To-do list

Multinomial probit I-prior models

Haziq Jamil
Department of Statistics
London School of Economics and Political Science

May 23, 2017

Abstract

Extension of I-prior models to multi-class classification with estimation using variational inference.

Keywords: some, keywords, go, here

1 Introduction

Consider multinomial response variables $y_1, \dots, y_n$, where each response y_i takes on one of the values $\{1, \dots, m\}$ from a set of m possible values. We model each response as following a categorical distribution (a special case of the multinomial distribution)

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

with probability mass function (pmf)

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

such that $p_{ij} \geq 0$ for each j and $\sum_{j=1}^m p_{ij} = 1$. It might also be convenient to think of the responses y_i as comprising of a binary vector of length m, with a single '1' at the position corresponding to the value that y_i takes. That is,

$$y_i = (y_{i1}, \dots, y_{im})$$

with

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

and $\sum_{j=1}^m y_{ij} = 1$. In this formulation, each y_{ij} is distributed as Bernoulli with probability p_{ij}.

Suppose for each observation y_i there is an associated p-dimensional vector of covariates $x_i = (x_{i1}, \dots, x_{ip})$ belonging to some set $\mathcal{X}$. We would like to model the multinomial outcomes y_i based on these vectors of covariates. Assume that for each $y_i = (y_{i1}, \dots, y_{im})$, there exists a

corresponding continuous, underlying, latent variable $y_i^* = (y_{i1}^*, \dots, y_{im}^*)$ such that

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

In other words,

$$y_{ij} = \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*].$$

We consider modelling the n latent variables in each class $j = 1, \dots, m$ according to the regression problem

$$\begin{gathered} y_{ij}^* = f_j(x_i) + \epsilon_{ij} \\ \epsilon_{ij} \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma_j^2) \\ i = 1, \dots, n \end{gathered} \tag{1}$$

with $f_j : \mathcal{X} \to \mathbb{R}$ being a regression function belonging to some reproducing kernel Hilbert space of functions $\mathcal{F}_j$ having the reproducing kernel $h_{\lambda_j} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Here, λ_j is the scale parameter for the reproducing kernel, so that $h_{\lambda_j}(\cdot) = \lambda_j h(\cdot)$. One advantage of the multinomial probit model is the ability to also model the correlations across choices, such that for each $j, k \in \{1, \dots, m\}$ and $j \neq k$,

$$\mathrm{Corr}[\epsilon_{ij}, \epsilon_{ik}] = \frac{\sigma_{jk}}{\sigma_j \, \sigma_k}.$$

This setting is suitable for modelling multinomial data where the independence axiom is not desired. Such cases arise frequently in economics and social science. The famous Red-Bus-Blue-Bus example is often used to illustrate independence of irrelevant alternatives (IIA). Suppose commuters face the decision between taking cars and red busses. The addition of blue busses to commuters' choice should not affect the choice between cars or busses (assuming commuters are indifferent about the colour of the bus). Yet, if IIA is imposed, then the three choices are distinct, and the fact that red and blue busses are substitutable is ignored.

The IIA assumption is realised by fixing $\sigma_{jk} = 0$, $j \neq k$, which is clearly a simplification of the model (and as we will see later, benefits us in the algebra when deriving some distributional results). As this may not be suitable for certain modelling purposes, we might want to consider how this assumption can be relaxed, but we leave this for future work.

Back to model (1): We wish to model the function f_j as having an I-prior. Denoting $f_{ij} = f_j(x_i)$ as the evaluation of the function $f_j(\cdot)$ at x_i, and also $\mathbf{f}_j = (f_{1j}, \dots, f_{nj})^\top$ as the vector containing all n evaluations pertaining to the jth alternative, an I-prior on f_j is

$$\mathbf{f}_j \sim \mathrm{N}(\mathbf{f}_j^0, \mathcal{I}_j)$$

where $\mathcal{I}_j$ is the $n \times n$ Fisher covariance kernel for the regression function f_j in model (1), which has (r, s) entries given by

$$\mathcal{I}_j\big(f_j(x_r), f_j(x_s)\big) = \sigma_j^{-2} \sum_{k=1}^{n} \sum_{l=1}^{n} h_{\lambda_j}(x_r, x_k) h_{\lambda_j}(x_l, x_s)$$

and $\mathbf{f}_j^0$ is a vector of prior means. By concatenating the vectors $\mathbf{f}_1, \ldots, \mathbf{f}_m$ into the vector $\mathbf{f}$ of length nm, it is easy to see that

$$\mathbf{f} \sim \mathrm{N}(\mathbf{f}^0, \boldsymbol{\Omega})$$

where $\mathbf{f}^0 = (\mathbf{f}_1^0, \ldots, \mathbf{f}_m^0)^\top$ and $\boldsymbol{\Omega} = \mathrm{diag}(\mathcal{I}_1, \ldots, \mathcal{I}_m)$. This stems from the fact that the error components are independent across choices (IIA), and as such, $\mathrm{Cov}[\mathbf{f}_j, \mathbf{f}_k] = \mathbf{0}$, $j \neq k$. In the more general case where $\sigma_{jk} \neq 0$, then extra care must be taken to ensure the covariances are represented in the I-prior covariance matrix.

We make two further simplifications to the model. Firstly, the choice model as stated is not identified in scale, i.e. multiplication of the latent variables by a positive constant does not make any difference to the outcome. This is a well known identification issue[1]with the multinomial probit model, and this is typically overcome by setting some restrictions. In our case, we set all $\sigma_j^2 = 1$.

1. But isn't identifiability resolved by setting all covariances to zero?

Secondly, we assume that the intercept functions (prior means) are constants, so that $\mathbf{f}_j^0 = \alpha_j \mathbf{1}_n$ for $j = 1, \ldots, m$, and the α_js are just additional hyperparameters to be estimated. With all of these in mind, the I-prior model simplifies to

$$\begin{aligned} y_{ij}^* &= \alpha_j + \sum_{k=1}^{n} h_{\lambda_j}(x_i, x_k) w_{kj} + \epsilon_{ij} \\ \epsilon_{ij} &\overset{\text{iid}}{\sim} \mathrm{N}(0, 1) \\ w_{kj} &\overset{\text{iid}}{\sim} \mathrm{N}(0, 1) \end{aligned}$$

for each $j \in \{1, \ldots, m\}$. Define $f_{ij} = \alpha_j + \sum_{k=1}^{n} h_{\lambda_j}(x_i, x_k) w_k$, so that each $y_{ij}^* | f_{ij} \sim \mathrm{N}(f_{ij}, 1)$. The probit link is seen as follows:

$$\begin{aligned} p_{ij} &= \mathrm{P}\left[y_{ij}^* = \max_k y_{ik}^* \right] \\ &= \mathrm{P}\left[y_{ij}^* > y_{ik}^* : \forall k \neq j \right] \\ &= \mathrm{P}\left[f_{ij} + \epsilon_{ij} > f_{ik} + \epsilon_{ik} : \forall k \neq j \right] \\ &= \mathrm{P}\left[\epsilon_{ik} - \epsilon_{ij} \leq f_{ij} - f_{ik} : \forall k \neq j \right] \\ &= \int \cdots \int \mathbb{1}\left[\epsilon_{ik} \leq \epsilon_{ij} + f_{ij} - f_{ik} : \forall k \neq j \right] \prod_{k=1}^{m} \left[\phi(\epsilon_{ik}) \, \mathrm{d}\epsilon_{ik} \right] \\ &= \int \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi(\epsilon_{ij} + f_{ij} - f_{ik}) \phi(\epsilon_{ij}) \, \mathrm{d}\epsilon_{ij} \qquad (2) \\ &= \mathrm{E}_Z \left[\prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi(Z + f_{ij} - f_{ik}) \right] \end{aligned}$$

where Z is a standard normal random variable, and $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a standard normal distribution respectively. It is well know that for $m > 3$ this has no closed form expression, which makes the probit model unattractive from a likelihood maximisation

[1] In the unrestricted case for a model with m alternatives, there would be $m(m+1)/2$ variance components to estimate. However, in general, only $m(m-1)/2$ can be freely estimated.

standpoint.

2. Citation?

The above describes m regression functions being estimated for each class, and each of the m regression functions estimated using an I-prior. It is possible for all m regression functions[2]to share a common I-prior $\mathbf{f}_1, \ldots, \mathbf{f}_m \overset{\text{iid}}{\sim} \text{N}(\alpha \mathbf{1}_n, \mathcal{I})$, so that only one set of intercept and RKHS scale parameters need to be estimated (instead of m sets, one for each class). There is also flexibility in using the same covariance kernel for instance, but different intercepts for the m I-priors, or vice-versa. The probit I-prior model can be represented by the following DAG.

