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 1

Preceding chapters

Appendix A

Functional derivative of the entropy

apx:funcder

We present the functional derivative of the entropy $H(p)$ in equation 3.6 (p. 17). Typically, this is tackled using calculus of variations, but it can also be obtained using the Fréchet and Gâteaux differentials. Both methods are presented.

A.1 The usual functional derivative

The functional derivative is defined as follows.

Definition A.1 (Functional derivative)**.** Given a manifold M representing continuous/smooth functions ρ with certain boundary conditions, and a functional $F : M \to \mathbb{R}$, the functional derivative of $F(\rho)$ with respect to ρ, denoted $\partial F/\partial \rho$, is defined by

$$\begin{aligned}\int \frac{\partial F}{\partial \rho}(x)\phi(x)\,\mathrm{d}x &= \lim_{\epsilon\to 0}\frac{F(\rho+\epsilon\phi)-F(\rho)}{\epsilon}\\ &= \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon}F(\rho+\epsilon\phi)\right]_{\epsilon=0},\end{aligned}$$

where ϕ is an arbitrary function. The function $\partial F/\partial\rho$ as the gradient of F at the point ρ, and

$$\partial F(\rho,\phi) = \int \frac{\partial F}{\partial \rho}(x)\phi(x)\,\mathrm{d}x$$

as the directional derivative at point ρ in the direction of ϕ. Analogous to vector calculus, the inner product with the gradient gives the directional derivative.

Now let X be a discrete random variable with probability mass function $p(x) \geq 0$, for $\forall x \in \Omega$, a finite set. The entropy is a functional of p, namely

$$H(p) = -\sum_{x \in \Omega} p(x) \log p(x).$$

Equivalently, using the counting measure ν on Ω, we can write

$$H(p) = -\int_{\Omega} p(x) \log p(x) \, \mathrm{d}\nu(x).$$

Using the definition of functional derivatives, we find that

$$\begin{aligned}
\int_{\Omega} \frac{\partial H}{\partial p}(x)\phi(x) \, \mathrm{d}x &= \left[\frac{\mathrm{d}}{\mathrm{d}\epsilon} H(p + \epsilon\phi) \right]_{\epsilon=0} \\
&= \left[-\frac{\mathrm{d}}{\mathrm{d}\epsilon} \big(p(x) + \epsilon\phi(x)\big) \log \big(p(x) + \epsilon\phi(x)\big) \right]_{\epsilon=0} \\
&= -\int_{\Omega} \left(\frac{p(x)\phi(x)}{p(x) + \epsilon\phi(x)} + \frac{\epsilon\phi(x)}{p(x) + \epsilon\phi(x)} + \phi(x) \log \big(p(x) + \epsilon\phi(x)\big) \right) \mathrm{d}x \\
&= -\int_{\Omega} \left(1 + \log p(x)\right) \phi(x) \, \mathrm{d}x.
\end{aligned}$$

Thus, $(\partial H / \partial p)(x) = -1 - \log p(x)$.

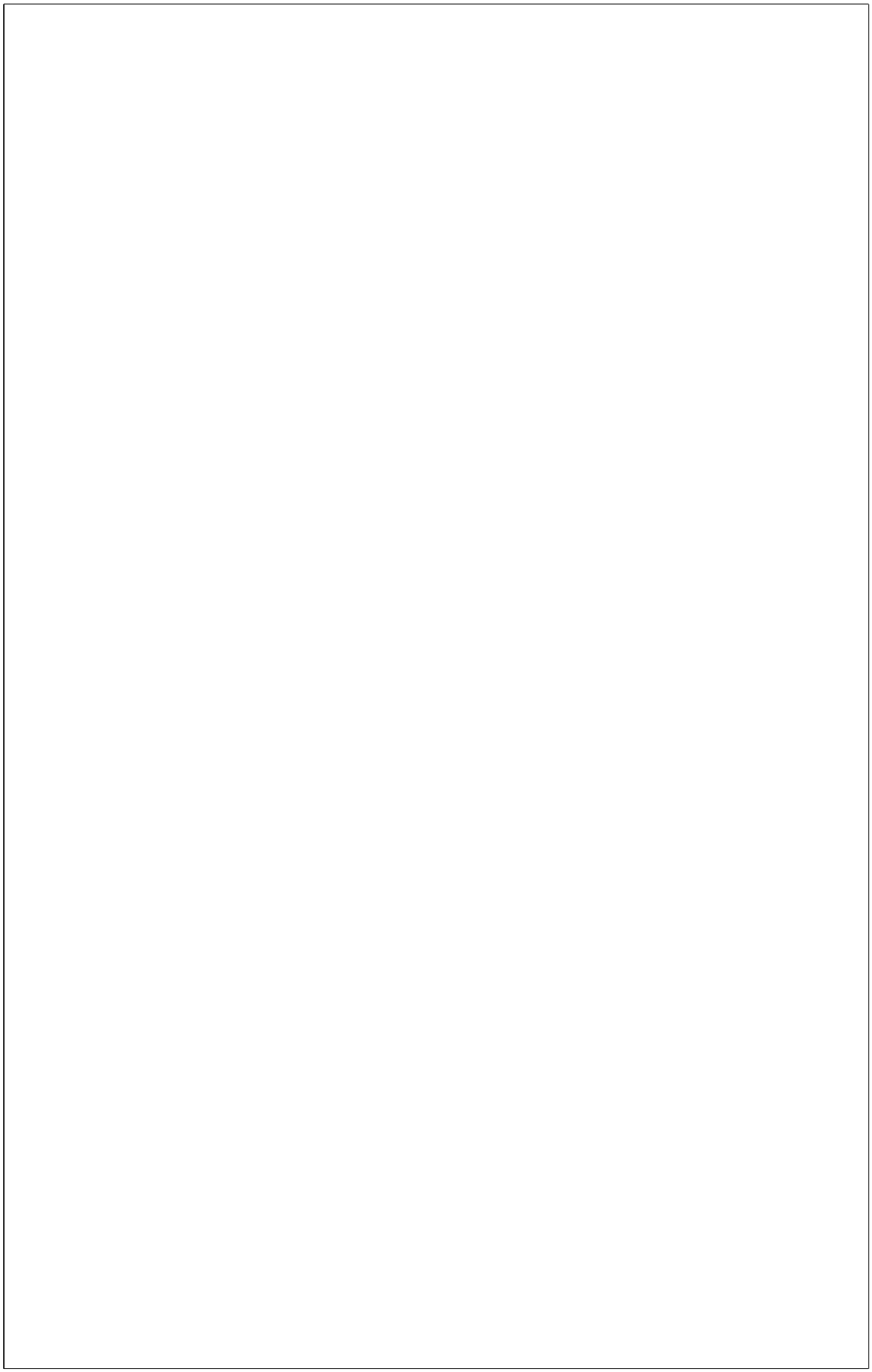
A.2 Fréchet differential of the entropy

Since we have already introduced concepts of Fréchet and Gâteaux derivatives earlier, we shall use those instead. Assume that the entropy H is Fréchet differentiable at p, and that the probability densities p under consideration belong to the Hilbert space of square integrable functions $\mathrm{L}^2(\Theta, \nu)$ with inner product $\langle p, p' \rangle_{L^2(\Theta,\nu)} = \int pp' \, \mathrm{d}\nu$. Now since the Fréchet derivative of H at p is assumed to exist, it is equal to the Gâteaux

derivative, which can be computed as follows:

$$
\begin{aligned}
\partial_q H(p) &= \frac{\mathrm{d}}{\mathrm{d}t} H(p + tq)\bigg|_{t=0} \\
&= \frac{\mathrm{d}}{\mathrm{d}t} \left\{ -\int_\Theta \big(p(\theta) + tq(\theta)\big) \log\big(p(\theta) + tq(\theta)\big)\, \mathrm{d}\nu(\theta) \right\} \bigg|_{t=0} \\
&= -\int_\Theta \left\{ \frac{\mathrm{d}}{\mathrm{d}t} \big(p(\theta) + tq(\theta)\big) \log\big(p(\theta) + tq(\theta)\big) \bigg|_{t=0} \right\} \mathrm{d}\nu(\theta) \\
&= -\int_\Theta \left(\frac{p(\theta)q(\theta)}{p(\theta) + tq(\theta)} + \frac{tq(\theta)^2}{p(\theta) + tq(\theta)} + q(\theta) \log\big(p(\theta) + tq(\theta)\big) \right) \bigg|_{t=0} \mathrm{d}\nu(\theta) \\
&= -\int_\Theta q(\theta)\big(1 + \log p(\theta)\big)\, \mathrm{d}\nu(\theta) \\
&= \big\langle -\big(1 + \log p\big), q \big\rangle_\Theta \\
&= \mathrm{d}H(p)(q).
\end{aligned}
$$

By definition, the gradient of H at p, denoted $\nabla H(p)$, is equal to $-1 - \log p$. This agrees with the usual functional derivative of the entropy obtained via standard calculus of variations.



