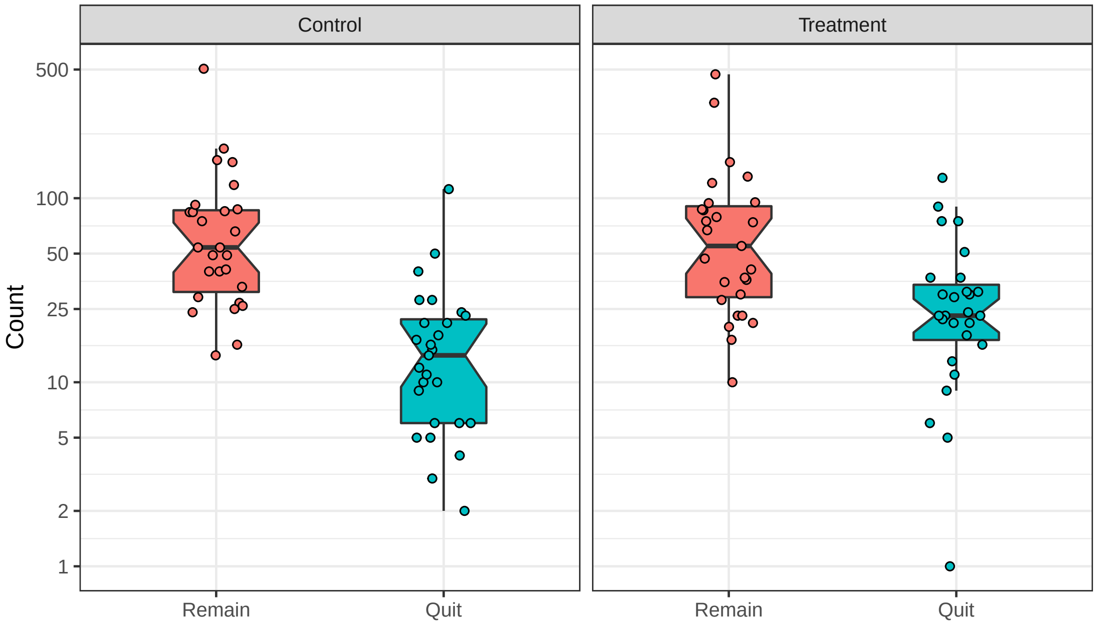


Figure 5.6: Comparative box-plots of the distribution of patients who successfully quit smoking and those who remained smokers, in the two treatment groups.

fig:plot.data.smoke

also common for the odds ratio to be reported on the log scale (usually as a remnant of logistic models). A value greater than one for the odds ratio (or equivalently, greater than zero for the log-odds ratio) indicates a significant treatment effect.

A random-effects analysis using a multilevel logistic model has been considered by Agresti and Hartzel (2000). Let $i = 1, \dots, n_j$ index the patients in study group $j \in \{1, \dots, 27\}$. For patient i in study j, p_{ij} denotes the probability that the patient has successfully quit smoking. Additionally, x_{ij} is the centred dummy variable indicating patient i's treatment group in study j. These take on two values: 0.5 for treated patients and -0.5 for control patients. The logistic random-effects model is

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{0j} + \beta_{1j}x_{ij}$$

with

$$\begin{pmatrix}\beta_{0j}\\ \beta_{1j}\end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 & \sigma_{01}\\ \sigma_{01} & \sigma_1^2\end{pmatrix}\right)$$

Agresti and Hartzel (2000) also made the additional assumption $\sigma_{01} = 0$, so that, coupled with the contrast coding used for x_{ij}, the total variance $\mathrm{Var}(\beta_{0j} + \beta_{1j}x_{ij})$ would be constant in both treatment groups. The overall log odds ratio is represented by β_1, and this is estimated as $0.57 = \log 1.76$.

In an I-prior model, the Bernoulli probabilities p_{ij} are regressed against the treatment group indicators x_{ij} and also the patients' study group j via the regression function f and a probit link:

$$\begin{aligned}\Phi^{-1}(p_{ij}) &= f(x_{ij}, j) \\ &= f_1(x_{ij}) + f_2(j) + f_{12}(x_{ij}, j).\end{aligned}$$

We have decomposed our function f into three parts: f_1 represents the treatment effect, f_2 represents the effect of the study groups, and f_{12} represents the interaction effect between the treatment and study group on the modelled probabilities. As both x_{ij} and j are nominal variables, the functions f_1 and f_2 both lie in the Pearson RKHS of functions $\mathcal{F}_1$ and $\mathcal{F}_2$, each with RKHS scale parameters λ_1 and λ_2. As such, it does not matter how the x_{ij} variables are coded (dummy coding 0, 1 vs. centred coding -0.5, 0.5) as the scaling of the function is determined by the RKHS scale parameters. The interaction effect f_{12} lies in the RKHS tensor product $\mathcal{F}_1 \otimes \mathcal{F}_2$. In I-prior modelling, there are only two parameters to estimate, while in the standard logistic random-effects model, there are six. The results of the I-prior fit are summarised in the table below.

tab:mod.com
pare.smoke

Table 5.4: Results of the I-prior model fit for three models.

Model	ELBO	Error rate (%)	Brier score	No. of parameters
f_1	-3210.76	23.65	0.179	1
$f_1 + f_2$	-3142.24	29.30	0.206	2
$f_1 + f_2 + f_{12}$	-3091.20	23.48	0.168	2

The approximated marginal log-likelihood value for the I-prior model (i.e. variational lower bound), the Brier score for each model and the number of RKHS scale parameters estimated in the model are reported in Table 5.4. Three models were fitted: 1) A model with only the treatment effect; 2) A model with a treatment effect and a study group effect; and 3) Model 2 with the additional assumption that treatment effect varies across study groups. Model 1 disregards the study group effects, while Model 2 assumes that the effectiveness of the nicotine gum treatment does not vary across study groups (akin to a varying-intercept model). A model comparison using the evidence lower bound indicates that Model 3 has the highest value, and the difference is significant from a Bayes factor standpoint ($\text{BF}(M_3, M_1)$ and $\text{BF}(M_3, M_2)$ are both greater than 150). The misclassification rate and Brier score indicates good predictive performance of the models, and there is not much to distinguish between the three although Model 3 is the best out of the three models.

Unlike in the logistic random-effects model, where the log odds ratio can be read off directly from the coefficients, with an I-prior probit model the log odds ratio needs to be calculated manually from the fitted probabilities. The probabilities of interest are the probabilities of quitting smoking under each treatment group for each study group j - call these $p_j(\text{treatment})$ and $p_j(\text{control})$. That is,

$$\begin{aligned}
p_j(\text{treatment}) &= \Phi\big(\tilde\mu(\text{treatment}, j)\big) \\
p_j(\text{control}) &= \Phi\big(\tilde\mu(\text{control}, j)\big),
\end{aligned}$$

where $\tilde\mu$ represents the posterior mean estimate for the regression function given in Which section?. These log odds ratio for each study group can then be calculated as usual. For the overall log odds ratio, the probabilities that are used are the averaged probabilities weighted according to the sample sizes in each group. This has been calculated as $0.55 = \log 1.73$, slightly lower than both the raw log odds ratio and the log odds ratio estimated by the logistic random-effects model. This can perhaps be attributed to some shrinkage of the estimated probabilities due to placing a prior with zero mean on the regression functions. The credibility intervals in Figure 5.7 for the log odds ratios under an I-prior are also noticeably narrower compared to the multilevel model fitted.

