

\_2;`2bbBQM KQ/2HHBM; mbBM; T`BQ`b /  
6Bb?2` BM7Q`K iBQM +Qp `B M+2 F2`M2

J/X > xB[ J/X C KBH

i?2bBb bm#KBii2/ iQ i?2 .2T `iK2Mi Q7 ai iBbiB+b Q7 i?2 GQM/C  
M/ SQHBiB+ Ha+B2M+2 7Q` i?2 /2;`22 Q7 .Q+iQ` Q7 S?

Ry P+iQ#2` kyR3

(;Bi) K bi2`!8+38yey  
+? M;2, kyR3@Ry@yN R9,jd,8d Yy3yy

k

































































































































































































































































































































































































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

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

machine learning; examples include support vector machines ([scholkopf2002learning](#)) and Gaussian process classification ([rasmussen2006gaussian](#)), with the latter being more closely related to the I-probit method. However, Gaussian process classification typically uses the logistic sigmoid function, and estimation most commonly performed using Laplace approximation, but other methods such as expectation propagation ([minka2001expectation](#)) and MCMC ([neal1999](#)) have been explored as well. Variational inference for Gaussian process probit models have been studied by [girolami2006variational](#), with their work providing a close reference to the variational algorithm employed by us.

Suggestions for future work include:

1. **Estimation of  $\Psi$ .** A limitation we had to face in this work was to treat  $\Psi$  as fixed. The discussion in [Section 5.6.3](#) shows that estimation of  $\Psi$  is possible, however, the specific nature of implementing this in computer code could not be explored in time. In particular, for the full I-probit model, the best method of imposing positive-definite constraints for  $\Psi$  in the M-step has not been fully researched.
2. **Inclusion of class-specific covariates.** Throughout the chapter, we assumed that covariates were unit-specific, rather than class-specific. To illustrate, consider modelling the choice of travel mode between two destinations (car, coach, train or aeroplane) as a function of disposable income and travel time. Individuals' income as a predictor of transportation choice is unit-specific, but clearly, travel time depends on the mode of transport. To incorporate class-specific covariates  $z_{ij}$ , the regression on the latent propensities in [\(5.2\)](#) could be extended as such:

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

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

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

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

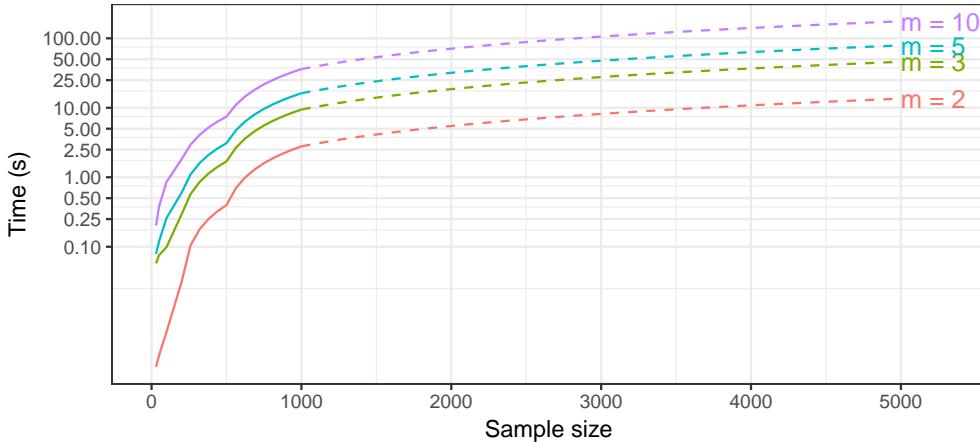


Figure 5.13: Time taken to complete a single variational inference iteration for varying sample sizes and number of classes  $m$ . The solid line represents actual timings, while the dotted lines are linear extrapolations.

obtaining posterior standard deviations, credibility intervals, and so on. However, a variational Bayes route has its cons:

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

among others, will need to be derived for the variational inference algorithm, and these can be tricky to compute.

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

course, one could easily specify vague or even diffuse priors. The concern is that the model could be sensitive to prior choices.

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



# Chapter 6

## Bayesian variable selection using I-priors

Earlier in Section 4.1 (p. 100), we saw that model (1.1) subject to normal assumptions (1.2), model assumptions A1–A3, and  $f$  belonging to the canonical RKHS of functions over  $\mathcal{X} \equiv \mathbb{R}^p$  yields the standard multiple regression model

$$y_i = \alpha + \sum_{k=1}^p x_{ik} \beta_k + \epsilon_i \quad (6.1)$$
$$\epsilon_i \stackrel{\text{iid}}{\sim} N_n(0, \sigma^2).$$

In this chapter, we use the notation  $\sigma^2 = \psi^{-1}$  to denote the error variance. Furthermore, an I-prior on the regression coefficient entails prescribing the following normal prior the  $\beta_k$ 's:

$$\boldsymbol{\beta} := (\beta_1, \dots, \beta_p)^\top \sim N_p(\mathbf{0}, \kappa \sigma^2 \mathbf{X}^\top \mathbf{X}).$$

This follows from (4.1) after a slight reparameterisation of the RKHS scale parameter  $\kappa \mapsto \lambda^2/\sigma^4$ . Throughout this chapter, we assume that the columns of the design matrix  $\mathbf{X} = (X_1, \dots, X_p)$  have been standardised, so that a single RKHS scale parameter is sufficient for the  $p$  covariates.

