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

Consider polytomous response variables $\mathbf{y} = \{y_1, \ldots, y_n\}$, where each y_i takes on exactly one of the values from the set of m possible choices $\{1, \ldots, m\}$. 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 frequently interested in discrete choice models to explain and predict choices between several alternatives, such as consumers' choices 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 model (1.1) subject to normality assumptions (1.2) is not entirely appropriate for polytomous variables $\mathbf{y}$. 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, $i = 1, \ldots, n$, denoted by

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

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

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

where the notation $[\cdot]$ refers to the Iverson bracket[1]. As a side note, when there are only two possibilities for each outcome y_i, i.e. $m = 2$, we have the Bernoulli distribution. The class probabilities are made to depend on the covariates through the relationship

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

where $g : [0,1]^m \to \mathbb{R}^m$ is some specified link function. As we will see later, an underlying 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 assumptions, 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 (c.f. equation 5.10). Similar problems are encountered in mixed logistic or probit multinomial models (Breslow and Clayton, 1993; McCulloch et al., 2000) and also in Gaussian process classification (Neal, 1999; Rasmussen and Williams, 2006). In these models, Laplace approximation for maximum likelihood (ML) estimation or Markov chain Monte Carlo (MCMC) methods for Bayesian estimation are used. We instead explore a *variational approximation* to the marginal log-likelihood, and by extension, to the posterior density of the regression functions. The main idea is to replace the difficult posterior distribution with an approximation that is tractable to be used within an EM framework. As such, the computational work derived in the previous section is applicable for the estimation of I-probit models as well.

As in the normal I-prior model, the I-probit model estimated using a *variational EM algorithm* is seen as an empirical Bayes method of estimation, since the model parameters are replaced with their (pseudo) ML estimates. It is emphasised again, that working in such a semi-Bayesian framework allows fast estimation of the model in comparison to traditional MCMC, 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 posterior distribution of the regression function, which, as we shall see, is approximated to be normally distributed.

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

We derive the I-probit model through a latent variable motivation. It is convenient, as we did in Section 4.1.4, to again think of the responses $y_i \in \{1, \dots, m\}$ 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,

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

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

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

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

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

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

with α being the grand intercept, α_j group or class intercepts, and $f_j : \mathcal{X} \to \mathbb{R}$ a regression function belonging to some RKKS $\mathcal{F}$ of functions over the covariate set $\mathcal{X}$ with reproducing kernel h_η. We can see some semblance of this model with the one in (4.7), and ultimately the aim is to assign I-priors to the regression function of these latent variables, which we shall describe shortly. For now, write $\boldsymbol{\mu}(x_i) \in \mathbb{R}^m$ whose j'th component is $\alpha + \alpha_j + f_j(x_i)$, and realise that each $\mathbf{y}^*_{i\cdot} = (y^*_{i1}, \dots, y^*_{im})^\top$ has the distribution $\text{N}_m(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1})$, conditional on the data x_i, the intercepts $\alpha, \alpha_1, \dots, \alpha_m$, the evaluations of the functions at x_i for each class $f_1(x_i), \dots, f_m(x_i)$, and the error covariance matrix $\boldsymbol{\Psi}^{-1}$.

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

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

where $\phi(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the density of the multivariate normal with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. 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 (pmf) for the classes. For reference, we define our *probit link function* $g_j^{-1}(\cdot|\boldsymbol{\Psi}) : \mathbb{R}^m \to [0, 1]$ by the mapping

$$\boldsymbol{\mu}(x_i) \mapsto \int_{\mathcal{C}_j} \phi(\mathbf{y}^* | \boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}) \, \mathrm{d}\mathbf{y}^*. \tag{5.4}$$

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 classification model (Section 4.1.4, p. 8), 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 (4.6). Instead of doing the same, we take a different approach. Treat the α_j's in (5.2) as intercept parameters to estimate with the additional requirement that $\sum_{j=1}^m \alpha_j = 0$. Further, let $\mathcal{F}$ be a (centred) RKHS/RKKS of functions over $\mathcal{X}$ with reproducing kernel h_η. Now, consider putting an I-prior on the regression functions $f_j \in \mathcal{F}$, $j = 1 \ldots, m$, defined by

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

with $\mathbf{w}_{i\cdot} := (w_{i1}, \ldots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}(0, \boldsymbol{\Psi})$. This is similar to the naïve I-prior specification (4.7), 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, rather than the I-prior specification described in the naïve classification, bodes well down the line computationally.

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

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

$$
\begin{aligned}
y_{ij}^* &= \alpha + \alpha_j + \overbrace{f_0(x_i, j) + \sum_{k=1}^{n} h_\eta(x_i, x_k) w_{ik}}^{f_j(x_i)} + \epsilon_{ij} \\
\boldsymbol{\epsilon}_{i\cdot} &:= (\epsilon_{i1}, \dots, \epsilon_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}^{-1}) \\
\mathbf{w}_{i\cdot} &:= (w_{i1}, \dots, w_{im})^\top \overset{\text{iid}}{\sim} \mathrm{N}_m(\mathbf{0}, \boldsymbol{\Psi}).
\end{aligned}
\tag{5.5}
$$

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

In the above, we have made use of matrix normal distributions,denoted by $\mathrm{MN}(\cdot, \cdot)$. The definition and properties of matrix normal distributions can be found in (Appendix C.2, p. 15).

Before proceeding with estimating the I-probit model (5.5), we lay out several standing assumptions:

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

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

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

The parameters in the standard linear multinomial probit model is well known to be unidentified (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 that assumptions A4 and A5 are imposed. Secondly, all latent variables can be scaled by some positive constant $c \in \mathbb{R}_{>0}$ without changing which latent variable is largest. This means that m-variate normal distribution $\mathrm{N}_m\left(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}\right)$ of the underlying latent variables $\mathbf{y}_{i\cdot}^*$ would yield the same class probabilities as the multivariate normal distribution $\mathrm{N}_m\left(a\mathbf{1}_m + c\boldsymbol{\mu}(x_i), c^2\boldsymbol{\Psi}^{-1}\right)$, according to (5.3). Therefore, the multinomial probit model is not identified as there exists more than one set of parameters for which the categorical likelihood $\prod_{i,j} p_{ij}$ is the same.

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

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

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

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

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

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

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

Remark 5.1. Identification is most commonly achieved by fixing the latent propensities of one of the classes to zero and fixing one element the covariance matrix (Bunch, 1991; Dansie, 1985). Fixing the last class, say, to zero, i.e. $y^*_{im} = 0, \forall i = 1, \dots, n$ has the effect of shrinking $\boldsymbol{\Psi}$ to an $(m-1)$ matrix, and thus one more restriction needs to be made (typically, $\boldsymbol{\Psi}_{11}$ is set to one). This speaks to the fact that the absolute values of the latent propensities themselves do not matter, and only their relative differences do. 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 unity yields

$$\begin{aligned} p_{i1} &= \mathrm{P}(y^*_{i1} > y^*_{i2} = 0) \\ &= \mathrm{P}\left(\alpha_1 + f_1(x_i) + \epsilon_{i1} > 0 \,|\, \epsilon_{i1} \overset{\text{iid}}{\sim} \mathrm{N}(0,1)\right) \\ &= \Phi\big(\alpha_1 + f_1(x_i)\big) \end{aligned} \tag{5.6}$$

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

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

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

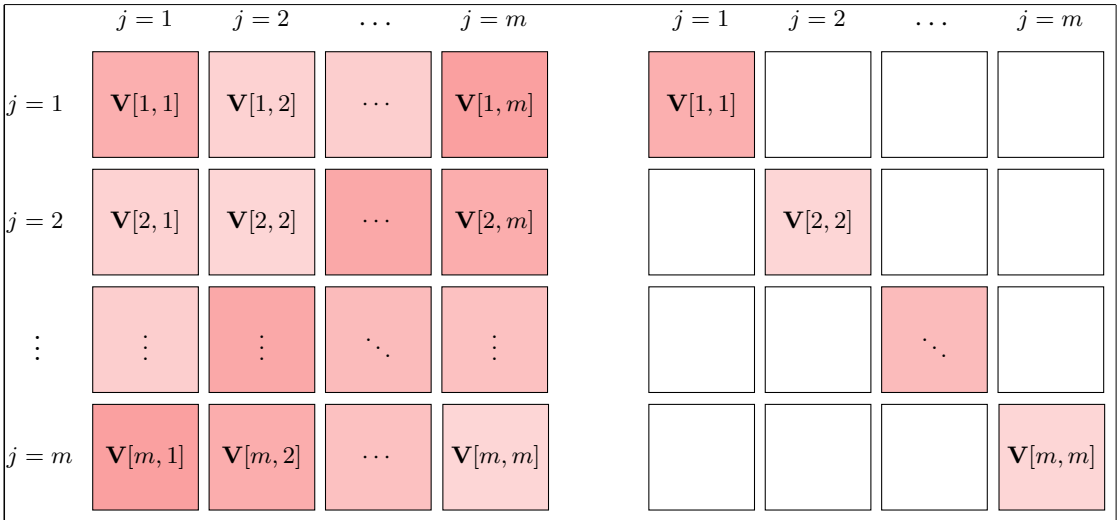


Figure 5.1: Illustration of the covariance structure of the full I-probit model (left) and the independent I-probit model (right). The full model has m^2 blocks of $n \times n$ symmetric matrices, and the blocks themselves are arranged symmetrically about the diagonal. The independent model, on the other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

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 $\boldsymbol{\Psi} = \text{diag}(\psi_1, \dots, \psi_m)$, 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 $y_{ij}^* \sim \text{N}(\mu_k(x_i), \sigma_j^2)$ independently 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

$$\begin{aligned} p_{ij} &= \int_{\{y_{ij}^* > y_{ik}^* | \forall k \neq j\}} \cdots \int \prod_{k=1}^{m} \left\{ \phi(y_{ik}^* | \mu_k(x_i), \sigma_k^2) \, \mathrm{d}y_{ik}^* \right\} \\ &= \int \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{y_{ij}^* - \mu_k(x_i)}{\sigma_k} \right) \phi(y_{ij}^* | \mu_j(x_i), \sigma_j^2) \, \mathrm{d}y_{ij}^* \\ &= \text{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\sigma_j}{\sigma_k} Z + \frac{\mu_j(x_i) - \mu_k(x_i)}{\sigma_k} \right) \right] \end{aligned} \tag{5.7}$$

where $Z \sim \text{N}(0, 1)$, $\Phi(\cdot)$ its cdf, and $\phi(\cdot | \mu, \sigma^2)$ is the pdf of $X \sim \text{N}(\mu, \sigma^2)$. 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.

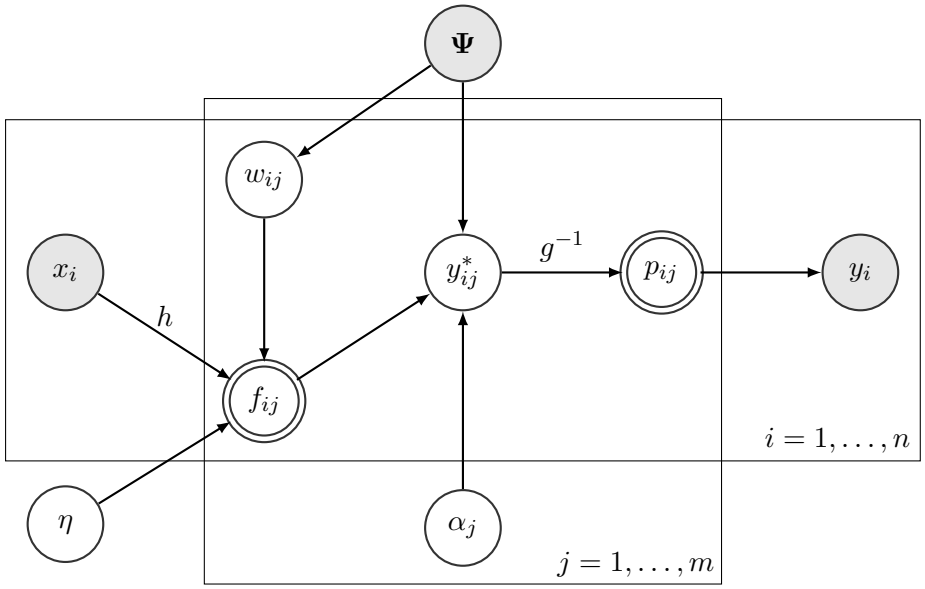


Figure 5.2: A directed acyclic graph (DAG) of the I-probit model. Observed or fixed nodes are shaded, while double-lined nodes represents calculable quantities.

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

$$L(\theta|\mathbf{y}) = \log \iint p(\mathbf{y}|\mathbf{y}^*, \theta) p(\mathbf{y}^*|\mathbf{w}, \theta) p(\mathbf{w}|\theta) \, \mathrm{d}\mathbf{y}^* \, \mathrm{d}\mathbf{w}. \tag{5.8}$$

Here, $p(\mathbf{w}|\theta)$ is the pdf of $\mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$, $p(\mathbf{y}^*|\mathbf{w}, \theta)$ is the pdf of $\mathrm{MN}_{n,m}(\mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}, \mathbf{I}_n, \boldsymbol{\Psi}^{-1})$, and $p(\mathbf{y}|\mathbf{y}^*, \theta) = \prod_{i=1}^n \prod_{j=1}^m \left[y^*_{ij} = \max \mathbf{y}^*_{i\cdot} \right]^{[y_i = j]}$, with $0^0 := 1$. Note that, given the corresponding latent propensities $\mathbf{y}^*_{i\cdot} = (y^*_{i1}, \ldots, y^*_{im})^\top$, the distribution $y_i | \mathbf{y}^*_{i\cdot}$ is tantamount to a degenerate categorical distribution: with knowledge of which latent propensities is largest, the outcome of the categorical response becomes a certainty.

The integral appearing in (5.8) 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 (5.8) gets us to either

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{y}^*, \theta) p(\mathbf{y}^*|\theta) \, \mathrm{d}\mathbf{y}^* \\ &= \log \int \left\{ \prod_{i=1}^{n} \prod_{j=1}^{m} \left[y_{ij}^* = \max \mathbf{y}_{i\cdot}^* \right]^{[y_i=j]} \right\} \phi(\mathbf{y}^*|\mathbf{1}_n \boldsymbol{\alpha}^\top, \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \, \mathrm{d}\mathbf{y}^* \\ &= \log \int_{\bigcap_{i=1}^{n} \{y_{iy_i}^* > y_{ik}^* | \forall k \neq y_i\}} \phi(\mathbf{y}^*|\mathbf{1}_n \boldsymbol{\alpha}^\top, \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \, \mathrm{d}\mathbf{y}^*, \end{aligned} \tag{5.9}$$

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

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y}|\mathbf{w}, \theta) \, p(\mathbf{w}|\theta) \, \mathrm{d}\mathbf{w} \\ &= \log \int \prod_{i=1}^{n} \left\{ \prod_{j=1}^{m} \left(g_j^{-1}(\overbrace{\boldsymbol{\alpha} + \mathbf{w}^\top \mathbf{h}_\eta(x_i)}^{\boldsymbol{\mu}(x_i)} \,|\, \boldsymbol{\Psi}) \right)^{[y_i=j]} \phi(\mathbf{w}_{i\cdot}|\mathbf{0}, \boldsymbol{\Psi}) \, \mathrm{d}\mathbf{w}_{i\cdot} \right\}, \end{aligned} \tag{5.10}$$

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 (5.9), the integral represents the probability of a mn-dimensional normal variate which is not straightforward to calculate, because its covariance matrix is dense. In (5.10), the integral has no apparent closed-form. The 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 (5.10). Note that the dependence of the pdfs on θ is implicit, but is dropped for clarity. Laplace's method (Kass and Raftery, 1995, §4.1.1, pp. 777–778) entails expanding a Taylor series for R about its posterior mode $\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(\mathbf{y}|\mathbf{w}) p(\mathbf{w})$, which gives the

relationship

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

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

The marginal distribution is then approximated by

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

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 is efficient—because the dimension of this integral is exactly the sample size.

