# To-do list



# Contents

Haziq Jamil
*Department of Statistics*
*London School of Economics and Political Science*
PhD thesis: 'Regression modelling using Fisher information covariance kernels (I-priors)'

# Chapter 5

# I-priors for categorical responses

chapter5

In a regression setting such as (1.1), consider polytomous response variables $y_1, \dots, y_n$, where each $y_i$ takes on exactly one of the values from the set of $m$ possible choices $\mathcal{M} = \{1, \dots, m\}$. Modelling categorical response variables is of profound interest in statistics, econometrics and machine learning, with applications aplenty. In the social sciences, categorical variables often arise from survey responses, and one may be interested in studying correlations between explanatory variables and the categorical response of interest. Economists are often interested in discrete choice models to explain and predict choices between several alternatives, such as consumer choice of goods or modes of transport. In this age of big data, machine learning algorithms are used for classification of observations based on what is usually a large set of variables or features.

The normality assumption (1.2) is not entirely appropriate anymore. As an extension to the I-prior methodology, we propose a flexible modelling framework suitable for regression of categorical response variables. In the spirit of generalised linear models (McCullagh and Nelder, 1989), we relate class probabilities of the observations to a normal I-prior regression model via a link function. Perhaps though, it is more intuitive to view it as machine learners do: since the regression function is ranged on the entire real line, it is necessary to "squash" it through some sigmoid function to conform it to the interval $[0, 1]$ suitable for probability ranges.

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

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

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

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

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

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

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

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

By choosing appropriate RKHSs/RKKSs for the regression functions, we are able to fit a multitude of binary and multinomial models, including multilevel or random-effects models, linear and non-linear classification models, and even spatio-temporal models. Examples of these models applied to real-world data is shown in Section 5.7. We find that the many advantages of the normal I-prior methodology transfer over quite well to the I-probit model for binary and multinomial regression.

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

## 5.1 A latent variable motivation: the I-probit model

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

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

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

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

{eq:latentmodel}

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

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

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

{eq:multinomial-latent}

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

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

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

{eq:pij}

where $\phi(\cdot|\mu, \Sigma)$ is the density of the multivariate normal with mean $\mu$ and variance $\Sigma$. 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 for the classes. While this does not have a closed-form expression and highlights one of the difficulties of working with probit models, the integral is by no means impossible to compute—see **??** for a note regarding this matter.

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

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

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

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

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

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

{eq:iprobit mod}

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 $\epsilon_{ij}$, 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 $\boldsymbol{\Psi} = \mathbf{I}_m$).

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

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

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

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

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

## 5.2 Identifiability and IIA

sec: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 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 of the underlying latent variables $\mathbf{y}_{i\cdot}^*$ with mean and variance $\{\boldsymbol{\mu}(x_i), \boldsymbol{\Psi}^{-1}\}$ would yield the same class probabilities as the multivariate normal distribution with mean and variance $\{a\mathbf{1}_m + c\boldsymbol{\mu}(x_i), c^2\boldsymbol{\Psi}^{-1}\}$, according to (5.3). Therefore, the multinomial probit model is not identified as there exists more than one set of parameters for which the categorical likelihood $\prod_{i,j} p_{ij}$ is the same.

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

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

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

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

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

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

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

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

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

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

Now, we turn to a discussion of the role of $\boldsymbol{\Psi}$ in the model. In decision theory, the independence axiom states that an agent's choice between a set of alternatives should not be affected by the introduction or elimination of a choice option. The probit model is suitable for modelling multinomial data where the independence axiom, which is also known as the *independence of irrelevant alternatives* (IIA) assumption, is not desired. Such cases arise frequently in economics and social science, and the famous Red-Bus-Blue-Bus example is often used to illustrate IIA: suppose commuters face the decision between taking cars and red busses. The addition of blue busses to commuters' choices

should, in theory, be more likely chosen by those who prefer taking the bus over cars. That is, assuming commuters are indifferent about the colour of the bus, commuters who are predisposed to taking the red bus would see the blue bus as an identical alternative. Yet, if IIA is imposed, then the three choices are distinct, and the fact that red and blue busses are substitutable is ignored.