Appendix B

Kronecker product and vectorisation

The Kronecker product crops up in the definition of matrix normal distributions, which is used in Chapter 5 for the I-probit model.

def:kroneckerprod

Definition B.1 (Kronecker product)**.** The Kronecker matrix product, denoted by $\otimes$, for two matrices $A \in \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{p\times q}$ is defined by

$$A \otimes B = \begin{pmatrix} A_{11}B & A_{12}B & \cdots & A_{1m}B \\ A_{21}B & A_{22}B & \cdots & A_{2m}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1}B & A_{n2}B & \cdots & A_{nm}B \end{pmatrix} \in \mathbb{R}^{np\times mq}.$$

The Kronecker product is a generalisation of the outer product for vectors to matrices. Of use will be these properties of the Kronecker product (Zhang and Ding, 2013):

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

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

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

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

The equivalence between matrix normal and multivariate normal distributions are established making use of vectorisation for matrices. This is defined below.

def:vectorisation

Definition B.2 (Vectorisation)**.** The vectorisation operation 'vec' stacks the columns of the matrices into one long vector, for instance, for the matrix $A \in \mathbb{R}^{n \times m}$

$$\operatorname{vec} A = (A_{11}, \ldots, A_{n1}, A_{12}, \ldots, A_{n2}, \ldots, A_{1m}, \ldots, A_{nm})^\top \in \mathbb{R}^{nm}.$$

Appendix C

Statistical distributions and their properties

This is a reference relating to the multivariate normal, matrix normal, truncated univariate and multivariate normal, which are collated from various sources for convenience. Of interest are probability density functions, first and second moments, and entropy (Definition 3.5, page 17). Note that in this part of the appendix, boldface notation for matrix and vectors are not used.

C.1 Multivariate normal distribution

apx:fishermultinormal

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

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

For $d = 1$, i.e. X is univariate, then its pdf is $p(X|\mu, \sigma^2) = \frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)$, and its cdf is $F(X|\mu, \sigma^2) = \Phi\left(\frac{X-\mu}{\sigma}\right)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of a univariate standard normal distribution. In the special case that $\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, then the components of $X = (X_1, \ldots, X_d)^\top$ are independently distributed according to $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$.

Lemma C.1 (Properties of multivariate normal)**.** *Assume that* $X \sim \mathrm{N}_d(\mu, \Sigma)$ *and* $Y \sim \mathrm{N}_d(\nu, \Psi)$*, where*

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

Then,

- ***Marginal distributions.***

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

- ***Conditional distributions.***

$$X_a | X_b \sim \mathrm{N}_{\dim X_a}(\tilde{\mu}_a, \tilde{\Sigma}_a) \quad \textit{and} \quad X_b \sim \mathrm{N}_{\dim X_b}(\tilde{\mu}_b, \tilde{\Sigma}_b),$$

  where

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

- ***Linear combinations.***

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

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

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

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

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

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

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

Lemma C.2. *Let $x, b \in \mathbb{R}^d$ be a vector, $X, B \in \mathbb{R}^{n\times d}$ a matrix, and $A \in \mathbb{R}^{d\times d}$ a symmetric, invertible matrix. Then,*

$$-\frac{1}{2}x^\top Ax + b^\top x = -\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) + \frac{1}{2}b^\top A^{-1}b$$
$$-\frac{1}{2}\operatorname{tr}(X^\top AX) + \operatorname{tr}(B^\top X) = -\frac{1}{2}\operatorname{tr}\left((X - A^{-1}B)^\top A(X - A^{-1}B)\right) + \frac{1}{2}\operatorname{tr}(B^\top A^{-1}B).$$

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

Lemma C.3. *Let $X \sim \mathrm{N}_p(\mu_\theta, \Sigma_\theta)$, that is, the mean vector μ_θ and covariance matrix Σ_θ depends on a real, q-dimensional vector θ. The Fisher information matrix $U \in \mathbb{R}^{q\times q}$ for θ has (i, j) entries given by*

$$U_{ij} = \frac{\partial \mu_\theta^\top}{\partial \theta_i}\Sigma_\theta^{-1}\frac{\partial \mu_\theta}{\partial \theta_j} + \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_i}\Sigma_\theta^{-1}\frac{\partial \Sigma_\theta}{\partial \theta_j}\right) \tag{C.1}$$

for $i, j = 1, \ldots, q$.

Proof. Define the derivative of a matrix $\Sigma \in \mathbb{R}^{p\times p}$ with respect to a scalar z, denoted $\partial\Sigma/\partial z \in \mathbb{R}^{p\times p}$, by $(\partial\Sigma/\partial z)_{ij} = \partial\Sigma_{ij}/\partial z$, i.e. derivatives are taken element-wise. The two identities below are useful:

$$\frac{\partial}{\partial z}\operatorname{tr}\Sigma = \operatorname{tr}\frac{\partial\Sigma}{\partial z} \tag{C.2}$$
$$\frac{\partial}{\partial z}\log|\Sigma| = \operatorname{tr}\left(\Sigma^{-1}\frac{\partial\Sigma}{\partial z}\right) \tag{C.3}$$
$$\frac{\partial\Sigma^{-1}}{\partial z} = -\Sigma^{-1}\frac{\partial\Sigma}{\partial z}\Sigma^{-1} \tag{C.4}$$

A useful reference for these identities is Petersen and Pedersen (2008).

Differentiating the log-likelihood for θ with respect to the i'th component of θ yields

$$\begin{aligned}
\frac{\partial}{\partial\theta_i}L(\theta|X) &= -\frac{1}{2}\frac{\partial}{\partial\theta_i}\log|\Sigma_\theta| - \frac{1}{2}\frac{\partial}{\partial\theta_i}\operatorname{tr}(\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top) \\
&= -\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\right) - \frac{1}{2}\operatorname{tr}\left(\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_i}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\quad - \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial}{\partial\theta_i}\left((X-\mu_\theta)(X-\mu_\theta)^\top\right)\right) \\
&= \overbrace{-\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\right)}^{(A)} \overbrace{-\frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right)}^{(B)} \\
&\quad \overbrace{+\operatorname{tr}\left(\Sigma_\theta^{-1}(X-\mu_\theta)\frac{\partial\mu_\theta^\top}{\partial\theta_i}\right)}^{(C)}.
\end{aligned}$$

Taking derivatives again, this time with respect to θ_j, of the three parts (A), (B) and (C) above, we get:

- **(A)**

$$\frac{1}{2}\frac{\partial}{\partial\theta_j}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\right) = \frac{1}{2}\operatorname{tr}\left(\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i} + \Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}\right)$$

- **(B)**

$$\begin{aligned}
&\frac{1}{2}\frac{\partial}{\partial\theta_j}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\quad= \frac{1}{2}\operatorname{tr}\left(\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\qquad+ \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}\Sigma_\theta^{-1}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\qquad+ \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}(X-\mu_\theta)(X-\mu_\theta)^\top\right) \\
&\qquad- \operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}\frac{\partial\mu_\theta}{\partial\theta_j}(X-\mu_\theta)^\top\right)
\end{aligned}$$

- **(C)**

$$\begin{aligned}
\frac{\partial}{\partial\theta_j}\operatorname{tr}\left(\Sigma_\theta^{-1}(X-\mu_\theta)\frac{\partial\mu_\theta^\top}{\partial\theta_i}\right) = \operatorname{tr}\Bigg(&\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}(X-\mu_\theta)\frac{\partial\mu_\theta^\top}{\partial\theta_i} - \Sigma_\theta^{-1}\frac{\partial\mu_\theta}{\partial\theta_j}\frac{\partial\mu_\theta^\top}{\partial\theta_i} \\
&- \Sigma_\theta^{-1}(X-\mu_\theta)\frac{\partial^2\mu_\theta}{\partial\theta_i\partial\theta_j}\Bigg)
\end{aligned}$$

The Fisher information matrix U contains (i,j) entries equal to the expectation of $-\frac{\partial^2}{\partial\theta_i\theta_j}L(\theta|X)$. Using the fact that 1) $\mathrm{E}[X-\mu_\theta]=0$; 2) $\mathrm{E}[\operatorname{tr}\Sigma]=\operatorname{tr}(\mathrm{E}\,\Sigma)$; 3) $\mathrm{E}[XX^\top]=\Sigma_\theta$; and 4) the trace is invariant under cyclic permutations, we get

$$\begin{aligned}
U_{ij} &= \operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\mu_\theta}{\partial\theta_j}\frac{\partial\mu_\theta^\top}{\partial\theta_i}\right) \\
&\quad+ \frac{1}{2}\operatorname{tr}\left(\cancel{\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}} + \cancel{\Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}} - \cancel{\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\frac{\partial\Sigma_\theta}{\partial\theta_i}} - \cancel{\Sigma_\theta^{-1}\frac{\partial^2\Sigma_\theta}{\partial\theta_i\theta_j}} - \frac{\partial\Sigma_\theta}{\partial\theta_i}\frac{\partial\Sigma_\theta^{-1}}{\partial\theta_j}\right) \\
&= \frac{\partial\mu_\theta^\top}{\partial\theta_i}\Sigma_\theta^{-1}\frac{\partial\mu_\theta}{\partial\theta_j} + \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_i}\Sigma_\theta^{-1}\frac{\partial\Sigma_\theta}{\partial\theta_j}\right)
\end{aligned}$$

as required. ■

C.2 Matrix normal distribution

apx:matrixnormal

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

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

Lemma C.4 (Equivalence between matrix and multivariate normal)**.** $X \sim \mathrm{MN}_{n,m}(\mu,\Sigma,\Psi)$ *if and only if* $\operatorname{vec} X \sim \mathrm{N}_{nm}(\operatorname{vec}\mu, \Psi\otimes\Sigma)$.

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

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

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

Some useful properties of the matrix normal distribution are listed:

- **Expected values**.

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

- **Transpose**. $X^\top \sim \mathrm{MN}_{m,n}(\mu^\top, \Psi, \Sigma)$.

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

C.3 Truncated univariate normal distribution

apx:truncun inorm

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

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

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

- **Entropy**.

