To-do list

Contents

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

Chapter 5

I-priors for categorical responses

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

As an extension to the I-prior methodology, we propose a flexible modelling framework suitable for regression of categorical response variables.

In the spirit of generalised linear models (**mccullagh1989**), we relate the class probabilities of the observations to the I-prior regression model via a link function. Perhaps though, it is more intuitive to view it as machine learners do: Since the regression function is ranged on the entire real line, it is necessary to "squash" it through some sigmoid function to conform it to the interval $[0, 1]$ suitable for probability measures. As in GLMs, the y_i's are assumed to follow an exponential family distribution, and in this case, the categorical distribution. We denote this by

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

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

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

$$p(y_i) = p_{i1}^{[y_i=1]} \cdots p_{im}^{[y_i=m]} \tag{5.1}$$ {eq:catdist}

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

$$g(p_{ij}) = \big(\alpha_j + f_j(x_i)\big)_{j=1}^{m}$$

where $g : [0, 1] \to \mathbb{R}^m$ is some specified link function. As we will see later, the normality assumption of the errors naturally implies a *probit* link function, i.e., g is the inverse cumulative distribution function (CDF) of a standard normal distribution (or more precisely, a function that *involves* the standard normal CDF). Normality is also a required assumption for I-priors to be specified on the regression functions. We call this method of probit regression using I-priors the *I-probit* regression model.

Note that the probabilities are modelled per class $j \in \{1, \ldots, m\}$ by individual regression curves f_j, and in the most general setting, m sets of intercepts α_j and kernel hyperparameters η_j must be estimated. The dependence of these m curves are specified through covariances $\sigma_{jk} := \text{Cov}[\epsilon_{ij}, \epsilon_{ik}]$, for each $j, k \in \{1, \ldots, m\}$ and $j \neq k$. While it may be of interest to estimate these covariances, this paper considers cases where the regression functions are class independent, i.e. $\sigma_{jk} = 0, \forall j \neq k$. This violates the independence of irrelevant alternatives (IIA) assumption (see Section 5.3 for details) crucial in choice models, but not so much necessary for classification when the alternatives are distinctively different.

The many advantages of the I-prior methodology of Jamil and Bergsma, 2017 transfer over quite well to the I-probit model for classification and inference. In particular, by choosing appropriate RKHSs for the regression functions, we are able to fit a multitude of binary and multinomial models, including multilevel or random-effects models, linear and non-linear classification models, and even spatio-temporal models. Examples of these models applied to real-world data is shown in Section **??**. Working in a Bayesian setting together with variational inference allows us to estimate the model much faster than traditional MCMC sampling methods, yet provides us with the conveniences that come with posterior estimates. For example, inferences around log-odds is usually cumbersome

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

for probit models, but a credibility interval can easily be obtained by resampling methods from the relevant posteriors, which are normally exponential family distributions in the I-probit model.

5.1 A naïve model

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

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

{eq:naivecl assmod}

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

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

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

{eq:anovacl ass}

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

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

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

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

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

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

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

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

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

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

{eq:naivecl assiprior}

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

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

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

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

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

and

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

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

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

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

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

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

5.2 A latent variable motivation: the I-probit model

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

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

{eq:latentmodel}

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

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

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

{eq:multinomial-latent}

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

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

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

{eq:pij}

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

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

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

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

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

{eq:pij2}

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

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

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

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

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

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

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

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

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

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

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

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

5.2.1 IIA

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

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

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

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

the class probability p_{ij} to be

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

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

5.3 Identifiability and IIA

sec:iia

The linear multinomial probit model is well known to be unidentified, and the reason for this is two-fold. Firstly, an addition of a constant to the latent variables y_{ij}^*'s in (5.7) will not change which latent variable is maximal, and therefore leaves the model unchanged. Secondly, all latent variables can be scaled by some positive constant without changing which latent variable is largest. Therefore, a *linear parameterisation* for the multinomial probit model is not identified as there can be more than one set of parameters for which the class probabilities are the same. To fix this issue, constraints are imposed on location and scale of the latent variables.

However, for the I-probit model, this is not the case, because the model is not related to the parameters θ linearly. One cannot simply add to or multiply θ by a constant and expect the model to be left unchanged. Thus, the I-probit model is identified in the parameter set θ without having to impose any restrictions, particularly on the precision matrix $\boldsymbol{\Psi}$.