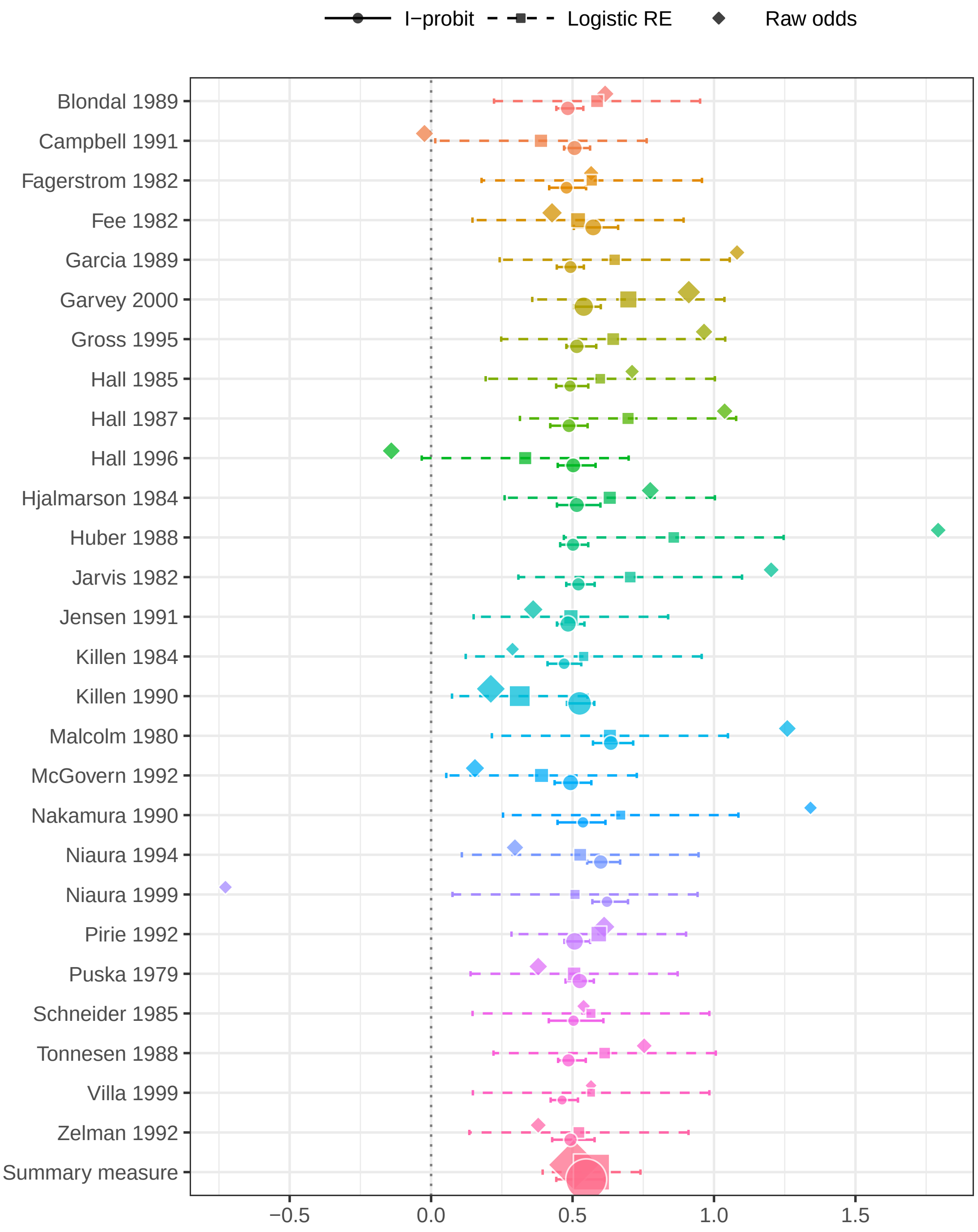


Figure 5.7: Forest plot of effect sizes (log odds ratios) in each group as well as the overall effect size together with their 95% confidence bands. The plot compares the raw log odds ratios, the logistic random-effect estimates, and the I-prior estimates. Sizes of the points indicate the relative sample sizes per study group.

fig:smoke.forest.plot

5.7.3 Multiclass classification: Vowel recognition data set

We illustrate multiclass classification using I-priors on a speech recognition data set[4] with $m = 11$ classes to be predicted from digitized low pass filtered signals generated from voice recordings. Each class corresponds to a vowel sound made when pronouncing a specific word. The words that make up the vowel sounds are shown in Table 5.5. Each word was uttered once by multiple speakers, and the data are split into a training and a test set. Four males and four female speakers contributed to the training set, while four male and three female speakers contributed to the test set. The recordings were manipulated using speech processing techniques, such that each speaker yielded six frames of speech from the eleven vowels, each with a corresponding 10-dimensional numerical input vector (the predictors). This means that the size of the training set is 528, while 462 data points are available for testing the predictive performance of the models. This data set is also known as Deterding's vowel recognition data (after the original collector, Deterding, 1989). Machine learning methods such as neural networks and nearest neighbour methods were analysed by Robinson (1989).

Table 5.5: The eleven words that make up the classes of vowels.

tab:vowel

Class	Label	Vowel	Word	Class	Label	Vowel	Word
1	hid	iː	heed	7	hOd	ɒ	hod
2	hId	ɪ	hid	8	hod	ɔː	hoard
3	hEd	ɛ	head	9	hUd	ʊ	hood
4	hAd	a	had	10	hud	uː	who'd
5	hYd	ʌ	hud	11	hed	əː	heard
6	had	ɑː	hard				

We will fit the data using an I-probit model with the canonical linear kernel, fBm-0.5 kernel, and the SE kernel with lengthscale $l = 1$. Each model took roughly 13 seconds per iteration in fitting the training data set ($n = 528$). In particular, the canonical kernel model took a long time to converge, with each variational inference iteration improving the lower bound only slighly each time. In contrast, both the fBm-0.5 and SE model were quicker to converge. Multiple restarts from different random seeds were conducted, and we found that they all converged to a similar lower bound value. This alleviates any concerns that the model might have converged to different multiple local optima.

[4] Data is publicaly available from the UCI Machine Learning Repository, URL: https://archive.ics.uci.edu/ml/datasets/Connectionist+Bench+(Vowel+Recognition+-+Deterding+Data).

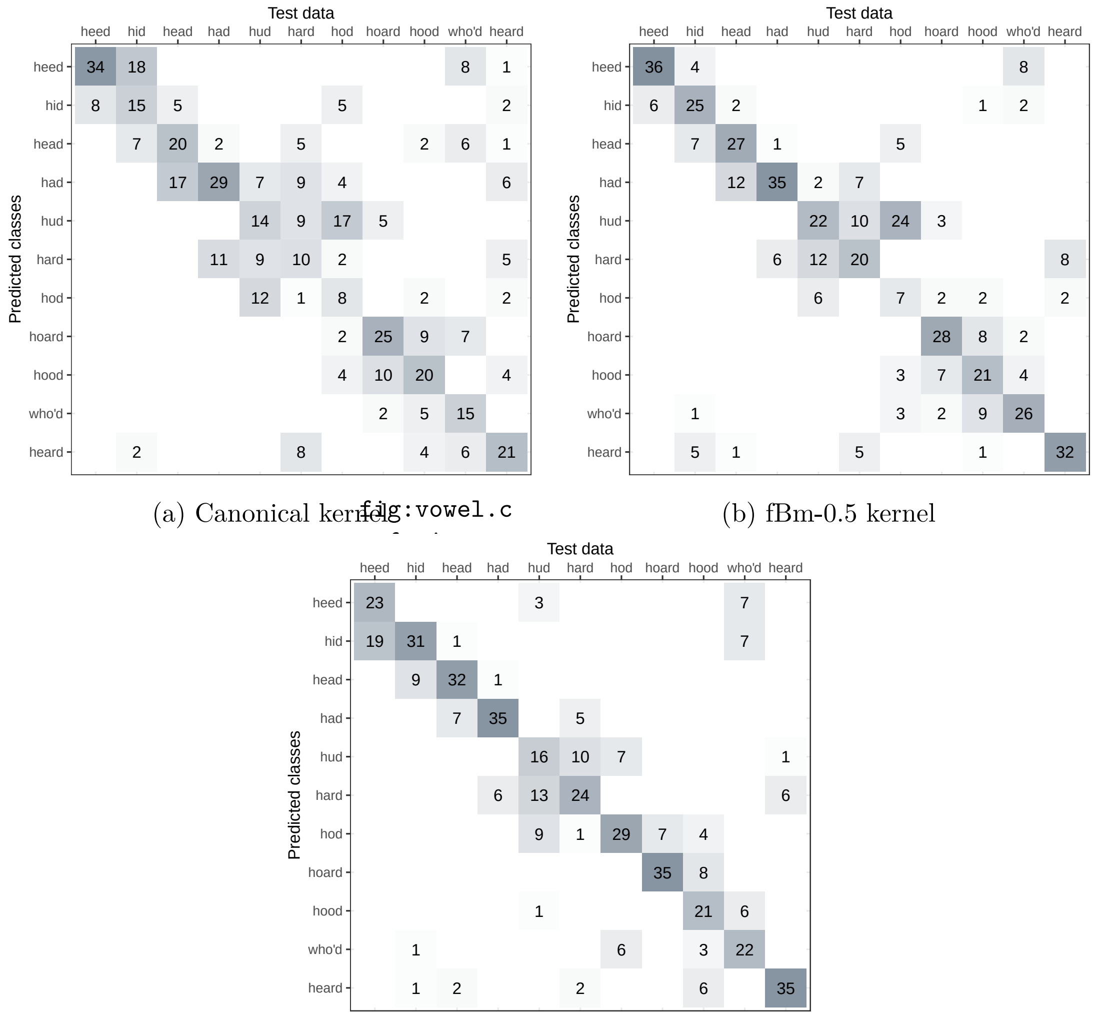
tab:vowel.tab

Table 5.6: Results of various classification methods for the vowel data set.

	Error rate (%)	
Model	Train	Test
I-probit		
Linear	29	54
Smooth (fBm-0.5)	22	40
Smooth (SE-1.0)	7	34
Others		
Linear regression	48	67
Logistic regression	22	51
Linear discriminant analysis	32	56
Quadratic discriminant analysis	1	53
Decision trees	5	54
Neural networks		45
k-nearest neighbours		44
FDA/BRUTO	6	44
FDA/MARS	13	39

A good way to visualise the performance of model predictions is through a confusion matrix, as shown in Figure 5.8. The numbers in each row indicate the instances of a predicted class, while the numbers in the column indicate instances of the actual classes, while nil values are indicated by blank cells.

Comparisons to other methods that had been used to analyse this data set is given in Table 5.6. In particular, the I-probit model is compared against 1) linear regression; 2) logistic linear regression; 3) linear and quadratic discriminant analysis; 4) decision trees; 5) neural networks; 6) k-nearest neighbours; and 7) flexible discriminant analysis. All of these methods are described in further detail in Friedman et al. (2001, Ch.4 & 12, Table 12.3). The I-probit model using both the fBm-0.5 and SE kernel offers one of the best out-of-sample classification error rates (34.4%) of all the methods compared. The linear I-probit model is seen to be comparable to logistic regression, linear and quadratic discrimant analysis, and also decision trees. It also provides significant improvement over multiple linear regression.



	heed	hid	head	had	hud	hard	hod	hoard	hood	who'd	heard
heed	34	18								8	1
hid	8	15	5				5				2
head		7	20	2		5			2	6	1
had			17	29	7	9	4				6
hud					14	9	17	5			
hard				11	9	10	2				5
hod					12	1	8		2		2
hoard							2	25	9	7	
hood							4	10	20		4
who'd								2	5	15	
heard		2				8			4	6	21

fig:vowel.confusion.matrix1

(a) Canonical kernfig:vowel.c

	heed	hid	head	had	hud	hard	hod	hoard	hood	who'd	heard
heed	36	4								8	
hid	6	25	2						1	2	
head		7	27	1			5				
had			12	35	2	7					
hud					22	10	24	3			
hard				6	12	20					8
hod					6		7	2	2		2
hoard								28	8	2	
hood							3	7	21	4	
who'd		1					3	2	9	26	
heard		5	1			5			1		32