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

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

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

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

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

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

C.4 Truncated multivariate normal distribution

apx:truncmultinorm

Consider the restriction of $X \sim \mathrm{N}_d(\mu, \Sigma)$ to a convex subset[1] $\mathcal{A} \subset \mathbb{R}^d$. Call this distribution the truncated multivariate normal distribution, and denote it $X \sim {}^{\mathrm{t}}\mathrm{N}_d(\mu, \Sigma, \mathcal{A})$. The pdf is $p(X|\mu, \Sigma, \mathcal{A}) = C^{-1}\phi(X|\mu, \Sigma)\,\mathbb{1}[X \in \mathcal{A}]$, where

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

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

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

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

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

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

distribution, the full conditionals of X_j given X_{-j} is

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

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

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

which means that we need only compute one global inverse Σ^{-1}. Therefore, the Gibbs sampler makes draws from truncated normal distributions in the following sequence, given initial values $X^{(0)}$:

- Draw $X_1^{(t)}|X_2^{(t)}, \ldots, X_d^{(t)} \sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_1, \tilde{\sigma}_1^2, x_1^-, x_1^+)$.
- Draw $X_2^{(t)}|X_1^{(t+1)}, X_3^{(t)}, \ldots, X_d^{(t)} \sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_2, \tilde{\sigma}_1^2, x_2^-, x_2^+)$.
- $\cdots$
- Draw $X_d^{(t)}|X_1^{(t+1)}, \ldots, X_{d-1}^{(t+1)} \sim {}^{\mathrm{t}}\mathrm{N}(\tilde{\mu}_d, \tilde{\sigma}_d^2, x_d^-, x_d^+)$.

In a later work, Damien and Walker (2001) introduce a latent variable $Y \in \mathbb{R}$ such that the joint pdf of X and Y is

$$p(X_1, \ldots, X_d, Y) \propto \exp(-Y/2)\, \mathbb{1}\left(Y > (X-\mu)^\top \Sigma^{-1}(X-\mu)\right) \mathbb{1}(X \in \mathcal{A}).$$

Now, the Gibbs conditional densities for the X_k's are given by

$$p(X_j|X_{-j}, Y) \propto \mathbb{1}(X_j \in \mathcal{B}_j)$$

where

$$\mathcal{B}_j \in (x_j^-, x_j^+) \cap \{X_j \,|\, (X-\mu)^\top \Sigma^{-1}(X-\mu) < Y\}.$$

Thus, given values for X_{-j} and Y, the bounds for X_j involves solving a quadratic equation in X_j. The Gibbs conditional density for $Y|X$ is a shifted exponential distribution, which can be sampled using the inverse-transform method. Thus, both X and Y can be sampled directly from uniform variates.

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

In the case where Σ is a diagonal matrix, the conically truncated multivariate normal distributions are easier to deal with due to the independence structure in the covariance matrix. In particular, most calculations of interest involve only a one dimensional inte-

gral of products of normal cdfs. We present some results that we have not previously seen before elsewhere.

thm:contruncn

Lemma C.5. *Let $X \sim {}^{\text{t}}\text{N}_d(\mu, \Sigma, \mathcal{C}_j)$, with $\mu = (\mu_1, \ldots, \mu_d)^\top$ and $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$, and $\mathcal{C}_j = \{X_j > X_k | k \neq j, \text{and } k = 1, \ldots, m\}$ a conical truncation of $\mathbb{R}^d$ such that the j'th component is largest. Then,*

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

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

where ϕ is the pdf of a standard normal distribution and

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

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

(ii) ***Moments****. The expectation $\text{E}\, X = \left(\text{E}\, X_1, \ldots, \text{E}\, X_d \right)^\top$ is given by*

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

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

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

where we had defined

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

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

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

Proof. See Appendix D for the proof. ■

C.5 Gamma distribution

apx:gammadist

For $X \in \mathbb{R}_{\geq 0}$, let X be distributed according to the gamma distribution with shape s and rate r, denoted $X \sim \Gamma(s, r)$. Then,

- **Pdf**. $p(X) = \Gamma(s)^{-1} r^s X^{s-1} e^{-rX}$.
- **Moments**. $\mathrm{E}\, X = s/r$, $\mathrm{Var}\, X = s/r^2$.
- **Entropy**. $H(p) = s - \log r + \log \Gamma(s) + (1-s)\psi(s)$.

In the above, $\Gamma(\cdot) = \Gamma_1(\cdot)$ and $\psi(\cdot) = \psi_1(\cdot)$ are the gamma and digamma functions, defined by

$$\Gamma(a) = \begin{cases} (a-1)! & \text{if } a \in \mathbb{Z}^+ \\ \int_0^\infty u^{a-1} e^{-u}\, \mathrm{d}u & \text{otherwise} \end{cases}$$

and

$$\psi(a) = \frac{\partial}{\partial a} \log \Gamma(a) = \frac{\partial \Gamma(a)/\partial a}{\Gamma(a)}.$$

Often, the gamma distribution is parameterised according to shape s and scale $\sigma = 1/r$ parameters, $X \sim \Gamma(s, \sigma)$.

C.6 Inverse gamma distribution

def:invgam

For $X \in \mathbb{R}_{\geq 0}$, a random variable X distributed according to an inverse gamma distribution with parameters s (shape) and σ (scale) is denoted by $X \sim \Gamma^{-1}(s, \sigma)$. Then,

- **Pdf**. $p(X) = \Gamma(s)^{-1} \sigma^s X^{-(s+1)} e^{-\sigma/X}$.
- **Moments**. $\mathrm{E}\, X = \sigma/(s-1)$, $\mathrm{Var}\, X = \sigma^2 \big((s-1)^2(s-2)\big)^{-1}$.
- **Entropy**. $H(p) = s + \log\big(\sigma \Gamma(s)\big) - (1+s)\psi(s)$.

with $\Gamma(\cdot)$ and $\psi(\cdot)$ representing the gamma and digamma functions respectively, as defined in Appendix C.5.

Lemma C.6. *If X has a Gamma distribution with shape and rate s and r, then $1/X \sim \Gamma^{-1}(s, r)$.*

Proof. Let $Y = 1/X$. Then the pdf of Y is

$$\begin{aligned} p_Y(Y) &= p_X(1/Y) \left| \frac{\partial}{\partial Y}(1/Y) \right| \\ &= \Gamma(s)^{-1} r^s (1/Y)^{s-1} e^{-r/Y} (1/Y^2) \\ &= \Gamma(s)^{-1} r^s Y^{-(s+1)} e^{-r/Y} \end{aligned}$$

which is the pdf of an inverse gamma with shape s and scale r. ∎

Appendix D

Proofs related to the conically truncated independent multivariate normal distribution

apx:contrun proof

We present the proof for Lemma C.5 related to the conically truncated independent multivariate normal distribution, which we had not encountered in the literature.

D.1 Proof of Lemma C.5: Pdf

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

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

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

the proof follows directly.

D.2 Proof of Lemma C.5: Moments

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

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

and in particular, we have

$$\mathrm{E}[Y - \mu] = -\sigma \frac{\phi(\beta)}{\Phi(\beta)} \tag{D.1}$$ {eq:tnXminMu}

$$\mathrm{E}[Y - \mu]^2 - \sigma^2 = -\sigma^2 \frac{\beta \phi(\beta)}{\Phi(\beta)} \tag{D.2}$$ {eq:tnXminMusq}

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

$$\begin{aligned}
\mathrm{E}\, g(X_i) &= C^{-1} \int \cdots \int [x_k < x_j, \forall k \neq j] \cdot g(x_i) \prod_{k=1}^{d} p(x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \frac{\Phi((x_j - \mu_j)/\sigma_j)}{\Phi((x_j - \mu_j)/\sigma_j)} \iint [x_i < x_j] \cdot g(x_i) p(x_i) p(x_j) \prod_{\substack{k=1 \\ k \neq i,j}}^{d} \Phi \left(\frac{x_j - \mu_k}{\sigma_k} \right) \mathrm{d}x_i\, \mathrm{d}x_j \\
&= C^{-1} \int \mathrm{E}_{X_i \sim {}^{\mathrm{t}}\mathrm{N}(\mu_i, \sigma_i^2, -\infty, x_j)} [g(X_i)] \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi \left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_j
\end{aligned} \tag{D.3}$$ {eq:tnproofi}

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

$$\begin{aligned}
\mathrm{E}\, g(X_j) &= C^{-1} \int \cdots \int [x_k < x_j, \forall k \neq j] \cdot g(x_j) \prod_{k=1}^{d} p(x_k)\, \mathrm{d}x_1 \cdots \mathrm{d}x_d \\
&= C^{-1} \int g(x_j) \prod_{\substack{k=1 \\ k \neq j}}^{d} \Phi \left(\frac{x_j - \mu_k}{\sigma_k} \right) p(x_j)\, \mathrm{d}x_d.
\end{aligned} \tag{D.4}$$ {eq:tnproofj}

Plugging in (D.1) for $g(X_i) = X_i - \mu_i$ in (D.3) we get

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

where Z is the distribution of $\mathrm{N}(0, 1)$, and we had used a change of variable $x_j = \sigma_j z + \mu_j$, so that $p(x_j) = \sigma_j^{-1} \phi(z)$ and $\mathrm{d}x_j = \sigma_j \mathrm{d}z$. For the j'th component, substitute $g(x_j) = x_j - \mu_j$ in (D.4) to get

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

where we have made use of Lemma D.1 in the second last step.