To see how the I-probit model is location identified, suppose a constant a is added to the latent variables. This would then imply the relationship

$$a + y_{ij}^* = \overbrace{a + \alpha_j}^{\alpha_j^*} + f_j(x_i) + \epsilon_{ij},$$

which is similar to adding the constant a to all of the intercept parameters α_j—denote these new intercepts by α_j^*. As a requirement of the functional ANOVA decomposition, the α_j^*'s need to sum to zero, but we already have that $\sum_{j=1}^m \alpha_j = 0$, so it must be that $a = 0$. This also highlights the reason why the grand intercept α is not included in the model.

As for identification in scale, consider multiplying the latent variables by $c > 0$. The argument usually goes like this: the scaled latent variables cy_{ij}^* must have been generated from the model $c\theta$. However, we have that

$$\begin{aligned} c\mathbf{V}_y^*(\theta) &= c(\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + c(\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \\ &= (c\boldsymbol{\Psi} \otimes \mathbf{H}_\eta^2) + (c\boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n) \\ &\neq \mathbf{V}_y^*(c\theta). \end{aligned}$$

Full I-probit model

To solve the location identification, we must anchor on one of the latent variables. In our case, set the last latent variable $y^*_{im} = 0, \forall i = 1, \dots, n$. This means that we only have $m - 1$ sets of intercepts and RKHS parameters to estimate. Although it is sufficient to just fix one of the intercepts to for location identification, setting the last latent variable to zero also implies some restrictions on the covariance matrix $\boldsymbol{\Sigma}$, which also fixes scale identification as we will see.

2. Is this correct terminology?

3. Is it?

To solve scale identification, some restrictions on the covariance matrix $\boldsymbol{\Sigma}$ needs to be made. In the unrestricted case, there are $m(m+1)/2$ covariance parameters to estimate (upper triangular covariances and diagonal variances) in total. However, only $m(m-1)/2 - 1$ are identified, therefore requiring $m + 1$ restrictions (**Keane1992**; **train2009discrete**). Getting rid of one latent variable to fix the location identification shrinks $\boldsymbol{\Sigma}$ to $(m-1) \times (m-1)$ in size, and now only $m(m-1)/2$ parameters need to be estimated. Therefore, one more restriction needs to be made, and the convention is for the first element in $\boldsymbol{\Sigma}$—the variance of ϵ_{i1}—to be set to one.

5.4 Estimation

The parameters to estimate in the probit I-prior model are $\theta = (\alpha_1, \dots, \alpha_{m-1}, \eta_1, \dots, \eta_{m-1})$, i.e. the $m - 1$ sets of intercepts and RKHS parameters after accounting for identifiability. The likelihood function $L(\cdot)$ for θ 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:

$$\begin{aligned} L(\theta) &= \int p(\mathbf{y}|\boldsymbol{\alpha}, \mathbf{f}, \boldsymbol{\Sigma})\, p(\mathbf{f}|\eta, \boldsymbol{\Sigma}) \mathrm{d}\mathbf{f} \\ &= \int \prod_{i=1}^{n} \prod_{j=1}^{m} \left(g^{-1}_{\boldsymbol{\Sigma}}\big(\alpha_j + f_j(x_i)\big)^{m}_{j=1} \right)^{[y_i = j]} \cdot \mathrm{N}_{nm}(\mathbf{0}, \mathbf{G}_\eta(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_n)\mathbf{G}_\eta) \mathrm{d}\mathbf{f} \end{aligned} \tag{5.13}$$

{eq:intractablelikelihood}

where we have denoted the probit relationship from (5.9) using the function $g^{-1}_{\boldsymbol{\Sigma}} : \mathbb{R}^m \to [0, 1]$. Unlike in the continuous response models, the I-prior cannot be easily integrated out due to the conditional categorical PMF of the y_i's, which they themselves involve integrals of normal densities. Thus, the intractable integral above presents a practical challenge which makes estimation via direct maximisation of the likelihood difficult to accomplish. Methods of approximating the integral in (5.13) 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 in particular, for large sample sizes and/or number of classes, is unfeasible for such methods.