(b) fBm-0.5 kernel

	heed	hid	head	had	hud	hard	hod	hoard	hood	who'd	heard
heed	23				3					7	
hid	19	31	1							7	
head		9	32	1							
had			7	35		5					
hud					16	10	7				1
hard				6	13	24					6
hod					9	1	29	7	4		
hoard								35	8		
hood					1				21	6	
who'd		1					6		3	22	
heard		1	2			2			6		35

fig:vowel.confusion.matrix3

(c) SE kernel

Figure 5.8: Confusion matrices for the vowel classification problem in which predicted values were obtained from the I-probit models. The maximum value for any one cell is 42. Blank cells indicate nil values.

fig:vowel.confusion.matrix

5.7.4 Spatio-temporal modelling of bovine tuberculosis in Cornwall

Data containing the number of breakdows of bovine tubercolosis (BTB) in Cornwall, the locations of the infected animals, and the year of occurence is analysed. The interest, as motivated by veterinary epidimiology, is to understand whether or not there is spatial segregation between the herds, and whether there is a time-element to presence or absence of this spatial segregation. There have been previous work done to analyse this data set: P. Diggle et al., 2005 developed a non-parametric method to estimate spatial segregation using a multivariate point process. The occurrences are modelled as Poisson point processes, and spatial segregation is said to have occured if the model-estimated type-specific breakdown probabilities at any given location are not significantly different from the sample proportions. The authors estimated the probabilities via kernel regression, and the test statistic of interest had to be estimated via Monte Carlo methods. Other work includes P. J. Diggle et al. (2013), who used a fully Bayes scheme for spatio-temporal multivariate log-Gaussian Cox processes, and implemented in the R package **lgcp** (Taylor et al., 2013).

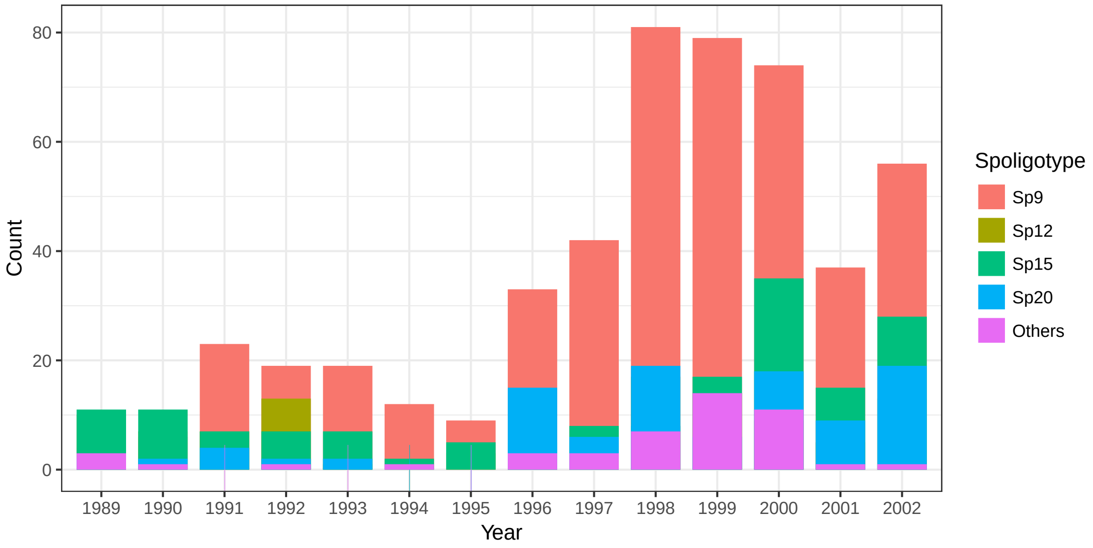


fig:plot.cow

Figure 5.9: Distribution of the different types (Spoligotypes) of bovine tubercolosis affecting herds in Cornwall over the period 1989 to 2002.

The data set contains $n = 919$ recorded cases over a span of 14 years. For each of the cases, spatial data pertaining to the location of the farm (Northings and Eastings, measured in kilometres) are available. Originally, 11 unique spoligotypes were recorded in the data, with the four most common spoligotypes being Sp9 ($m = 1$), Sp12 ($m = 2$), Sp15 ($m = 3$) and Sp20 ($m = 4$), as shown by the histogram in Figure 5.9. We had grouped the remaining seven spoligotypes into an 'Others' category ($m = 5$), so that the problem becomes a multinomial regression with five distinct outcomes.

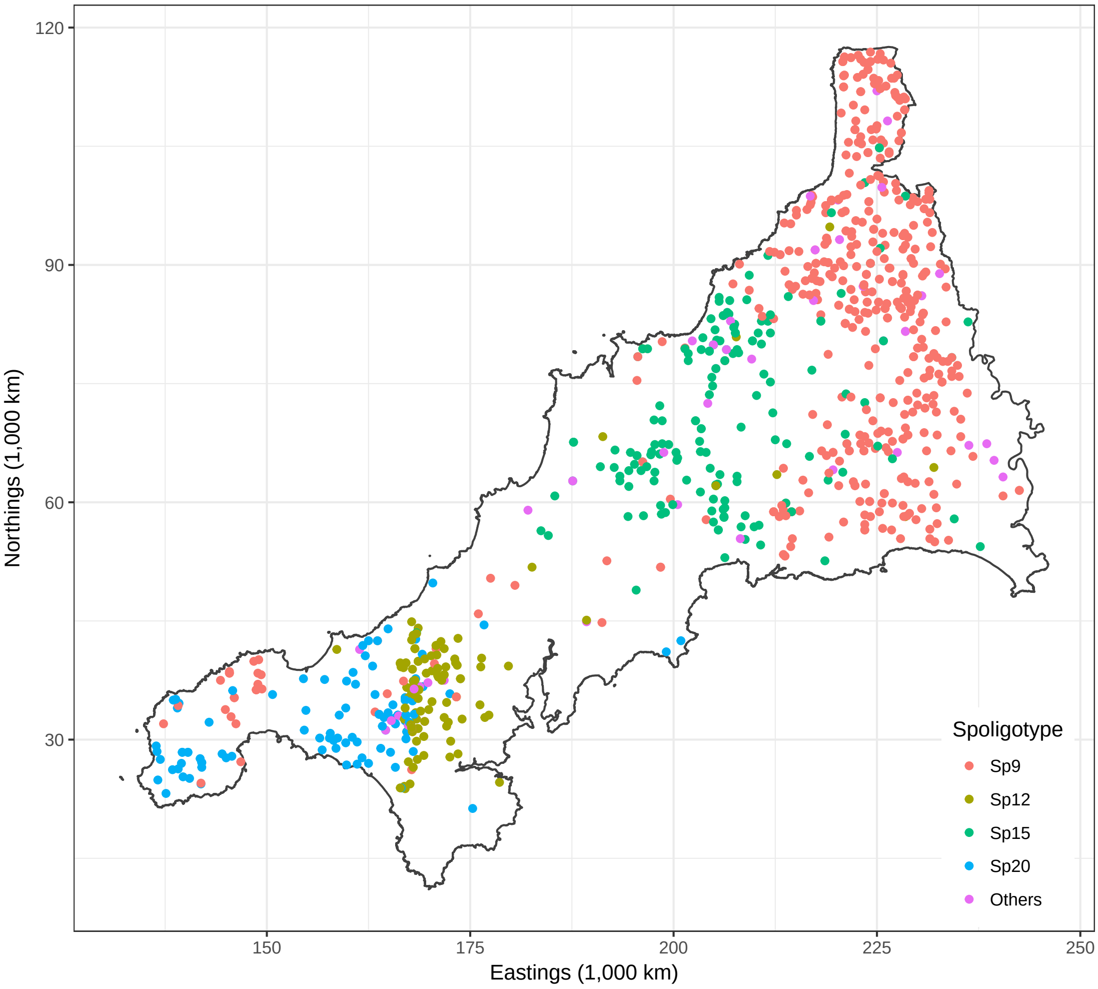


fig:plot.cornwall

Figure 5.10: Spatial distribution of all cases over the 14 years.

We are able to investigate any spatio-temporal patterns of infection using I-priors rather simply. Let p_{ij} denote the probability that a particular farm i is infected with a BTB disease with spoligotype $j \in \{1, \dots, 5\}$. We model the transformed probabilities $g(p_{ij})$ (as described in the categorical response chapter) as following a smooth function f which takes two covariates: the spatial data $x_1 \in \mathbb{R}^2$, and the temporal data x_2 (year of infection):

$$\begin{aligned} g(p_{ij}) &= f_j(x_1, x_2) \\ &= f_{1j}(x_1) + f_{2j}(x_2) + f_{12j}(x_1, x_2) \end{aligned}$$

We assume a smooth effect of space and time on the probabilities, and appropriate RKHSs for the functions $f_1 \in \mathcal{F}_1$ and $f_2 \in \mathcal{F}_2$ are the fBm-0.5 RKHS. Alternatively, as per P. Diggle et al. (2005), divide the data into four distinct time periods: 1) 1996 and earlier; 2) 1997 to 1998; 3) 1999 to 2000; and finally 4) 2001 to 2002. In this case, x_2 would indicate which period the infection took place in, and thus would have a nominal effect on the probabilities. An appropriate RKHS for f_2 in such a case would be the Pearson RKHS. In either case, the function $f_{12} \in \mathcal{F}_1 \otimes \mathcal{F}_2$ would be the "interaction effect", meaning that with such an effect present, the spatial distribution of the diseases are assumed to vary across the years.

We fitted four different models:

- ***M*₀: Intercept only**.

$$p_{ij} = g_j^{-1}\big(\alpha_k\big)_{k=1}^m$$

- ***M*₁: Spatial segregation**.

$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i)\big)_{k=1}^m$$

  $f_{1k} \in \mathcal{F}_1$ Pearson RKHS.