For the second moments, plug in (D.2) for $g(X_i) = (X_i - \mu_i)^2 - \sigma_i^2$ in (D.3) to get

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

And similarly, for the j'th component

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

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

lem:EZgZ

Lemma D.1. *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. Introduce the following notation for convenience:

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

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

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

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

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

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

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

is found to be

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

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

D.3 Proof of Lemma C.5: Entropy

As a direct consequence of the definition of entropy,

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

Appendix E

I-prior interpretation of the g-prior

misc:gprior

The I-prior for $\boldsymbol{\beta}$ in a standard linear model resembles the objective g-prior (Zellner, 1986) for regression coefficients,

$$\boldsymbol{\beta} \sim \mathrm{N}_p\left(\mathbf{0}, g(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1}\right),$$

although they are quite different objects. The g-prior for $\boldsymbol{\beta}$ has the *inverse* (scaled) Fisher information matrix as its covariance matrix. This, in itself, has a much different and arguably counterintuitive meaning: large amounts of Fisher information about $\boldsymbol{\beta}$ corresponds to a small prior variance, and hence less deviation away from the prior mean of zero in estimating $\boldsymbol{\beta}$. The choice of the hyperparameter g has been the subject of much debate, with choices ranging from fixing $g = n$ (corresponding to the concept of *unit Fisher information*), to fully Bayesian and empirical Bayesian methods of estimating g from the data.

On the other hand, we note that the g-prior has an I-prior interpretation when argued as follows. Assume that the regression function f lies in the continual dual space of $\mathbb{R}^p$ equipped with the inner product $\langle \mathbf{x}, \mathbf{x}' \rangle_{\mathcal{X}} = \mathbf{x}^\top (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}$. With this inner product and from (3.3) (p. 11), the Fisher information for $\boldsymbol{\beta}$ is

$$\begin{aligned}
\mathcal{I}_g(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \psi_{ij} (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}_i \otimes (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \mathbf{x}_j \\
&= (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1} \cancel{(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})} \cancel{(\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1}} \\
&= (\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X})^{-1},
\end{aligned}$$

and this, rather than the usual $\mathbf{X}^\top \boldsymbol{\Psi} \mathbf{X}$ as the prior covariance matrix for $\boldsymbol{\beta}$, means that the I-prior is in fact the standard g-prior.

The metric induced by the inner product is actually the *Mahalanobis distance*, a scale-invariant natural distance if the covariates are measured on different scales. To expand on this idea, circle back to the regression function and write it as $f(\mathbf{x}) = \langle \mathbf{x}, \boldsymbol{\beta} \rangle_{\mathcal{X}}$. In usual least squares regression, the choice of inner product is irrelevant, so the usual dot product is commonly used (however, as we have seen above, the choice of inner product determines the form of the Fisher information for $\boldsymbol{\beta}$). In particular, suppose that all the x_{ik}'s, $k = 1, \ldots, p$ for each unit $i = 1, \ldots, n$ are measured on the same scale; for instance, these could be measurements in centimetres. In this case, the dot product is reasonable, because $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \sum_{k=1}^{p} x_{ik} x_{jk}$ and the inner product has a coherent unit, namely the squared unit of the x_{ik}'s. However, if they were a mix of various scaled measurements, then obviously the inner product's unit is incoherent—one would be resorted to adding measurements in different units, for example, cm^2 and kg^2 and so on. In such a case, a unitless inner product is appropriate, like the Mahalonobis inner product, which technically rescales the x_{ik}'s to unity. In summary, if the covariates are all measured on the same scale, then the I-prior is appropriate, and if not, the g-prior is appropriate.

Appendix F

Additional details for various I-prior regression models

These are additional details relating to discussion on various I-prior regression models in Section 4.1 of Chapter 4 (p. 4). These details relate to the standard linear multilevel model and the naïve classification model.

F.1 The I-prior for standard multilevel models

misc:multilevelmodels

We show the corresponding I-prior for the regression coefficients of the standard linear multilevel model (4.3). Write $\alpha = \beta_0$, and for simplicity, assume iid errors, i.e., $\boldsymbol{\Psi} = \psi \mathbf{I}_n$. The form of $f \in \mathcal{F}$ is now $f(\mathbf{x}_i^{(j)}, j) = \sum_{i'=1}^{n_{j'}} \sum_{j'=1}^{m} h_\lambda\big((\mathbf{x}_i^{(j)}, j), (\mathbf{x}_{i'}^{(j')}, j')\big) w_{i'j'}$, where each $w_{i'j'} \sim \mathrm{N}(0, \psi^{-1})$.

Now, functions in the scaled RKHS $\mathcal{F}_2$ have the form

$$
\begin{aligned}
f_2(j) &= \sum_{i=1}^{n_{j'}} \sum_{j'=1}^{m} \lambda_2 \left(\frac{\delta_{jj'}}{p_j} - 1 \right) w_{ij'} \\
&= \lambda_2 \left(\frac{w_{+j}}{p_j} - w_{++} \right),
\end{aligned}
$$

where a '+' in the index of w_{ik} indicates a summation over that index, and p_j is the empirical distribution over $\mathcal{M}$, i.e. $p_j = n_j/n$. Clearly $f_2(j)$ is a variable depending on

j, so write $f_2(j) = \beta_{0j}$. The distribution of β_{0j} is normal with zero mean and variance

$$\begin{aligned}\operatorname{Var}\beta_{0j} &= \lambda_2^2\left(\frac{n_j\psi}{n_j^2/n^2} + n\psi\right)\\ &= n\psi\lambda_2^2\left(\frac{1}{p_j}+1\right).\end{aligned}$$

The covariance between any two random intercepts β_{0j} and $\beta_{0j'}$ is

$$\begin{aligned}\operatorname{Cov}(\beta_{0j},\beta_{0j'}) &= \operatorname{Cov}\left(\lambda_2\left(\frac{w_{+j}}{p_j}-w_{++}\right),\lambda_2\left(\frac{w_{+j'}}{p_{j'}}-w_{++}\right)\right)\\ &= \frac{\lambda_2^2}{p_jp_{j'}}\underbrace{\operatorname{Cov}(w_{+j},w_{+j'})}_{\to 0} - \frac{\lambda_2^2}{p_j}\operatorname{Cov}(w_{+j},w_{++}) - \frac{\lambda_2^2}{p_{j'}}\operatorname{Cov}(w_{++},w_{+j'})\\ &\quad + \lambda_2^2\operatorname{Cov}(w_{++},w_{++})\\ &= -\frac{\lambda_2^2}{n_j/n}n_j\psi - \frac{\lambda_2^2}{n_{j'}/n}n_{j'}\psi + \lambda_2^2 n\psi\\ &= -n\psi\lambda_2^2.\end{aligned}$$

Functions in $\mathcal{F}_{12}$, on the other hand, have the form

$$\begin{aligned}f_{12}(\mathbf{x}_i,j) &= \sum_{i'=1}^{n_{j'}}\sum_{j'=1}^{m}\lambda_1\lambda_2\cdot\tilde{\mathbf{x}}_i^{(j)\top}\tilde{\mathbf{x}}_{i'}^{(j')}\cdot\left(\frac{\delta_{jj'}}{p_j}-1\right)w_{i'j'}\\ &= \tilde{\mathbf{x}}_i^{(j)\top}\underbrace{\left(\frac{\lambda_1\lambda_2}{p_j}\sum_{i'=1}^{n_j}\tilde{\mathbf{x}}_{i'}^{(j)}w_{i'j} - \lambda_1\lambda_2\sum_{i'=1}^{n_{j'}}\sum_{j'=1}^{m}\tilde{\mathbf{x}}_{i'}^{(j')}w_{i'j'}\right)}_{\boldsymbol{\beta}_{1j}},\end{aligned}$$

and this is, as expected, a linear form dependent on cluster j. We can calculate the variance for β_{1j} to be

$$\begin{aligned}\operatorname{Var}\boldsymbol{\beta}_{1j} &= \lambda_1^2\lambda_2^2\operatorname{Var}\left(\frac{1}{p_j}\tilde{\mathbf{X}}_j^\top\mathbf{w}_j - \tilde{\mathbf{X}}^\top\mathbf{w}\right)\\ &= \lambda_1^2\lambda_2^2\left(\frac{\psi}{n_j^2/n^2}\tilde{\mathbf{X}}_j^\top\tilde{\mathbf{X}}_j + \psi\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \frac{1}{p_j}\tilde{\mathbf{X}}_j^\top\operatorname{Cov}(\mathbf{w}_j,\mathbf{w})\tilde{\mathbf{X}}^\top\right)\\ &= n\psi\lambda_1^2\lambda_2^2\left(\frac{1}{p_j}\mathbf{S}_j + \mathbf{S} - \mathbf{S}_j\right)\\ &= n\psi\lambda_1^2\lambda_2^2\left(\left(\frac{1}{p_j}-1\right)\mathbf{S}_j + \mathbf{S}\right)\end{aligned}$$

where $\mathbf{S}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})^\top(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})$, $\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{m}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})^\top(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})$, and $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{m}\mathbf{x}_i^{(j)}$. The covariance between two vectors of the random slopes is

$$\begin{aligned}
\operatorname{Cov}(\boldsymbol{\beta}_{1j}, \boldsymbol{\beta}_{1j'}) &= \lambda_1^2\lambda_2^2 \operatorname{Cov}\left(\frac{1}{p_j}\tilde{\mathbf{X}}_j^\top \mathbf{w}_j - \tilde{\mathbf{X}}^\top\mathbf{w}, \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top\mathbf{w}_{j'} - \tilde{\mathbf{X}}^\top\mathbf{w}\right) \\
&= \psi\lambda_1^2\lambda_2^2\left(\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} - \frac{1}{p_j}\tilde{\mathbf{X}}_j^\top\tilde{\mathbf{X}}_j - \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top\tilde{\mathbf{X}}_{j'}\right) \\
&= n\psi\lambda_1^2\lambda_2^2\left(\mathbf{S} - \mathbf{S}_j - \mathbf{S}_{j'}\right).
\end{aligned}$$