Standard errors for the parameters can be obtained from diagonal entries of the information matrix involving the second derivatives of $\log p(\mathbf{y})$. However, it is not known whether the asymptotic variance of the parameters are affected by a Laplace approximation to the likelihood.

Lastly, as a comment, Laplace's method only approximates the true marginal likelihood well if the true posterior density function is small far away from the mode. In other words, a second order approximation of $R(\mathbf{w})$ must be be reliable for Laplace's method to be successful. This is typically the case if the posterior distribution is symmetric about the mode and falls quickly in the tails.

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

$$\begin{aligned} L(\theta|\mathbf{y}, \mathbf{y}^*, \mathbf{w}) &= \log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta) \\ &= \sum_{i=1}^{n} \log p(y_i|\mathbf{y}_{i\cdot}^*) + \log p(\mathbf{y}^*|\mathbf{w}) + \log p(\mathbf{w}) \\ &= \text{const.} + \cancel{\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})\Big) \\ &\quad \cancel{-\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}^{-1}\mathbf{w}^\top\mathbf{w}\Big) \end{aligned} \tag{5.11}$$

which looks like the complete data log-likelihood seen previously in (4.15) (**??**, p. 15), except that here, together with $\mathbf{w}$, the $\mathbf{y}_{i\cdot}^*$'s are not observed.

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

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

where $\tilde{q}$ satisfies $\tilde{q} = \arg\min_q \mathrm{D}_{\mathrm{KL}}(q\|p) = \arg\min_q \int \log \frac{q(\mathbf{y}^*, \mathbf{w})}{p(\mathbf{y}^*, \mathbf{w}|\mathbf{y}, \theta)} q(\mathbf{y}^*, \mathbf{w})\, \mathrm{d}\mathbf{z}$, subject to certain constraints. The constraint considered by us in this thesis is that q satisfies a *mean-field* factorisation

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

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 any variational density q, the marginal log-likelihood is an upper-bound for the quantity $\mathcal{L}_q(\theta) := \mathcal{L}(q, \theta)$ defined by

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

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

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

$$\begin{aligned} Q(\theta) = {} & \mathrm{E}_{\mathbf{y}^*,\mathbf{w}\sim q^{(t+1)}}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}|\theta)] \\ = {} & \text{const.} - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}\,\mathrm{E}[\mathbf{w}^\top\mathbf{H}_\eta^2\mathbf{w}] + \boldsymbol{\Psi}^{-1}\,\mathrm{E}[\mathbf{w}^\top\mathbf{w}]\Big) \\ & - \frac{1}{2}\operatorname{tr}\Big(\boldsymbol{\Psi}\big\{\,\mathrm{E}[\mathbf{y}^{*\top}\mathbf{y}^*] + n\boldsymbol{\alpha}\boldsymbol{\alpha}^\top - 2\boldsymbol{\alpha}\mathbf{1}_n^\top\,\mathrm{E}[\mathbf{y}^*] - 2\,\mathrm{E}[\mathbf{w}^\top]\mathbf{H}_\eta[\mathrm{E}[\mathbf{y}^*] - \mathbf{1}_n\boldsymbol{\alpha}^\top]\big\}\Big), \end{aligned} \tag{5.12}$$

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). Albert and Chib (1993) showed that a data augmentation approach, i.e. 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 (5.8) 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.

A downside to the data augmentation scheme for probit models in a 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 (5.9) or (5.10) after integrating out $\mathbf{w}$ and $\mathbf{y}^*$ respectively, is less demanding for MCMC sampling than the model with likelihood (5.8). However, as mentioned already, (5.9) 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 (5.10), 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 (c.f. equation 5.6), 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 (5.6), the binary I-probit model that is fitted is

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

where h_λ is the (scaled) kernel of the fBm-0.5 RKHS $\mathcal{F}$ to which f belongs.

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

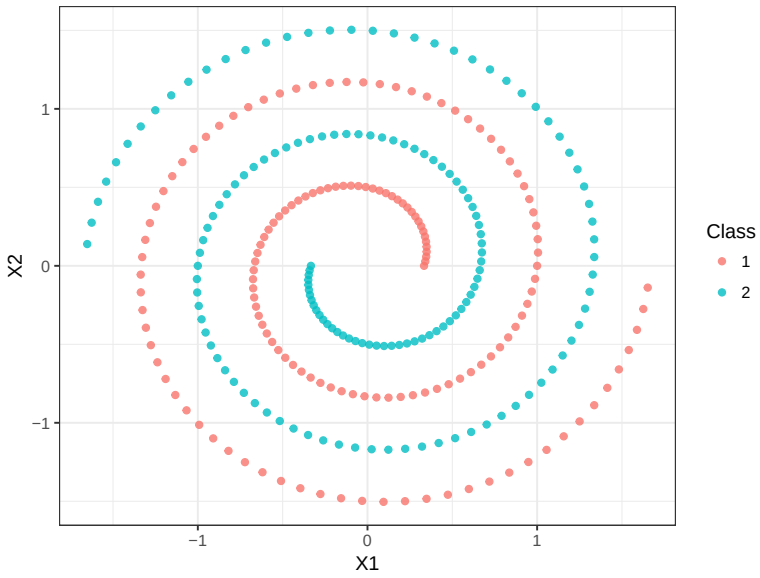


Figure 5.3: A scatter plot of simulated spiral data set.

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 calculated at the parameter estimates 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.

	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	-171.8	-43.2	-8.5
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	>1800

A plot of the log-likelihood (or ELBO) 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

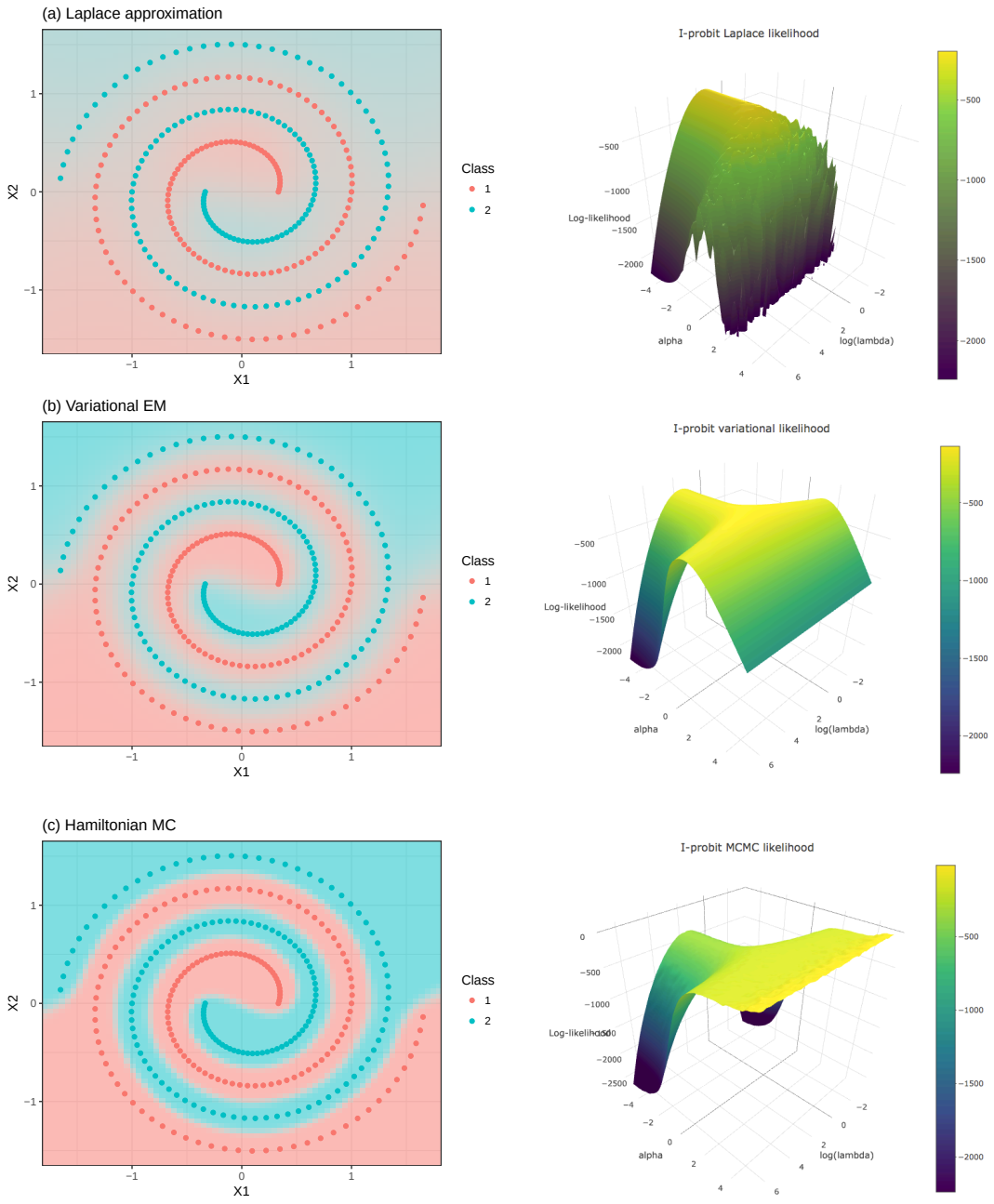


Figure 5.4: Plots showing predicted probabilities (shaded region) for belonging to class '1' or '2' indicated by colour and intensity, and log-likelihood/ELBO surface plots for (a) Laplace's method, (b) variational EM, and (c) Hamiltonian MC. For the Hamiltonian MC plot, parameters are treated as fixed, and the mean log-density of the I-probit model recorded.

to probit model likelihoods (Kuss and Rasmussen, 2005). On the other hand, the log-likelihood calculated using a Hamiltonian MC sampler (treating parameters as fixed values) yields a slightly different graph: the log-likelihood increases as values of α become larger, resulting in the upwards inflection of the log-likelihood surface (as opposed to a downward inflection seen in the variational and Laplace likelihood).

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

Finally, on the computational side, variational inference was by far the fastest method to fit the model. Sampling using Hamiltonian MC was very slow, because the parameter space is in effect $O(n+2)$ (parameters are $\{w_1, \ldots, w_n, \alpha, \lambda\}$ under the model with likelihood (5.10), 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 minutes or even hours to. The predictive performance, while not as good as HMC, is certainly an acceptable compromise in favour of speed.

5.4 The variational EM algorithm for I-probit models

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

The algorithm cycles through a variational inference E-step, in which the variational density $q(\mathbf{y}^*, \mathbf{w}) = \prod_{i=1}^n q(\mathbf{y}_{i\cdot}^*)q(\mathbf{w})$ is optimised with respect to the Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}\big(q(\mathbf{y}^*, \mathbf{w}) \| p(\mathbf{y}^*, \mathbf{w}|\mathbf{y})\big)$, and an M-step, in which the approximate expected joint density (5.12) 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.

[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 $\hat{\mathrm{P}}(y_i = j)$. 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.

5.4.1 The variational E-step

Let $\tilde{q}(\mathbf{y}^*, \mathbf{w})$ be the pdf that minimises the Kullback-Leibler divergence $\mathrm{D}_{\mathrm{KL}}\big(q \| p\big)$ 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

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

$$\log \tilde{q}(\mathbf{w}) = \mathrm{E}_{\mathbf{y}^*\sim\tilde{q}}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w})] + \text{const.} \tag{5.14}$$

where $p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}) = p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{w})p(\mathbf{w})$ is as per (5.8). 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 Appendix H (p. 39).

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

The fact that the rows $\mathbf{y}^*_{i\cdot} \in \mathbb{R}^m$, $i = 1, \dots, n$ of $\mathbf{y}^* \in \mathbb{R}^{n\times m}$ are independent can be exploited, and this results in a further induced factorisation $q(\mathbf{y}^*) = \prod_{i=1}^n q(\mathbf{y}^*_i)$. Define the set $\mathcal{C}_j = \{y^*_{ij} > y^*_{ik} \,|\, \forall k \neq j\}$. Then $q(\mathbf{y}^*_{i\cdot})$ is the density of a multivariate normal distribution with mean $\tilde{\boldsymbol{\mu}}_{i\cdot} = \boldsymbol{\alpha} + \tilde{\mathbf{w}}^\top \mathbf{h}_\eta(x_i)$, where $\tilde{\mathbf{w}} = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}[\mathbf{w}]$, and variance $\boldsymbol{\Psi}^{-1}$, subject to a truncation of its components to the set $\mathcal{C}_{y_i}$. That is, for each $i = 1, \dots, n$ 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

$$\mathbf{y}^*_{i\cdot} \overset{\text{iid}}{\sim} \begin{cases} \mathrm{N}_m(\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}) & \text{if } y^*_{iy_i} > y^*_{ik}, \forall k \neq y_i \\ 0 & \text{otherwise.} \end{cases} \tag{5.15}$$

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 Gibbs-based approach (Robert, 1995) for sampling from a truncated multivariate normal can be implemented, and this is detailed in Appendix C.4.

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

$$\tilde{y}_{ik}^* = \begin{cases} \tilde{\mu}_{ik} - \sigma_k C_i^{-1} \int \phi_{ik}(z) \prod_{l \neq k, y_i} \Phi_{il}(z)\phi(z)\,\mathrm{d}z & \text{if } k \neq y_i \\ \tilde{\mu}_{iy_i} - \sigma_{y_i} \sum_{k \neq y_i} \left(\tilde{y}_{ik}^* - \tilde{\mu}_{ik}\right) & \text{if } k = y_i \end{cases} \tag{5.16}$$

with

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

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

Variational distribution for the I-prior random effects w

Given that both $\operatorname{vec}\mathbf{y}^* | \operatorname{vec}\mathbf{w}$ and $\operatorname{vec}\mathbf{w}$ are normally distributed as per the model (5.5), 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

$$\operatorname{vec}\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta)\operatorname{vec}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = \boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n = \mathbf{V}_{y^*}. \tag{5.17}$$

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 Section 5.6.2 for details. Storage requirement is $O(n^2m^2)$, as a result of the covariance matrix in (5.17).