The topic of interest for this chapter is model selection for linear regression models. That is, from a set of  $p$  covariates or predictors  $\{X_1, \dots, X_p\}$ , the task is to determine the best choice of subset(s) of variables that should be included in a regression model used to explain the variation in the response variable. As such, the term *variable selection* is synonymous to model selection for linear regression models. Fundamental to this notion of variable selection is an inherent belief in sparseness of the true data generative process surrounding the response variable, i.e. not all of the variables need be used to predict the response. Model selection is indeed a huge topic to cover fully. We broadly

classify variable selection into three categories: 1) (pairwise) model comparison using some criterion; 2) shrinkage to induce sparsity; and 3) Bayesian model selection. We understand that different categorisations and hence categories of model selection exist in the literature, but our focus is on the discussion of the three types as mentioned.

Model selection criteria, both from a frequentist and Bayesian standpoint, can either be of a predictive nature ( $R^2$ , mean squared error of prediction (MSEP),  $C_p$  ([mallows1973some](#)),  $k$ -fold cross-validation MSEP, etc.), or a likelihood-based information criterion (likelihood ratios, Bayes factors, Akaike information criterion (AIC, [akaike1973](#)), Bayesian information criterion (BIC, [schwarz1978estimating](#)), etc.). Selecting a model based on either of these criteria requires comparison of all  $2^p$  criteria, which is not feasible for large  $p$ . Typically, these criteria are used in conjunction with stepwise procedures such as forward-selection or backward-deletion to restrict attention to a smaller number of potential subsets ([George1993](#); [miller2002subset](#)).

On the other hand, regularised least squares regression (ridge regression ([hoerl1970ridge](#)), Lasso ([tibshirani1996regression](#)), or a convex combination of the two via elastic nets ([zou2005regularization](#)), etc.) provide additional information to the regression model in order to provide a sparse solution to linear system of equations in  $\beta$ . These methods are proven to be popular as they are fast and perform exceptionally well in many situations, even in cases where  $p > n$ . Additionally, the Lasso produces solutions for  $\beta$  which are exactly zero. However, the Lasso in general produces estimates which are biased towards zero, are inconsistent, and have no valid standard errors ([friedman2001elements](#); [kyung2010penalized](#)). Further criticisms of the Lasso include its inability to select more than  $n$  predictors in a  $p > n$  situation, and poor performance when multicollinearity exists among the covariates.

From a Bayesian perspective, regularisation is akin to placing priors on the  $\beta_k$ 's to shrink the effects of the  $\beta_k$ 's: the ridge regression has a Bayesian interpretation of placing normal priors on the regression coefficients, while the Lasso a Laplace or double exponential prior ([park2008bayesian](#)). The term adaptive shrinkage has been used for the method in which hyperpriors are placed on the scale parameter of the prior for the  $\beta_k$ 's. The idea is to adaptively shape the prior depending on the importance of the variable in the regression model. Bayesian shrinkage includes the task of specifying tuning parameters, which could potentially affect chain mixing in a Markov chain Monte Carlo method (MCMC) procedure (which is often used).

Bayesian model selection is probabilistic in nature: a priori, one assigns probabilities over the set of models, and then after observing the data, posterior model probabilities (PMPs) are used to discern which of the models was likeliest to have been behind the data generative process of the observed responses. Of course, with large  $p$  then calculation of all  $2^p$  posterior model probabilities to ascertain which is highest will be

a challenge, if not impossible. But, as with most Bayesian applications, MCMC can be applied as a practical means of overcoming this intractability. This stochastic approach to variable selection was pioneered by [George1993](#), and studied by others such as [Kuo1998](#); [dellaportas2002bayesian](#); [Ntzoufras2008](#). Unlike shrinkage methods, Bayesian model selection is able to quantify the amount of times a variable “enters the model” (inclusion probabilities), and thereby measuring its worth as a predictor.

Note that, in addition to model probabilities and inclusion probabilities, estimates of regression coefficients are obtained simultaneously in Bayesian variable selection. When several competing models have high posterior probabilities, regression coefficients from each model, or indeed any quantity of interest, may be combined and weighted using their posterior model probabilities, a technique known as *Bayesian model averaging* ([madigan1994model](#); [hoeting1999bayesian](#)). Averaging over a set of models takes into account the uncertainty surrounding model selection, which other standard statistical procedures ignore upon selection of a single model from which to do inference. It is known to be the case that predictive accuracy of the model-averaged quantity is improved, as measured by a logarithmic scoring rule ([raftery1997bayesian](#)).