Another quantity of interest is the covariance between the random intercepts and random slopes:

$$\begin{aligned}
\operatorname{Cov}(\beta_{0j}, \boldsymbol{\beta}_{1j}) &= \lambda_1\lambda_2^2 \operatorname{Cov}\left(\frac{1}{p_j}\mathbf{1}_{n_j}^\top\mathbf{w}_j - \mathbf{1}_n^\top\mathbf{w}, \frac{1}{p_j}\tilde{\mathbf{X}}_j^\top\mathbf{w}_j - \tilde{\mathbf{X}}^\top\mathbf{w}\right) \\
&= \psi\lambda_1\lambda_2^2\left(\cancelto{0}{\mathbf{1}_n^\top\tilde{\mathbf{X}}} + \frac{1}{p_j^2}\mathbf{1}_{n_j}^\top\tilde{\mathbf{X}}_j - \frac{2}{p_j}\mathbf{1}_{n_j}^\top\tilde{\mathbf{X}}_j\right) \\
&= n\psi\lambda_1\lambda_2^2\left(\left(\frac{1}{p_j} - 2\right)\frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}})\right) \\
&= n\psi\lambda_1\lambda_2^2\left(\frac{1}{p_j} - 2\right)(\bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}})
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{Cov}(\beta_{0j}, \boldsymbol{\beta}_{1j'}) &= \lambda_1\lambda_2^2 \operatorname{Cov}\left(\frac{1}{p_j}\mathbf{1}_{n_j}^\top\mathbf{w}_j - \mathbf{1}_n^\top\mathbf{w}, \frac{1}{p_{j'}}\tilde{\mathbf{X}}_{j'}^\top\mathbf{w}_{j'} - \tilde{\mathbf{X}}^\top\mathbf{w}\right) \\
&= \psi\lambda_1\lambda_2^2\left(\cancelto{0}{\mathbf{1}_n^\top\tilde{\mathbf{X}}} + \frac{1}{p_jp_{j'}}\mathbf{1}_{n_j}^\top\cancelto{0}{\operatorname{Cov}(\mathbf{w}_j, \mathbf{w}_{j'})}\tilde{\mathbf{X}}_{j'} - \frac{1}{p_j}\mathbf{1}_{n_j}^\top\tilde{\mathbf{X}}_j - \frac{1}{p_{j'}}\mathbf{1}_{n_{j'}}^\top\tilde{\mathbf{X}}_{j'}\right) \\
&= n\psi\lambda_1\lambda_2^2\left(-\frac{1}{n_j}\sum_{i=1}^{n_j}(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}) - \frac{1}{n_{j'}}\sum_{i=1}^{n_{j'}}(\mathbf{x}_i^{(j')} - \bar{\mathbf{x}})\right) \\
&= n\psi\lambda_1\lambda_2^2\left(2\bar{\mathbf{x}} - \bar{\mathbf{x}}^{(j)} - \bar{\mathbf{x}}^{(j')}\right).
\end{aligned}$$

F.2 The I-prior for naïve classification

For the naïve I-prior classification model (4.7), the I-prior is derived as follows. Firstly, 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 2.36, but also, as per the definition of the Pearson RKHS and centred identity kernel RKHS). 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,

$$\begin{aligned} \mathrm{E}\left[\sum_{j=1}^m y_{ij}\right] &= \sum_{j=1}^m (\alpha + \alpha_j + f_j(x_i)) \\ &= m\alpha + \sum_{j=1}^m \alpha_j \overset{0}{\cancel{}} + \sum_{j=1}^m f_j(x_i) \overset{0}{\cancel{}} \end{aligned}$$

and since $\sum_{j=1}^m y_{ij} = 1$, we get the ML estimate $\hat{\alpha} = 1/m$, and thus the grand intercept can be fixed to resolve identification.

It is much more 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 $\mathbf{w}$ follows a matrix normal distribution $\mathrm{MN}_{n,m}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Psi})$, 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 $\mathbf{B}_\eta$ the $n \times n$ kernel matrix with entries supplied by kernel $1 + b_\eta$ over $\mathcal{X} \times \mathcal{X}$, and $\mathbf{A}$ the $m \times m$ matrix with entries supplied by a over $\mathcal{M} \times \mathcal{M}$. From (4.7), we have that

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

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

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

and

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

By the results of Chapter 4, the posterior distribution of the I-prior random effects is $\operatorname{vec} \mathbf{w} | \mathbf{y} \sim \mathrm{N}(\operatorname{vec} \tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$, where

$$\operatorname{vec} \tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w (\boldsymbol{\Psi} \otimes \mathbf{H}_\eta) \operatorname{vec}(\mathbf{y} - \mathbf{1}_n \boldsymbol{\alpha}^\top) \quad \text{and} \quad \tilde{\mathbf{V}}_w^{-1} = \mathbf{A}\boldsymbol{\Psi}\mathbf{A} \otimes \mathbf{B}_\eta^2 + \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n = \mathbf{V}_y. \tag{F.3}$$

{eq:naiveclasspostw}

Suppose hypothetically, one uses the uncentered identity kernel $a(j, j') = \delta_{jj'}$, in which case centring of the intercepts α_j must be handled separately. 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

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

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

Appendix G

Posterior distribution of the I-prior regression function

We derive the posterior distribution for the I-prior random effects $\mathbf{w} = (w_1, \ldots, w_n)^\top$, which is related to the I-prior regression function via $f(x_i) = \sum_{k=1}^n h_\eta(x_i, x_k) w_k$, or in matrix terms, $\mathbf{f} := \big(f(x_1), \ldots, f(x_n)\big)^\top = \mathbf{H}_\eta \mathbf{w}$, and $f \in \mathcal{F}$ an RKHS with kernel h_η. A closely related distribution of interest is the posterior predictive distribution of y_{new}, the prediction at a new data point x_{new}. We note the similarity of these results with the posterior distributions of Gaussian process regressions (Rasmussen and Williams, 2006).

G.1 Deriving the posterior distribution for w

apx:posterior

In the following derivation, we implicitly assume the dependence on $\mathbf{f}_0$ and θ. The distribution of $\mathbf{y}|\mathbf{w}$ is $\mathrm{N}_n(\boldsymbol{\alpha}+\mathbf{f}_0+\mathbf{H}_\eta\mathbf{w}, \boldsymbol{\Psi}^{-1})$, where $\boldsymbol{\alpha} = \alpha \mathbf{1}_n$, while the prior distribution for $\mathbf{w}$ is $\mathrm{N}_n(\mathbf{0}, \boldsymbol{\Psi})$. Since $p(\mathbf{w}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{w})p(\mathbf{w})$, we have that

$$
\begin{aligned}
\log p(\mathbf{w}|\mathbf{y}) &= \log p(\mathbf{y}|\mathbf{w}) + \log p(\mathbf{w}) \\
&= \text{const.} + \cancel{\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}(\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0 - \mathbf{H}_\eta\mathbf{w})^\top \boldsymbol{\Psi} (\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0 - \mathbf{H}_\eta\mathbf{w}) \\
&\quad - \cancel{\frac{1}{2}\log|\boldsymbol{\Psi}|} - \frac{1}{2}\mathbf{w}^\top \boldsymbol{\Psi}^{-1}\mathbf{w} \\
&= \text{const.} - \frac{1}{2}\mathbf{w}^\top(\mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1})\mathbf{w} + (\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0)^\top \boldsymbol{\Psi}\mathbf{H}_\eta\mathbf{w}.
\end{aligned}
$$

Setting $\mathbf{A} = \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta + \boldsymbol{\Psi}^{-1}$, $\mathbf{a}^\top = (\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0)^\top\boldsymbol{\Psi}\mathbf{H}_\eta$, and using the fact that

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

we have that $\mathbf{w}|\mathbf{y}$ is normally distributed with the required mean and variance.

Alternatively, one could have shown this using standard results of multivariate normal distributions. Noting that the covariance between $\mathbf{y}$ and $\mathbf{w}$ is

$$\begin{aligned}\operatorname{Cov}[\mathbf{y}, \mathbf{w}] &= \operatorname{Cov}[\boldsymbol{\alpha} + \mathbf{f}_0 + \mathbf{H}_\eta \mathbf{w} + \boldsymbol{\epsilon}, \mathbf{w}] \\ &= \mathbf{H}_\eta \operatorname{Cov}[\mathbf{w}, \mathbf{w}] \\ &= \mathbf{H}_\eta \boldsymbol{\Psi}\end{aligned}$$

and that $\operatorname{Cov}[\mathbf{w}, \mathbf{y}] = \boldsymbol{\Psi}\mathbf{H}_\eta = \mathbf{H}_\eta\boldsymbol{\Psi} = \operatorname{Cov}[\mathbf{y}, \mathbf{w}]$ by symmetry, the joint distribution $(\mathbf{y}, \mathbf{w})$ is

$$\begin{pmatrix}\mathbf{y} \\ \mathbf{w}\end{pmatrix} \sim \mathrm{N}_{n+n}\left(\begin{pmatrix}\boldsymbol{\alpha} + \mathbf{f}_0 \\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{V}_y & \mathbf{H}_\eta\boldsymbol{\Psi} \\ \boldsymbol{\Psi}\mathbf{H}_\eta & \boldsymbol{\Psi}\end{pmatrix}\right).$$

Thus,

$$\begin{aligned}\mathrm{E}[\mathbf{w}|\mathbf{y}] &= \mathrm{E}\,\mathbf{w} + \operatorname{Cov}(\mathbf{w}, \mathbf{y})(\operatorname{Var}\mathbf{y})^{-1}(\mathbf{y} - \mathrm{E}\,\mathbf{y}) \\ &= \boldsymbol{\Psi}\mathbf{H}_\eta\mathbf{V}_y^{-1}(\mathbf{y} - \boldsymbol{\alpha} - \mathbf{f}_0),\end{aligned}$$

and

$$\begin{aligned}\operatorname{Var}[\mathbf{w}|\mathbf{y}] &= \operatorname{Var}\mathbf{w} - \operatorname{Cov}(\mathbf{w}, \mathbf{y})(\operatorname{Var}\mathbf{y})^{-1}\operatorname{Cov}(\mathbf{y}, \mathbf{w}) \\ &= \boldsymbol{\Psi} - \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{V}_y^{-1}\mathbf{H}_\eta\boldsymbol{\Psi} \\ &= \boldsymbol{\Psi} - \boldsymbol{\Psi}\mathbf{H}_\eta\left(\boldsymbol{\Psi}^{-1} + \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta\right)^{-1}\mathbf{H}_\eta\boldsymbol{\Psi} \\ &= \left(\boldsymbol{\Psi}^{-1} + \mathbf{H}_\eta\boldsymbol{\Psi}\mathbf{H}_\eta\right)^{-1} \\ &= \mathbf{V}_y^{-1}\end{aligned}$$

as a direct consequence of the Woodbury matrix identity (Petersen and Pedersen, 2008, eq. 156, §3.2.2).