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

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

where $\mathrm{N}_d(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and

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

We note the similarity between (5.17) above and the posterior distribution for the I-prior random effects in a normal model (4.11) seen in the previous chapter, with the difference being (5.17) uses the continuous latent propensities $\mathbf{y}^*$ instead of the the observations $\mathbf{y}$. The consequence of this is that the posterior regression functions are class independent, the exact intended effect by specifying a diagonal precision matrix $\boldsymbol{\Psi}$. Storage requirement is $O(n^2m)$, since we need $\mathbf{V}_{w_1}, \dots, \mathbf{V}_{w_m}$.

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

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

where $r(\cdot)$ is some function of $\mathbf{w}$, and expectations are taken under the posterior distribution of $\mathbf{y}^*$. Hypothetically, if the true pdf $p(\mathbf{y}^*|\mathbf{y})$ were tractable, then the E-step can be computed using the true conditional distribution. Since it is not tractable, we resort to an approximation, and in the case of a variational approximation, (5.17) is obtained.

5.4.2 The M-step

From (5.12), the function to be maximised in the M-step is

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

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

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

by the results of the derivations in Appendix H.1.2 (p. 43). In the above, we had defined ψ_{ij}^- to be the (i,j)'th element of $\boldsymbol{\Psi}^{-1}$, and

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

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

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

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

$$\frac{\partial}{\partial\eta}Q(\eta|\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,j=1}^{m}\psi_{ij}\operatorname{tr}\left(\frac{\partial\mathbf{H}_\eta^2}{\partial\eta}\tilde{\mathbf{W}}_{ij}\right) + \operatorname{tr}\left(\boldsymbol{\Psi}\tilde{\mathbf{w}}^\top\frac{\partial\mathbf{H}_\eta}{\partial\eta}(\tilde{\mathbf{y}}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top)\right) \tag{5.18}$$

$$\frac{\partial}{\partial\boldsymbol{\alpha}}Q(\boldsymbol{\alpha}|\eta) = 2n\boldsymbol{\Psi}\boldsymbol{\alpha} - 2\sum_{i=1}^{n}\boldsymbol{\Psi}\big(\mathbf{y}_{i\cdot}^* - \tilde{\mathbf{w}}^\top\mathbf{h}_\eta(x_i)\big) \tag{5.19}$$

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 (5.18) can be found in closed-form, much like in the exponential family EM algorithm described in Section 4.3.3 (p. 24). Under the same setting as in that subsection, assume that only $\eta = \{\lambda_1, \ldots, \lambda_p\}$ need be estimated, and for each $k = 1, \ldots, p$, we can decompose the kernel matrix as $\mathbf{H}_\eta = \lambda_k\mathbf{R}_k + \mathbf{S}_k$ and its square as $\mathbf{H}_\eta^2 = \lambda_k^2\mathbf{R}_k^2 + \lambda_k\mathbf{U}_k + \mathbf{S}_k^2$. As a follow-on from (5.18), the conditional solution for λ_k given the rest of the parameters is obtained by solving

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

equals zero. 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 $Q(\eta)$).

Update for intercepts $\boldsymbol{\alpha}$

It is easy to see that the unique solution to (5.19) is

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

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

5.4.3 Summary

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

$$\mathcal{L}_q(\theta) = \frac{nm}{2} + \sum_{i=1}^n \log C_i(\theta) + \frac{1}{2}\log|\tilde{\mathbf{V}}_w| - \frac{n}{2}\log|\boldsymbol{\Psi}| - \frac{1}{2}\sum_{i,j=1}^m \psi_{ij}^- \operatorname{tr}(\tilde{\mathbf{W}}_{ij}), \tag{5.20}$$

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

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

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 an 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, $\mathbf{f}(x_{\text{new}}) = \mathbf{w}^\top \mathbf{h}_\eta(x_{\text{new}})$, where $\mathbf{h}_\eta(x_{\text{new}})$ is the vector of length n containing entries $h_\eta(x_i, x_{\text{new}})$, 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 (5.17). By writing $\operatorname{vec}\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}_{\cdot 1}, \ldots, \tilde{\mathbf{w}}_{\cdot m})^\top$ to separate out the I-prior random effects per class, we have that $\mathbf{w}_{\cdot j} | \hat{\theta} \sim \text{N}_n(\tilde{\mathbf{w}}_{\cdot j}, \tilde{\mathbf{V}}_w[j,j])$, 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 matrix structured matrix $\mathbf{V}_w$. Thus, for each class $j = 1, \ldots, m$ and any $x \in \mathcal{X}$,

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

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

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

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 $\mathbf{y}^*$
      $\tilde{q}^{(t+1)}(\mathbf{y}^*_{i\cdot}) \leftarrow {}^{\text{t}}\text{N}_m\left(\tilde{\boldsymbol{\alpha}}^{(t)} + \tilde{\mathbf{w}}^{(t)\top}\mathbf{h}_{\eta^{(t)}}(x_i), \boldsymbol{\Psi}, \mathcal{C}_{y_i}\right)$
      $\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} \leftarrow \text{E}_{q^{(t+1)}}(\mathbf{y}^*_{i\cdot})$
    end for

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

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

    $\mathbf{a} \leftarrow \frac{1}{n}\sum_{i=1}^n \left(\tilde{\mathbf{y}}^{*(t+1)}_{i\cdot} - \tilde{\mathbf{w}}^{(t+1)\top}\tilde{\mathbf{h}}_{\eta^{(t+1)}}(x_i)\right)$ ▷ Update $\boldsymbol{\alpha}$
    $\boldsymbol{\alpha}^{(t+1)} \leftarrow \mathbf{a} - \frac{1}{m}\sum_{j=1}^m a_j$
  end procedure

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

  $\{\tilde{q}(\mathbf{y}^*), \tilde{q}(\mathbf{w}), \hat{\theta}\} \leftarrow \{\tilde{q}^{(t)}(\mathbf{y}^*), \tilde{q}^{(t)}(\mathbf{w}), \theta^{(t)}\}$
  return Variational densities $\{\tilde{q}(\mathbf{y}^*), \tilde{q}(\mathbf{w})\}$
  return Estimates $\{\hat{\boldsymbol{\alpha}}, \hat{\eta}\}$
  return ELBO $\mathcal{L}_q(\theta) = \mathcal{L}^{(t)}$
end while

Then, in step two, using the results obtained in the previous chapter in Section 4.4 (p. 27), we have that the latent propensities $y^*_{\text{new},j}$ for each class are normally distributed with mean, variance, and covariances

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

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

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

The predicted class is inferred from the latent variables using

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

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

$$\hat{p}_{\text{new},j} = \int_{\{y^*_j > y^*_k | \forall k \neq j\}} \cdots \int \phi(y^*_1, \dots, y^*_m | \hat{\boldsymbol{\mu}}_{\text{new}}, \hat{\mathbf{V}}_{\text{new}}) \, \mathrm{d}y^*_1 \cdots \mathrm{d}y^*_m. \tag{5.21}$$

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 (5.7), 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 $O(n^2 m)$ 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 $\operatorname{vec} \mathbf{w}^{(t)} | \hat{\theta} \sim \mathrm{N}_{nm}(\operatorname{vec} \tilde{\mathbf{w}}, \mathbf{V}_w)$. Then, obtain samples of class probabilities $\{p^{(1)}_{xj}, \dots, p^{(T)}_{xj}\}_{j=1}^m$, for a given data point $x \in \mathcal{X}$ by

evaluating

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

where $\hat{\boldsymbol{\mu}}^{(t)}(x) = \hat{\boldsymbol{\alpha}} + \mathbf{w}^{(t)\top}\mathbf{h}_{\hat{\eta}}(x)$, and $\hat{\mathbf{V}}(x)_{jk}$ equals $\hat{\sigma}_j^2(x)$ 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 $\big\{r(p_{xj}^{(1)}),\ldots,r(p_{xj}^{(T)})\big\}$.

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., 2018):

1. Obtain $\hat{\theta} = \arg\max_\theta \mathcal{L}_q(\theta)$ using $\mathcal{S} = \{(y_1,x_1),\ldots,(y_n,x_n)\}$.
2. For $t = 1,\ldots,T$, do
   (a) Obtain $\mathcal{S}^{(t)} = \{(y_1^{(t)},x_1^{(t)}),\ldots,(y_n^{(t)},x_n^{(t)})\}$ 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

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

The obvious downside to this bootstrap scheme is computational time.

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

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

since the Kullback-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 $\mathrm{BF}(M_1,M_0)$ for comparing model M_1 against model M_0, where $\mathrm{BF}(M_1,M_0)$ is approx-

imated by

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

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

Table 5.2: Guidelines for interpreting Bayes factors (Kass and Raftery, 1995).

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

Remark 5.5. In the previous chapter on normal I-prior models, the I-prior could be integrated out of the model completely, resulting in a normal log-likelihood for the parameters. Model comparison can be validly done using likelihood ratio tests and asymptotic chi-square distributions. Here however, we only have a lower bound to the log-likelihood, and most likely the asymptotic results of likelihood ratio tests do not hold. Then, the concept of approximate (empirical) Bayes factors seem most intuitive, even if not rooted in theory.

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 $\boldsymbol{\Psi}$ is desired, and thus pave the way for future work.

5.6.1 Efficient computation of class probabilities

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 $\mathrm{N}_m\left(\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}\right)$, 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, Hajivassiliou et al. (1996) concluded

that the Geweke-Hajivassiliou-Keane (GHK) probability simulator (Geweke, 1989; Hajivassiliou and McFadden, 1998; 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 $\mathbf{z} \sim \mathrm{N}_{m-1}(\boldsymbol{\nu}_{(j)}, \boldsymbol{\Omega}_{(j)})$, 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 5.6. Incidentally, the probit model in (5.1) is equivalently represented by

$$y_i = \begin{cases} 1 & \text{if } \max(y_{i2}^* - y_{i1}^*, \dots, y_{im}^* - y_{i1}^*) < 0 \\ j & \text{if } \max(y_{i2}^* - y_{i1}^*, \dots, y_{im}^* - y_{i1}^*) = y_{ij}^* - y_{i1}^* \geq 0, \end{cases} \tag{5.22}$$

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 5.2, 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 $z_{j'} < 0$, $j' = 1, \ldots, m-1$, can be expressed as

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

where

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

The GHK algorithm entails making draws from one-sided right truncated standard normal distributions (for instance, using an inverse transform method detailed in Appendix C.3, p. 16):

- Draw $\tilde{\zeta}_1 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, u_1)$.
- Draw $\tilde{\zeta}_2 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_2)$, where $\tilde{u}_2 = u_2(\tilde{\zeta}_1)$.
- Draw $\tilde{\zeta}_3 \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_3)$, where $\tilde{u}_3 = u_3(\tilde{\zeta}_1, \tilde{\zeta}_2)$.
- $\cdots$
- Draw $\tilde{\zeta}_{m-1} \sim {}^{\mathrm{t}}\mathrm{N}(0, 1, -\infty, \tilde{u}_{m-2})$, where $\tilde{u}_{m-1} = u_m(\tilde{\zeta}_1, \ldots, \tilde{\zeta}_{m-2})$.

These values are then used in the following manner:

- Use $\tilde{\zeta}_1$ to obtain a "draw" of $\mathrm{P}(\zeta_2 < u_2 | \zeta_1 < \zeta_1)$,

$$\begin{aligned} \widetilde{\mathrm{P}}(\zeta_2 < u_2 | \zeta_1 < \zeta_1) &= \mathrm{P}(\zeta_2 < u_2 | \zeta_1 = \tilde{\zeta}_1) \\ &= \Phi\Big(- \big(\nu_{(j)2} + L_{12} \tilde{\zeta}_1\big) / L_{22} \Big) \end{aligned}$$

- Use $\tilde{\zeta}_1$ and $\tilde{\zeta}_2$ to obtain a "draw" of $\mathrm{P}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2)$,

$$\begin{aligned} \widetilde{\mathrm{P}}(\zeta_3 < u_3 | \zeta_1 < u_1, \zeta_2 < u_2) &= \mathrm{P}(\zeta_3 < u_3 | \zeta_1 = \tilde{\zeta}_1, \zeta_2 = \tilde{\zeta}_2) \\ &= \Phi\Big(- \big(\nu_{(j)3} + L_{13} \tilde{\zeta}_1 + L_{23} \tilde{\zeta}_2\big) / L_{33} \Big) \end{aligned}$$

- And so on.

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

$$\tilde{p}_j = \Phi\left(-\nu_{(j)1}/L_{11}\right) \prod_{j'=2}^{m-1} \Phi\left(-\left(\nu_{(j)j'} + \textstyle\sum_{k=1}^{j'-1} L_{kj'}\tilde{\zeta}_k\right)/L_{j'j'}\right). \tag{5.23}$$

By performing the above scheme T number of times to obtain T such simulated probabilities $\{p_j^{(1)}, \ldots, p_j^{(T)}\}$, 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 (5.23) 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 (5.23) turn into sums, and revert back by exponentiating.

Remark 5.7. The GHK algorithm provides reasonably fast and accurate calculations of class probabilities when $\boldsymbol{\Psi}$ is dense. As we alluded to earlier in the chapter, the class probabilities condense to a unidimensional integral involving products of normal cdfs (c.f. equation 5.7) if $\boldsymbol{\Psi}$ is diagonal. Note that if $\boldsymbol{\Psi}$ is diagonal, then the transformed $\boldsymbol{\Omega}_{(j)} = \mathbf{Q}\boldsymbol{\Psi}^{-1}\mathbf{Q}^\top$ is certainly not: the components of $\mathbf{z}$ are correlated because they are all anchored on the same random variable. Thus, direct evaluation of (5.7) using quadrature methods avoids the Cholesky step and random sampling employed by the GHK method.

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

$$\mathbf{V}_w = \left[(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n)\right]^{-1}. \qquad \text{(from 5.17)}$$

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 $\operatorname{vec} \mathbf{w}$.

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 (5.12) conditional on the rest of the parameters can be written as

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

with $\boldsymbol{\mu} = \mathbf{1}_n\boldsymbol{\alpha}^\top + \mathbf{H}_\eta\mathbf{w}$. This can be solved using numerical methods, though it must be ensured that the identifiability constraints and positive-definiteness are satisfied. Specifically in the case where $\boldsymbol{\Psi}$ is a diagonal matrix $\operatorname{diag}(\psi_1, \ldots, \psi_m)$, then

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

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

$$\hat{\psi}_j = \left(\frac{\mathrm{E}(\mathbf{w}_{\cdot j}^\top\mathbf{w}_{\cdot j})}{\mathrm{E}\left((\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j})^\top(\mathbf{y}^*_{\cdot j} - \boldsymbol{\mu}_{\cdot j})\right)}\right)^{\frac{1}{2}},$$

independently of the rest of the other ψ_k's, $k \neq j$. As per the derivations in Appendix H.1.2 (p. 43), the numerator of this expression is equal to $\text{tr}(\tilde{\mathbf{V}}_{w_j} + \tilde{\mathbf{w}}_{\cdot j}\tilde{\mathbf{w}}_{\cdot j}^\top) = \text{tr}(\tilde{\mathbf{W}}_{jj})$. The denominator on the other hand is

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

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, we show that the form for the second moments involve integration of standard normal cdfs and pdfs (Lemma C.5, p. 18), much like the formula for the first moments.