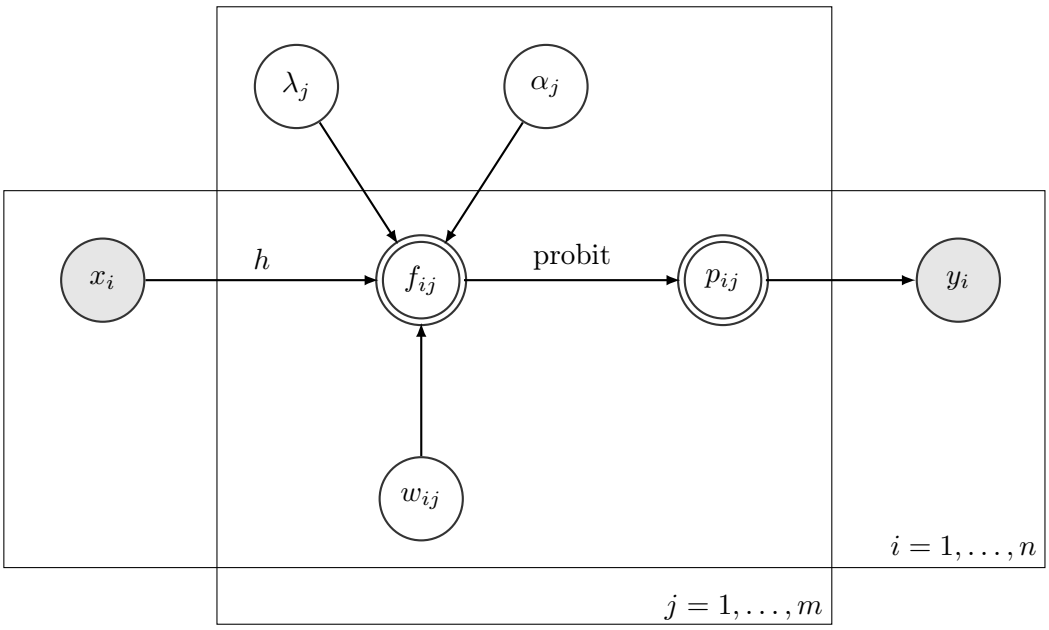


Figure 1: A DAG of the probit I-prior model. Observed nodes are shaded, while double-lined nodes represented known or calculable quantities. The latent variables y_{ij}^* have been marginalised and absorbed into the probit. The w_{ij}s are the standard normal random-effects associated with the I-prior. There are at most m sets of intercept (α_j) and scale (λ_j) parameters to estimate.

2 Estimation

The parameters to estimate in the probit I-prior model are $\boldsymbol{\theta} = (\alpha_1, \ldots, \alpha_m, \lambda_1, \ldots, \lambda_m)$. The likelihood function $L(\cdot)$ for $\boldsymbol{\theta}$ is obtained by integrating out the I-prior from the multinomial likelihood, as follows:

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \int \prod_{i=1}^{n} \prod_{j=1}^{m} p(y_i | f_{ij})\, p(f_{ij} | \alpha_j, \lambda_j)\, \mathrm{d}\mathbf{f} \\
&= \int \cdots \int \prod_{i=1}^{n} \prod_{j=1}^{m} p_{ij}^{\mathbb{1}[y_i = j]}\, p(\mathbf{f}_j | \alpha_j, \lambda_j)\, \mathrm{d}\mathbf{f}_1 \ldots \mathrm{d}\mathbf{f}_m.
\end{aligned}
$$

[2]It is also possible to reparameterise the model (anchoring on one latent variable as the reference class and working with the latent differences) so that only $m - 1$ I-priors are required.

From (2), we know that p_{ij} is related to the f_{ij} via the integral involving the CDF and PDF of a standard normal. Thus, the intractable integral above presents a practical challenge which makes estimation via direct maximisation of the likelihood difficult to accomplish. Several approximations are considered, and discussed below.

2.1 Laplace approximation

Add section for Laplace's method

2.2 Markov Chain Monte Carlo

Add section for MCMC

2.3 Variational inference

Add section for variational inference

3 Estimation using variational inference

The variational approximation to the probit I-prior model is made tractable by the inclusion of the latent variables y^*_{ij} in the analysis as an intermediary step. The unknown quantities to estimate via variational inference are y^*_{ij}, w_{ij}, λ_j and α_j.

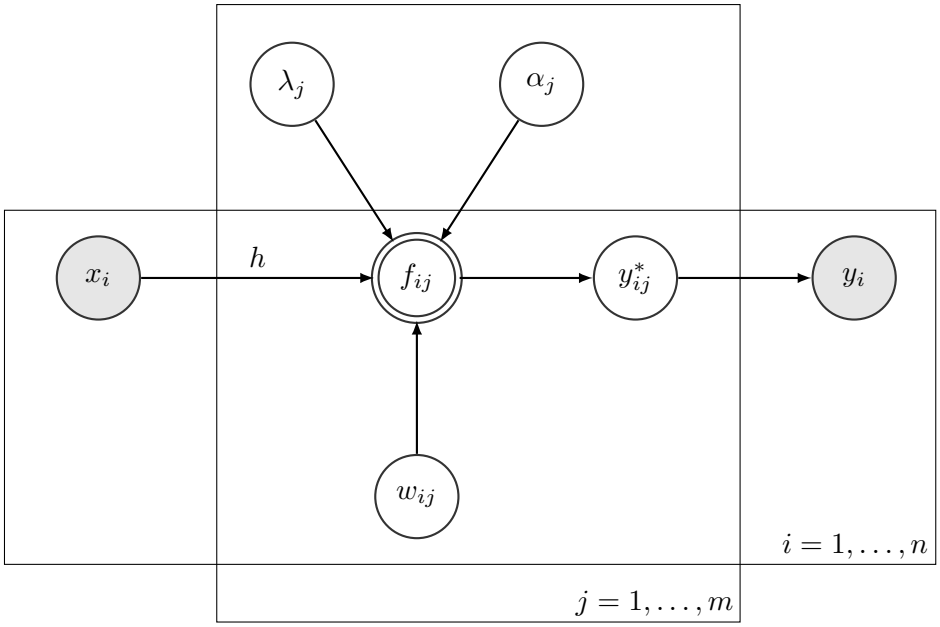


Figure 2: A DAG of the variational probit I-prior model.

3.1 The joint likelihood

$$\begin{aligned} p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \lambda, \alpha) &= p(\mathbf{y}|\mathbf{y}^*, \mathbf{w}, \alpha, \lambda) p(\mathbf{y}^*, \mathbf{f}, \mathbf{w}, \alpha, \lambda) \\ &= p(\mathbf{y}|\mathbf{y}^*) p(\mathbf{y}^*|\mathbf{f}) p(\mathbf{w}) p(\lambda) p(\alpha) \end{aligned}$$

3.2 Relevant distributions

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

$$\begin{aligned} p(\mathbf{y}^*|\mathbf{f}) &= \prod_{i=1}^{n} \prod_{j=1}^{m} \mathrm{N}(f_{ij}, 1) \\ &= \exp\left[-\frac{nm}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_{ij}^* - f_{ij})^2 \right] \\ &= \exp\left[-\frac{nm}{2} \log 2\pi - \frac{1}{2} \|\mathbf{y}^* - \mathbf{f}\|^2 \right] \end{aligned}$$

$$f_{ij} = \alpha_j + \sum_{k=1}^{n} h_{\lambda_j}(x_i, x_k) w_{kj}$$

$$\begin{aligned} p(\mathbf{w}) &= \prod_{i=1}^{n} \prod_{j=1}^{m} p(w_{ij}) \\ &= \prod_{i=1}^{n} \prod_{j=1}^{m} \mathrm{N}(0, 1) \\ &= \exp\left[-\frac{nm}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^2 \right] \\ &= \exp\left[-\frac{nm}{2} \log 2\pi - \frac{1}{2} \mathbf{w}^\top \mathbf{w} \right] \end{aligned}$$

$p(\lambda, \alpha) \propto \text{const.}$

3.3 Mean field approximation

$$\begin{aligned} p(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda|\mathbf{y}) &\equiv q(\mathbf{y}^*) q(\mathbf{w}) q(\lambda) q(\alpha) \\ &\equiv \prod_{i,j} q(y_{ij}^*) q(\mathbf{w}) q(\lambda) q(\alpha) \end{aligned}$$

The first line is by assumption, while the second line follows from an induced factorisation, as we will see later. Denote by $\tilde{q}$ the distributions which minimise the KL divergence (maximises the lower bound). Then, for each of $\xi \in \{\mathbf{y}^*, \mathbf{w}, \alpha, \lambda\}$, $\tilde{q}$ satisfies

$$\log \tilde{q}(\xi) = \mathrm{E}_{-\xi}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda)] + \text{const.}$$

3.3.1 $\tilde{q}(\mathbf{y}^*)$

In this subsection, we use the notation $y_i^* = (y_{i1}^*, \ldots, y_{im}^*)$ to denote the vector of length m containing the latent variables for response i. The joint distribution for $\mathbf{y}^* = (y_1^*, \ldots, y_n^*)^\top$ is a product of the distribution for each of the components y_i^* - this is a consequence of the independence structure across observations. Therefore, we can consider the variational density for each y_i^* separately.