G.2 Deriving the posterior predictive distribution

apx:postpred

The posterior predictive distribution is obtained in an empirical Bayesian manner, in which the parameters of the model are replaced with their ML estimates (denoted with hats).

A priori, assume that $y_{\text{new}} \sim \mathrm{N}(\hat{\alpha}, v_{\text{new}})$, where $v_{\text{new}} = \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top\hat{\boldsymbol{\Psi}}\mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1}$. Consider the joint distribution of $(y_{\text{new}}, \mathbf{y}^\top)^\top$, which is multivariate normal (since both y_{new} and $\mathbf{y}$ are. Write

$$\begin{pmatrix}y_{\text{new}} \\ \mathbf{y}\end{pmatrix} \sim \mathrm{N}_{n+1}\left(\begin{pmatrix}\hat{\alpha} \\ \hat{\alpha}\mathbf{1}_n\end{pmatrix}, \begin{pmatrix}v_{\text{new}} & \operatorname{Cov}[y_{\text{new}}, \mathbf{y}] \\ \operatorname{Cov}[y_{\text{new}}, \mathbf{y}]^\top & \hat{\mathbf{V}}_y\end{pmatrix}\right),$$

where

$$\begin{aligned}
\mathrm{Cov}[y_{\text{new}}, \mathbf{y}] &= \mathrm{Cov}[f_{\text{new}} + \epsilon_{\text{new}}, \mathbf{f} + \boldsymbol{\epsilon}] \\
&= \mathrm{Cov}[f_{\text{new}}, \mathbf{f}) + \mathrm{Cov}(\epsilon_{\text{new}}, \boldsymbol{\epsilon}] \\
&= \mathrm{Cov}\left[\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \tilde{\mathbf{w}}, \mathbf{H}_{\hat{\eta}} \tilde{\mathbf{w}}\right] + (\sigma_{\text{new},1}, \ldots, \sigma_{\text{new},n}) \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}}.
\end{aligned}$$

The vector of covariances $\boldsymbol{\sigma}_{\text{new}}$ between observations $y_1, \ldots, y_n$ and the predicted point y_{new} would need to be prescribed a priori (treated as extra parameters), or estimated again, which seems excessive. Under an iid assumption of the error precisions, then $\boldsymbol{\sigma}_{\text{new}} = \mathbf{0}$ would be acceptable.

In any case, using standard multivariate normal results, we get that $y_{\text{new}}|\mathbf{y}$ is also normally distributed with mean

$$\begin{aligned}
E[y_{\text{new}}|\mathbf{y}] &= \hat{\alpha} + (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}}) \hat{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}} \\
&= \hat{\alpha} + \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \overbrace{\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} \hat{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}}}^{\hat{\mathbf{w}}} + \boldsymbol{\sigma}_{\text{new}} \hat{\mathbf{V}}_y^{-1} \tilde{\mathbf{y}} \\
&= \hat{\alpha} + \mathrm{E}\left[f(x_{\text{new}})|\mathbf{y}\right] + \text{mean correction term}
\end{aligned}$$

and variance

$$\begin{aligned}
\mathrm{Var}[y_{\text{new}}|\mathbf{y}] &= v_{\text{new}} - (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}}) \hat{\mathbf{V}}_y^{-1} (\mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} + \boldsymbol{\sigma}_{\text{new}})^\top \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \hat{\mathbf{h}}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} - \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} \hat{\mathbf{V}}_y^{-1} \mathbf{H}_{\hat{\eta}} \hat{\boldsymbol{\Psi}} \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) \\
&\quad + \text{variance correction term} \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \left(\hat{\boldsymbol{\Psi}} - \hat{\boldsymbol{\Psi}} \mathbf{H}_{\hat{\eta}} \hat{\mathbf{V}}_y^{-1} \mathbf{H}_{\hat{\eta}} \hat{\boldsymbol{\Psi}}\right) \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} \\
&\quad + \text{variance correction term} \\
&= \mathbf{h}_{\hat{\eta}}(x_{\text{new}})^\top \hat{\mathbf{V}}_y^{-1} \mathbf{h}_{\hat{\eta}}(x_{\text{new}}) + \psi_{\text{new}}^{-1} + \text{variance correction term} \\
&= \mathrm{Var}\left[f(x_{\text{new}})|\mathbf{y}\right] + \psi_{\text{new}}^{-1} + \text{variance correction term}.
\end{aligned}$$

Appendix H

Variational EM algorithm for I-probit models

apx:varemip robit

The two sections that follow detail the derivation of the variational densities used in the E-step of the variational EM algorithm, and also the lower bound (ELBO) used to monitor convergence.

H.1 Derivation of the variational densities

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

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

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

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

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

The joint likelihood is given by

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

For reference, the three relevant distributions are listed below.

- $\boldsymbol{p(\mathbf{y}|\mathbf{y}^*)}$. For each observation $i \in \{1, \dots, n\}$, given the corresponding latent propensities $\mathbf{y}_i^* = (y_{i1}^*, \dots, y_{im}^*)$, the distribution for y_i is a degenerate distribution which depends on the j'th component of $\mathbf{y}_i^*$ being largest, where the value observed for y_i was j. Since each of the y_i's are independent, everything is multiplicative.

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

- $\boldsymbol{p(\mathbf{y}^*|\mathbf{w})}$. Given values for the parameters and I-prior random effects, the distribution of the latent propensities is matrix normal

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

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

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

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

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

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

H.1.1 Derivation of $\tilde{q}(\mathbf{y}^*)$

The rows of $\mathbf{y}^*$ are independent, and thus we can consider the variational density for each $\mathbf{y}_i^*$ separately. Consider the case where y_i takes one particular value $j \in \{1, \dots, m\}$. In such cases, we have that $y_{ij}^* > y_{ik}$ for all $k \neq j$, and that

$$\begin{aligned} \log\tilde{q}(\mathbf{y}_{i\cdot}^*) &= \mathrm{E}_{\mathbf{w}\sim\tilde{q}}\left[-\frac{1}{2}(\mathbf{y}_i^*-\boldsymbol{\mu}_i)^\top\boldsymbol{\Psi}(\mathbf{y}_i^*-\boldsymbol{\mu}_i)\right] + \text{const.} \\ &= \left[-\frac{1}{2}(\mathbf{y}_i^*-\tilde{\boldsymbol{\mu}}_i)^\top\boldsymbol{\Psi}(\mathbf{y}_i^*-\tilde{\boldsymbol{\mu}}_i)\right] + \text{const.} \qquad (\star) \end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_{i\cdot} = \boldsymbol{\alpha} + \tilde{\mathbf{w}}\mathbf{h}_\eta(x_i)$, $\tilde{\mathbf{w}} = \mathrm{E}_{\mathbf{w}\sim\tilde{q}}[\mathbf{w}]$. This is recognised as the logarithm of a multivariate normal pdf with mean $\tilde{\boldsymbol{\mu}}_{i\cdot}$ and variance $\boldsymbol{\Psi}^{-1}$. On the other hand, when $y_i \neq j$, the pdf is zero. Thus,

$$\tilde{q}(\mathbf{y}^*_{i\cdot}) = \begin{cases} \phi(\mathbf{y}^*_{i\cdot}|\tilde{\boldsymbol{\mu}}_{i\cdot}, \boldsymbol{\Psi}^{-1}) & \text{if } y^*_{ij} > y^*_{ik}, \forall k \neq j \\ 0 & \text{otherwise,} \end{cases}$$

implying a truncated multivariate normal distribution for $\mathbf{y}^*_{i\cdot}$. The required moments from the truncated multivariate normal distribution can be obtained using the methods described in Appendix C.4 (p. 17).

Remark H.1. In the above derivation, we needn't consider the second order terms in the expectations because they do not involve $\mathbf{y}^*_{i\cdot}$, and thus, these terms can be absorbed into the constant. To see this,

$$\begin{aligned} \mathrm{E}[(\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot})^\top \boldsymbol{\Psi}(\mathbf{y}^*_{i\cdot} - \boldsymbol{\mu}_{i\cdot})] &= \mathrm{E}[\mathbf{y}^{*\top}_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} + \boldsymbol{\mu}^\top_{i\cdot}\boldsymbol{\Psi}\boldsymbol{\mu}_{i\cdot} - 2\boldsymbol{\mu}^\top_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot}] \\ &= \mathbf{y}^{*\top}_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} - 2\,\mathrm{E}[\boldsymbol{\mu}^\top_{i\cdot}]\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} + \text{const.} \\ &= \mathbf{y}^{*\top}_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} - 2\tilde{\boldsymbol{\mu}}^\top_{i\cdot}\boldsymbol{\Psi}\mathbf{y}^*_{i\cdot} + \text{const.} \\ &= (\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot})^\top \boldsymbol{\Psi}(\mathbf{y}^*_{i\cdot} - \tilde{\boldsymbol{\mu}}_{i\cdot}) + \text{const.} \end{aligned}$$

The square is then completed to get the final line, which is the expression for the term $(\star)$ multiplied by a half.