5.7 Examples

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 in R using the in-development **iprobit** package written by us. Code for replication is provided at http://myphdcode.haziqj.ml. 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

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

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).

Following (5.6), 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

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

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 (Friedman et al., 2001). 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.

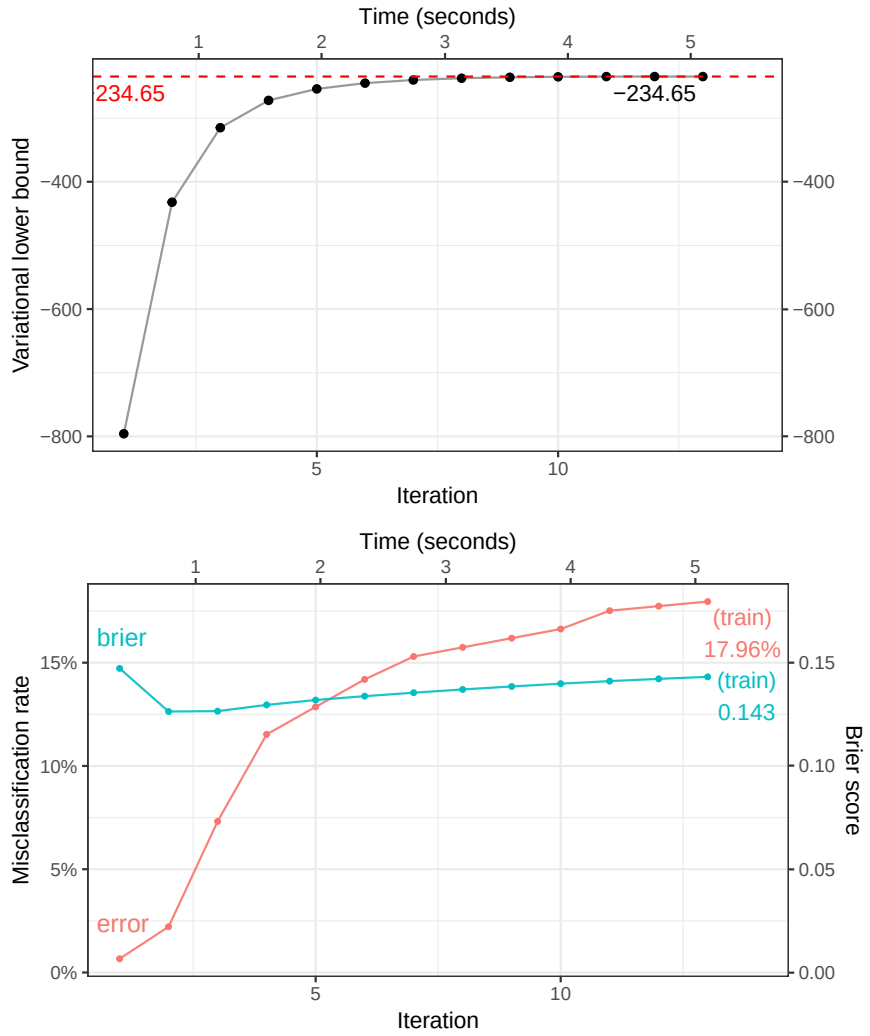


Figure 5.5: Plot of variational lower bound over time (top), and plot of training error rate and Brier scores over time (bottom).

Results for the methods listed above were extracted from the in-depth study by Cannings and Samworth (2017), who also conducted identical experiments using their random projection (RP) ensemble classification method. These are all 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

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.

	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)

ensemble learning method, which is also generally faster to train than Gaussian process-like regressions such as I-prior models. The time complexity of a random forest algorithm is $O(pqn\log(n))$ (Louppe, 2014), 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.

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 for quitting 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.

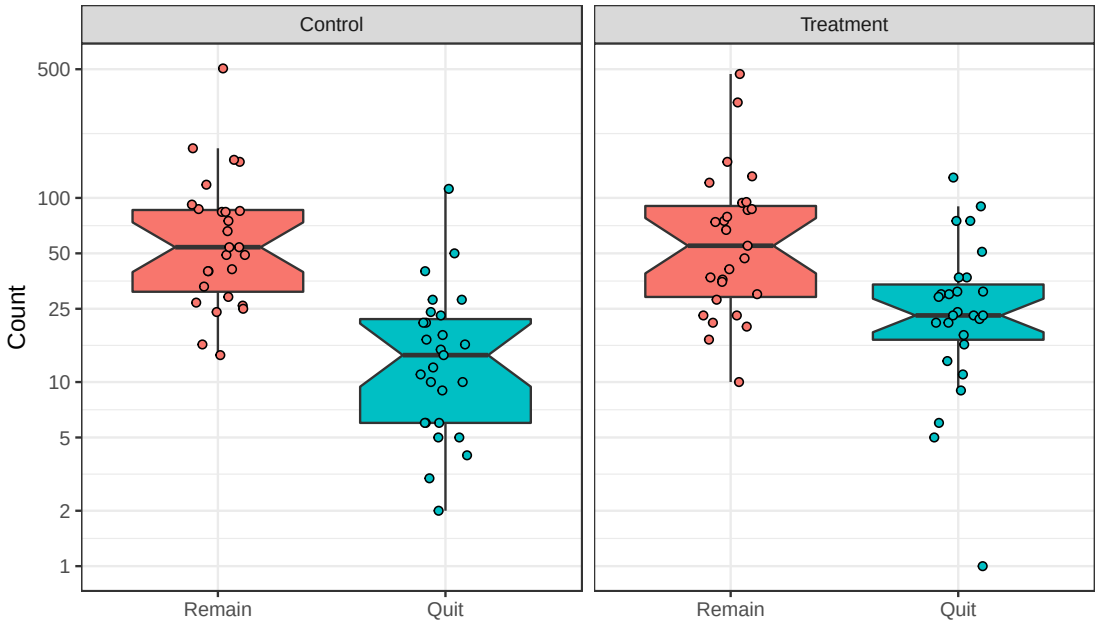


Figure 5.6: Comparative box-plots of the distribution of patients who successfully quit smoking and those who remained smokers, in the two treatment groups. It is evident that there are more successful patients quitting smoking in the treatment group than in the control group. The raw odds ratio of quitting smoking (treatment vs. control) is 1.66.

A summary of the data is displayed by the box-plot in Figure 5.6. On the whole, there are a total of 5,908 patients, and they are distributed roughly equally among the control and treatment groups (46.3% and 53.7% 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

$$\text{odds} = \frac{\text{P(quit smoking)}}{1 - \text{P(quit smoking)}},$$

and these probabilities, odds and ultimately 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 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 $i = 1, \ldots, n_k$ index the patients in study group $k \in \{1, \ldots, 27\}$. For patient i in study j, p_{ik} denotes the probability that the patient has successfully quit smoking. Additionally, x_{ik} is the centred dummy variable indicating patient i's treatment group in study k. 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_{ik}, the total variance $\mathrm{Var}(\beta_{0k} + \beta_{1j}x_{ik})$ would be constant in both treatment groups. The overall log odds ratio is represented by β_1, and this is estimated as $0.57 \approx \log 1.76$.

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

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

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_{ik} and k 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_{ik} 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 the I-probit model, 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.

Table 5.4: Results of the I-probit 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 k—call these $p_k(\text{treatment})$ and $p_k(\text{control})$. That is,

$$p_k(\text{treatment}) = \Phi\big(\hat{\nu}(\text{treatment}, k)\big)$$
$$p_k(\text{control}) = \Phi\big(\hat{\nu}(\text{control}, k)\big),$$

where $\hat{\nu}$ represents the standardised posterior mean estimate for the regression functions which are distributed according to

$$f(x_{ik}, k)|\mathbf{y}, \hat{\theta} \sim \text{N}\left(\hat{\mu}(x_{ik}, k), \hat{\sigma}^2(x_{ij}, k)\right),$$

with $x_{ik} \in \{\text{treatment}, \text{control}\}$ and $k \in \{1, \ldots, 27\}$ (see details in Section 5.5). The 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 $0.51 \approx \log 1.66$, slightly lower than both the raw log odds ratio and the log odds ratio estimated by the

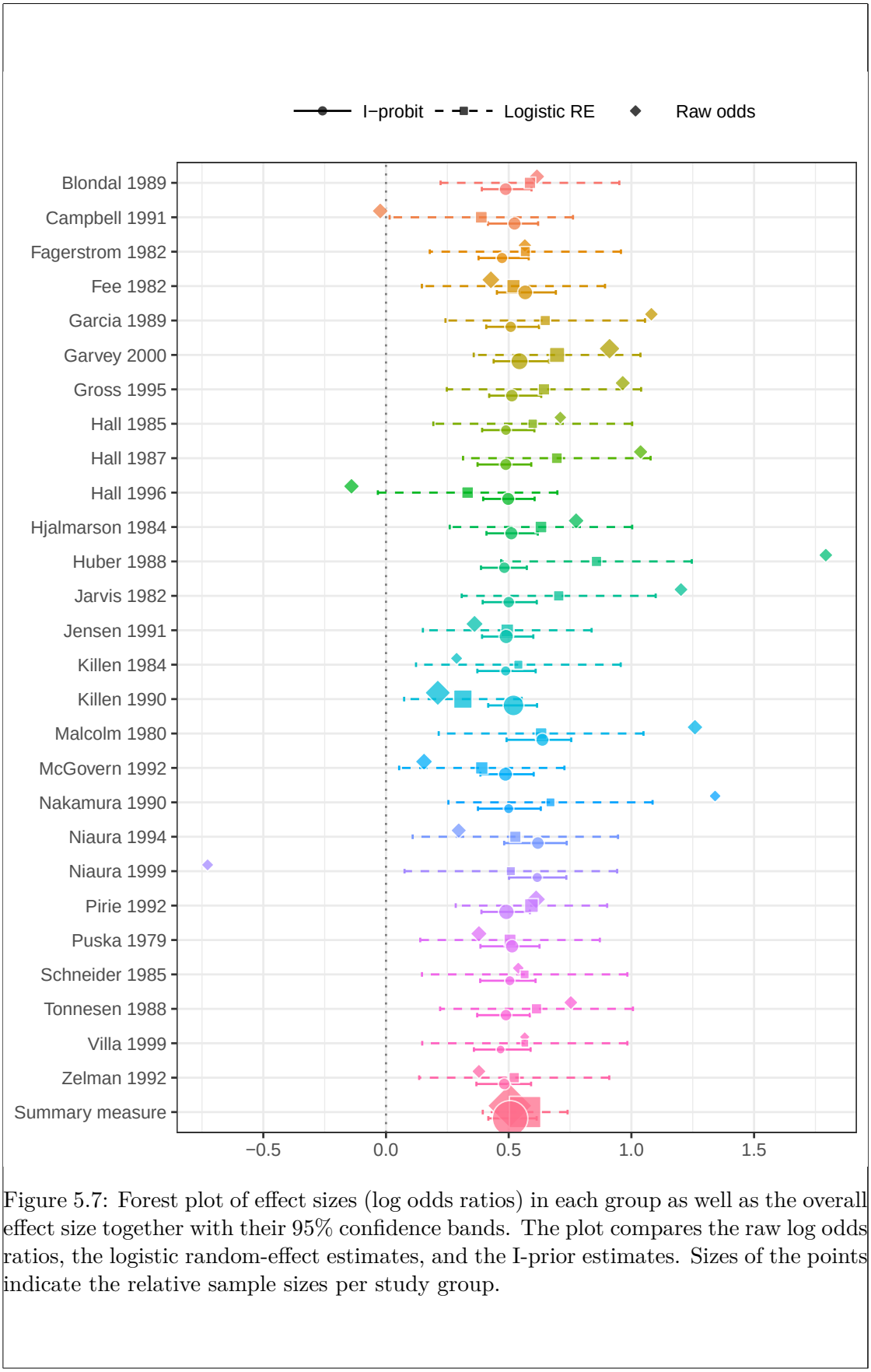


Figure 5.7: Forest plot of effect sizes (log odds ratios) in each group as well as the overall effect size together with their 95% confidence bands. The plot compares the raw log odds ratios, the logistic random-effect estimates, and the I-prior estimates. Sizes of the points indicate the relative sample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

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 for the log odds ratios in the forest plot of Figure 5.7 are also noticeably narrower under an I-prior compared to the fitted multilevel model. One explanation is that empirical Bayes estimates, such as the I-probit estimates under a variational EM framework, tend to underestimate the variability in the estimates because the variability in the parameters are ignored when point estimates are used, compared to distributions in a true Bayesian estimation framework.

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 $8 \times 6 \times 11 = 528$, while $7 \times 6 \times 11 = 462$ 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, 1990). Machine learning methods such as neural networks and nearest neighbour methods were analysed by Robinson (1989).

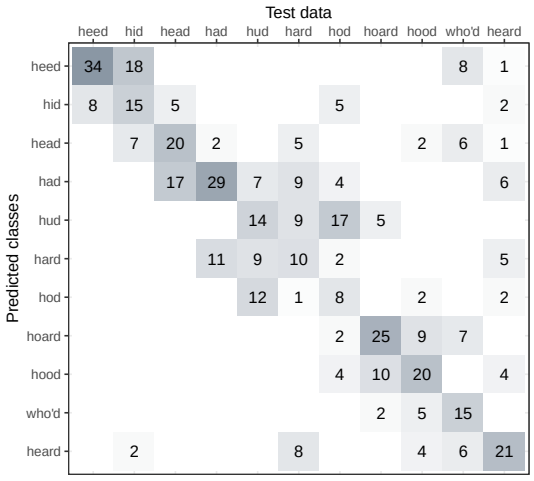
Table 5.5: The eleven words that make up the classes of vowels.

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	ɜː	heard
6	had	ɑː	hard				

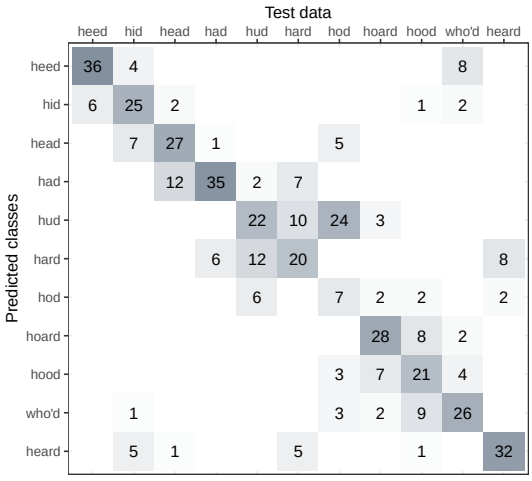
We will fit the data using an I-probit model with the canonical linear kernel, fBm-0.5 kernel, and the SE kernel with lengthscale $l = 1$. Each model took roughly 13 seconds per iteration in fitting the training data set ($n = 528$). In particular, the canonical kernel

[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).