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

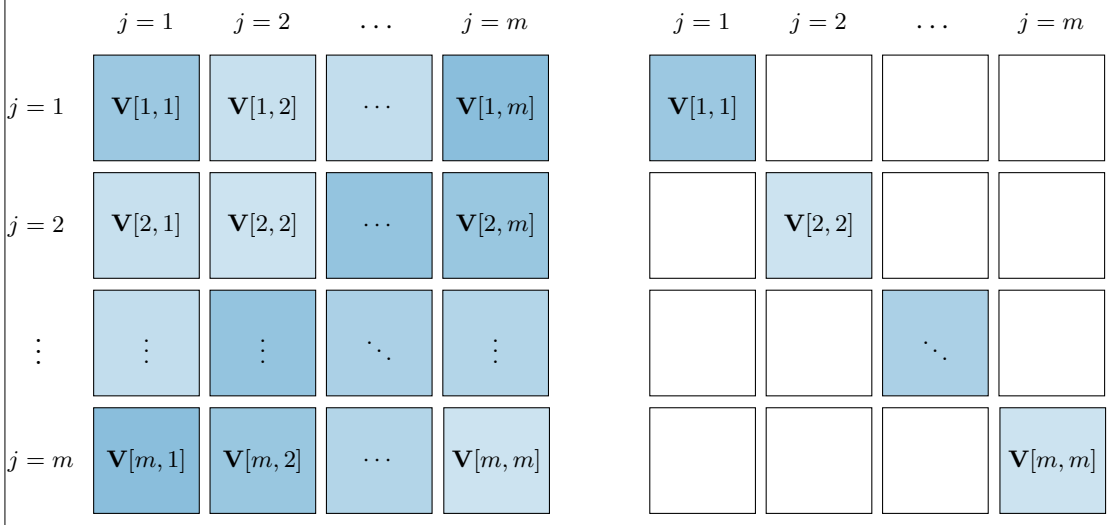


Figure 5.1: Illustration of the covariance structure of the full I-probit model (left) and the independent I-probit model (right). The full model has $m^2$ blocks of $n \times n$ symmetric matrices, and the blocks themselves are arranged symmetrically about the diagonal. The independent model, on the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

fig:iprobcovstr

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

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

$$\begin{aligned} p_{ij} &= \int_{\{y_{ij}^* > y_{ik}^* | \forall k \neq j\}} \cdots \int \prod_{k=1}^{m} \left\{ \phi(y_{ik}^* | \mu_k(x_i), \sigma_k^2) \, \mathrm{d}y_{ik}^* \right\} \\ &= \int \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left( \frac{y_{ij}^* - \mu_k(x_i)}{\sigma_k} \right) \cdot \phi(y_{ij}^* | \mu_j(x_i), \sigma_j^2) \, \mathrm{d}y_{ij}^* \\ &= \text{E}_Z \left[ \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left( \frac{\sigma_j}{\sigma_k} Z + \frac{\mu_j(x_i) - \mu_k(x_i)}{\sigma_k} \right) \right] \end{aligned} \tag{5.5}$$

{eq:pij2}

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. The probit link function is evidently seen in the above equation.

## 5.3 Estimation

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

$$\begin{aligned} L(\theta) &= \log \int p(\mathbf{y} | \mathbf{w}, \theta) \, p(\mathbf{w} | \theta) \, \mathrm{d}\mathbf{w} \\ &= \log \int \prod_{i=1}^{n} \prod_{j=1}^{m} \left( g_j^{-1}\big(\alpha_k + \overbrace{f_k(x_i)}^{\sum_{i'=1}^{n} h_\eta(x_i, x_{i'}) w_{i'k}}\big)_{k=1}^{m} \right)^{[y_i = j]} \cdot \text{MN}_{n,m}(\mathbf{w} | \mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi}) \, \mathrm{d}\mathbf{w} \end{aligned} \tag{5.6}$$

{eq:intractablelikelihood}

where we have denoted the probit relationship from (5.3) using the function $g_j^{-1} : \mathbb{R}^m \to [0, 1]$. Unlike in the continuous response models, the integral does not present itself in

closed form due to the conditional categorical PMF of the $y_i$'s, which they themselves involve integrals of multivariate normal densities. For binary response models, $g^{-1}$ is simply the probit function, but for multinomial responses, this can be quite challenging to evaluate—more on this in **??**.

Furthermore, the posterior distribution of the regression function, which requires the density of $\mathbf{w}|\mathbf{y}$, depends on the marginalisation provided by (5.6). The challenge of estimation is then to first overcome this intractability by means of a suitable approximation of the integral. We present three possible avenues to achieve this aim, namely the Laplace approximation, variational Bayes, and Markov chain Monte Carlo (MCMC) methods.