- ***M*₂: Spatio-temporal**.

$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i) + f_{2k}(t_i) + f_{12k}(x_i, t_i)\big)_{k=1}^m$$

  $f_{1k} \in \mathcal{F}_1$ Pearson RKHS, $f_{2k} \in \mathcal{F}_2$ fBm-0.5 RKHS, and $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

- ***M*₃: Spatio-period**.

$$p_{ij} = g_j^{-1}\big(\alpha_k + f_{1k}(x_i) + f_{2k}(t_i) + f_{12k}(x_i, t_i)\big)_{k=1}^m$$

  $f_{1k} \in \mathcal{F}_1$ Pearson RKHS, $f_{2k} \in \mathcal{F}_2$ Pearson RKHS, and $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

where g^{-1} is the link function described earlier. Model M_0 corresponds to a model which ignores any spatial or temporal effects (the baseline intercept only model). Model M_1

tab:table.b
tb

Table 5.7: Results of the fitted I-probit models.

	M_0: Intercepts only		M_1: Spatial only		M_2: Spatio-temporal		M_3: Spatio-period	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Intercept (Sp9)	0.948	0.033	1.364	0.033	1.401	0.033	1.395	0.033
Intercept (Sp12)	-0.173	0.033	-0.435	0.033	-0.506	0.033	-0.463	0.033
Intercept (Sp15)	0.103	0.033	-0.020	0.033	-0.008	0.033	-0.010	0.033
Intercept (Sp20)	-0.202	0.033	-0.775	0.033	-0.795	0.033	-0.783	0.033
Intercept (Others)	-0.676	0.033	-0.134	0.033	-0.091	0.033	-0.139	0.033
Scale (spatial)			0.194	0.003	-0.176	0.003	0.172	0.003
Scale (temporal)					-0.006	0.000	-0.004	0.000
Log-likelihood	-1187.47		-564.33		-537.23		-543.94	
Error rate (%)	46.25		19.26		18.06		18.50	
Brier score	0.249		0.143		0.136		0.138	

takes into account only spatial effects. Both models M_2 and M_3 account for spatio-temporal effects, but M_2 assumes a smooth effect of time, while M_3 segregates the points into four distinct time periods for analysis. Model comparison is easily done, and Table 5.7 indicates that model M_2 has the highest log-likelihood of the four models, making it the preferable model.

Alternatively, spatio-temporal effects of the BTB breakdowns can easily be inferred through the RKHS scale parameters. Let h_k, $k \in \{1,2\}$ denote the reproducing kernel of the spatial and temporal RKHSs respectively. Then, an I-prior on $f_j = f_{1j} + f_{2j} + f_{12j}$, $j = 1, \dots, 5$, takes the form

$$f_j(x_1, x_2) = \sum_{i=1}^n \big(\lambda_1 h_1(x_1, x_{i1}) + \lambda_2 h_2(x_2, x_{i2}) + \lambda_1\lambda_2 h_1(x_1, x_{i1}) h_2(x_2, x_{i2})\big) w_{ij}$$

where it is assumed $(w_{i1}, \dots, w_{i5})^\top \overset{\text{iid}}{\sim} \text{N}(\mathbf{0}, \mathbf{I}_5)$. The hypothesis of temporal significance is the same as testing the significance of the λ_2 parameter, while the test of both spatial and temporal effects are conducted on λ_1 and λ_2 simultaneously. From Chapter X, we know that these scale parameters follow a normal posterior distribution, so we can calculate the Z-scores by dividing the mean by its corresponding standard deviation. Absolute values greater than three would satisfy a Bayesian hypothesis test of significance at the 0.01 level. The conclusion from Table 5.7 is that the data supports a hypothesis for a spatio-temporal or spatio-period model.

For a more visual approach, we can look at the plots of the surface probabilities. To obtain these probabilities, we first determined the spatial points (Northings and Eastings) which fall inside the polygon which makes up Cornwall. We then obtained predicted probabilities for each class of disease at each location. Figure 5.11 was obtained using the model with spatial covariates only, thus ignoring any temporal effects. In the case of the spatio-temporal model, we used the model which had the period formulation for time (Model 3). This way, we can display the surface probabilities of the time periods in four plots only, which is more economical to exhibit within the margins of this thesis. Note that there is no issue with using the continuous time model—we have produced an animated gif image at `http://phd.haziqj.ml/examples/`, showing the evolution of the surface probabilities over time.

As the model suggests, there is indeed spatial segregation for the four most common spoligotypes, and this is also very prominently seen from Figure 5.11. In comparing the distribution of the spoligotypes over the years, we may refer to Figure 5.12. For

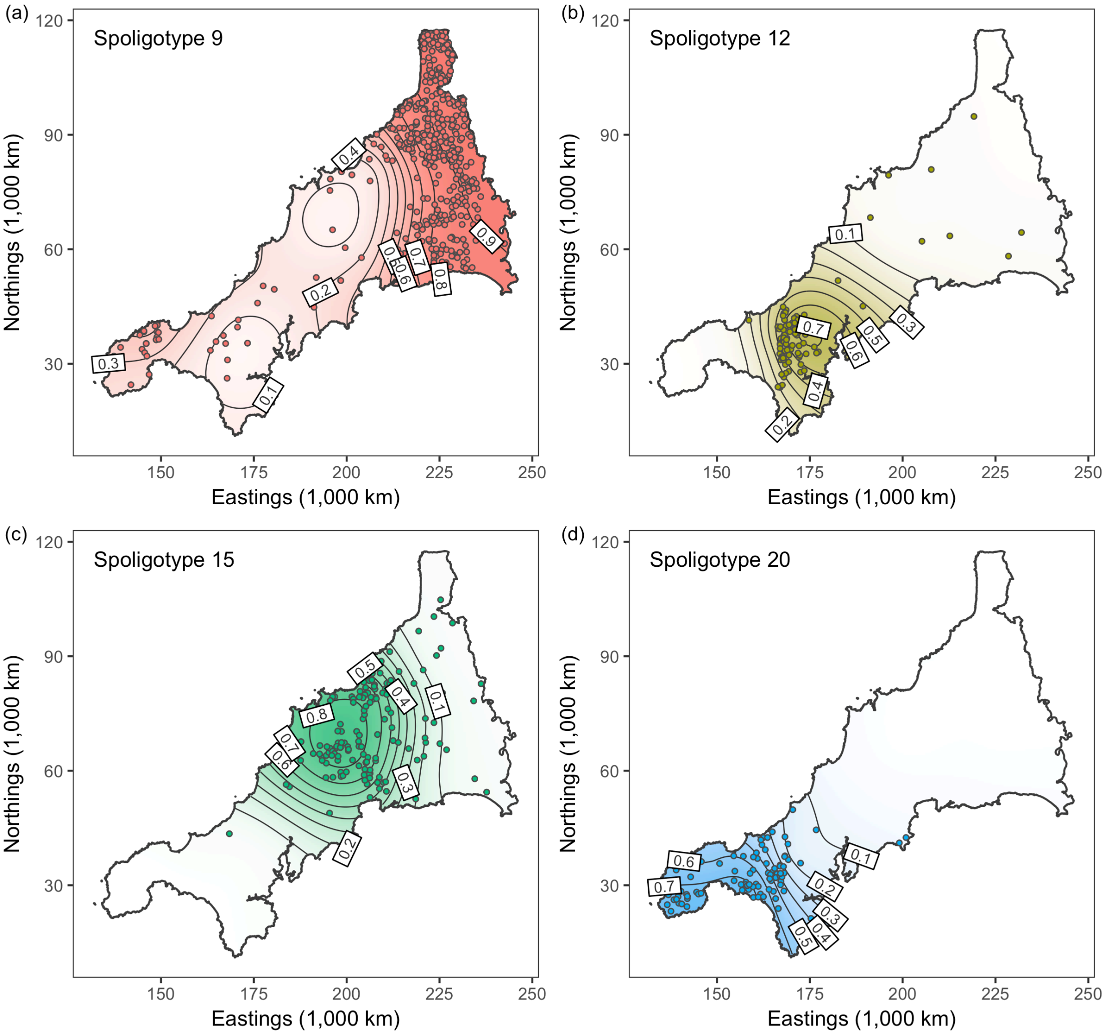


Figure 5.11: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium Mycobacterium bovis over the entire time period using model M_1.