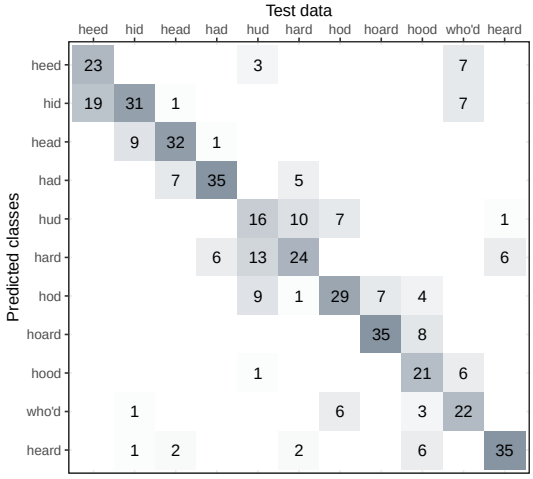
model took a long time to converge, with each variational inference iteration improving the lower bound only slighly each time. In contrast, both the fBm-0.5 and SE model were quicker to converge. Multiple restarts from different random seeds were conducted, and we found that they all converged to a similar lower bound value. This alleviates any concerns that the model might have converged to different multiple local optima.



(a) Canonical kernel



(b) fBm-0.5 kernel



(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 cell is 42 (seven speakers delivered six frames of speech per vowel). Blank cells indicate nil values.

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.

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

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.

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 of the infection of the herds, and whether there is a time-element to the presence or absence of this spatial segregation. There has been previous work done to analyse this data set. 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 works include Diggle et al. (2013), who used a fully Bayesian approach for spatio-temporal multivariate log-Gaussian Cox processes, which is implemented in the R package **lgcp** (Taylor et al., 2013).

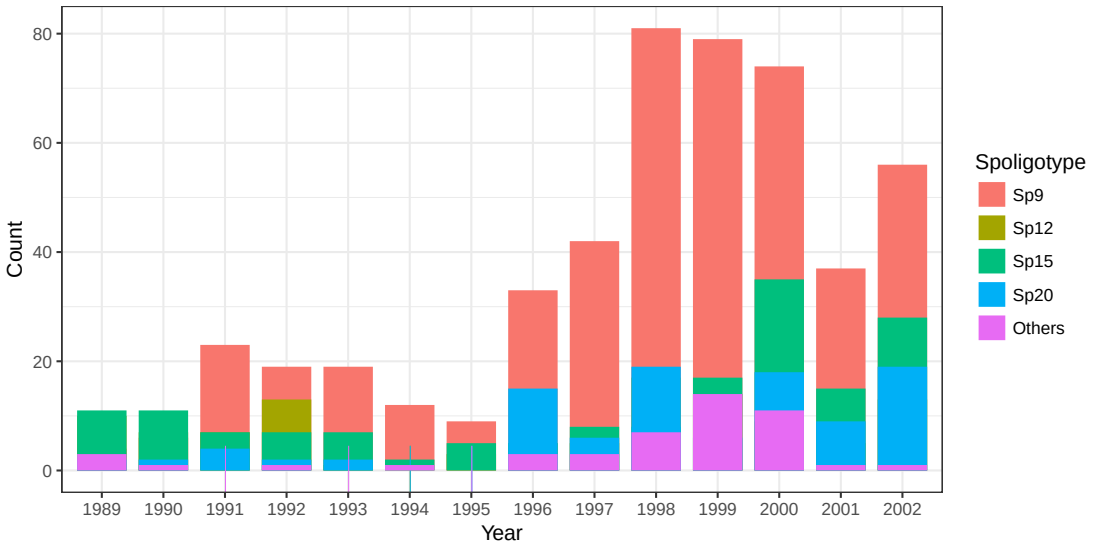


Figure 5.9: Distribution of the different types (Spoligotypes) of bovine tubercolosis affecting herds in Cornwall over the period 1989 to 2002.

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.

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 $g_j(p_{ij})$ as following a function which takes two covariates, i.e. the spatial data $x_1 \in \mathbb{R}^2$, and the temporal data x_2 (year of infection):

$$\begin{aligned} p_{ij} &= g_j^{-1}\big(f_k(x_1, x_2)\big)_{k=1}^m \\ &= g_j^{-1}\big(f_{1k}(x_1) + f_{2k}(x_2) + f_{12k}(x_1, x_2)\big)_{k=1}^m, \end{aligned}$$

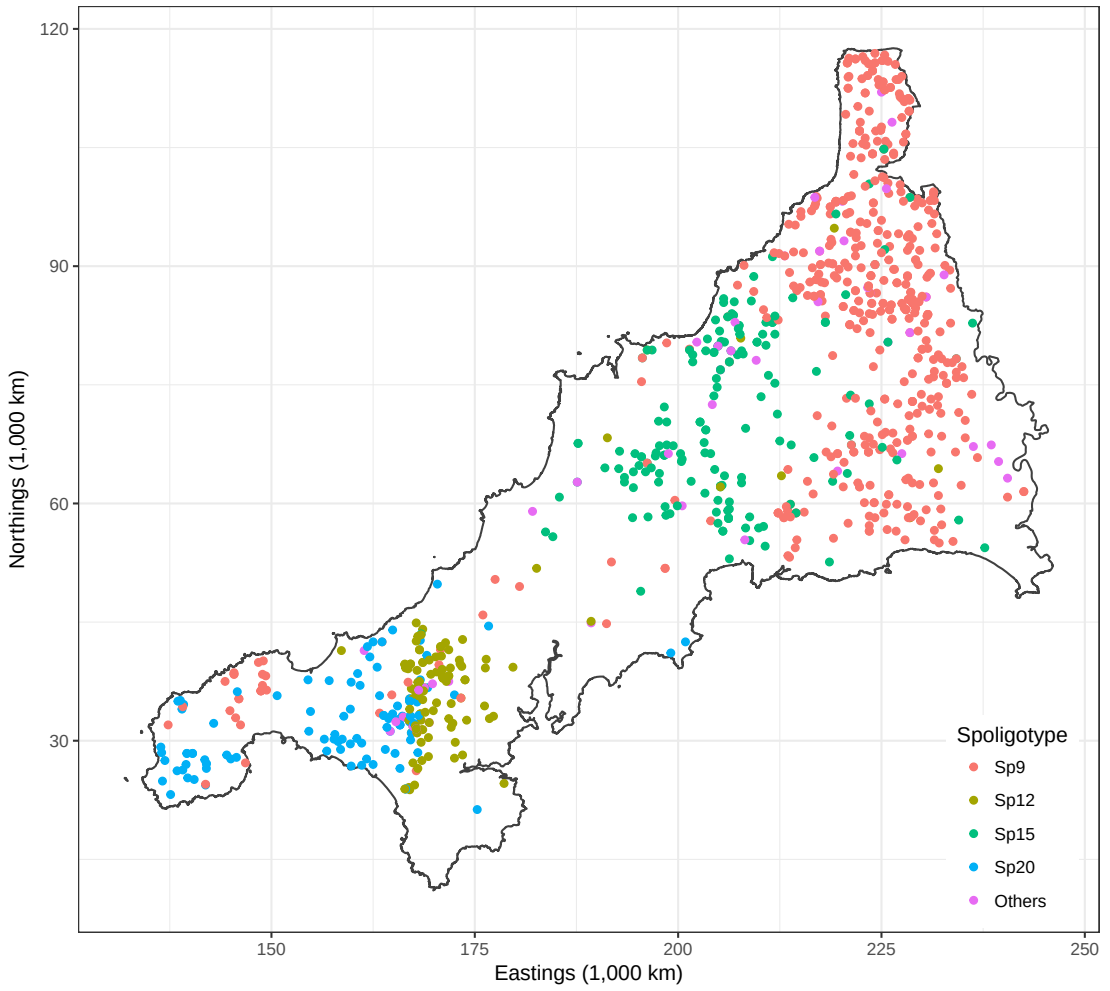


Figure 5.10: Spatial distribution of all cases over the 14 years.

where the function $g_j^{-1} : \mathbb{R}^m \to [0,1]$ is the same squashing function used in equation (5.10). 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 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:

- $\boldsymbol{M_0}$**: Intercept only**.

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

Table 5.7: Results of the fitted I-probit models. Estimates of the class intercepts and scale parameters, together with their respective bootstrap standard errors, are presented. For model comparison, we can look at ELBOs, error misclassification rates, and Brier scores.

	M_0: Intercepts only		M_1: Spatial only		M_2: Spatio-temporal		M_3: Spatio-period	
	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.	Estimate	S.E.
Intercept (Sp9)	0.948	0.000	1.364	0.015	1.401	0.079	1.395	0.103
Intercept (Sp12)	-0.173	0.000	-0.435	0.013	-0.506	0.017	-0.463	0.045
Intercept (Sp15)	0.103	0.000	-0.020	0.011	-0.008	0.059	-0.010	0.094
Intercept (Sp20)	-0.202	0.000	-0.775	0.051	-0.795	0.223	-0.783	0.343
Intercept (Others)	-0.676	0.000	-0.134	0.016	-0.091	0.077	-0.139	0.104
Scale (spatial)			0.194	0.008	-0.176	0.178	0.172	0.169
Scale (temporal)					-0.006	0.003	-0.004	0.006
ELBO	-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	

- ***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 $f_{12k} \in \mathcal{F}_1 \otimes \mathcal{F}_2$

Model M_0 corresponds to a model which ignores any spatial or temporal effects (the baseline intercept only model). Model M_1 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 ELBO of the four models, making it the preferable 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 M_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 yearly evolution of the surface probabilities between 1989 and 2002.

As the plots 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, a series of predicted probability surface plots over the four time periods obtained from model M_3. For each time period, we also superimposed the actual observations onto the predicted surface probabilities. In addition, coloured dotted lines are displayed to indicate the

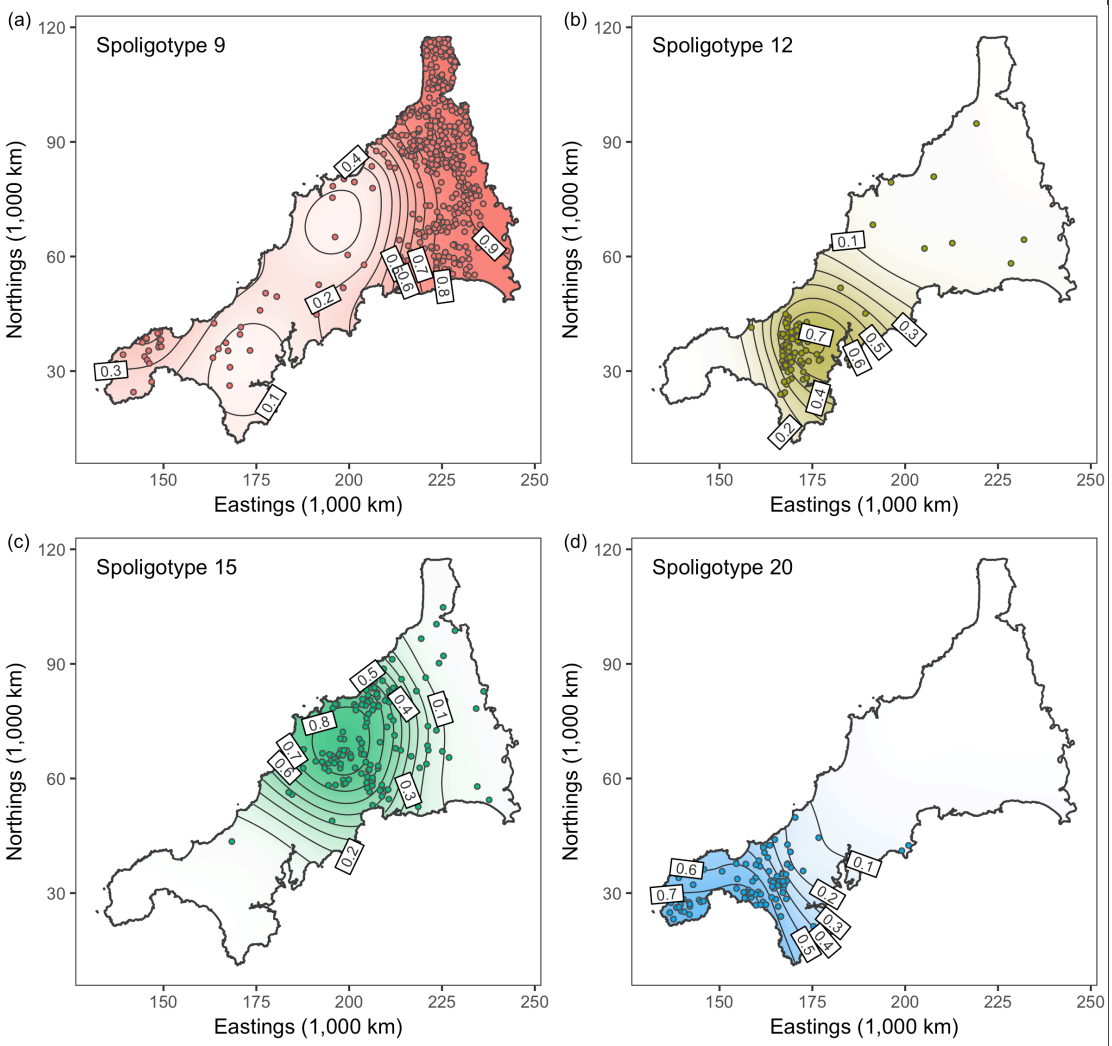


Figure 5.11: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium *Mycobacterium bovis* over the entire time period using model M_1.