### 5.3.1 Laplace approximation

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

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

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

The marginal distribution is then approximated by

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

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

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

### 5.3.2 Variational inference

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

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

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

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

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

Under this scheme, the posterior for $\mathbf{y}^*$ is found to be a *conically truncated multivariate normal* distribution, and for $\mathbf{w}$, a multivariate normal distribution. The posterior density $q(\theta)$ is often of a recognisable form, and usually one of the exponential family densities (normal, Wishart or gamma). This is useful, because point estimates of the

hyperparameters can be taken to be either the mean or mode of these well-known distributions. In cases where $q(\theta)$ does not conform to an exponential family type density, then inference can still be done by sampling methods.

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

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

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

The algorithm to obtain $\tilde{q}$ which maximises the ELBO is known as the *coordinate ascent variational inference* (CAVI) algorithm. The algorithm itself typically condenses to that of a simple, sequential updating scheme, akin to the expectation-maximisation (EM) algorithm for exponential families we saw in Chapter 4, which is very fast to implement compared to the other methods described in the previous subsections. A full derivation of the variational algorithm used by us will be described in Section 5.4.

### 5.3.3 Markov chain Monte Carlo methods

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

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

Hamiltonian Monte Carlo is another possibility, since it does not require conjugacy. For binary models, this is a feasible approach because the class probabilities are a function of the normal CDF, which means that it is doable using off-the-shelf software such as Stan. However, with multinomial responses, the 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 5.2. The predictor variables are $X_1$ and $X_2$, each of which are scaled similarly.

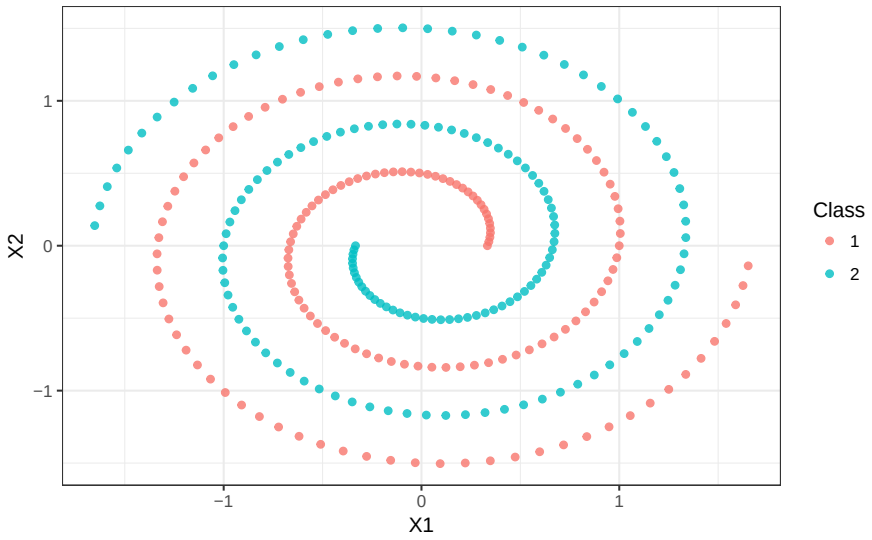


fig:example iprobit

Figure 5.2: A plot of simulated spiral data set.

The I-probit model that is fitted is

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

This binary model follows from the more general multinomial I-probit model by fixing all latent propensities in one of the classes to zero, and setting $\boldsymbol{\Psi} = \mathbf{I}_m$. This is possible because only differences in latent propensities are of interest, and not the actual values themselves, and thus only $m - 1$ sets of posterior regression functions need to be estimated—see **??** for further details.

The three estimation methods described aim to overcome the intractable integral by means of either a deterministic approximation (Laplace and variational inference) or a stochastic approximation (Hamiltonian MC). For the Bayesian methods, i.e. variational inference and Hamiltonian MC, vague priors were used on $\alpha$ and $\lambda$, namely $\mathrm{N}(0, 100)$ and $\mathrm{N}_+(0, 100)$ respectively. Restriction of $\lambda$ to the positive orthant is required for identifiability. The Laplace and variational methods were performed in the **iprobit** package, while Stan was used to code the Hamiltonian MC sampler. The results are presented in Table 5.1.