Consider the case where y_i takes one particular value $j \in \{1, \ldots, m\}$. The mean-field density $q(y_i^*)$ for each $i = 1, \ldots, n$ is found to be

$$\begin{aligned}
\log \tilde{q}(y_i^*) &= \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \cdot \mathrm{E}_{\mathbf{w},\alpha,\lambda}\left[-\frac{1}{2}\sum_{k=1}^{m}(y_{ik}^* - f_{ik})^2\right] + \text{const.} \\
&= \mathbb{1}[y_{ij}^* = \max_k y_{ik}^*] \cdot \left[-\frac{1}{2}\sum_{k=1}^{m}(y_{ik}^* - \tilde{f}_{ik})^2\right] + \text{const.} \\
&\equiv \begin{cases} \prod_{k=1}^{m} \mathrm{N}(\tilde{f}_{ik}, 1) & \text{if } y_{ij}^* > y_{ik}^*, \forall k \neq j \\ 0 & \text{otherwise} \end{cases}
\end{aligned}$$

where $\tilde{f}_{ik} = \mathrm{E}[\alpha_k] + \sum_{l=1}^{m} h_{\mathrm{E}[\lambda_k]}(x_i, x_l)\,\mathrm{E}[w_{il}]$, and expectations are taken under the optimal mean-field distribution $\tilde{q}$. The distribution for $q(y_i^*)$ is a truncated m-variate normal distribution such that the jth component is always largest. It is worth investigating the properties of this distribution, and we now present some relevant definitions and results.

Definition 1 (Conically-truncated multivariate normal distribution)**.** *Let $\mathbf{X} = (X_1, \ldots, X_d)$ be a d-dimensional random variable with pdf defined as*

$$p(\mathbf{x}) = \begin{cases} \prod_{i=1}^{d} \mathrm{N}(\mu_i, \sigma_i) & \text{if } X_j > X_i, \forall i \neq j \\ 0 & \text{otherwise} \end{cases}$$

for some $j \in \{1, \ldots, d\}$. We denote the distribution of $\mathbf{X}$ by $\mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$. The pdf of $\mathbf{X}$ has support on the set $\{\mathbb{R}^d \,|\, x_j > x_i, \forall i \neq j\}$ and the following functional form:

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

where ϕ is the pdf of a standard normal distribution and

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

where $Z \sim \mathrm{N}(0,1)$. *In the case where all variances are unity, the pdf of* $\mathbf{X} \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \mathbf{I}_d)$ *is*

$$p(\mathbf{x}) = \left\{ (2\pi)^{d/2} \, \mathrm{E}_Z \left[\prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(Z + \mu_j - \mu_i\right) \right] \right\}^{-1} \exp\left[-\frac{1}{2} \sum_{i=1}^{d} (x_i - \mu_i)^2 \right].$$

Proof. A derivation of the functional form for the pdf of $X \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is given. Using the fact that $\int p(x)\,\mathrm{d}x = 1$, and that

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

the proof follows directly. □

Lemma 1. *Let* $X \sim \mathrm{N}^{(j)}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *with pdf* $p(\mathbf{x})$ *as defined in Definition 1. Then*

(i) The expectation $\mathrm{E}[\mathbf{X}] = \big(\mathrm{E}[X_1], \ldots, \mathrm{E}[X_d]\big)$ *is given by*

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

(ii) The differential entropy $\mathcal{H}(p)$ *is given by*

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

where $C = \mathrm{E}\left[\prod_{i \neq j} \Phi_i\right]$*, and we had defined*

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

with $Z \sim \mathrm{N}(0,1)$*, and* $\phi(\cdot)$ *and* $\Phi(\cdot)$ *the pdf and cdf of* Z *respectively.*

As we know, y_i takes on any one value from the set $\{1, \ldots, m\}$. Thus, we have that the distribution of $(y_{i1}^*, \ldots, y_{im}^*)$ is $\mathrm{N}^{(y_i)}(\boldsymbol{\mu}_i, \mathbf{I}_m)$, where $\boldsymbol{\mu}_i = (\tilde{f}_{i1}, \ldots, \tilde{f}_{im})$. The expectation is given by

$$\mathrm{E}[y_{ik}^*] = \begin{cases} \tilde{f}_{ik} - C_i^{-1} \mathrm{E}_Z\left[\phi_{ik}(Z) \prod_{l \neq k, y_i} \Phi_{il}(Z)\right] & \text{if } k \neq y_i \\ \tilde{f}_{iy_i} - \sum_{k \neq y_i} \left(\mathrm{E}[y_{ik}^*] - \tilde{f}_{ik}\right) & \text{if } k = y_i \end{cases}$$

where

$$\phi_{ik}(Z) = \phi(Z + \tilde{f}_{iy_i} - \tilde{f}_{ik})$$
$$\Phi_{ik}(Z) = \Phi(Z + \tilde{f}_{iy_i} - \tilde{f}_{ik})$$
$$C_i = \mathrm{E}_Z\left[\prod_{\substack{i=1 \\ i \neq j}}^{d} \Phi\left(Z + \tilde{f}_{iy_i} - \tilde{f}_{ik}\right)\right]$$

and $Z \sim \mathrm{N}(0,1)$ with PDF and CDF $\phi(\cdot)$ and $\Phi(\cdot)$ respectively. In order to calculate these expectations, we need to compute the following integrals:

$$\mathrm{E}_Z\left[\phi_{ik}(Z) \prod_{l \neq k, j} \Phi_{il}(Z)\right] = \int \phi_{ik}(z) \prod_{l \neq k, j} \Phi_{il}(z)\phi(z)\,\mathrm{d}z, \quad \forall k \neq y_i$$
$$C_i = \mathrm{E}_Z\left[\prod_{l \neq j} \Phi_{il}(Z)\right] = \int \prod_{l \neq j} \Phi_{il}(z)\phi(z)\,\mathrm{d}z$$

Since these are functions of a Gaussian pdf, these can be computed rather efficiently using quadrature methods.

3.3.2 $\tilde{q}(\mathbf{w})$

For each $j = 1, \ldots, m$, denote $\mathbf{y}_j^* = (y_{1j}^*, \ldots, y_{nj}^*)^\top$ as the vector of length n containing all latent observations for each class. Then,

$$\log \tilde{q}(\mathbf{w}) = \mathrm{E}_{\mathbf{y}^*, \alpha, \lambda}\left[-\frac{1}{2}\sum_{j=1}^{m} \|\mathbf{y}_j^* - \alpha_j \mathbf{1}_n - \mathbf{H}_{\lambda_j}\mathbf{w}_j\|^2 - \frac{1}{2}\sum_{j=1}^{m} \|\mathbf{w}_j\|^2\right] + \text{const.}$$

$$= -\frac{1}{2}\sum_{j=1}^{m} \mathrm{E}_{\mathbf{y}^*,\alpha,\lambda}\left[\mathbf{w}_j^\top \mathbf{H}_{\lambda_j}^2 \mathbf{w}_j + \mathbf{w}_j^\top \mathbf{w}_j - 2(\mathbf{y}_j^* - \alpha_j \mathbf{1}_n)^\top \mathbf{H}_{\lambda_j} \mathbf{w}_j\right] + \text{const.}$$

$$= -\frac{1}{2}\sum_{j=1}^{m} \left(\mathbf{w}_j^\top (\mathrm{E}[\mathbf{H}_{\lambda_j}^2] + \mathbf{I}_n)\mathbf{w}_j - 2(\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathrm{E}[\mathbf{H}_{\lambda_j}]\mathbf{w}_j\right) + \text{const.}$$

Let $\mathbf{A}_j = \mathrm{E}[\mathbf{H}_{\lambda_j}^2] + \mathbf{I}_n$ and $\mathbf{a}_j = \mathrm{E}[\mathbf{H}_{\lambda_j}](\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)$. Then, using the fact that

$$\mathbf{w}_j^\top \mathbf{A}_j \mathbf{w}_j - 2\mathbf{a}_j^\top \mathbf{w}_j = (\mathbf{w}_j - \mathbf{A}_j^{-1}\mathbf{a}_j)^\top \mathbf{A}_j (\mathbf{w}_j - \mathbf{A}_j^{-1}\mathbf{a}_j),$$

we see the $\log \tilde{q}(\mathbf{w})$ is a sum of quadratic terms in $\mathbf{w}_j$, and we recognise this as the kernel of the product of indepdendent multivariate normal densities. Therefore, for each $j = 1, \ldots, m$,

$$\tilde{q}(\mathbf{w}_j) \equiv \mathrm{N}(\mathbf{A}_j^{-1}\mathbf{a}_j, \mathbf{A}_j^{-1}),$$

and $\tilde{q}(\mathbf{w}) = \prod_{j=1}^m \tilde{q}(\mathbf{w}_j)$. Because of this induced factorisation, we can obtain mean-field densities for each $\mathbf{w}_j$ separately. For convenience later in deriving the lower bound, we note that the second moment of $\tilde{q}(\mathbf{w}_j)$ is equal to $\mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top] = \mathbf{A}_j^{-1}(\mathbf{I}_n + \mathbf{a}_j \mathbf{a}_j^\top \mathbf{A}_j^{-1}) =: \widetilde{\mathbf{W}}_j$.