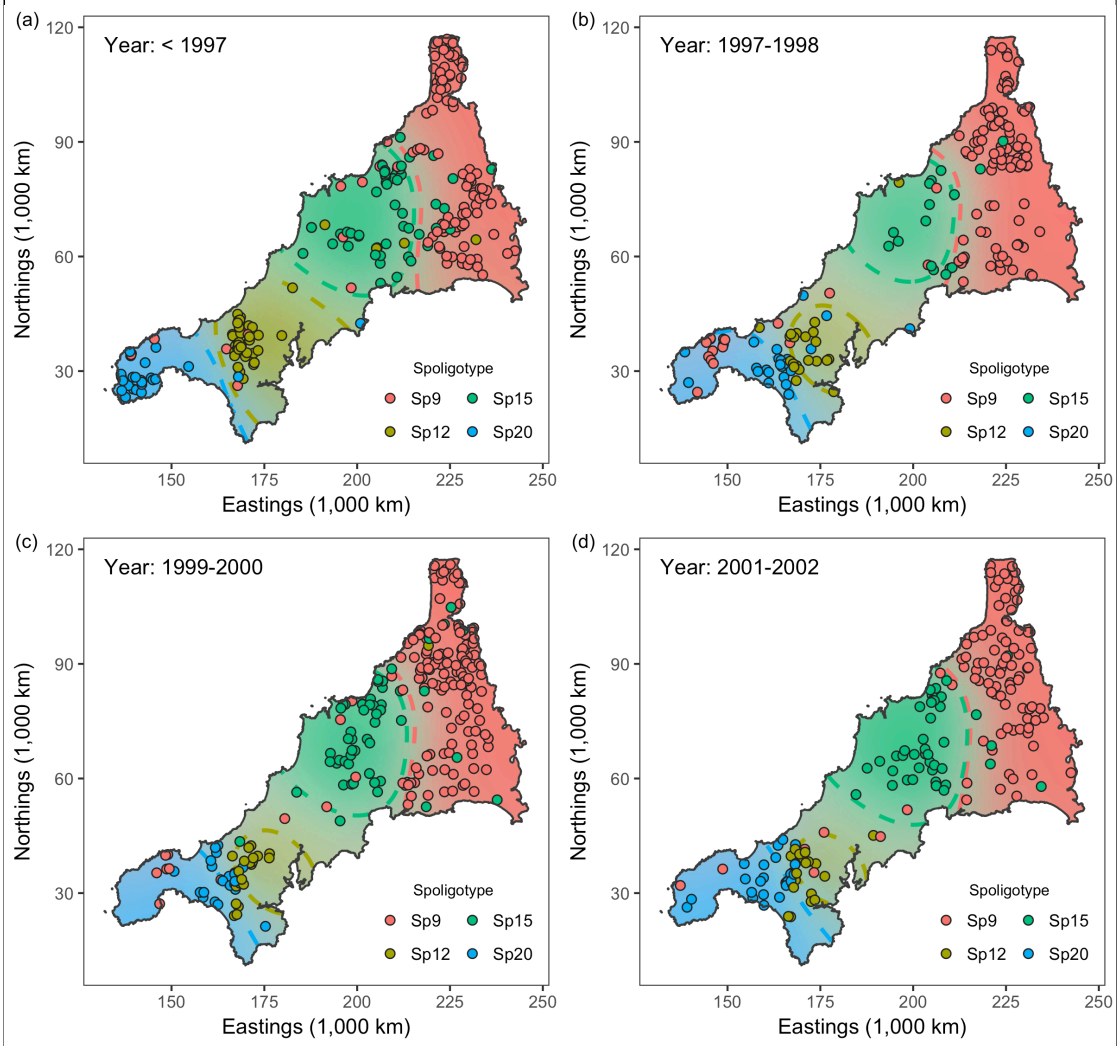


Figure 5.12: Predicted probability surfaces for BTB contraction in Cornwall for the four largest spoligotypes of the bacterium *Mycobacterium bovis* over four different time periods using model M_3.

"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.

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 (5.8). Techniques such as quadrature methods, Laplace approximation and MCMC tend to fail, or are unsatisfactorily slow to accomplish. The main reason for this is the dimension of this integral, which is nm, and thus for large sample sizes and/or number of classes, is unfeasible with such methods. We turned to variational inference in the face of an intractable posterior density that hampers an EM algorithm, and the result is a sequential updating scheme, similar in time and storage requirements to the EM algorithm.

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

Suggestions for future work include:

1. **Estimation of $\boldsymbol{\Psi}$**. A limitation we had to face in this work was to treat $\boldsymbol{\Psi}$ as fixed. The discussion in Section 5.6.3 shows that estimation of $\boldsymbol{\Psi}$ is possible, however, the specific nature of implementing this in computer code could not be explored in time. In particular, for the full I-probit model, the best method of imposing positive-definite constraints for $\boldsymbol{\Psi}$ in the M-step has not been fully researched.

2. **Inclusion of class-specific covariates**. Throughout the chapter, we assumed that covariates were unit-specific, rather than class-specific. To illustrate, consider modelling the choice of travel mode between two destinations (car, coach, train or aeroplane) as a function of disposable income and travel time. Individuals' income as a predictor of transportation choice is unit-specific, but clearly, travel time depends on the mode of transport. To incorporate class-specific covariates z_{ij}, the regression on the latent propensities in (5.2) could be extended as such:

$$y_{ij}^* = \overbrace{\alpha_j + f_j(x_i) + e(z_{ij})}^{f(x_i, z_{ij}, j)} + \epsilon_{ij}$$

   An I-prior would then be applied as usual, with careful consideration of the RKKS used to model f.

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

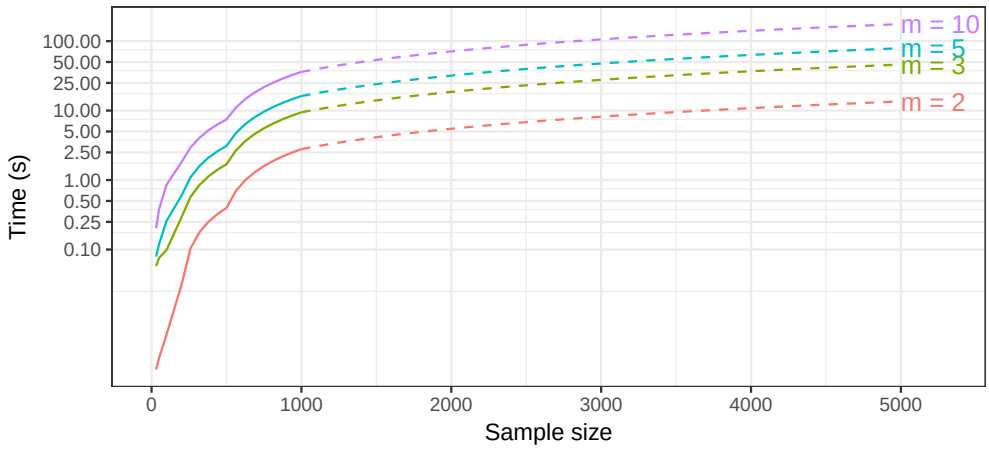


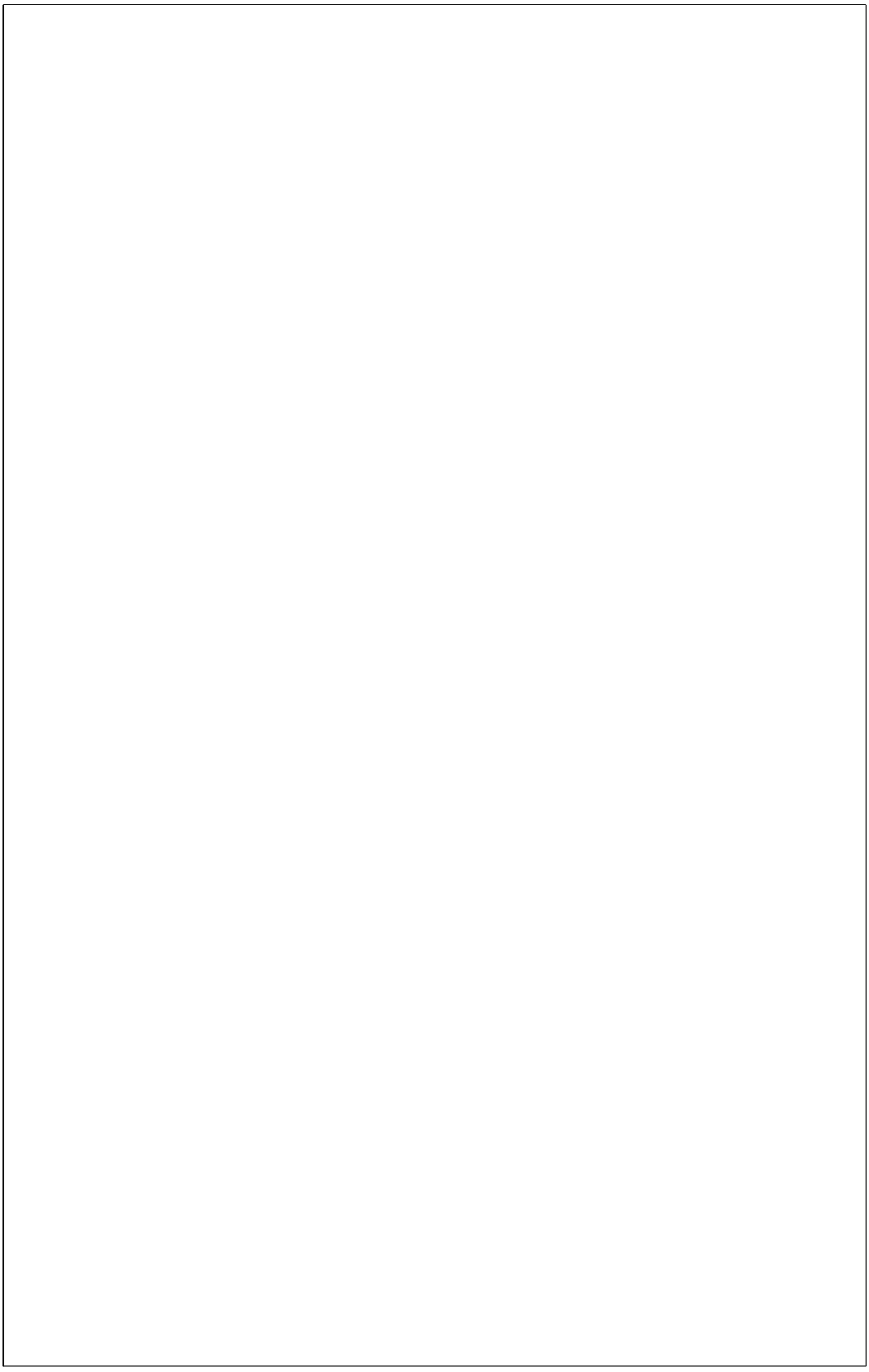
Figure 5.13: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes m. The solid line represents actual timings, while the dotted lines are linear extrapolations.

As a final remark, we note that variational Bayes, which entails a fully Bayesian treatment of the model (setting priors on model parameters θ), is a viable alternative to variational EM. The output of such a variational inference algorithm would be approximate posterior densities for θ, in addition to $q(\mathbf{y}^*)$ and $q(\mathbf{w})$, instead of point estimates for θ. Posterior inferences surrounding the parameters would then be possible, such as obtaining posterior standard deviations, credibility intervals, and so on. However, a variational Bayes route has its cons:

1. **Tedious derivations**. As the parameters now have a distribution $\theta = \{\boldsymbol{\alpha}, \eta, \boldsymbol{\Psi}\} \sim q(\boldsymbol{\alpha}, \eta, \boldsymbol{\Psi})$, quantities such as
   - $\mathrm{E}[\log|\boldsymbol{\Psi}|]$;
   - $\mathrm{E}[\mathbf{H}_\eta^2]$; and
   - $\operatorname{tr}\mathrm{E}\left[(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})\boldsymbol{\Psi}(\mathbf{y}^* - \mathbf{1}_n\boldsymbol{\alpha}^\top - \mathbf{H}_\eta\mathbf{w})^\top\right]$,

   among others, will need to be derived for the variational inference algorithm, and these can be tricky to compute.
2. **Suited only to conjugate exponential family models**. When conjugate exponential family models are considered, the approximate variational densities (under a mean-field assumption) are easily recognised, as they themselves belong to the same exponential family as the model or prior. However, I-prior does not always admit conjugacy for the kernel parameters η (only for ANOVA RKKSs scale parameters), and most certainly not for $\boldsymbol{\Psi}$ (at least not in the current parameterisation). When this happens, techniques such as importance sampling or Metropolis algorithms need to be employed to obtain the posterior means required for the variational algorithm to proceed.
3. **Prior specification and sensitivity**. It is not clear how best to specify prior information (from a subjectivist's standpoint) for the RKHS scale parameters, intercepts, and perhaps the error precision, because these are parameters relating to the latent propensities which are not very meaningful or interpretable. Of course, one could easily specify vague or even diffuse priors. The concern is that the model could be sensitive to prior choices.

In consideration of the above, we opted to employ a variational EM algorithm for estimation of I-probit models, instead of a full variational Bayes estimation. In any case, computational complexity is expected to be the same between the two methods. An interesting point to note is that the RKHS scale parameters and intercept would admit a normal posterior under a variational Bayes scheme. This means that the posterior mode and the posterior mean coincide, so point estimates under a variational EM algorithm are exactly the same as the posterior mean estimates under a variational Bayes framework when a diffuse prior is used.



Bibliography

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.

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. DOI: 10.2307/2290350.

Beal, Matthew James and Zoubin 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 University Press, pp. 453–464.

Bishop, Christopher (2006). *Pattern Recognition and Machine Learning*. Springer-Verlag. ISBN: 978-0-387-31073-2.

Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe (2017). "Variational Inference: A Review for Statisticians". In: *Journal of the American Statistical Association* 112.518, pp. 859–877. DOI: 10.1080/01621459.2017.1285773.

Breiman, Leo (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/A:1010933404324.

Breslow, Norman E. and David G. Clayton (1993). "Approximate Inference in Generalized Linear Mixed Models". In: *Journal of the American Statistical Association* 88.421, pp. 9–25. DOI: 10.2307/2290687.

Bunch, David S. (1991). "Estimability in the multinomial probit model". In: *Transportation Research Part B: Methodological* 25.1, pp. 1–12. DOI: 10.1016/0191-2615(91)90009-8.

Cannings, Timothy I. and Richard J. Samworth (2017). "Random-projection ensemble classification". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, pp. 959–1035. DOI: 10.1111/rssb.12228.

Chen, Yen-Chi, Y. Samuel Wang, and Elena A. Erosheva (2018). "On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example". In: *Annals of Applied Statistics* to appear. arXiv: 1711.11057 [stat.ME].

Dansie, Brenton R. (1985). "Parameter estimability in the multinomial probit model". In: *Transportation Research Part B: Methodological* 19.6, pp. 526–528. DOI: 10.1016/0191-2615(85)90047-5.

Deterding, David Henry (1990). "Speaker Normalization for Automatic Speech Recognition". PhD thesis. University of Cambridge.

Diggle, Peter, Paula Moraga, Barry Rowlingson, and Benjamin Taylor (2013). "Spatial and Spatio-Temporal Log-Gaussian Cox Processes: Extending the Geostatistical Paradigm". In: *Statistical Science* 28.4, pp. 542–563. DOI: 10.1214/13-STS441.

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. DOI: 10.1111/j.1467-9876.2005.05373.x.

Friedman, Jerome H., Trevor Hastie, and Robert Tibshirani (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag. ISBN: 978-0-387-84857-0. DOI: 10.1007/978-0-387-84858-7.

Geweke, John (1989). "Bayesian Inference in Econometric Models Using Monte Carlo Integration". In: *Econometrica* 57.6, pp. 1317–1339. DOI: 10.2307/1913710.

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* 76.4, pp. 609–632. DOI: 10.2307/2109766.

Girolami, Mark and Simon Rogers (2006). "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Computation* 18.8, pp. 1790–1817. DOI: 10.1162/neco.2006.18.8.1790.