fig:plot.btb

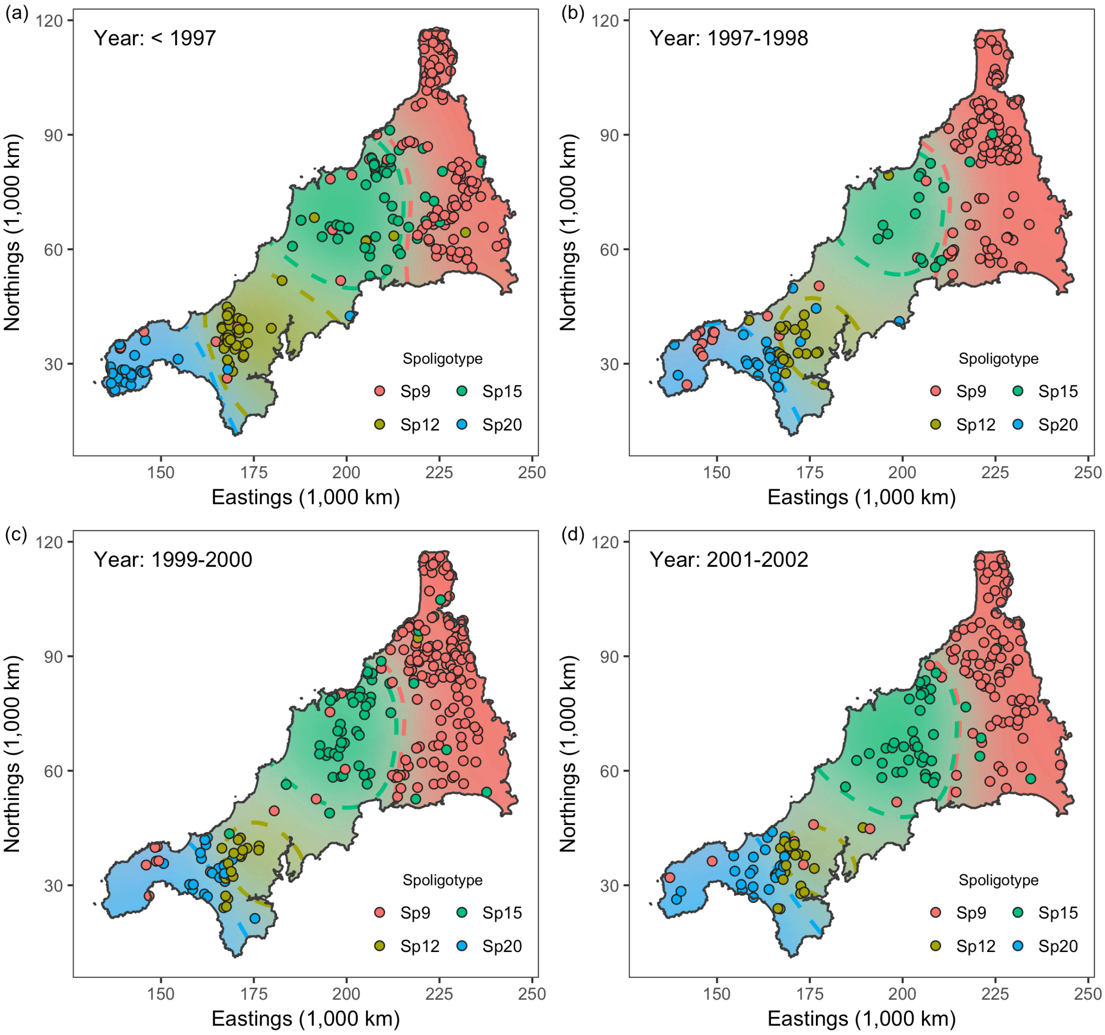


Figure 5.12: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium Mycobacterium bovis over four different time periods using model M_3.

fig:plot.temporal.btb

each time period, we superimpose the actual observations onto the predicted surface probabilities. In addition, coloured dotted lines are displayed to indicate the "decision boundaries" for each of the four spoligotypes. The most evident change is seen to the spatial distribution of spoligotype 12, with the decision boundary giving it a large area in years 1996 and earlier, but this steadily shrunk over the years. Spoligotype 9, which is most commonly seen in the east of Cornwall, seems to have made its way down to the south-west over the years. The other two spoligotypes seem to be rather constant over the years. This is supported also by the spatio-period model results in Table 5.7, where the test of nullity for the scale parameters of these two spoligotypes are not rejected.

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 $\mathbf{\Psi}$**. A limitation we had to face in this work was to treat $\mathbf{\Psi}$ as fixed. This limitation was in part due to the non-conjugate nature of the variational density for $\mathbf{\Psi}$. 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 Section 5.9.2 and **??** 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 $e : \{z_{ij} | i = 1, \dots, n, \ j = 1, \dots, m\} \to \mathbb{R}$. 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 $O(n^3 m)$ 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 compu-

tational efficiency could lie in sparse variational methods, a suggestion that was made to improve normal I-prior models as well.

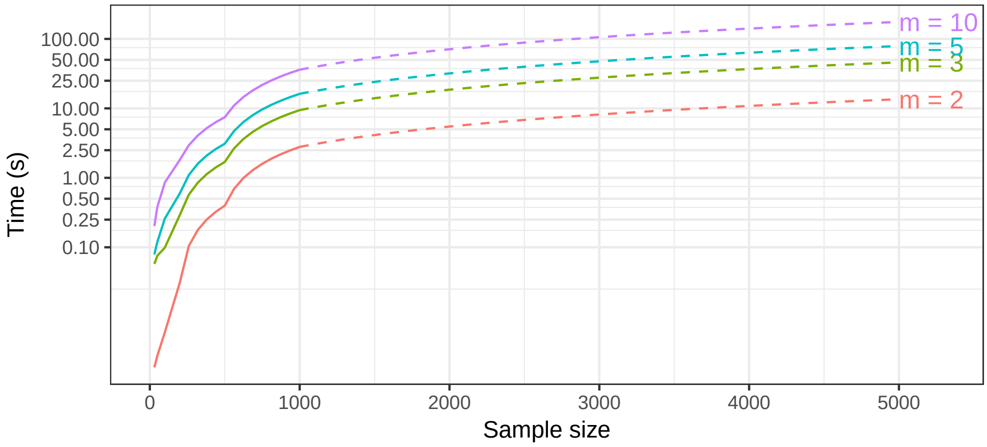


Figure 5.13: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes m. The solid line represents actual timings, while the dotted lines are linear extrapolations.

5.9 Miscellanea

5.9.1 A brief introduction to variational inference

sec:varintro

Consider a statistical model for which we have observations $\mathbf{y} := \{y_1, \dots, y_n\}$, but also some latent variables $\mathbf{z} := \{z_1, \dots, z_n\}$. Typically, in such models, there is a want to to evaluate the integral

$$\mathcal{I} = \int p(\mathbf{y}|\mathbf{z})p(\mathbf{z}) \, \mathrm{d}\mathbf{z}. \tag{5.23}$$

{eq:varint}

Models that include latent variables are plenty, for example: Gaussian mixture models, latent class analysis, factor models, random coefficient models, and so on. Marginalising out the latent variables in (5.23) is usually a precursor to obtaining a log-likelihood function to be maximised, in a frequentist setting. In Bayesian analysis, the $\mathbf{z}$'s are parameters which are treated as random, and the integral corresponds to the marginal density for $\mathbf{y}$, on which the posterior depends.

In many instances, for one reason or another, evaluation of $\mathcal{I}$ is difficult, in which case inference is halted unless a way of overcoming the intractable integral (5.23) is found. Here, we discus *variational inference* (VI), a fully Bayesian treatment of the statistical model with a deterministic algorithm, i.e. does not involve sampling from posteriors. The crux of variational inference is this: find a suitably close distribution function $q(z)$ that approximates the true posterior $p(\mathbf{z}|\mathbf{y})$, where closeness here is defined in the Kullback-Leibler divergence sense,

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

Posterior inference is then conducted using $q(\mathbf{z})$ in lieu of $p(\mathbf{z}|\mathbf{y})$. Advantages of this method are that 1) it is fast to implement computationally (compared to MCMC); 2) convergence is assessed simply by monitoring a single convergence criterion; and 3) it works well in practice, as attested to by the many studies implementing VI.

Briefly, we present the motivation behind variational inference and the minimisation of the KL divergence. Denote by $q(\cdot)$ some density function of $\mathbf{z}$. One may show that log marginal density (the log of the intractable integral (5.23)) holds the following bound:

$$\begin{aligned}
\log p(y) &= \log p(\mathbf{y},\mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \quad \text{(Bayes' theorem)} \\
&= \int \left\{ \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, \mathrm{d}z \quad \text{(expectations both sides)} \\
&= \mathcal{L}(q) + \mathrm{KL}(q\|p) \\
&\geq \mathcal{L}(q) \qquad (5.24)
\end{aligned}$$

{eq:varbound}

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

$$\begin{aligned}
\mathcal{L}(q) &= \int \log \frac{p(\mathbf{y},\mathbf{z})}{q(\mathbf{z})} q(\mathbf{z}) \, \mathrm{d}\mathbf{z} \\
&= \mathrm{E}_{\mathbf{z}\sim q} \log p(\mathbf{y},\mathbf{z}) + H(q), \qquad (5.25)
\end{aligned}$$

{eq:elbo1}

where H is the entropy functional, is known as the *evidence lower bound* (ELBO). Evidently, the closer q is to the true p, the better, and this is achieved by maximising $\mathcal{L}$, or equivalently, minimising the KL divergence from p to q. Note that the bound (5.24) achieves equality if and only if $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y})$, but of course the true form of the posterior is unknown to us—see Section 5.9.2 for a discussion. Maximising $\mathcal{L}(q)$ or minimising $\mathrm{KL}(q\|p)$ with respect to the density q is a problem of calculus of variations, which incidentally, is where variational inference takes its name. The astute reader will realise that $\text{KL}(q||p)$ is impossible to compute, since one does not know the true distribution $p(\mathbf{z}|\mathbf{y})$. Efforts are concentrated on maximising the ELBO instead.

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

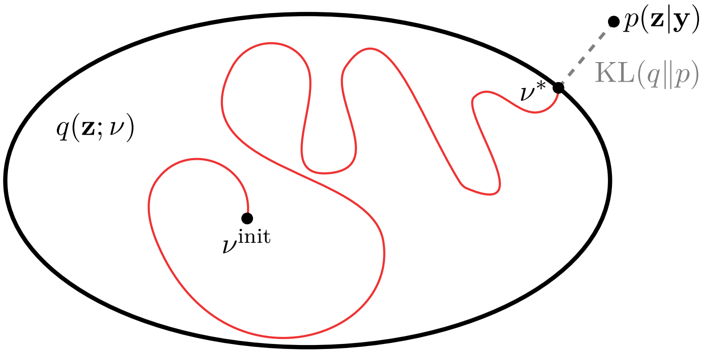


Figure 5.14: Schematic view of variational inference[5]. The aim is to find the closest distribution q (parameterised by a variational parameter ν) to p in terms of KL divergence within the set of variational distributions, represented by the ellipse.