3.3.3 $\tilde{q}(\lambda)$

For $j = 1, \ldots, m$,

$$\log \tilde{q}(\lambda_j) = \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\alpha}\left[-\frac{1}{2}\sum_{j=1}^{m} \|\mathbf{y}_j^* - \alpha_j \mathbf{1}_n - \lambda_j \mathbf{H}\mathbf{w}_j\|^2\right] + \text{const.}$$

$$= -\frac{1}{2}\sum_{j=1}^{m} \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\alpha}\left[\lambda_j^2 \mathbf{w}_j^\top \mathbf{H}^2 \mathbf{w}_j - 2\lambda_j (\mathbf{y}_j^* - \alpha_j \mathbf{1}_n)^\top \mathbf{H}\mathbf{w}_j\right] + \text{const.}$$

$$= -\frac{1}{2}\sum_{j=1}^{m} \left(\lambda_j^2 \operatorname{tr}\left(\mathbf{H}^2 \mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top]\right) - 2\lambda_j (\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathbf{H}\, \mathrm{E}[\mathbf{w}_j]\right) + \text{const.}$$

By completing the squares, we recognise this is as the kernel of the product of independent univariate normal densities. Thus, each $\lambda_j \sim \mathrm{N}(d_j/c_j, 1/c_j)$, where

$$c_j = \operatorname{tr}\left(\mathbf{H}^2 \mathrm{E}[\mathbf{w}_j \mathbf{w}_j^\top]\right) \quad \text{and} \quad d_j = (\mathrm{E}[\mathbf{y}_j^*] - \mathrm{E}[\alpha_j]\mathbf{1}_n)^\top \mathbf{H}\, \mathrm{E}[\mathbf{w}_j].$$

Supposing we use the same covariance kernel (and therefore scale parameter) for each regression class, the distribution for λ is easily seen as

$$\lambda \sim \mathrm{N}\left(\frac{\sum_{j=1}^m d_j}{\sum_{j=1}^m c_j}, \frac{1}{\sum_{j=1}^m c_j}\right).$$

3.3.4 $\tilde{q}(\alpha)$

For $j = 1, \ldots, m$, denote $\mathbf{H}_i$ as the row vector of the kernel matrix $\mathbf{H}$. Then,

$$\begin{aligned}
\log \tilde{q}(\alpha) &= \mathrm{E}_{\mathbf{y}^*,\mathbf{w},\lambda}\left[-\frac{1}{2}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{ij}^* - \alpha_j - \lambda_j \textstyle\sum_{k=1}^{n} h(x_i, x_k) w_{kj}\right)^2\right] + \text{const.} \\
&= -\frac{1}{2}\sum_{j=1}^{m}\mathrm{E}_{\mathbf{y}^*,\mathbf{w},\lambda}\left[n\alpha_j^2 - 2\alpha_j \sum_{i=1}^{n}(y_{ij}^* - \lambda_j \mathbf{H}_i \mathbf{w}_j)\right] + \text{const.} \\
&= -\frac{n}{2}\sum_{j=1}^{m}\left[\left(\alpha_j - \frac{1}{n}\sum_{i=1}^{n}(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i \mathbf{w}_j)\right)^2\right] + \text{const.}
\end{aligned}$$

which is of course the kernel of the product of m univariate normal densities, each with mean and variance

$$\tilde{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\,\mathrm{E}[\mathbf{w}_j]\right) \quad \text{and} \quad v_{\alpha_j} = \frac{1}{n}.$$

Suppose that we use a single intercept parameter α. In this case, α is is also normally distributed with mean and variance

$$\tilde{\alpha} = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\mathrm{E}[y_{ij}^*] - \mathrm{E}[\lambda_j]\mathbf{H}_i\,\mathrm{E}[\mathbf{w}_j]\right) \quad \text{and} \quad v_{\alpha} = \frac{1}{nm}.$$

3.4 Monitoring the lower bound

A convergence criterion would be when there is no more significant increase in the lower bound $\mathcal{L}$, as defined by

$$\begin{aligned}
\mathcal{L} &= \int q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha) \log\left[\frac{p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \lambda, \alpha)}{q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha)}\right] \mathrm{d}\mathbf{y}^*\,\mathrm{d}\mathbf{w}\,\mathrm{d}\lambda\,\mathrm{d}\alpha \\
&= \mathrm{E}[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \lambda, \alpha)] - \mathrm{E}[\log q(\mathbf{y}^*, \mathbf{w}, \lambda, \alpha)] \\
&= \cancel{\mathrm{E}\left[\log \prod_{i=1}^{n}\prod_{j=1}^{m} p(y_i|y_{ij}^*)\right]} + \mathrm{E}\left[\log p(\mathbf{y}^*|\mathbf{f})\right] + \mathrm{E}\left[\log p(\mathbf{w})\right] + \cancel{\mathrm{E}\left[\log p(\lambda)\right]} + \cancel{\mathrm{E}\left[\log p(\alpha)\right]} \\
&\quad - \mathrm{E}\left[\log q(\mathbf{y}^*)\right] - \mathrm{E}\left[\log q(\mathbf{w})\right] - \mathrm{E}\left[\log q(\lambda)\right] - \mathrm{E}\left[\log q(\alpha)\right]
\end{aligned}$$

Note that the categorical pmf $p(y_i|y_{ij}^*)$ becomes degenerate once the latent variables are known, so this term is cancelled out. With the exception of $q(\mathbf{y}^*)$, all of the distributions are Gaussian. The following results will be helpful.

Definition 2 (Differential entropy)**.** *The differential entropy $\mathcal{H}$ of a pdf $p(x)$ is given by*

$$\mathcal{H}(p) = -\int p(x)\log p(x)\,\mathrm{d}x = -\mathrm{E}_p[\log p(x)].$$

Lemma 2. *Let $p(x)$ be the pdf of a random variable x. Then if*

(i) p is a univariate normal distribution with mean μ and variance σ^2,

$$\mathcal{H}(p) = \frac{1}{2}(1 + \log 2\pi) + \frac{1}{2}\log \sigma^2$$

(ii) p is a d-dimensional normal distribution with mean μ and variance Σ,

$$\mathcal{H}(p) = \frac{d}{2}(1 + \log 2\pi) + \frac{1}{2}\log |\Sigma|$$

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

$$\begin{aligned}
\mathrm{E}\left[\log p(\mathbf{y}^*|\mathbf{f})\right] - \mathrm{E}\left[\log q(\mathbf{y}^*)\right] &= \sum_{i=1}^{n}\sum_{j=1}^{m} \mathrm{E}\left[\log p(y_{ij}^*|f_{ij})\right] + \sum_{i=1}^{n} \mathcal{H}\big(q(y_i^*)\big) \\
&= \cancel{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\mathrm{E}[y_{ij}^* - f_{ij}]^2\right)} \\
&\quad + \cancel{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{1}{2}\log 2\pi + \frac{1}{2}\mathrm{E}[y_{ij}^* - f_{ij}]^2\right)} + \sum_{i=1}^{n}\log C_i
\end{aligned}$$

3.4.2 Terms involving distributions of $\mathbf{w}$

$$\begin{aligned}
\mathrm{E}\left[\log p(\mathbf{w})\right] - \mathrm{E}\left[\log q(\mathbf{w})\right] &= \sum_{j=1}^{m}\Big(\mathrm{E}\left[\log p(\mathbf{w}_j)\right] - \mathrm{E}\left[\log q(\mathbf{w}_j)\right]\Big) \\
&= \sum_{j=1}^{m}\left(-\frac{n}{2}\log 2\pi - \frac{1}{2}\mathrm{E}[\mathbf{w}_j^\top \mathbf{w}_j] + \mathcal{H}\big(q(\mathbf{w}_j)\big)\right) \\
&= \sum_{j=1}^{m}\left(\cancel{-\frac{n}{2}\log 2\pi} - \frac{1}{2}\operatorname{tr}\left(\mathrm{E}[\mathbf{w}_j\mathbf{w}_j^\top]\right) + \frac{n}{2}(1 + \cancel{\log 2\pi}) - \frac{1}{2}\log|\mathbf{A}_j|\right) \\
&= \frac{nm}{2} - \frac{1}{2}\sum_{j=1}^{m}\Big(\operatorname{tr}\widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j|\Big)
\end{aligned}$$