Guvenir, H. Altay, Burak Acar, Gulsen Demiroz, and Ayhan Cekin (1997). "A supervised machine learning algorithm for arrhythmia analysis". In: *Computers in Cardiology 1997*. Lund, Sweden, pp. 433–436. DOI: 10.1109/CIC.1997.647926.

Hajivassiliou, Vassilis and Daniel McFadden (1998). "The Method of Simulated Scores for the Estimation of LDV Models". In: *Econometrica* 66.4, pp. 863–896. DOI: 10.2307/2999576.

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. DOI: 10.1016/0304-4076(94)01716-6.

Hastie, Trevor and Robert Tibshirani (1986). "Generalized Additive Models". In: *Statistical Science* 1.3, pp. 297–310. DOI: 10.1214/ss/1177013604.

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors". In: *Journal of the American Statistical Association* 90.430, pp. 773–795. DOI: 10.2307/2291091.

Keane, Michael (1992). "A Note on Identification in the Multinomial Probit Model". In: *Journal of Business & Economic Statistics* 10.2, pp. 193–200. DOI: 10.2307/1391677.

Keane, Michael and Kenneth 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* 76.4, pp. 648–672. DOI: 10.2307/2109768.

Kuss, Malte and Carl Edward Rasmussen (2005). "Assessing Approximate Inference for Binary Gaussian Process Classification". In: *Journal of Machine Learning Research* 6, pp. 1679–1704.

Louppe, Gilles (Oct. 2014). "Understanding Random Forests: From Theory to Practice". PhD thesis. University of Liege, Belgium. arXiv: 1407.7502 [stat.ML].

McCullagh, Peter and John A. Nelder (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall/CRC. ISBN: 978-0-412-31760-6.

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. DOI: 10.1016/S0304-4076(00)00034-8.

Minka, Thomas P. (Aug. 2001). "Expectation propagation for approximate Bayesian inference". In: *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI 2001)*. Ed. by Daphne Koller John Breese. San Francisco, CA, pp. 362–369. ISBN: 1-55860-800-1. arXiv: 1301.2294 [cs.AI].

Neal, Radford M. (1999). "Regression and Classification using Gaussian Process Priors". In: *Bayesian Statistics 6*. Proceedings of the Sixth Valencia International Meeting. Ed. by José M. Bernardo, James O. Berger, A. Philip Dawid, and Adrian F. M. Smith. Oxford University Press, pp. 475–501. ISBN: 978-0-19-850485-6.

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. DOI: 10.1023/A:10089053.

Rasmussen, Carl Edward and Christopher K. I. Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press. ISBN: 0-262-18253-X. URL: http://www.gaussianprocess.org/gpml/.

Robert, Christian (1995). "Simulation of truncated normal variables". In: *Statistics and Computing* 5.2, pp. 121–125. DOI: 10.1007/BF00143942.

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

Schölkopf, Bernhard and Alexander J. Smola (2002). *Learning with Kernels*. Support Vector Machines, Regularization, Optimization, and Beyond. The MIT Press. ISBN: 978-0-262-19475-4.

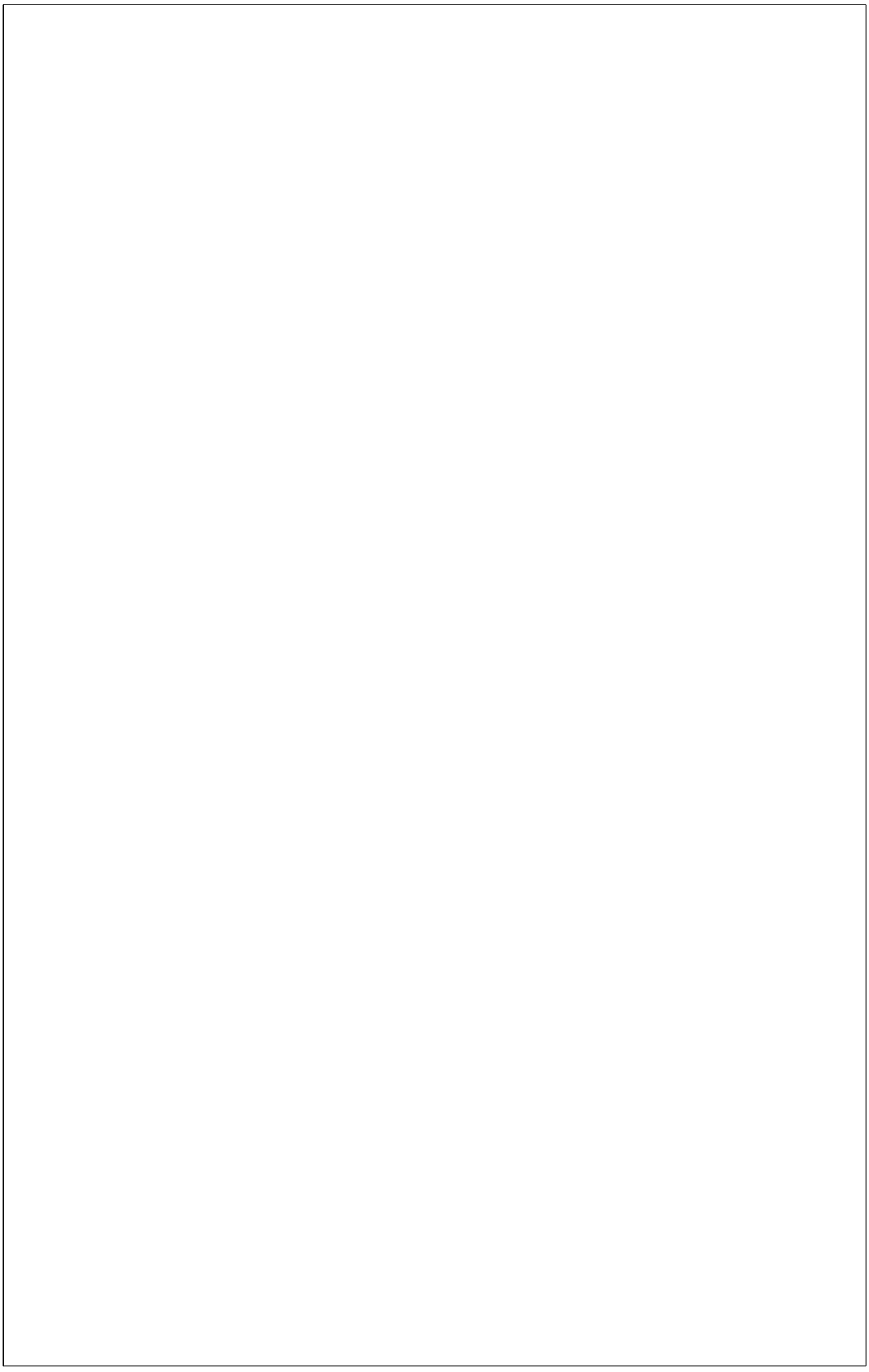
Skrondal, Anders and Sophia Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. Multilevel, Longitudinal, and Structural Equation Models. Chapman & Hall/CRC. ISBN: 978-1-58488-000-4.

Steinwart, Ingo and Andreas Christmann (2008). *Support Vector Machines*. New York: Springer-Verlag. ISBN: 978-0-387-77241-7. DOI: 10.1007/978-0-387-77242-4.

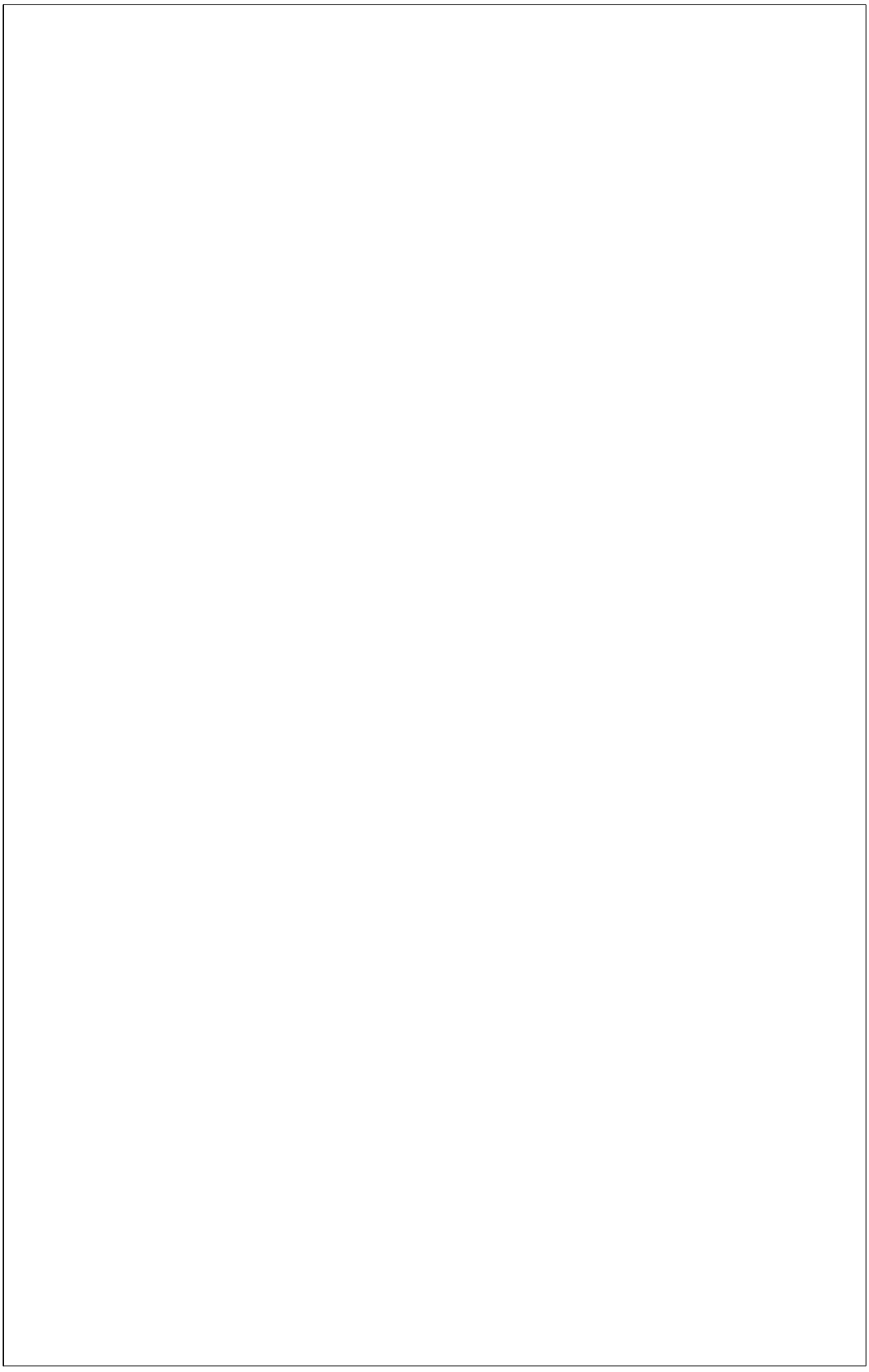
Taylor, Benjamin, Tilman Davies, Barry Rowlingson, and Peter Diggle (2013). "**lgcp**: An R Package for Inference with Spatial and Spatio-Temporal Log-Gaussian Cox Processes". In: *Journal of Statistical Software* 52.4, pp. 1–40. DOI: 10.18637/jss.v052.i04.

Tibshirani, Robert, Trevor Hastie, Balasubramanian Narasimhan, and Gilbert Chu (May 2002). "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression". In: *Proceedings of the National Academy of Sciences of the United States of America (PNAS 2002)*. Vol. 99. 10, pp. 6567–6572. DOI: 10.1073/pnas.082099299.

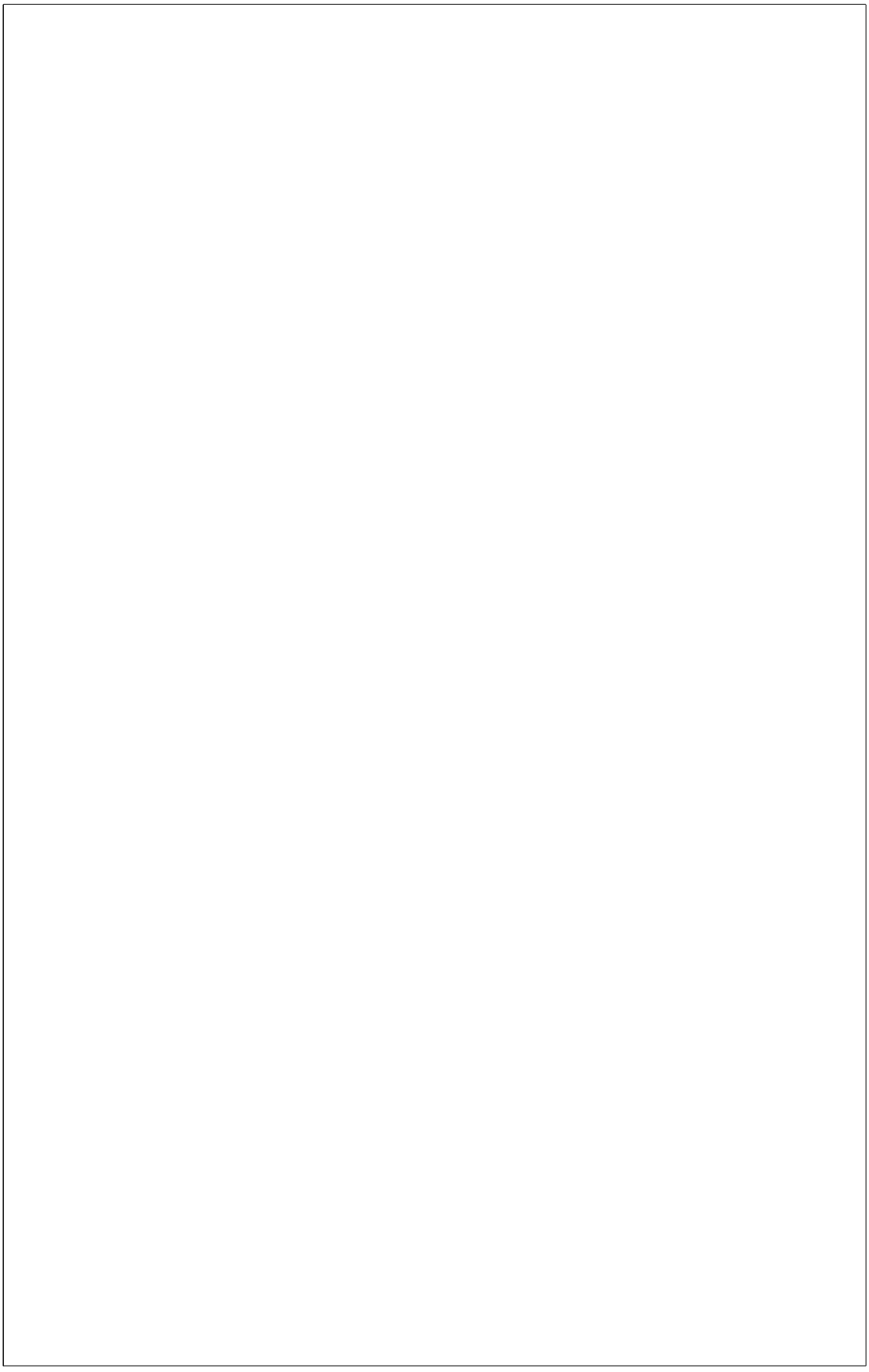
Train, Kenneth (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press. ISBN: 978-0-511-80527-1. DOI: 10.1017/CBO9780511805271.