H.1.2 Derivation of $\tilde{q}(\mathbf{w})$

apx:qw

The terms involving $\mathbf{w}$ in the joint likelihood (5.14) are the $p(\mathbf{y}^*|\mathbf{w})$ and $p(\mathbf{w})$ terms, so the rest are absorbed into the constant. The easiest way to derive $\tilde{q}(\mathbf{w})$ is to vectorise $\mathbf{y}^*$ and $\mathbf{w}$. We know that

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

and

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

using properties of matrix normal distributions.

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

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

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

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

and

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

We will often refer to $\tilde{\mathbf{w}}$ as the $n \times m$ matrix constructed by filling in its entries with $\operatorname{vec} \tilde{\mathbf{w}}$ column-wise (akin to the opposite of vectorisation). This way, the $\tilde{\mathbf{w}}$ contains posterior mean values arranged by class $j = 1, \ldots, m$ column-wise, and by observations $i = 1, \ldots, n$ row-wise. Ideally, we do not want to work with the $nm \times nm$ matrix $\mathbf{V}_w$, since its inverse is expensive to compute. Refer to Section 5.6.2 for details.

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

$$\begin{aligned}\tilde{\mathbf{V}}_w &= \big(\operatorname{diag}(\psi_1, \ldots, \psi_m) \otimes \mathbf{H}_\eta^2 + \operatorname{diag}(\psi_1, \ldots, \psi_m) \otimes \mathbf{I}_n \big)^{-1} \\ &= \operatorname{diag}\Big(\big(\psi_1 \mathbf{H}_\eta^2 + \psi_1^{-1}\mathbf{I}_n\big)^{-1}, \cdots, \big(\psi_m \mathbf{H}_\eta^2 + \psi_m^{-1}\mathbf{I}_n\big)^{-1} \Big) \\ &=: \operatorname{diag}(\tilde{\mathbf{V}}_{w_1}, \ldots, \tilde{\mathbf{V}}_{w_m}).\end{aligned}$$

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

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

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

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

A quantity that we will be requiring time and again will be $\operatorname{tr}(\mathbf{C} \operatorname{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}])$, where $\mathbf{C} \in \mathbb{R}^{m \times m}$ and $\mathbf{D} \in \mathbb{R}^{n \times n}$ are both square and symmetric matrices. Using the definition of the trace directly, we get

$$\begin{aligned}\operatorname{tr}(\mathbf{C} \operatorname{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}]) &= \sum_{i,j=1}^m \mathbf{C}_{ij} \operatorname{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}]_{ij} \\ &= \sum_{i,j=1}^m \mathbf{C}_{ij} \operatorname{E}[\mathbf{w}_{\cdot i}^\top \mathbf{D} \mathbf{w}_{\cdot j}]. \qquad \text{(H.1)}\end{aligned}$$ {eq:trCEwDw}

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

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

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

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

where δ is the Kronecker delta. Continuing on (H.1) leads us to

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

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

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

H.2 Deriving the ELBO expression

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

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

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

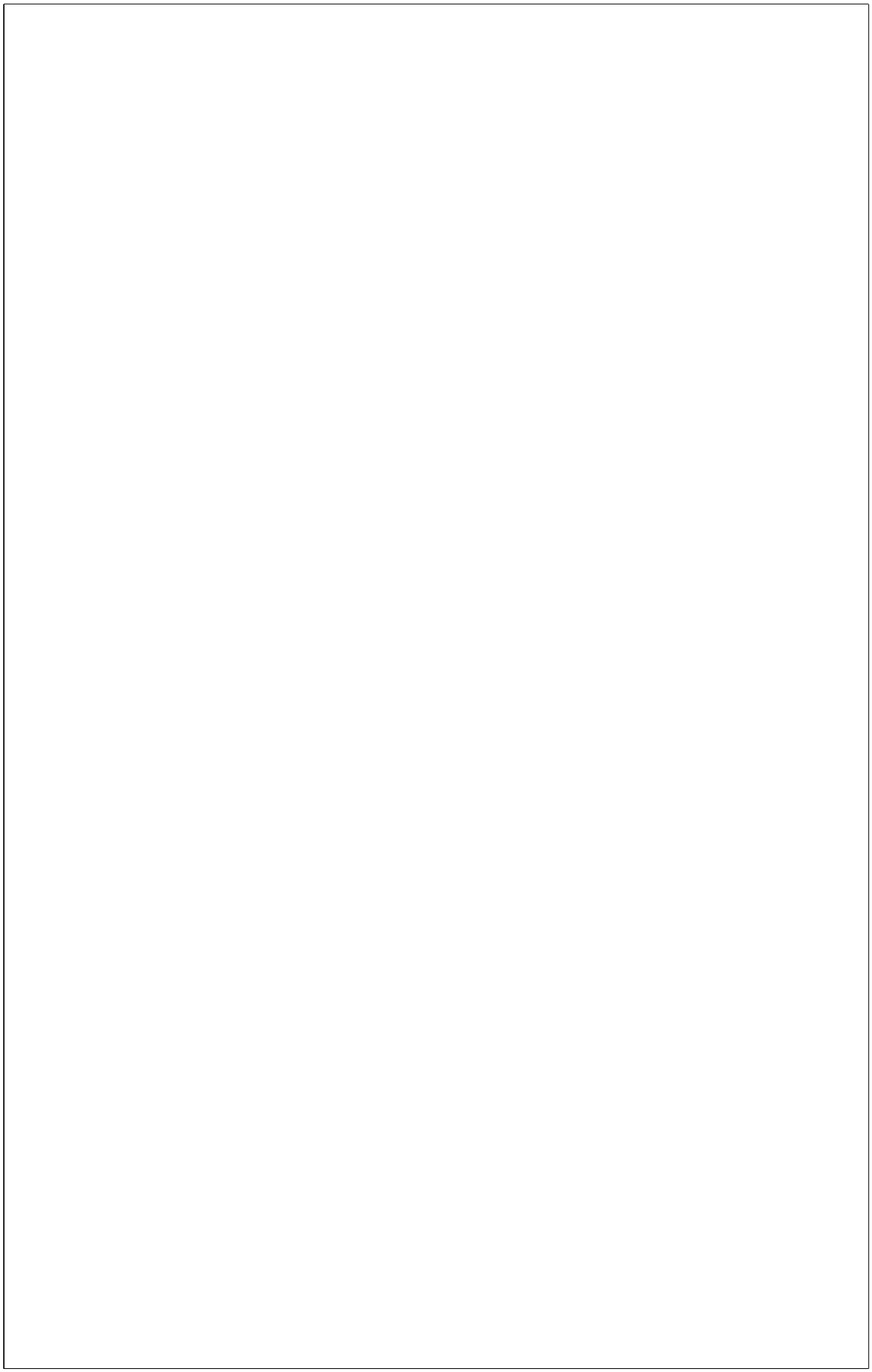
H.2.1 Terms involving distributions of $\mathbf{y}^*$

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

where C_i is the normalising constant for the distribution of multivariate truncated normal $\mathbf{y}_{i\cdot}^* \sim {}^{\text{t}}\text{N}(\tilde{\boldsymbol{\mu}}(x_i), \boldsymbol{\Psi}^{-1}, \mathcal{C}_{y_i})$, with $\tilde{\boldsymbol{\mu}}(x_i) = \boldsymbol{\alpha} + \tilde{\mathbf{w}}\mathbf{h}_\eta(x_i)$.

H.2.2 Terms involving distributions of w

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



Appendix I

The Gibbs sampler for the I-prior Bayesian variable selection model

apx:gibbsbvs

The I-prior Bayesian variable selection model has the following hierarchical form:

$$\begin{gathered}
\mathbf{y}|\alpha, \boldsymbol{\beta}, \gamma, \sigma^2, \kappa \sim \mathrm{N}_n(\alpha \mathbf{1}_n + \mathbf{X}\boldsymbol{\theta}, \sigma^2 I_n) \\
\boldsymbol{\theta} = (\gamma_1\beta_1, \ldots, \gamma_p\beta_p)^\top \\
\boldsymbol{\beta}|\sigma^2, \kappa \sim \mathrm{N}_p(\mathbf{0}, \sigma^2\kappa\mathbf{X}^\top\mathbf{X}) \\
\alpha|\sigma^2 \sim \mathrm{N}(0, \sigma^2 A) \\
\sigma^2, \kappa \sim \Gamma^{-1}(c, d) \\
\gamma_j \sim \mathrm{Bern}(\pi_j) \quad j = 1, \ldots, p
\end{gathered}$$

In the simulations and real-data examples, we used $\pi_j = 0.5, \forall j$, $A = 100$, and $c = d = 0.001$, and the columns of the matrix $\mathbf{X}$ are standardised.

The first line of the set of equations above is the likelihood, while the joint prior density is given by

$$p(\alpha, \beta, \gamma, \sigma^2, \kappa) = p(\beta|\sigma^2)p(\alpha|\sigma^2)p(\sigma^2)p(\kappa)p(\gamma_1)\cdots p(\gamma_p).$$

For simplicity, in the following subsections we shall denote by Θ the entire set of parameters, while $\Theta_{-\xi}$ implies the set of parameters excluding the parameter ξ.