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

tab:comprei probit

| | Laplace approximation | Variational inference | Hamiltonian MC |
|---|---|---|---|
| Intercept ($\alpha$) | -0.02 (0.03) | 0.00 (0.06) | 0.00 (0.58) |
| Scale ($\lambda$) | 0.85 (0.01) | 5.67 (0.23) | 29.3 (5.21) |
| Log density | -202.7 | -140.7 | -163.8 |
| Error rate (%) | 44.7 | 0.00 | 2.24 |
| Brier score | 0.20 | 0.02 | 0.01 |
| Iterations | 20 | 56 | 2000 |
| Time taken (s) | >3600 | 5.32 | >3600 |

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

A plot of the log-likelihood surface for three methods in Figure 5.3 reveals some insight. The variational likelihood reveals two ridges, with the maxima occurring around the intersection of these two ridges. The Laplace likelihood seems to indicate a similar shape—in both the Laplace and variational method, the posterior distribution of $\mathbf{w}$ is approximated by a Gaussian distribution, with different means and variances. However,

parts of the Laplace likelihood are poorly approximated resulting in a loss of fidelity around the supposed maxima, which might have contributed to the set of values that were estimated. Laplace's method is known to yield poor approximations to probit-type likelihoods, as studied by Kuss and Rasmussen (2005). On the other hand, the log-likelihood using the posterior distribution of the Hamiltonian MC sampler (treating parameters as fixed values) yields a completely different shape compared to the other two methods.

In terms of predictive abilities, both the variational and Hamiltonian MC methods, even though the posteriors are differently estimated, has good predictive performance as indicated by their error rates and Brier scores. Figure 5.3 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 stronger than VI). 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 $\{w_1, \ldots, w_n, \alpha, \lambda\}$), and unlike in the normal model, we are not able to easily marginalise out the I-prior. As for Laplace, each Newton step involves obtaining posterior modes of the $w$'s, and this contributed to the slowness of this method. The reality is that variational inference takes seconds to complete what either the Laplace or full MCMC methods would take hours to. The predictive performance, while not as good as HMC, is certainly an acceptable compromise in favour of speed.