3.4.3 Terms involving distribution of $q(\lambda)$

$$\begin{aligned}
-\mathrm{E}\left[\log q(\lambda)\right] &= \sum_{j=1}^{m}\mathcal{H}\big(q(\lambda_j)\big) \\
&= \sum_{j=1}^{m}\left(\frac{1}{2}(1 + \log 2\pi) - \frac{1}{2}\log c_j\right)
\end{aligned}$$

$$= \frac{m}{2}(1 + \log 2\pi) - \frac{1}{2}\sum_{j=1}^{m} \log c_j$$

or if using single λ

$$-\operatorname{E}[\log q(\lambda)] = \frac{1}{2}(1 + \log 2\pi) - \frac{1}{2}\log \sum_{j=1}^{m} c_j.$$

3.4.4 Terms involving distribution of $q(\alpha)$

$$\begin{aligned} -\operatorname{E}[\log q(\alpha)] &= \sum_{j=1}^{m} \mathcal{H}\big(q(\alpha_j)\big) \\ &= \frac{m}{2}(1 + \log 2\pi - \log n) \end{aligned}$$

or if using single α

$$-\operatorname{E}[\log q(\alpha)] = \frac{1}{2}(1 + \log 2\pi - \log nm).$$

3.4.5 The lower bound

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{n} \log C_i + \frac{nm}{2} - \frac{1}{2}\sum_{j=1}^{m} \Big(\operatorname{tr} \widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j|\Big) \\ &\quad + \frac{m}{2}(1 + \log 2\pi) - \frac{1}{2}\sum_{j=1}^{m} \log c_j + \frac{m}{2}(1 + \log 2\pi - \log n) \\ &= \frac{m}{2}\big(n + 2(1 + \log 2\pi) - \log n\big) - \frac{1}{2}\sum_{j=1}^{m} \Big(\operatorname{tr} \widetilde{\mathbf{W}}_j + \log|\mathbf{A}_j| + \log c_j\Big) + \sum_{i=1}^{n} \log C_i \end{aligned}$$

Of course, if using either single α or single λ, then the formula needs to be adjusted accordingly.

3.5 Prediction

For a new data point x_{new}, we calculate the predicted latent values $\tilde{f}_{\text{new}} = (\tilde{f}_{\text{new},1}, \ldots, \tilde{f}_{\text{new},m})$ for each of the classes, using the variational estimates of the posterior means for the unknown quantities (denoted with tildes), as follows:

$$\tilde{f}_{\text{new},j} = \tilde{\alpha}_j + \sum_{k=1}^{n} h_{\tilde{\lambda}_j}(x_{\text{new}}, x_k)\tilde{w}_{kj}, \qquad j = 1, \ldots, m.$$

The predicted class is equal to

$$y_{\text{new}} = \arg\max_j \tilde{f}_{\text{new},j}.$$

To get the fitted probabilities for each class, the following integrals needs to be computed:

$$\begin{aligned}\tilde{p}_{\text{new},j} &= \mathrm{E}_Z\left[\prod_{\substack{k=1\\k\neq j}}^{m}\Phi\left(Z+\tilde{f}_{\text{new},j}-\tilde{f}_{\text{new},k}\right)\right]\\ &= \int\prod_{\substack{k=1\\k\neq j}}^{m}\Phi\left(z+\tilde{f}_{\text{new},j}-\tilde{f}_{\text{new},k}\right)\phi(z)\,\mathrm{d}z\end{aligned}$$

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

3.6 The variational Bayes EM algorithm

6. This isn't the variational EM... CAVI?

Since there is a cyclic dependence of the parameters on each other, we employ a sequential update algorithm. In what follows, a tilde on the parameters indicate that these are the expectations of the parameters given the optimal factorised distributions $\tilde{q}$ derived earlier.

STEP 1: Update $\tilde{\mathbf{y}}^{*(t+1)}$ given $\tilde{\mathbf{w}}^{(t)}$, $\tilde{\lambda}^{(t)}$, and $\tilde{\alpha}^{(t)}$

STEP 2: Update $\tilde{\mathbf{w}}^{(t+1)}$ given $\tilde{\mathbf{y}}^{*(t+1)}$, $\tilde{\lambda}^{(t)}$, and $\tilde{\alpha}^{(t)}$

STEP 3: Update $\tilde{\lambda}^{(t+1)}$ given $\tilde{\mathbf{y}}^{*(t+1)}$, $\tilde{\mathbf{w}}^{(t+1)}$, and $\tilde{\alpha}^{(t)}$

STEP 4: Update $\tilde{\alpha}^{(t+1)}$ given $\tilde{\mathbf{y}}^{*(t+1)}$, $\tilde{\mathbf{w}}^{(t+1)}$, and $\tilde{\lambda}^{(t+1)}$

Algorithm 1 VB-EM algorithm for the probit I-prior model

procedure INITIALISE
  for $j = 1,\dots,m$ **do**
    $\tilde{\mathbf{w}}_j^{(0)} \leftarrow \mathbf{0}_n$
    $\tilde{\alpha}_j^{(0)} \leftarrow \mathrm{N}(0,1)$
    $\tilde{\lambda}_j^{(0)} \leftarrow \mathrm{N}(0,1)$
    $\tilde{\lambda}_j^{sq(0)} \leftarrow (\tilde{\lambda}_j^{(0)})^2$ ▷ this is $\mathrm{E}[\lambda_j^2]$
    $\mathbf{H}_{\lambda_j}^{(0)} \leftarrow \tilde{\lambda}_j^{(0)}\mathbf{H}$
    $\mathbf{H}_{\lambda_j}^{sq(0)} \leftarrow \tilde{\lambda}_j^{sq(0)}\mathbf{H}^2$
  end for
end procedure

procedure UPDATE FOR $\tilde{\mathbf{f}}$ (time t)
  for $j = 1,\dots,m$ **do**
    $\tilde{\mathbf{f}}_j^{(t+1)} \leftarrow \tilde{\alpha}_j^{(t)}\mathbf{1}_n + \mathbf{H}_{\lambda_j}\tilde{\mathbf{w}}_j^{(t)}$
  end for
  $\tilde{\mathbf{f}}^{(t+1)} \leftarrow \left(\tilde{\mathbf{f}}_1^{(t+1)},\dots,\tilde{\mathbf{f}}_m^{(t+1)}\right)^\top$
end procedure

procedure UPDATE FOR y_{ij}^* (time t)
  for $i = 1, \ldots, n$ **do**
    $j \leftarrow y_i$
    $C_i^{(t+1)} \leftarrow \prod_{k \neq j} \Phi\left((\tilde{f}_{ij}^{(t+1)} - \tilde{f}_{ik}^{(t+1)})/\sqrt{2}\right)$
    for $k = 1, \ldots, j-1, j+1, \ldots, m$ **do**
      $D_{ik} \leftarrow \mathrm{E}_Z\left[\phi_k(Z + \tilde{f}_{ij}^{(t+1)} - \tilde{f}_{ik}^{(t+1)}) \prod_{l \neq k,j} \Phi_l(Z + \tilde{f}_{ij}^{(t+1)} - \tilde{f}_{ik}^{(t+1)})\right]$
      $\tilde{y}_{ik}^{*(t+1)} \leftarrow \tilde{f}_{ik}^{(t+1)} - D_{ik}/C_i^{(t+1)}$
    end for
    $\tilde{y}_{ij}^{*(t+1)} \leftarrow \tilde{f}_{ij}^{(t+1)} - \sum_{k \neq j} \left(\tilde{y}_{ik}^{*(t+1)} - \tilde{f}_{ik}^{(t+1)}\right)$
  end for
end procedure

procedure UPDATE FOR $\mathbf{w}_j$ (time t)
  for $j = 1, \ldots, m$ **do**
    $\tilde{\mathbf{y}}_j^{*(t+1)} \leftarrow (\tilde{y}_{1j}^{(t+1)}, \ldots, \tilde{y}_{nj}^{(t+1)})^\top$
    $\mathbf{A}_j \leftarrow \mathbf{H}_{\lambda_j}^{sq(t)} + \mathbf{I}_n$
    $\mathbf{a}_j \leftarrow \mathbf{H}_\lambda(\tilde{\mathbf{y}}_j^{*(t+1)} - \tilde{\alpha}_j^{(t)} \mathbf{1}_n)$
    $\tilde{\mathbf{w}}_j^{(t+1)} \leftarrow \mathbf{A}_j^{-1} \mathbf{a}_j$
    $\widetilde{\mathbf{W}}_j^{(t+1)} \leftarrow \mathbf{A}_j^{-1}(\mathbf{I}_n + \mathbf{a}_j \mathbf{a}_j^\top \mathbf{A}_j^{-1})$
    $\mathrm{logdetA}_j^{(t+1)} \leftarrow \log|\mathbf{A}_j|$
  end for
end procedure