Figures

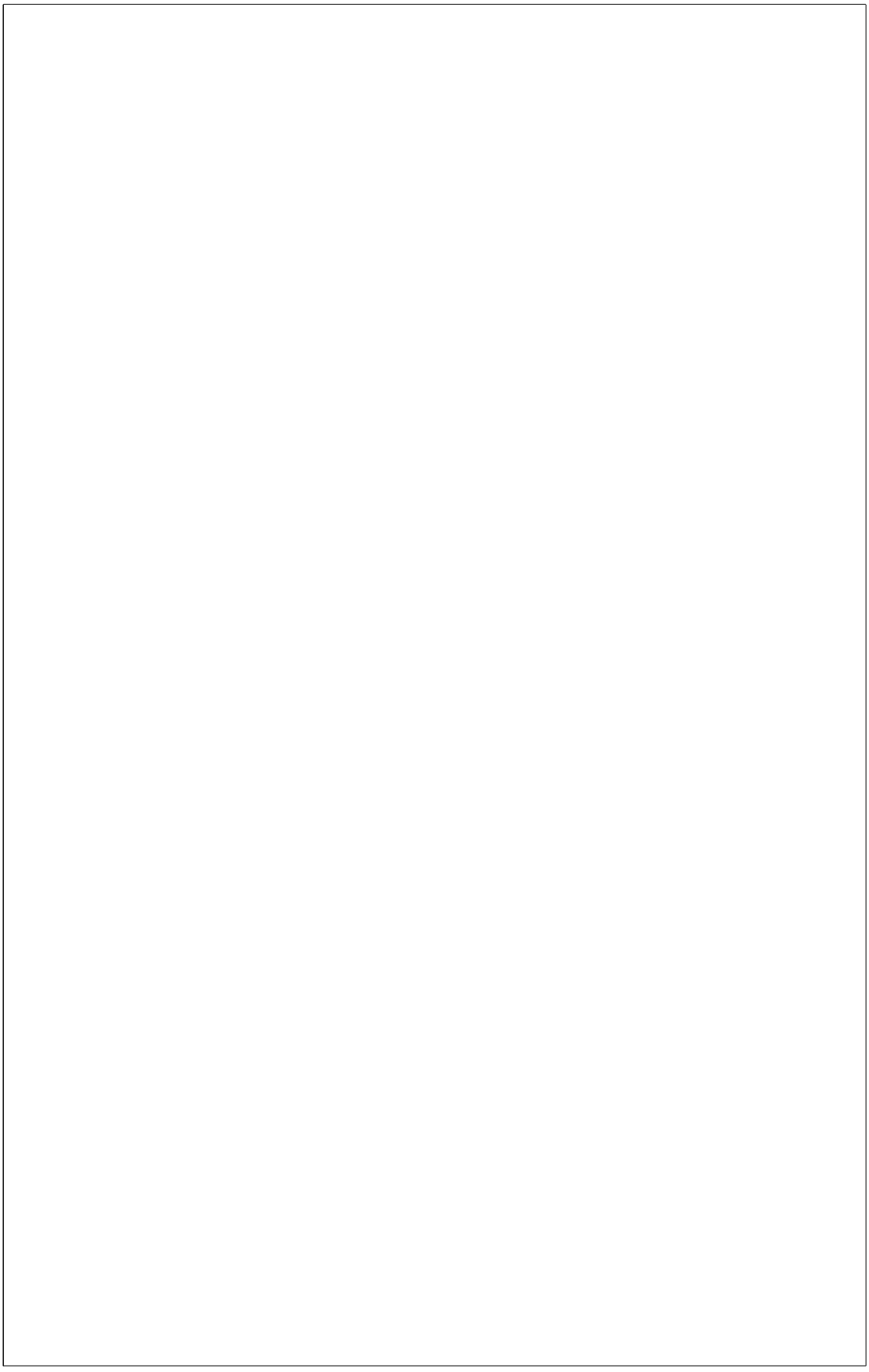


Tables



Theorems

Definitions



Nomenclature

As much as possible, and unless otherwise stated, the following conventions are used throughout this thesis.

Conventions

$\mathbf{a}, \mathbf{b}, \mathbf{c}, \ldots$	Boldface lower case letters denote real vectors
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$	Boldface upper case letters denote real matrices
$\mathcal{A}, \mathcal{B}, \mathcal{C}, \ldots$	Calligraphic upper case letters denote sets
x'	Primes are used to distinguish elements (not indicate derivatives)
$\hat{\theta}$	Hats are used to denote estimators of parameters

Indexing

$\mathbf{A}_{ij}$, A_{ij}, a_{ij}	The (i,j)'th element of the matrix $\mathbf{A}$
$\mathbf{A}_{i\cdot}$	The i'th row of the matrix $\mathbf{A}$ as a tall vector (transposed row vector)
$\mathbf{A}_{\cdot j}$	The j'th column vector of the matrix $\mathbf{A}$

Symbols

$\mathbb{N}$	The set of natural numbers (excluding zero)
$\mathbb{Z}$	The set of integers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}_{>0}$	The set of positive real numbers, $\{x \in \mathbb{R} \vert x > 0\}$
$\mathbb{R}_{\geq 0}$	The set of non-negative real numbers, $\{x \in \mathbb{R} \vert x \geq 0\}$
$\mathbb{R}^d$	The d-dimensional Euclidean space
$\mathcal{A}^c$	The complement of a set $\mathcal{A}$
$\mathcal{P}(\mathcal{A})$	The power set of the set $\mathcal{A}$
$\{\}, \emptyset$	The empty set
$\mathbf{0}$	A vector of zeroes
$\mathbf{1}_n$	A length n vector of ones
$\mathbf{I}_n$	The $n \times n$ identity matrix
$\exists$	(short hand) There exists
$\forall$	(short hand) For all
$\lim_{n\to\infty}$	The limit as n tends to infinity
$\xrightarrow{\text{dist.}}$	Convergence in distribution
$O(n)$	Computational complexity (time or storage)
Δx	A quantity representing a change in x

Relations

$a \approx b$	a is approximately or almost equal to b
$a \propto b$	a is equivalent to b up to a constant of proportionality
$a \equiv b$	a is identical to b
$A \Rightarrow B$	The statement B being true is predicated on A being true
$A \Leftrightarrow B$	The statement A is true if and only if B is true
$a \in \mathcal{A}$	a is an element of the set $\mathcal{A}$
$\mathcal{A} \subseteq \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which may include itself
$\mathcal{A} \subset \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$ which does not include itself
$a := b$, $a \leftarrow b$	a is assigned the value b
$X \sim p(X)$	The random variable X is distributed according to the pdf $p(X)$
$X \sim D$	The random variable X is distributed according to the pdf specified by the distribution D, e.g. $D \equiv \mathrm{N}(0, 1)$
$X_1,...,X_n \overset{\text{iid}}{\sim} D$	Each random variable X_i, $i = 1, \ldots, n$ is independently and identically distributed according to the pdf specified by the distribution D
$X \vert Y$	The (random) variable X given/conditional on Y

Functions

$\inf \mathcal{A}$	The infimum of a set $\mathcal{A}$
$\sup \mathcal{A}$	The supremum of a set $\mathcal{A}$
$\min \mathcal{A}$	The minimum value of a set $\mathcal{A}$
$\max \mathcal{A}$	The maximum value of a set $\mathcal{A}$
$\arg\min_x f(x)$	The value of x which minimises the function $f(x)$
$\arg\max_x f(x)$	The value of x which maximises the function $f(x)$
$\vert a \vert$ with $a \in \mathbb{R}$	The absolute value of a; $\vert a \vert = a$ if a is positive, and $-a$ if a is negative, and $\vert 0 \vert = 0$
$\delta_{xx'}$	The Kronecker delta; $\delta_{xx'} = 1$ if $x = x'$, and 0 otherwise
$[A]$	The Iverson bracket; $[A] = 1$ if the logical proposition A is true, and 0 otherwise
$\mathbb{1}_{\mathcal{A}}(x)$	The indicator function; $\mathbb{1}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$, and 0 otherwise
e^x, $\exp(x)$	The natural exponential function
$\log(x)$	The natural logarithmic function
$\frac{\mathrm{d}}{\mathrm{d}x} f(x)$, $\dot{f}(x)$	The derivative of f with respect to x
$f \circ g$	Composition of functions, i.e. g following f

Abstract vector space operations and notations

$\mathcal{V}^{\perp}$	The orthogonal complement of the space $\mathcal{V}$
$\mathcal{V}^{\vee}$	The algebraic dual space of $\mathcal{V}$
$\mathcal{V}^{*}$	The continuous dual space of $\mathcal{V}$
$\overline{\mathcal{V}}$	The closure of the space $\mathcal{V}$
$\mathcal{B}(\mathcal{V})$	The Borel σ-algebra of $\mathcal{V}$
$\mathrm{L}^p(\mathcal{X}, \nu)$	The set of p-integrable functions over the space $\mathcal{X}$ with measure ν
$\mathrm{L}(\mathcal{V}; \mathcal{W})$	The set of bounded, linear operators from $\mathcal{V}$ to $\mathcal{W}$
$\dim(\mathcal{V})$	The dimensions of the vector space $\mathcal{V}$
$\langle x, y \rangle_{\mathcal{V}}$	The inner product between x and y in the vector space $\mathcal{V}$

$\|x\|_{\mathcal{V}}$	The norm of x in the vector space $\mathcal{V}$
$D(x, y)$	The distance between x and y
$x \otimes y$	The tensor product of x and y which are elements of a vector space
$\mathcal{F} \otimes \mathcal{G}$	The tensor product space of two vector spaces
$\mathcal{F} \oplus \mathcal{G}$	The direct sum (or tensor sum) of two vector spaces
$\mathrm{d}f(x)$	The first Fréchet differential of f at x
$\mathrm{d}^2 f(x)$	The second Fréchet differential of f at x
$\partial_v f(x)$	The first Gâteaux differential of f at x in the direction v
$\partial_v^2 f(x)$	The second Gâteaux differential of f at x in the direction v
$\nabla f(x)$	The gradient of f at x (f is a mapping between Hilbert spaces)
$\nabla^2 f(x)$	The Hessian of f at x (f is a mapping between Hilbert spaces)

Matrix and vector operations

$\mathbf{a}^\top$, $\mathbf{A}^\top$	The transpose of a vector $\mathbf{a}$ or matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	The inverse of a square matrix $\mathbf{A}$
$\|\mathbf{a}\|^2$	The squared 2-norm the vector $\mathbf{a}$, equivalent to $\mathbf{a}^\top\mathbf{a}$
$\|\mathbf{A}\|$	The determinant of a matrix $\mathbf{A}$
$\mathrm{tr}(\mathbf{A})$	The trace of a square matrix $\mathbf{A}$
$\mathrm{diag}(\mathbf{A})$	The diagonal elements of a square matrix $\mathbf{A}$
$\mathrm{rank}(\mathbf{A})$	The rank of a matrix $\mathbf{A}$
$\mathrm{vec}(\mathbf{A})$	The column-wise vectorisation of a matrix $\mathbf{A}$
$\mathbf{a} \otimes \mathbf{b}$	The outer product of two vectors $\mathbf{a}$ and $\mathbf{b}$
$\mathbf{A} \otimes \mathbf{B}$	The Kronecker product of matrix $\mathbf{A}$ with matrix $\mathbf{B}$
$\mathbf{A} \circ \mathbf{B}$	The Hadamard product two matrices $\mathbf{A}$ and $\mathbf{B}$

Statistical functions

$\mathrm{P}(A)$	The probability of event A occurring
$p(X\|\theta)$	The probability density function of X given parameters θ
$L(\theta\|X)$	The log-likelihood of θ given data X, sometimes simply $L(\theta)$
$\mathrm{BF}(M, M')$	Bayes factor for comparing two models M and M'
$\mathcal{I}(\theta)$	The Fisher information for θ
$\mathrm{E}[X]$, $\mathrm{E}\,X$	The expectation[5] of the random element X
$\mathrm{Var}[X]$, $\mathrm{Var}\,X$	The variance[5] of the random element X
$\mathrm{Cov}[X, Y]$	The covariance[5] between two random elements X and Y
$H(p)$	The entropy of the distribution $p(X)$
$\mathrm{D}_{\mathrm{KL}}\big(q(x)\|p(x)\big)$	The Kullback-Leibler divergence from $p(x)$ to $q(x)$, denoted also by $\mathrm{D}_{\mathrm{KL}}(q\|p)$ for short

Statistical distributions

$\mathrm{N}(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

[5] When there is ambiguity as to which random element the expectation or variance is taken under or what its distribution is, this is explicated by means of subscripting, e.g. $\mathrm{E}_{X\sim\mathrm{N}(0,1)}\,X$ to denote the expectation of a standard normal random variable.

$\phi(z)$	The standard normal pdf
$\Phi(z)$	The standard normal cdf
$\phi(x\|\mu, \sigma^2)$	The pdf of $\mathrm{N}(\mu, \sigma^2)$
$\phi(\mathbf{x}\|\boldsymbol{\mu}, \boldsymbol{\Sigma})$	The pdf of $\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\mathrm{MN}_{n,m}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$	Matrix normal distribution with mean $\boldsymbol{\mu}$ and row variances $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ and column variances $\boldsymbol{\Psi} \in \mathbb{R}^{m\times m}$
${}^{\mathrm{t}}\mathrm{N}(\mu, \sigma^2, a, b)$	Truncated univariate normal distribution with mean μ and variance σ^2 restricted to the interval (a, b)
$\mathrm{N}_+(0, 1)$	The half-normal distribution
$\mathrm{N}_+(0, \sigma^2)$	The folded-normal distribution with variance σ^2
${}^{\mathrm{t}}\mathrm{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathcal{A})$	Truncated d-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ restricted to the set $\mathcal{A}$
$\Gamma(s, r)$	Gamma distribution with shape s and rate r parameters
$\Gamma^{-1}(s, \sigma)$	Inverse gamma distribution with shape s and scale σ parameters
χ^2_d	Chi-squared distribution with d degrees of freedom
$\mathrm{Bern}(p)$	Bernoulli distribution with probability of success p
$\mathrm{Cat}(p_1, \ldots, p_m)$	Categorical distribution with m categories, and each category has probability of success p_j