The second type of constraint, and the one considered in this thesis, is simply an assumption that the approximate posterior q factorises into M disjoint factors. Partition $\mathbf{z}$ into M disjoint groups $\mathbf{z} = (z_{[1]}, \dots, z_{[M]})$. Note that each factor $z_{[k]}$ may be multidimensional. Then, the structure

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

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

Let us denote the distributions which minimise the Kullback-Leibler divergence (maximise the variational lower bound) by the use of tildes. By appealing to Bishop (2006,

[5] Reproduced from the talk by David Blei entitled ‘Variational Inference: Foundations and Innovations’, 2017. URL: https://simons.berkeley.edu/talks/david-blei-2017-5-1.

equation 10.9, p. 466), we find that for each $z_{[k]}$, $k = 1, \dots, M$, $\tilde q_k$ satisfies

$$\log \tilde q_k(z_{[k]}) = \operatorname{E}_{-k} \log p(\mathbf{y}, \mathbf{z}) + \text{const.} \tag{5.26}$$

{eq:qtilde}

where expectation of the joint log density of $\mathbf{y}$ and $\mathbf{z}$ is taken with respect to all of the unknowns $\mathbf{z}$, except the one currently in consideration $z_{[k]}$, under their respective $\tilde q_k$ densities.

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (5.26) to recognise it as a known log-density function, which is the case when exponential family distributions are considered. That is, suppose that each complete conditional $p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y})$, where $\mathbf{z}_{-k} = \{z_{[i]} | i \neq k\}$, follows an exponential family distribution

$$p(z_{[k]}|\mathbf{z}_{-k}, \mathbf{y}) = B(z_{[k]}) \exp\big(\langle \zeta_k(\mathbf{z}_{-k}, \mathbf{y}), z_{[k]} \rangle - A(\zeta_k)\big).$$

Then, from (5.26),

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

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

One notices that the optimal mean-field variational densities for each component are coupled with one another, in the sense that the distribution $\tilde q_k$ depends on the moments of the rest of the components $\mathbf{z}_{-k}$. For very simple problems, an exact solution for each $\tilde q_k$ can be found, but usually, the way around this is to employ an iterative procedure. The *coordinate ascent mean-field variational inference* (CAVI) algorithm cycles through each of the distributions in turn, updating them in sequence starting with arbitrary distributions as initial values.

Algorithm 2 The CAVI algorithm

alg:cavi

```
initialise Variational factors $q_k(z_{[k]})$
while ELBO $\mathcal{L}(q)$ not converged do
    for $k = 1, \dots, M$ do
        $\tilde q_k(z_{[k]}) \leftarrow \text{const.} \times \exp \operatorname{E}_{-k} \log p(\mathbf{y}, \mathbf{z})$        ▷ from (5.26)
    end for
    $\mathcal{L}(q) \leftarrow \operatorname{E}_{\mathbf{z} \sim \prod_k \tilde q_k} \log p(\mathbf{y}, \mathbf{z}) + \sum_{k=1}^m H\big[q_k(z_{[k]})\big]$        ▷ Update ELBO
end while
return $\tilde q(\mathbf{z}) = \prod_{k=1}^M \tilde q_j(z_{[k]})$
```

Each iteration of the CAVI brings about an improvement in the ELBO (hence the name coordinate ascent). The algorithm terminates when there is no more significant improvement in the ELBO, indicating a convergence of the CAVI. Blei et al. (2017) notes that the ELBO is typically a non-convex function, in which case convergence may be to (one of possibly many) local optima. A simple solution would be to restart the CAVI at multiple initial values, and the solution giving the highest ELBO is the distribution that is closest to the true posterior.

5.9.2 Variational methods and the EM algorithm

sec:varEM

Consider again the latent variable setup described in Section 5.9.1, but suppose the goal now is to maximise the (marginal) log-likelihood of the parameters θ of the model. We will see how the EM algorithm relates to minimising the KL divergence between a density $q(\mathbf{z})$ and the posterior of $\mathbf{z}$, and connect this idea to variational methods.

As we did in deriving (5.24), we decompose the marginal log-likelihood as

$$\log p(y|\theta) = \operatorname{E}\left[\log \frac{p(\mathbf{y},\mathbf{z}|\theta)}{q(\mathbf{z})}\right] - \operatorname{E}\left[\log \frac{p(\mathbf{z}|\mathbf{y},\theta)}{q(\mathbf{z})}\right] = \mathcal{L}(q) + \operatorname{KL}(q\|p).$$

This decomposition is shown in Figure 5.15. We realise that the KL divergence non-negative, and is zero exactly when $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y},\theta)$. Substituting this into the above equation yields the relationship

$$\begin{aligned}\log p(y|\theta) &= \operatorname{E}\left[\log \frac{p(\mathbf{y},\mathbf{z}|\theta)}{p(\mathbf{z}|\mathbf{y},\theta)}\right] - \operatorname{E}\left[\cancel{\log \frac{p(\mathbf{z}|\mathbf{y},\theta)}{p(\mathbf{z}|\mathbf{y},\theta)}}\right] \\ &= \operatorname{E} \log p(\mathbf{y},\mathbf{z}|\theta) - \operatorname{E} p(\mathbf{z}|\mathbf{y},\theta).\end{aligned}$$

[6] Reproduced from Bishop (2006, Figure 9.11).

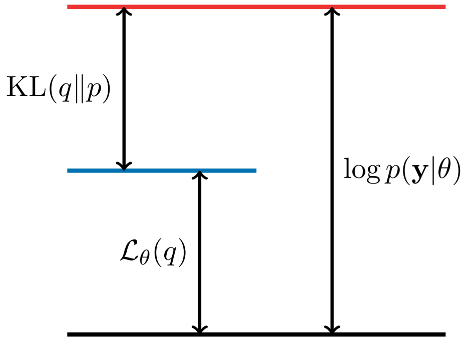


Figure 5.15: Illustration[6] of the decomposition of the log-likelihood into $\mathcal{L}_\theta(q)$ and $\text{KL}[q(\mathbf{z})\|p(\mathbf{z}|\mathbf{y})]$. The quantity $\mathcal{L}_\theta(q)$ is a lower bound for the log-likelihood.

fig:loglikdecomp

By taking expectations under the posterior distribution with known parameter values $\theta^{(t)}$, the term on the left becomes the Q function of the E-step

$$Q(\theta) = Q(\theta|\theta^{(t)}) = \text{E}_{\mathbf{z}}\left[\log p(\mathbf{y},\mathbf{z}|\theta) \,\middle|\, \mathbf{y}, \theta^{(t)}\right],$$

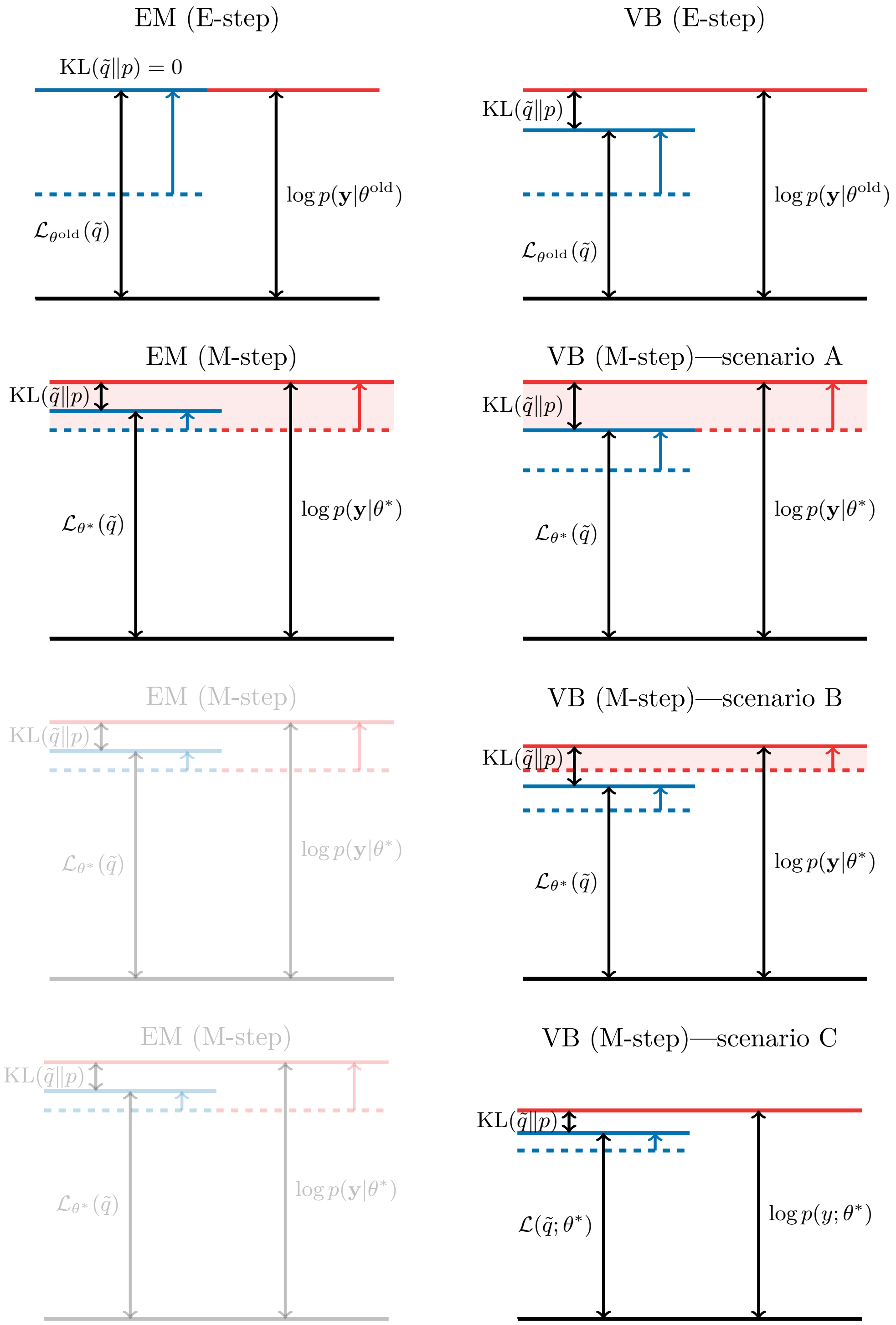
while the term on the left is an entropy term. Thus, minimising the KL divergence corresponds to the E-step in the EM algorithm. As a side fact, for any θ, we find that

$$\begin{aligned} \log p(\mathbf{y}|\theta) - \log p(\mathbf{y}|\theta^{(t)}) &= Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + \Delta\,\text{entropy} \\ &\geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}). \end{aligned}$$

because entropy differences are positive by Gibbs' inequality. We see that maximising Q with respect to θ (the M-step) brings about an improvement to the log-likelihood value. To summarise, the EM algorithm is seen as

- **E-step**. Maximise $\mathcal{L}_\theta\big[q(\mathbf{z})\big]$ with respect to q, keeping θ fixed. This is equivalent to minimising $\text{KL}(q\|p)$.
- **M-step**. Maximise $\mathcal{L}\big[q(\mathbf{z}|\theta)\big]$ with respect to θ, keeping q fixed.