procedure UPDATE FOR λ (time t)
  for $j = 1, \ldots, m$ **do**
    $c_j^{(t+1)} \leftarrow \mathrm{tr}\left(\mathbf{H}^2 \widetilde{\mathbf{W}}_j\right)$
    $d_j \leftarrow (\tilde{\mathbf{y}}_j^{*(t+1)} - \tilde{\alpha}_j^{(t)} \mathbf{1}_n)^\top \mathbf{H} \tilde{\mathbf{w}}_j^{(t+1)}$
    $\tilde{\lambda}_j^{(t+1)} \leftarrow d_j / c_j^{(t+1)}$
    $\tilde{\lambda}_j^{sq(t+1)} \leftarrow 1/c_j^{(t)} + (d_i/c_i^{(t+1)})^2$
  end for
  if single λ **then** $\forall j$
    $\tilde{\lambda}_j^{(t+1)} \leftarrow \sum_j d_j \Big/ \sum_j c_j^{(t+1)}$
    $\tilde{\lambda}_j^{sq(t+1)} \leftarrow 1 \Big/ \sum_j c_j^{(t+1)} + \left(\sum_j d_j \Big/ \sum_j c_j^{(t+1)}\right)^2$
  end if
  call UPDATE KERNEL MATRICES
end procedure

procedure UPDATE KERNEL MATRICES (time t)
  for $j = 1, \ldots, m$ **do**
    $\mathbf{H}_{\lambda_j}^{(t+1)} \leftarrow \tilde{\lambda}_j^{(t+1)} \mathbf{H}$
    $\mathbf{H}_{\lambda_j}^{sq(t+1)} \leftarrow \tilde{\lambda}_j^{sq(t+1)} \mathbf{H}^2$
  end for
end procedure

procedure UPDATE FOR α (time t)
  if single α **then**
    $\tilde{\alpha}^{(t+1)} \leftarrow \frac{1}{nm} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\tilde{y}_{ij}^{*(t+1)} - \tilde{\lambda}_j^{(t+1)} \mathbf{H}_i \tilde{\mathbf{w}}_j^{(t+1)} \right)$
  else
    for $j = 1, \ldots, m$ **do**
      $\tilde{\alpha}_j^{(t+1)} \leftarrow \frac{1}{n} \sum_{i=1}^{n} \left(\tilde{y}_{ij}^{*(t+1)} - \tilde{\lambda}_j^{(t+1)} \mathbf{H}_i \tilde{\mathbf{w}}_j^{(t+1)} \right)$
    end for
  end if
end procedure

procedure CALCULATE LOWER BOUND (time t)
  $\mathcal{L}^{(t)} \leftarrow \frac{1}{2}\left(nm - \log nm + 3(1 + \log 2\pi)\right) - \frac{1}{2}\left(\log\det \mathrm{A}^{(t)} + \operatorname{tr} \widetilde{\mathbf{W}}^{(t)} + \sum_{i=1}^{2} \log c_i^{(t)} \right) + \sum_{i=1}^{n} \log C_i^{(t)}$
end procedure

procedure THE VB-EM ALGORITHM
  $t \leftarrow 0$
  while $\mathcal{L}^{(t+1)} - \mathcal{L}^{(t)} > \delta$ **or** $t < t_{max}$ **do**
    call UPDATE FOR $\mathbf{y}^*$
    call UPDATE FOR $\mathbf{w}$
    call UPDATE FOR λ
    call UPDATE FOR α
    call CALCULATE LOWER BOUND
    $t \leftarrow t + 1$
  end while
end procedure

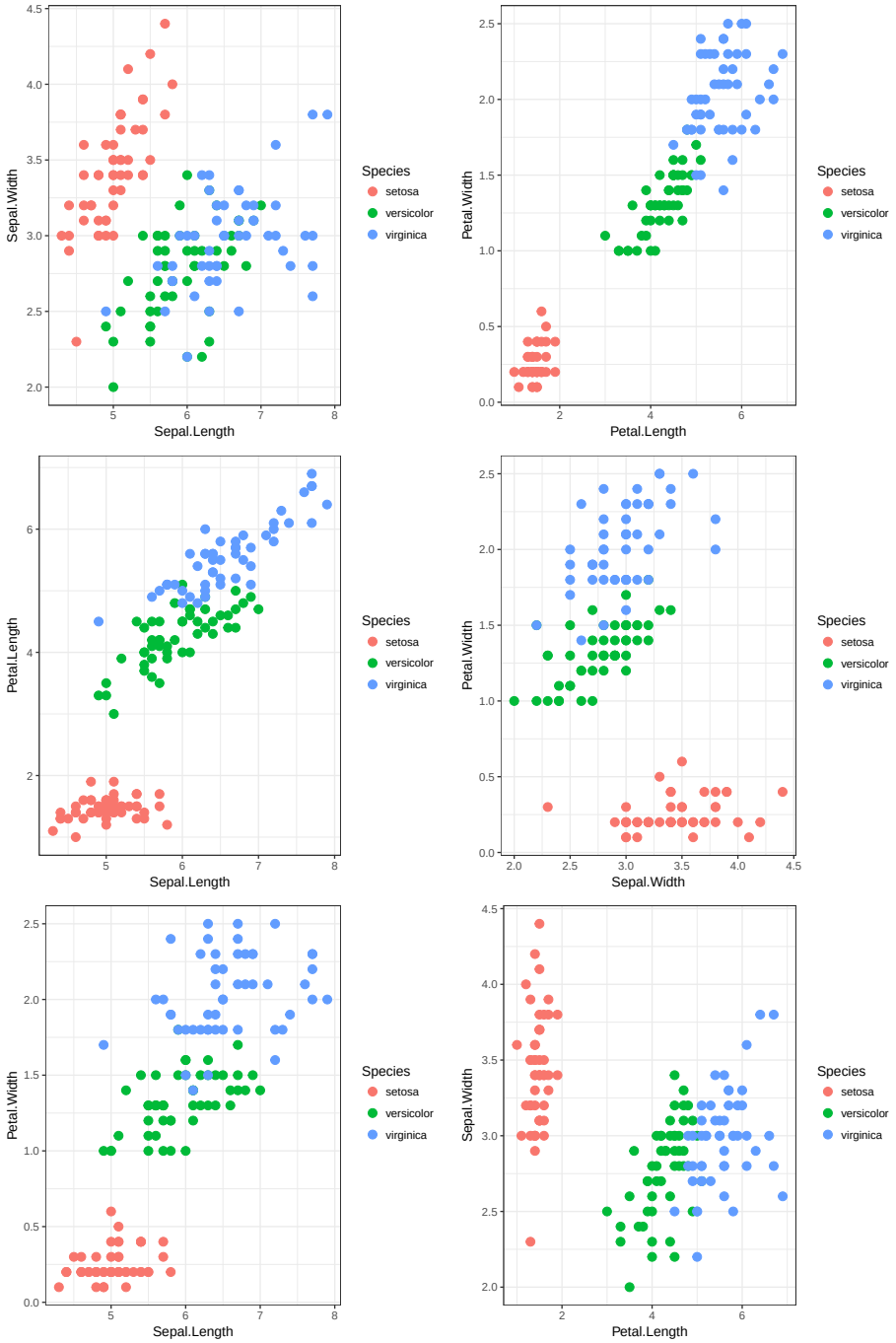
return $(\hat{\mathbf{y}}^*, \hat{\mathbf{w}}, \hat{\lambda}, \hat{\alpha}) \leftarrow (\tilde{\mathbf{y}}^{*(t)}, \tilde{\mathbf{w}}^{(t)}, \tilde{\lambda}^{(t)}, \tilde{\alpha}^{(t)})$ ▷ converged parameter estimates

return $(\hat{y}_1, \ldots, \hat{y}_n) \leftarrow \left(\arg\max_{k=1}^{m} \hat{y}_{1k}^*, \ldots, \arg\max_{k=1}^{m} \hat{y}_{nk}^* \right)$ ▷ predicted classes

for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, m$ **do**
    return $\hat{p}_{ij} \leftarrow \prod_{\substack{k=1 \\ k \neq j}}^{m} \Phi\left(\frac{\hat{y}_{ij}^* - \hat{y}_{ik}^*}{\sqrt{2}} \right)$ ▷ predicted probabilities
  end for
end for