5.4.1 Laplace approximation

5.4.2 Markov chain Monte Carlo methods

5.4.3 Variational inference

We turn to variational inference as a method of estimation. Variational methods are widely discussed in the machine learning literature, and there have been efforts to popularise it in statistics (**blei2017variational**). Suppose that, in a fully Bayesian setting, we append the unknown model parameters to the vector $\mathbf{f}_j$ to form $\mathbf{z}_j = (\mathbf{f}_j, \alpha_j, \eta_j)$. The crux of variational inference is this: Find a suitably close distribution function $q(\mathbf{z}_j)$ that approximates the true posterior $p(\mathbf{z}_j|\mathbf{y})$, where closeness here is defined in the Kullback-Leibler divergence sense,

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

One may then show that log marginal density (the log of the intractable integral) holds the following bound:

$$\log p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{z}_j) p(\mathbf{z}_j) \mathrm{d}\mathbf{z}_j = \mathcal{L}(q) + \mathrm{KL}(q||p) \geq \mathcal{L}(q),$$

where $\mathcal{L}$ is some known functional of q. The bound achieves equality if and only if $q \equiv p$, but of course the true form of the posterior is unknown to us. Minimising $\mathrm{KL}(q||p)$ with respect to some constraints is a problem of calculus of variations, which incidentally is where variational inference takes its name. The constraint considered in this paper is simply that the approximate posterior q factorises into M disjoint factors. That is, suppose the elements of $\mathbf{z}_j$ is partitioned into M disjoint groups $\mathbf{z}_j = (z_j^{(1)}, \ldots, z_j^{(M)})$, then the restriction

$$q(\mathbf{z}_j) = \prod_{k=1}^{m} q_k(z_j^{(k)})$$

for q is considered. This factorised form of variational inference is known in the statistical physics literature as the *mean-field theory* (**itzykson1991statistical**). By factorising appropriately, we can obtain approximated posteriors for the regression function and the parameters of the I-prior model. The algorithm itself typically condenses to that of a simple, sequential updating scheme, akin to the expectation-maximisation (EM) algorithm for exponential families (**mclachlan2007algorithm**), which is very fast to implement compared to the other methods described in the previous paragraph.

5.4.4 Comparison of estimation methods

5.5 A variational algorithm

5.6 Post-estimation

5.7 Examples

5.8 Discussion

I-prior extended to non-normal data. Naive works good, but can be better. Simply transform the normal model through a squashing function. All the nice things about I-prior can be applied here too. Probit model variety of binary and multinomial regression models.

Laplace slow, unreliable modes. MCMC also slow. Variational has similarity to EM, but advantageous: easier to calculate posterior s.d., ability to do inference on transformed parameters.

As with the normal model, storage and time requirements slow. again, look to machine learning. improvements in variational algorithm.

Extend to include class-specific covariates.

improvement in calculating the normal integral? Need to see timing where takes longest

In terms of similarity to other works, the generalised additive models (GAMs) of **hastie1986** comes close. The setup of GAMs is near identical to the I-probit model, al-

though 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 (**scholkopf2002learning**) and Gaussian process classification (Rasmussen and Williams, 2006), with the latter being more closely related to the I-probit method. I-priors differ from Gaussian process priors in the specification of the covariance kernel. Gaussian process classification typically uses the logistic link function (or squashing function, to use machine learning nomenclature), and estimation is done most commonly using the Laplace approximation, but other methods such as expectation propagation (**minka2001expectation**) and MCMC (**neal1999**) have been explored as well. Variational inference for Gaussian process probit models have been studied by **girolami2006variational**, with their work providing a close reference to the variational algorithm employed by us.

5.9 Miscellanea

5.9.1 A note on computing the multivariate normal integral

misc:mnint

How is this calculated? Simulation usually, but also quadrature methods not too bad if m not too large. Stata sheet useful? Talk about if iid errors.

Much research has been devoted into developing efficient computational methods for computing these integral, and MCMC methods seem to be the tool of choice in Bayesian analysis(**mcculloch1994exact**; **nobile1998hybrid**; **mcculloch2000bayesian**). Things get more tractable if $\boldsymbol{\Sigma}$ is assumed to be diagonal (which corresponds to abandoning the independence of irrelevant alternatives assumption) and much more so if we assume that $\boldsymbol{\Sigma} = \mathbf{I}_m$. The latter yields the *normalised I-probit model*, and a discussion of the merits of this model is given later.

5. can use Hamiltonian Monte Carlo?

5.9.2 Similarity of EM algorithm and variational Bayes

Appendix

Bibliography

jamil2017 Jamil, Haziq and Wicher Bergsma (2017). "iprior: An R Package for Regression Modelling using I-priors". In: *Manuscript in submission*.

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