When the true posterior distribution $p(\mathbf{z}|\mathbf{y})$ is not tractable, then the E-step becomes intractable as well. By constraining the maximisation in the E-step to consider q belonging to a family of tractable densities, the E-step yields a variational approximation $\tilde q$ to the true posterior. In Section 5.9.1, we saw that constraining q to be of a factorised form, then $\tilde q$ is a mean-field density. This form of the EM is known as *variational Bayes EM algorithm* (VB-EM) (Beal and Ghahramani, 2003).



EM (E-step)

VB (E-step)

EM (M-step)

VB (M-step)—scenario A

EM (M-step)

VB (M-step)—scenario B

EM (M-step)

VB (M-step)—scenario C

$\mathrm{KL}(\tilde{q}\|p) = 0$
$\log p(\mathbf{y}|\theta^{\mathrm{old}})$
$\mathcal{L}_{\theta^{\mathrm{old}}}(\tilde{q})$
$\mathrm{KL}(\tilde{q}\|p)$
$\log p(\mathbf{y}|\theta^{\mathrm{old}})$
$\mathcal{L}_{\theta^{\mathrm{old}}}(\tilde{q})$
$\mathrm{KL}(\tilde{q}\|p)$
$\log p(\mathbf{y}|\theta^{*})$
$\mathcal{L}_{\theta^{*}}(\tilde{q})$
$\mathrm{KL}(\tilde{q}\|p)$
$\log p(\mathbf{y}|\theta^{*})$
$\mathcal{L}_{\theta^{*}}(\tilde{q})$
$\mathrm{KL}(\tilde{q}\|p)$
$\log p(\mathbf{y}|\theta^{*})$
$\mathcal{L}_{\theta^{*}}(\tilde{q})$
$\mathrm{KL}(\tilde{q}\|p)$
$\log p(\mathbf{y}|\theta^{*})$
$\mathcal{L}_{\theta^{*}}(\tilde{q})$
$\mathrm{KL}(\tilde{q}\|p)$
$\log p(\mathbf{y}|\theta^{*})$
$\mathcal{L}_{\theta^{*}}(\tilde{q})$
$\mathrm{KL}(\tilde{q}\|p)$
$\log p(y;\theta^{*})$
$\mathcal{L}(\tilde{q};\theta^{*})$

Figure 5.16: Illustration of EM vs VB-EM. Whereas the EM guarantees an increase in log-likelihood value (red shaded region), the VB-EM does not.

In variational inference, a fully Bayesian treatment of the parameters is considered, with the aim of obtaining approximation to their posterior distributions. In VB-EM, the variational approximation is only performed on the latent, or 'missing' variables, to use the EM nomenclature. After a variational E-step, the M-step proceeds as usual, and as such, all of the material relating to the EM in the previous chapter is applicable. The VB-EM can also be seen as obtaining (approximate) maximum a posteriori estimates with diffuse priors on the parameters.

variational inference, EM algorithm, variational Bayes EM, differences, pros cons, MAP vs MLE, MAP vs fully Bayes

Appendix

5.10 Some distributions and their properties

This is a reference relating to the multivariate normal, matrix normal, truncated univariate and multivariate normal, Wishart, and gamma distributions which are collated from various sources for convenience. Of interest are probability density functions, first and second moments, and entropy (Definition 3.5, page 19).

5.10.1 Multivariate normal distribution

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

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

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

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

Then,

- ***Marginal distributions.***

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

- ***Conditional distributions.***

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

where

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

- ***Linear combinations.***

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

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

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

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

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

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

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

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

$$-\frac{1}{2}x^\top Ax + b^\top x = -\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) + \frac{1}{2}b^\top A^{-1}b$$

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

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

5.10.2 Matrix normal distribution

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

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

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

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

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

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

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

- **Non-commutative.** In general, $A \otimes B \neq B \otimes A$, but they are *permutation equivalent*, i.e. $A \otimes B \neq P(B \otimes A)Q$ for some permutation matrices P and Q.
- **The mixed product property.** $(A \otimes B)(C \otimes D) = AC \otimes BD$.

- **Inverse**. $A \otimes B$ is invertible if and only if A and B are both invertible, and $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$.
- **Transpose**. $(A \otimes B)^\top = A^\top \otimes B^\top$.
- **Determinant**. If A is $n \times n$ and B is $m \times m$, then $|A \otimes B| = |A|^m |B|^n$. Note that the exponent of $|A|$ is the order of B and vice versa.
- **Trace**. Suppose A and B are square matrices. Then $\operatorname{tr}(A \otimes B) = \operatorname{tr} A \operatorname{tr} B$.
- **Rank**. $\operatorname{rank}(A \otimes B) = \operatorname{rank} A \operatorname{rank} B$.
- **Matrix equations**. $AXB = C \Leftrightarrow (B^\top \otimes A) \operatorname{vec} X = \operatorname{vec}(AXB) = \operatorname{vec} C$.

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

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

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

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

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

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

Some useful properties of the matrix normal distribution are listed:

- **Expected values.**

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

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

5.10.3 Truncated univariate normal distribution

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

- **Pdf**. $p(X|\mu,\sigma,a,b) = C^{-1}(2\pi\sigma^2)^{-1/2}e^{-\frac{1}{2\sigma^2}(X-\mu)^2} = \sigma C^{-1}\phi(\frac{x-\mu}{\sigma})$.
- **Moments**.

$$
\begin{aligned}
\mathrm{E}\, X &= \mu + \sigma\frac{\phi(\alpha)-\phi(\beta)}{C} \\
\mathrm{E}\, X^2 &= \sigma^2 + \mu^2 + \sigma^2\frac{\alpha\phi(\alpha)-\beta\phi(\beta)}{C} + 2\mu\sigma\frac{\phi(\alpha)-\phi(\beta)}{C} \\
\operatorname{Var} X &= \sigma^2\left[1 + \frac{\alpha\phi(\alpha)-\beta\phi(\beta)}{C} - \left(\frac{\phi(\alpha)-\phi(\beta)}{C}\right)^2\right]
\end{aligned}
$$

- **Entropy**.

$$\begin{aligned} H(p) &= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{2C} \\ &= \frac{1}{2}\log 2\pi e\sigma^2 + \log C + \frac{1}{2\sigma^2} \cdot \overbrace{\sigma^2 \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{C}}^{\operatorname{Var} X - \sigma^2 + (\operatorname{E} X - \mu)^2} \\ &= \frac{1}{2}\log 2\pi\sigma^2 + \log C + \frac{1}{2\sigma^2}\operatorname{E}[X-\mu]^2 \end{aligned}$$

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

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

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

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

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

5.10.4 Truncated multivariate normal distribution

Consider the restriction of $X \sim \text{N}_d(\mu, \Sigma)$ to a convex subset[7] $\mathcal{A} \subset \mathbb{R}^d$. Call this distribution the truncated multivariate normal distribution, and denote it $X \sim {}^{\text{t}}\text{N}_d(\mu, \Sigma, \mathcal{A})$.

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

The pdf is $p(X|\mu, \Sigma, \mathcal{A}) = C^{-1}\phi(X|\mu, \Sigma)\, \mathbb{1}[X \in \mathcal{A}]$, where

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

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

Sampling from a truncated multivariate normal distribution is described by Robert (1995), who used a Gibbs-based approach, which we now describe. Assume that the one-dimensional slices of $\mathcal{A}$

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

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

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

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

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

which means that we need only compute one global inverse Σ^{-1}. Therefore, the Gibbs sampler makes draws from truncated normal distributions in the following sequence:

1. Draw ...

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

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

thm:contruncn

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

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

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

where ϕ *is the pdf of a standard normal distribution and*

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

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

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

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

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

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

where we had defined

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

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

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

Proof. See Section 5.11 for the proof. □

5.11 Proofs related to conically truncated multivariate normal distribution

apx:contrun proof

5.11.1 Proof of Lemma 5.4: Pdf

Using the fact that $\int p(x)\,\mathrm{d}x = 1$, and that

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

the proof follows directly.