4 Examples

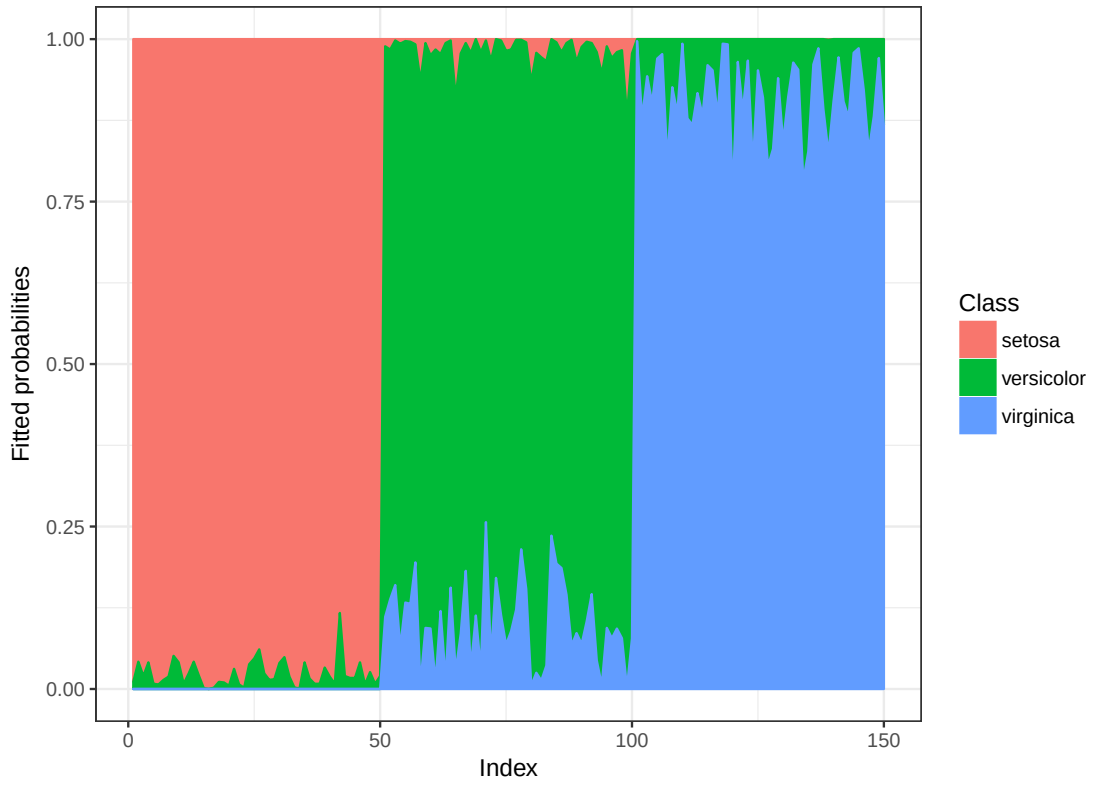
4.1 Iris data set



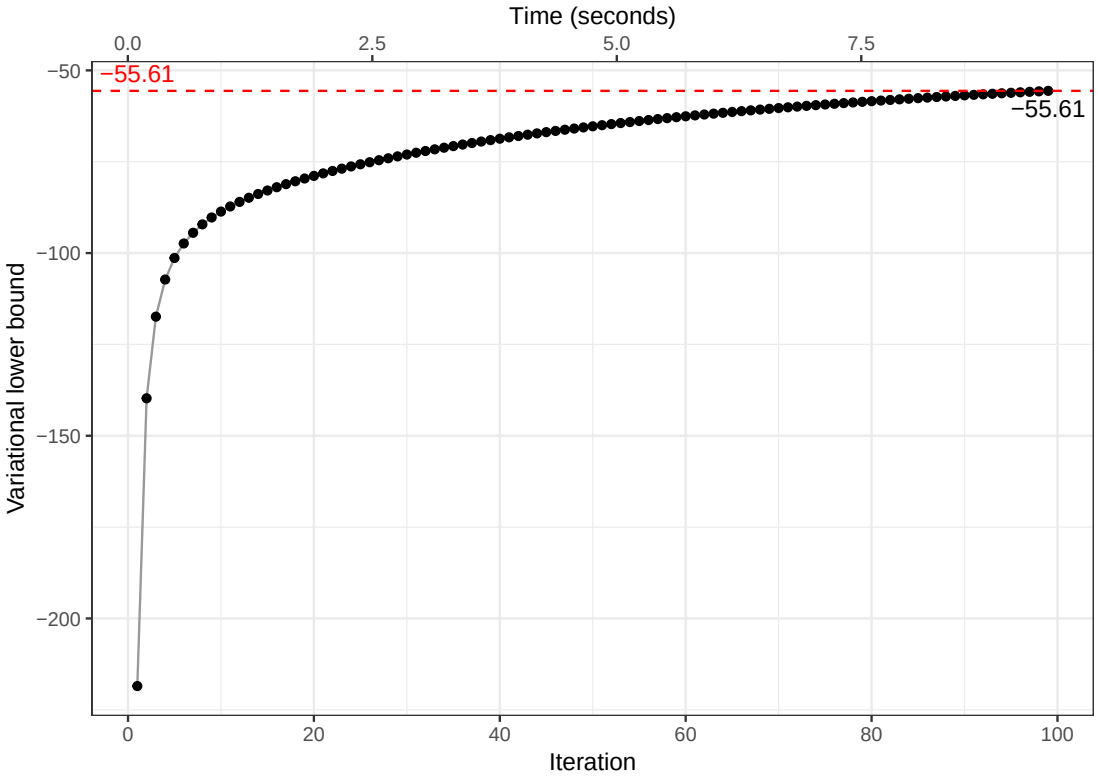
```
R> (mod <- iprobit_mult(y, X, silent = TRUE))

## Lower bound value =  -55.60715
## Iterations =  100
##
##           Class = 1 Class = 2 Class = 3
## Intercept  -0.21708   1.20009  -1.13551
## lambda     -0.35226  -0.35226  -0.35226

R> plot(mod)
```




```
R> iplot_lb(mod)
```



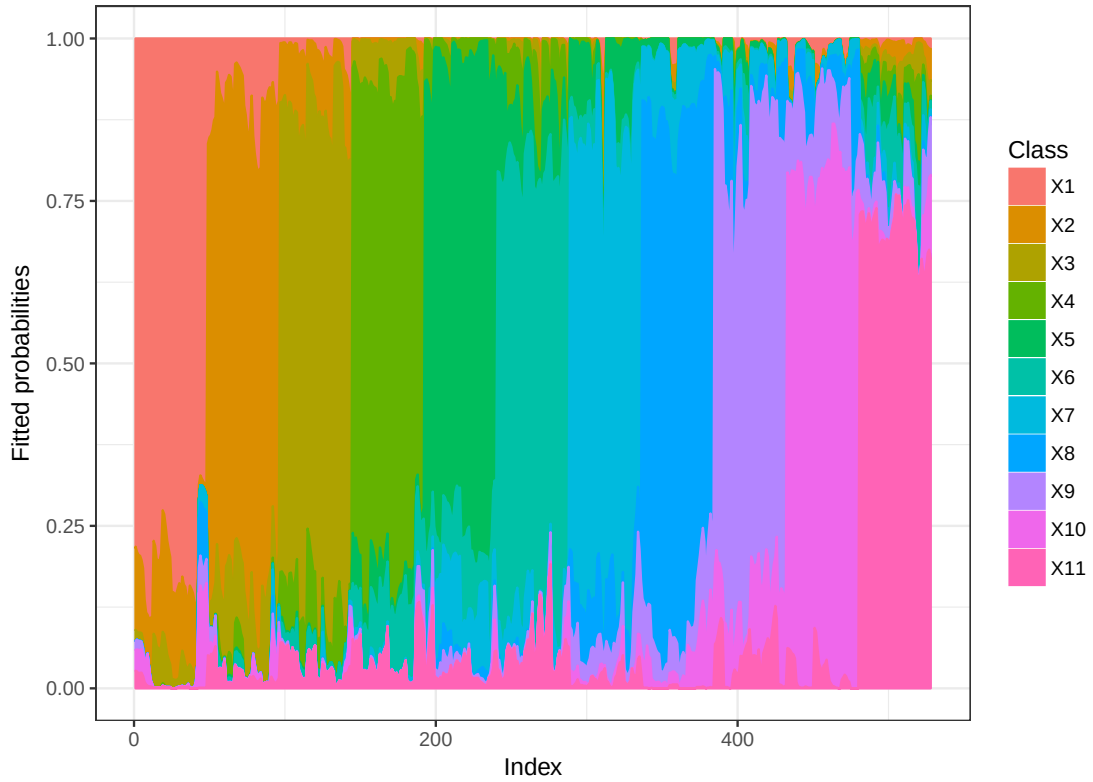
4.2 Vowel recognition data set

```
##   class    x.1   x.2    x.3   x.4    x.5   x.6    x.7    x.8    x.9   x.10
## 1     1 -3.639 0.418 -0.670 1.779 -0.168 1.627 -0.388  0.529 -0.874 -0.814
## 2     2 -3.327 0.496 -0.694 1.365 -0.265 1.933 -0.363  0.510 -0.621 -0.488
## 3     3 -2.120 0.894 -1.576 0.147 -0.707 1.559 -0.579  0.676 -0.809 -0.049
## 4     4 -2.287 1.809 -1.498 1.012 -1.053 1.060 -0.567  0.235 -0.091 -0.795
## 5     5 -2.598 1.938 -0.846 1.062 -1.633 0.764  0.394 -0.150  0.277 -0.396
## 6     6 -2.852 1.914 -0.755 0.825 -1.588 0.855  0.217 -0.246  0.238 -0.365
```

```
R> set.seed(123)
R> (mod <- iprobit_mult(vow.tr$class, vow.tr[, -1], kernel = "FBM", silent = TRUE))

## Lower bound value =  -736.8918
## Iterations =  100
##
##           Class = 1 Class = 2 Class = 3 Class = 4 Class = 5 Class = 6
## Intercept  -0.11514   0.13838   0.04304   0.07129   0.21767   0.46536
## lambda     -0.13430  -0.13430  -0.13430  -0.13430  -0.13430  -0.13430
##           Class = 7 Class = 8 Class = 9 Class = 10 Class = 11
## Intercept   0.40117   -0.3387   0.47458   -0.06605    0.67874
## lambda     -0.13430   -0.1343  -0.13430   -0.13430   -0.13430

R> plot(mod)
```




```
R> predict(mod, X.test = vow.ts[, -1], y.test = vow.ts[, 1])

## Test error rate: 41 %
```

	Error rates	
	Training	Test
k-Nearest neighbours	NA	44
Linear regression	48	67
Linear discriminant analysis	32	56
Neural network	NA	45
FDA/BRUTO	6	44
FDA/MARS	13	39
I-probit (FBM-0.5)	0	41

5 Discussion

5.1 Estimating a scaled probit model

For future work, can consider estimating the covariances/correlations across choices. More suitable for social science data.