I.1 Conditional posterior for $\boldsymbol{\beta}$

$$\begin{aligned}
\log p(\boldsymbol{\beta}|\mathbf{y}, \Theta_{-\boldsymbol{\beta}}) &= \text{const.} + \log p(\mathbf{y}|\Theta) + \log p(\boldsymbol{\beta}|\sigma^2) \\
&= \text{const.} - \frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}_\gamma\boldsymbol{\beta}\|^2 - \frac{1}{2\sigma^2}\boldsymbol{\beta}^\top(\kappa\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta} \\
&= \text{const.} - \frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^\top(\mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1})\boldsymbol{\beta} - 2(\mathbf{y} - \alpha\mathbf{1}_n)^\top\mathbf{X}_\gamma\boldsymbol{\beta}\right) \\
&= \text{const.} - \frac{1}{2\sigma^2}\big(\boldsymbol{\beta} - \tilde{\mathbf{B}}(\mathbf{y} - \alpha\mathbf{1}_n)\big)^\top\tilde{\mathbf{B}}^{-1}\big(\boldsymbol{\beta} - \tilde{\mathbf{B}}(\mathbf{y} - \alpha\mathbf{1}_n)\big)
\end{aligned}$$

where $\tilde{\mathbf{B}} = \mathbf{X}_\gamma^\top\mathbf{X}_\gamma + (\kappa\mathbf{X}^\top\mathbf{X})^{-1}$, and $\mathbf{X}_\gamma = (\gamma_1 X_1 \cdots \gamma_p X_p)$ is the $n \times p$ design matrix $\mathbf{X}$ with each of the p columns multiplied by the indicator variable γ. This is of course recognised as the log density of a p-variate normal distribution with mean and variance

$$\mathrm{E}[\boldsymbol{\beta}|\Theta_{-\boldsymbol{\beta}}] = \tilde{\mathbf{B}}(y - \alpha\mathbf{1}_n) \text{ and } \mathrm{Var}[\boldsymbol{\beta}|\Theta_{-\boldsymbol{\beta}}] = \sigma^2\tilde{\mathbf{B}}.$$

I.2 Conditional posterior for γ

Consider each γ_j in turn. For $j \in \{1, \ldots, p\}$,

$$\begin{aligned}
p(\gamma_j|\mathbf{y}, \Theta_{-\gamma_j}) &\propto p(\mathbf{y}|\Theta)p(\gamma_j) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2\right)\pi_j^{\gamma_j}(1 - \pi_j)^{1-\gamma_j}
\end{aligned}$$

Since the support of γ_j is $\{0, 1\}$, the above is a probability mass function which can be normalised easily. When $\gamma_j = 1$, we have

$$p(\gamma_j|\mathbf{y}, \Theta_{-\gamma_j}) \propto \pi_j \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[1]}\|^2\right) := u_j$$

while for $\gamma_j = 0$, we have

$$p(\gamma_j|\mathbf{y}, \Theta_{-\gamma_j}) \propto (1 - \pi_j) \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}_j^{[0]}\|^2\right) := v_j.$$

For $j = 1, \ldots, p$, we have used the notation $\theta_j^{[\omega]}$ to mean

$$\boldsymbol{\theta}_j^{[\omega]} = \begin{cases} (\theta_1, \ldots, \theta_{j-1}, \beta_j, \theta_{j+1}, \ldots, \theta_p) & \omega = 1 \\ (\theta_1, \ldots, \theta_{j-1}, 0, \theta_{j+1}, \ldots, \theta_p) & \omega = 0. \end{cases}$$

Therefore, the conditions distribution for γ_j is Bernoulli with success probability

$$\tilde{\pi}_j = \frac{u_j}{u_j + v_j}.$$

I.3 Conditional posterior for α

We can obtain the conditional posterior for α in a similar fashion we obtained the conditional posterior for $\boldsymbol{\beta}$. That is,

$$\begin{aligned}
\log p(\alpha|\mathbf{y}, \Theta_{-\alpha}) &= \text{const.} + \log p(\mathbf{y}|\Theta) + \log p(\alpha|\sigma^2) \\
&= \text{const.} - \frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2 - \frac{\alpha^2}{2\sigma^2 A} \\
&= \text{const.} - \frac{1}{2\sigma^2}\left((n + A^{-1})\alpha^2 - 2\alpha\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta})\right) \\
&= \text{const.} - \frac{1}{2\sigma^2(n + A^{-1})}\left(\alpha - \frac{\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\top\boldsymbol{\theta})}{n + A^{-1}}\right)^2.
\end{aligned}$$

Thus, the conditional posterior for α is normal with mean and variance which can be easily read off the final line above.

I.4 Conditional posterior for σ^2

The conditional density for σ^2 is

$$\begin{aligned}
\log p(\sigma^2|\mathbf{y}, \Theta_{-\sigma^2}) &= \text{const.} + \log p(\mathbf{y}|\Theta) + \log p(\sigma^2) \\
&= \text{const.} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2 - (c+1)\log\sigma^2 - d/\sigma^2 \\
&= \text{const.} - (n/2 + c + 1)\log\sigma^2 - \frac{\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2/2 + d}{\sigma^2}
\end{aligned}$$

which is an inverse gamma distribution with shape $\tilde{c} = n/2 + c + 1$ and scale $\tilde{d} = \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2/2 + d$.

I.5 Conditional posterior for κ

Interestingly, since κ is a hyperparameter to be estimated, it does not actually make use of any data, apart from the appearance of $\mathbf{X}$ in the covariance matrix for $\boldsymbol{\beta}$.

$$\begin{aligned}
\log p(\kappa|\mathbf{y}, \Theta_{-\kappa}) &= \text{const.} + \log p(\beta|\sigma^2, \kappa) + \log p(\kappa) \\
&= \text{const.} - \frac{p}{2}\log\kappa - \frac{1}{\kappa}\cdot\frac{1}{2\sigma^2}\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta} - (c+1)\log\kappa - d/\kappa \\
&= \text{const.} - (p/2 + c + 1)\log\kappa - \frac{\boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta}/\sigma^2 + d}{\kappa}
\end{aligned}$$

This is an inverse gamma distribution with shape $\tilde{c} = p/2 + c + 1$ and scale $\tilde{d} = \boldsymbol{\beta}^\top(\mathbf{X}^\top\mathbf{X})^{-1}\boldsymbol{\beta}/\sigma^2 + d$.

I.6 Computational note

From the above, we see that all of the Gibbs conditionals are of recognisable form, making Gibbs sampling a straightforward MCMC method to implement. We built an R package **ipriorBVS** that uses JAGS (Plummer, 2003), a variation of WinBUGS, internally for the Gibbs sampling, and wrote a wrapper function which takes formula based inputs for convenience. The **ipriorBVS** also performs two-stage BVS, and supported priors are the I-prior, g-prior, and independent prior, as used in this thesis. Although a Gibbs sampler could be coded from scratch, JAGS has the advantage of being tried and tested and has simple controls for tuning (burn-in, adaptation, thinning, etc.). Furthermore, the output from JAGS can be inspected using a myriad of multipurpose MCMC tools to diagnose convergence problems. The **ipriorBVS** package is available at https://github.com/haziqj/ipriorBVS.

In all examples, a default setting of 4,000 burn-in samples, 1,000 adaptation size, and 10,000 samples with no thinning seemed adequate. There were no major convergence issues encountered.

Computational complexity is dominated by the inversion of a $p \times p$ matrix, and matrix multiplications of order $O(np^2)$. These occur in the conditional posterior for $\boldsymbol{\beta}$. Overall, if $n \gg p$, then time complexity is $O(np^2)$. Storage requirements are $O(np)$.

Bibliography

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

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

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

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

plummer2003jags Plummer, Martyn (2003). "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. Vol. 124. Vienna, Austria, p. 125.

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

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

zellner1986assessing Zellner, Arnold (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions". In: *Bayesian inference and decision techniques*.

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

Figures

Tables

Theorems

Definitions

Nomenclature

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

Conventions

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

Indexing

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

Symbols

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

Relations

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

Functions

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

Abstract vector space operations and notations

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

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

Matrix and vector operations

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

Statistical functions

$\mathrm{P}(A)$	The probability of event A occurring
$p(X\|\theta)$	The probability density function of X given parameters θ
$L(\theta\|X)$	The log-likelihood of θ given data X, sometimes simply $L(\theta)$
$\mathrm{BF}(M, M')$	Bayes factor for comparing two models M and M'
$\mathcal{I}(\theta)$	The Fisher information for θ
$\mathrm{E}[X]$, $\mathrm{E}\,X$	The expectation[1] of the random element X
$\mathrm{Var}[X]$, $\mathrm{Var}\,X$	The variance[1] of the random element X
$\mathrm{Cov}[X, Y]$	The covariance[1] between two random elements X and Y
$H(p)$	The entropy of the distribution $p(X)$
$\mathrm{D}_{\mathrm{KL}}\big(q(x)\|p(x)\big)$	The Kullback-Leibler divergence from $p(x)$ to $q(x)$, denoted also by $\mathrm{D}_{\mathrm{KL}}(q\|p)$ for short

Statistical distributions

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

foot:exp

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

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

Abbreviations

This document is incomplete. The external file associated with the glossary 'abbreviations' (which should be called `appendix.gls-abr`) hasn't been created.

Check the contents of the file `appendix.glo-abr`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "appendix"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "appendix"
  ```

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.