5.11.2 Proof of Lemma 5.4: Moments

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

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

and in particular, we have

$$\operatorname{E}[Y - \mu] = -\sigma \frac{\phi(\beta)}{\Phi(\beta)} \tag{5.27}$$

{eq:tnXminMu}

$$\operatorname{E}[Y - \mu]^2 - \sigma^2 = -\sigma^2 \frac{\beta\phi(\beta)}{\Phi(\beta)} \tag{5.28}$$

{eq:tnXminMusq}

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

$$\begin{aligned}
\operatorname{E} g(X_i) &= C^{-1} \int \cdots \int g(x_i) \, \mathbb{1}[x_k < x_j, \forall k \neq j] \prod_{k=1}^d p(x_k) \, \text{d}x_1 \cdots \text{d}x_d \\
&= C^{-1} \frac{\Phi((x_j - \mu_j)/\sigma_j)}{\Phi((x_j - \mu_j)/\sigma_j)} \iint g(x_i) \, \mathbb{1}[x_i < x_j] p(x_i) p(x_j) \prod_{\substack{k=1 \\ k \neq i,j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \text{d}x_i \, \text{d}x_j \\
&= C^{-1} \int \operatorname{E}_{X_i \sim {}^{\text{t}}\text{N}(\mu_i, \sigma_i^2, -\infty, x_j)} \, g(X_i) \prod_{\substack{k=1 \\ k \neq j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j) \, \text{d}x_j
\end{aligned} \tag{5.29}$$

{eq:tnproofi}

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

$$\begin{aligned}
\operatorname{E} g(X_j) &= C^{-1} \int \cdots \int g(x_j) \, \mathbb{1}[x_k < x_j, \forall k \neq j] \prod_{k=1}^d p(x_k) \, \text{d}x_1 \cdots \text{d}x_d \\
&= C^{-1} \int g(x_j) \prod_{\substack{k=1 \\ k \neq j}}^d \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) p(x_j) \, \text{d}x_d.
\end{aligned} \tag{5.30}$$

{eq:tnproofj}

Plugging in (5.27) for $g(X_i) = X_i - \mu_i$ in (5.29) we get

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

where Z is the distribution of $\mathrm{N}(0,1)$, and we had used a change of variable $x_j = \sigma_j z + \mu_j$, so that $p(x_j) = \sigma_j^{-1} \phi(z)$ and $\mathrm{d}x_j = \sigma_j \mathrm{d}z$. For the j'th component, substitute $g(x_j) = x_j - \mu_j$ in (5.30) to get

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

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

For the second moments, plug in (5.28) for $g(X_i) = (X_i - \mu_i)^2 - \sigma_i^2$ in (5.29) to get

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

And similarly, for the j'th component

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

Lastly, we use the following result in the derivation above.

lem:EZgZ

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

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

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

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

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

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

$$\begin{aligned} g_2(z) &= \Phi_2 \\ \Rightarrow \dot{g}_2(z) &= \sigma_2 \phi_2 = \sum_{\substack{i=1 \\ i \neq 1}}^{2} \left[\sigma_i \phi_i \sum_{\substack{k=1 \\ k \neq 1,2}}^{2} \Phi_k \right]. \end{aligned}$$

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

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

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

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

is found to be

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

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

5.11.3 Proof of Lemma 5.4: Entropy

As a direct consequence of the definition of entropy,

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

5.12 Derivation of the variational densities

In what follows, the implicit dependence of the densities on the parameters of the model θ are dropped. We derive a mean-field variational approximation of

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

The first line is by assumption, while the second line follows from an induced factorisation on the latent propensities, as we will see later. Recall that the optimal mean-field variational density $\tilde q$ satisfy

$$\log \tilde q(\mathbf{y}^*) = \operatorname{E}_{\mathbf{w}\sim\tilde q}\big[\log p(\mathbf{y},\mathbf{y}^*,\mathbf{w})\big] + \text{const.} \qquad \text{(from ??)}$$

$$\log \tilde q(\mathbf{w}) = \operatorname{E}_{\mathbf{y}^*\sim\tilde q}\big[\log p(\mathbf{y},\mathbf{y}^*,\mathbf{w})\big] + \text{const.} \qquad \text{(from ??)}$$

The joint likelihood $p(\mathbf{y},\mathcal{Z})$ is given by

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

For reference, the three 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

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

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

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

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

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

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

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

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

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

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 Lemma 5.4 provides a simplification.

Remark 5.7. 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.

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

The terms involving $\mathbf{w}$ in (5.26) are the $p(\mathbf{y}^*|\boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi})$ and $p(\mathbf{w}|\boldsymbol{\Psi})$ terms, and the rest are absorbed into the constant. The easiest way to derive $\tilde{q}(\mathbf{w})$ is to vectorise $\mathbf{y}^*$ and $\mathbf{w}$. We know that

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

and

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

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

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

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

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

and making a first-order approximation $(\mathrm{E}\,\mathbf{A})^{-1} \approx \mathrm{E}[\mathbf{A}^{-1}]$[8],

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

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.

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

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

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

$$\begin{aligned}
\operatorname{vec}\tilde{\mathbf{w}} &= \tilde{\mathbf{V}}_w(\operatorname{diag}(\tilde\psi_1, \dots, \tilde\psi_m) \otimes \tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top) \\
&= \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \dots, \tilde{\mathbf{V}}_{w_m}) \operatorname{diag}(\tilde\psi_1\tilde{\mathbf{H}}_\eta, \dots, \tilde\psi_m\tilde{\mathbf{H}}_\eta) \operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top) \\
&= \operatorname{diag}(\tilde\psi_1\tilde{\mathbf{V}}_{w_1}\tilde{\mathbf{H}}_\eta, \dots, \tilde\psi_m\tilde{\mathbf{V}}_{w_m}\tilde{\mathbf{H}}_\eta)(\tilde{\mathbf{y}}^* - \mathbf{1}_n\tilde{\boldsymbol{\alpha}}^\top) \\
&= \begin{pmatrix} \overset{\tilde{\mathbf{w}}_{\cdot 1}}{\tilde\psi_1\tilde{\mathbf{V}}_{w_1}\tilde{\mathbf{H}}_\eta(\tilde{\mathbf{y}}^*_{\cdot 1} - \tilde\alpha_1\mathbf{1}_n)} & \overset{\cdots}{\cdots} & \overset{\tilde{\mathbf{w}}_{\cdot m}}{\tilde\psi_m\tilde{\mathbf{V}}_{w_m}\tilde{\mathbf{H}}_\eta(\tilde{\mathbf{y}}^*_{\cdot m} - \tilde\alpha_m\mathbf{1}_n)} \end{pmatrix}^\top.
\end{aligned}$$

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

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

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

[8] 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.

of the trace directly, we get

$$\begin{aligned}\operatorname{tr}(\mathbf{C}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]) &= \sum_{i,j=1}^m \mathbf{C}_{ij}\,\mathrm{E}[\mathbf{w}^\top\mathbf{D}\mathbf{w}]_{ij} \\ &= \sum_{i,j=1}^m \mathbf{C}_{ij}\,\mathrm{E}[\mathbf{w}_{\cdot i}^\top\mathbf{D}\mathbf{w}_{\cdot j}]. \end{aligned}\tag{5.31}$$

{eq:trCEwDw }

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

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

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

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

where δ is the Kronecker delta. Continuing on (5.31) leads us to

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

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

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

5.13 Deriving the ELBO expression

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

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

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

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

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

where C_i is the normalising constant for the distribution of multivariate truncated normal $\mathbf{y}_{i\cdot}^* \sim {}^{\mathrm{t}}\mathrm{N}(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$.

5.13.2 Terms involving distributions of w

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

Bibliography

agresti2000tutorial Agresti, Alan and Jonathan Hartzel (2000). “Tutorial in biostatistics: Strategies comparing treatment on binary response with multi-centre data”. In: *Statistics in medicine* 19, pp. 1115–1139.

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.

breiman2001random Breiman, Leo (2001). “Random forests”. In: *Machine learning* 45.1, pp. 5–32.

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

cannings2017random Cannings, Timothy I and Richard J Samworth (2017). “Random-projection ensemble classification”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology) with discussion* 79.4, pp. 959–1035.

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*.

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

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

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

deterding1989speaker Deterding, David Henry (1989). “Speaker normalization for automatic speech recognition”. PhD thesis. University of Cambridge.

diggle2013spatial Diggle, Peter J, Paula Moraga, Barry Rowlingson, and Benjamin M Taylor (2013). “Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm”. In: *Statistical Science*, pp. 542–563.

diggle2005nonparametric Diggle, Peter, Pingping Zheng, and Peter Durr (2005). “Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 54.3, pp. 645–658.

friedman2001elements Friedman, Jerome, Trevor Hastie, and Robert Tibshirani (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics New York.

geweke1989bayesian Geweke, John (1989). “Bayesian inference in econometric models using Monte Carlo integration”. In: *Econometrica: Journal of the Econometric Society*, pp. 1317–1339.

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.

guvenir1997supervised Guvenir, H Altay, Burak Acar, Gulsen Demiroz, and Ayhan Cekin (1997). “A supervised machine learning algorithm for arrhythmia analysis”. In: *Computers in Cardiology 1997*. IEEE, pp. 433–436.

hajivassiliou1998method Hajivassiliou, Vassilis A and Daniel L McFadden (1998). “The method of simulated scores for the estimation of LDV models”. In: *Econometrica*, pp. 863–896.

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.

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

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

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.

keane1994solution Keane, Michael P and Kenneth I Wolpin (1994). “The solution and estimation of discrete choice dynamic programming models by simulation and interpolation: Monte Carlo evidence”. In: *the Review of economics and statistics*, pp. 648–672.

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.

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

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

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

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

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

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

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

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

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

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

robinson1989dynamic Robinson, Anthony John (1989). “Dynamic error propagation networks”. PhD thesis. University of Cambridge.

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

skrondal2004generalized Skrondal, Anders and Sophia Rabe-Hesketh (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

steinwart2008support Steinwart, Ingo and Andreas Christmann (2008). *Support vector machines*. Springer Science & Business Media.

taylor2013lgcp Taylor, Benjamin M, Tilman M Davies, Barry S Rowlingson, Peter J Diggle, et al. (2013). “lgcp: an R package for inference with spatial and spatio-temporal log-Gaussian Cox processes”. In: *Journal of Statistical Software* 52.4, pp. 1–40.

tibshirani2002diagnosis Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu (2002). "Diagnosis of multiple cancer types by shrunken centroids of gene expression". In: *Proceedings of the National Academy of Sciences* 99.10, pp. 6567–6572.

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

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