5.2 Reduction to binary models

Describe the model when there are only two alternatives.

5.3 Multiple scale parameters with multiple kernels

It's definitely possible to extend to multiple scale parameters. It's just a matter of algebra.

5.4 Modelling ordinal data

A different model using thresholds, but it might be possible to model these and estimate using variational inference.

A Proofs

A.1 Proof of Lemma 1

Proof. (i) Due to the independence structure in the pdf of $\mathbf{X}$, it is easy to consider the expectations of each of the components separately and marginalising out the rest of the components. For $i \neq j$, we have

$$
\begin{aligned}
\mathrm{E}[x_i] &= C^{-1} \int \cdots \int \mathbb{1}[x_k < x_j, \forall k \neq j] \cdot x_i \prod_{k=1}^{d} \frac{1}{\sigma_k} \phi\left(\frac{x_k - \mu_k}{\sigma_k}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \iint \mathbb{1}[x_i < x_j] \frac{x_i}{\sigma_i} \phi\left(\frac{x_i - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_i \, \mathrm{d}x_j \\
&= C^{-1} \iint \mathbb{1}[\sigma_i z_i + \mu_i < \sigma_j z_j + \mu_j] (\sigma_i z_i + \mu_i) \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j \\
&= \mu_i C^{-1} \iint \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j \\
&\quad + \sigma_i C^{-1} \iint \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] z_i \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j \\
&= \mu_i C^{-1} \overbrace{\int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_j}^{C} \\
&\quad + \sigma_i C^{-1} \int \mathbb{1}[z_i < (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i] z_i \phi(z_i) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j) \, \mathrm{d}z_i \, \mathrm{d}z_j
\end{aligned}
$$

The integral involving z_i in the second part of the sum is recognised as the (unnormalised) expectation of the lower-tail of a univariate standard normal distribution truncated at

$\tau_{ij} = (\sigma_j z_j + \mu_j - \mu_i)/\sigma_i$. That is,

$$\mathrm{E}[Z_i | Z_i < \tau_{ij}] = \big[\Phi(\tau_{ij})\big]^{-1} \int \mathbb{1}[z_i < \tau_{ij}] z_i \phi(z_i)\,\mathrm{d}z_i = -\frac{\phi(\tau_{ij})}{\Phi(\tau_{ij})}$$

Plugging this expression back into the derivation of this expectation, we get

$$\begin{aligned}
\mathrm{E}[X_i] &= \mu_i - \sigma_i C^{-1} \int \phi\left(\frac{\sigma_j z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \phi(z_j)\,\mathrm{d}z_j \\
&= \mu_i - \sigma_i C^{-1}\,\mathrm{E}\left[\phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{k \neq i,j} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right)\right].
\end{aligned}$$

The expectation for the jth component is

$$\begin{aligned}
\mathrm{E}[X_j] &= C^{-1} \int \cdots \int \mathbb{1}[x_k < x_j, \forall k \neq j] \cdot x_j \prod_{k=1}^{d} \frac{1}{\sigma_k} \phi\left(\frac{x_k - \mu_k}{\sigma_k}\right) \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \int x_j \prod_{k \neq j} \Phi\left(\frac{x_j - \mu_k}{\sigma_k}\right) \cdot \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \mathrm{d}x_j \\
&= C^{-1} \int (\sigma_j z_j + \mu_j) \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot \phi(z_j)\,\mathrm{d}z_j \\
&= \mu_j C^{-1} \overbrace{\int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot \phi(z_j)\,\mathrm{d}z_j}^{C} \\
&\quad + \sigma_j C^{-1} \int \prod_{k \neq j} \Phi\left(\frac{\sigma_j z_j + \mu_j - \mu_k}{\sigma_k}\right) \cdot z_j \phi(z_j)\,\mathrm{d}z_j \\
&= \mu_j + \sigma_j C^{-1}\,\mathrm{E}\left[Z_j \prod_{k \neq j} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right)\right] \\
&= \mu_j + \sigma_j \sum_{\substack{i=1 \\ i \neq j}}^{d} \sigma_i C^{-1}\,\mathrm{E}\left[\phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_i}{\sigma_i}\right) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi\left(\frac{\sigma_j Z_j + \mu_j - \mu_k}{\sigma_k}\right)\right] \\
&= \mu_j - \sigma_j \sum_{i \neq j} \big(\mathrm{E}[X_i] - \mu_i\big)
\end{aligned}$$

where we have made use of Lemma 3 in the second last step of the above.

(ii) The differential entropy is given by

$$\begin{aligned}
\mathcal{H}(p) &= -\int p(\mathbf{x}) \log p(\mathbf{x})\,\mathrm{d}\mathbf{x} = -\,\mathrm{E}\left[\log p(\mathbf{x})\right] \\
&= -\,\mathrm{E}\left[-\log C - \frac{d}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{d} \log \sigma_i^2 - \frac{1}{2}\sum_{i=1}^{d} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2\right]
\end{aligned}$$

$$= \log C + \frac{d}{2}\log 2\pi + \frac{1}{2}\sum_{i=1}^{d}\log\sigma_i^2 + \frac{1}{2}\sum_{i=1}^{d}\frac{1}{\sigma_i^2}\,\mathrm{E}[x_i-\mu_i]^2.$$

□

Lemma 3. *Let $Z \sim \mathrm{N}(0,1)$. Then for all $m \in \{\mathbb{N} \,|\, m > 1\}$ and $(\mu,\sigma) \in \mathbb{R}\times\mathbb{R}^+$,*

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

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

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

We adopt the following notation for convenience:

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

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

$$g_2(z) = \Phi_2$$
$$\Rightarrow g_2'(z) = \sigma_2\phi_2 = \sum_{\substack{i=1\\i\neq 1}}^{2}\left[\sigma_i\phi_i\sum_{\substack{k=1\\k\neq 1,2}}^{2}\Phi_k\right].$$

Now assume that the inductive hypothesis holds for some $m \in \{\mathbb{N} \,|\, m > 1\}$. That is, the derivative of

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

which is

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

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

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

is found to be

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

Thus, by induction and linearity of expectations, the proof is complete. □

References

Albert, J. and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data". In: *Journal of the American Statistical Association* 88.422, pp. 669–679. ISSN: 0162-1459. DOI: 10.2307/2290350.

Bolduc, D. (1999). "A practical technique to estimate multinomial probit models in transportation". In: *Transportation Research Part B: Methodological* 33.1, pp. 63–79. ISSN: 01912615. DOI: 10.1016/S0191-2615(98)00028-9.

Dow, J. K. and J. W. Endersby (2004). "Multinomial probit and multinomial logit: A comparison of choice models for voting research". In: *Electoral Studies* 23.1, pp. 107–122. ISSN: 02613794. DOI: 10.1016/S0261-3794(03)00040-4.

Girolami, M. and S. Rogers (2006). "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors". In: *Neural Computation* 18.8, pp. 1790–1817. ISSN: 0899-7667. DOI: 10.1093/bioinformatics/btm535.

Natarajan, R., C. E. McCulloch, and N. M. Kiefer (1995). "Maximum likelihood for the multinomial probit model". In: *Unpublished manuscript*.

Rogers, S. and M. Girolami (2007). "Multi-class Semi-supervised Learning With The ϵ-truncated Multinomial Probit Gaussian Process". In: *JLMR: Workshop and Conference Proceedings* 2006, pp. 17–32.

Sheffi, Y., R. Hall, and C. Daganzo (1982). "On the estimation of the multinomial probit model". In: *Transportation Research Part A: General* 16.5-6, pp. 447–456. ISSN: 01912607. DOI: 10.1016/0191-2607(82)90071-1.