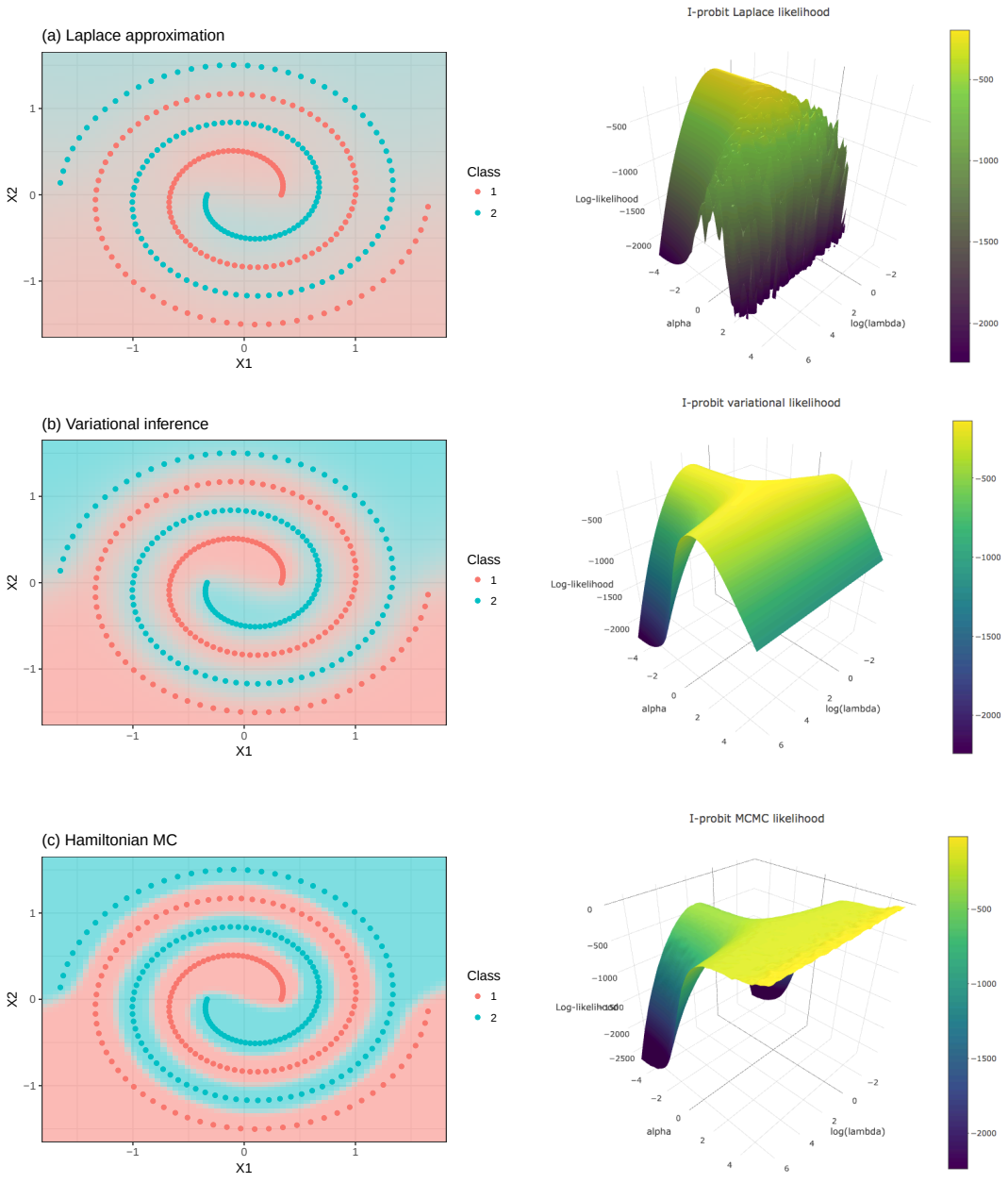


Figure 5.3: Plots showing predicted probabilities (shaded region) for belonging to class 1 or 2 indicated by colour and intensity, and likelihood surface plots for (a) Laplace's method, (b) variational inference, and (c) Hamiltonian MC.

fig:example iprobitfit

sec:iprobit var

## 5.4 A variational algorithm

## 5.5 Post-estimation

## 5.6 Computational consideration

## 5.7 Examples

sec:iprobit eg

## 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 (5.6). 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. How-

ever, Gaussian process classification typically uses the logistic sigmoid function, and estimation most commonly performed using Laplace approximation, but other methods such as expectation propagation (Minka, 2001) and MCMC (Neal, 1999) have been explored as well. Variational inference for Gaussian process probit models have been studied by Girolami and Rogers (2006), with their work providing a close reference to the variational algorithm employed by us.

Suggestions for future work include:

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

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

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

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

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

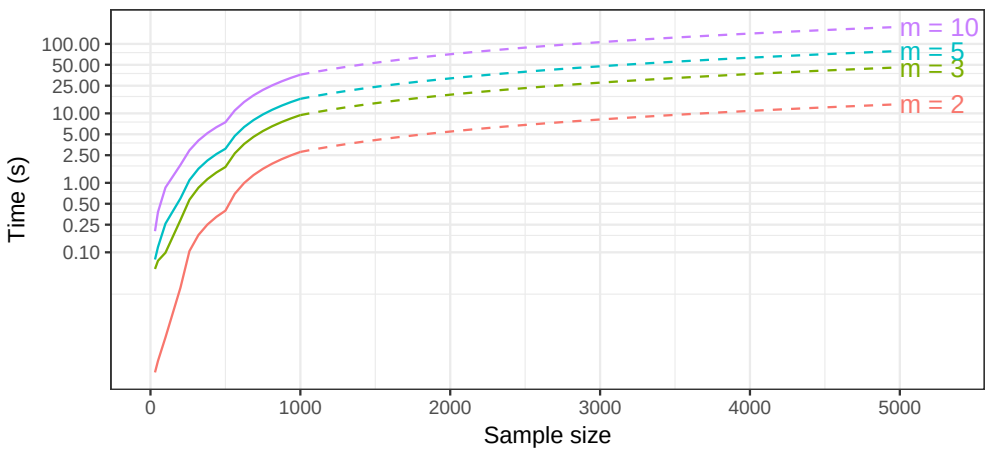


Figure 5.4: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes $m$. The solid line represents actual timings, while the dotted lines are linear extrapolations.

## 5.9 Miscellanea

# Appendix

# Bibliography

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

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

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

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

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

hastie1986 Hastie, Trevor and Robert Tibshirani (Aug. 1986). "Generalized Additive Models". In: *Statist. Sci.* 1.3, pp. 297–310. DOI: 10.1214/ss/1177013604. URL: https://doi.org/10.1214/ss/1177013604.

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

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

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

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

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

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

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

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

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