Bayesian model selection is not without criticism, however. For complex models with many predictors or samples, MCMC is slow and may mix poorly ([OHara2009](#)). Often, there are a lot of tuning parameters that need to be set correctly for the problem at hand. Also, the choice of priors for model parameters affects consistency of Bayesian model selection procedures. Specifically, improper priors cannot be used to calculate posterior model probabilities ([casella2009consistency](#))—otherwise, one risks running into Lindley’s paradox<sup>1</sup> ([lindley1957statistical](#)).

The plan for this chapter is to describe a fully Bayesian model for variable selection using I-priors. The approach that we take is a stochastic search of the model space due to [Kuo1998](#), realised through a simple Gibbs sampling procedure. The main motivation behind using I-priors in Bayesian variable selection is its suitability in accommodating to datasets with strong multicollinearity and being able to run with little to no prior information about the parameters. A simulation study is conducted and several real-world examples presented to demonstrate this fact.

---

<sup>1</sup>Briefly, in testing a point null hypothesis of the mean of a normally distributed parameter, the null hypothesis is increasingly accepted as the prior variance of the parameter approaches infinity, regardless of evidence for or against the null. The paradox is also termed Jeffreys-Lindley paradox ([robert2014jeffreys](#)).

## 6.1 Preliminary: model probabilities, model evidence and Bayes factors

The paradigm of model selection is as follows. From a finite set of models  $\mathcal{M} = \{M_1, \dots, M_K\}$ , pairs of data  $\{(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)\}$ ,  $y_i \in \mathbb{R}$  and  $\mathbf{x}_i \in \mathcal{X} \equiv \mathbb{R}^p$ , had been generated according to the generative process dictated by one of the models  $M_k \in \mathcal{M}$  and its respective parameters  $\Theta_k$ . Having observed only this data set, the goal is to infer which of the models had generated the data, and consequently obtain estimates for the parameters. It is perhaps most natural to ponder which of the models is most likely to be the “true” one given the data presented, and thus, this natural way of thinking leads one to the concept of *model probabilities*. From a Bayesian perspective in particular, *posterior model probabilities* allow us to quantify the certainty to which any model is behind the data generative process, after taking into account relevant evidence (observation of the data) and prior beliefs about model and parameter uncertainty.

Let  $p(M_1), \dots, p(M_K)$  be prior probabilities assigned to the model space  $\mathcal{M}$ , and  $p(\Theta_k|M_k)$  be the prior on the parameters of model  $M_k$ . For any model  $M_k \in \mathcal{M}$ , the posterior model probability for model  $m$  is

$$p(M_k|\mathbf{y}) = \frac{p(\mathbf{y}|M_k)p(M_k)}{\sum_{k=1}^K p(\mathbf{y}|M_k)p(M_k)} \quad (6.2)$$

where

$$p(\mathbf{y}|M_k) = \int p(\mathbf{y}|M_k, \Theta_k)p(\Theta_k|M_k) d\Theta_k \quad (6.3)$$

is known as the marginal likelihood, or *evidence*, for model  $M_k$ . As a remark, the prior distributions for the parameters do not necessarily need to depend on the model, so we might have that  $p(\Theta_k|M_k) = p(\Theta_k)$ . A natural strategy for model selection is to select the model such that  $p(M_k|\mathbf{y})$  is largest (the *highest probability model*, HPM), but several models rather than just a single one may be reported to convey model uncertainty ([Chipman2001](#)).

Note, that models may be pairwise compared based on these posterior model probabilities, for which the posterior odds

$$\frac{p(M_k|\mathbf{y})}{p(M_0|\mathbf{y})} = \underbrace{\frac{p(\mathbf{y}|M_k)}{p(\mathbf{y}|M_0)}}_{\text{Bayes factor}} \times \underbrace{\frac{p(M_k)}{p(M_0)}}_{\text{prior odds}} \quad (6.4)$$

provide a point summary for comparing model  $M_k$  against model  $M_0$ . The first term on the right-hand side is the Bayes factor for comparing any model  $M_k \in \mathcal{M}$  to another model  $M_0 \in \mathcal{M}$ , and is denoted by  $\text{BF}(M_k, M_0)$ . Thus, model selection based on

posterior model probabilities can be formalised as the Bayesian alternative to classical hypothesis testing using Bayes factors ([kass1995bayes](#)).

The issue that is faced with Bayesian model selection is that all posterior model probabilities must be calculated in order for a full comparison to be made. When the model space is very large, this can prove to be an insurmountable task. In the case of linear regression, where each of the  $p$  variables may be selected or not, the size of the model space is  $2^p$ . Even for moderate sized  $p$  this can already be a challenge computationally. In the coming sections, we shall see that this problem is alleviated by the use of MCMC methods to evaluate posterior model probabilities.

## 6.2 The Bayesian variable selection model

We shall loosely refer to a model as a subset of variables selected from the full set of variables  $\{X_1, \dots, X_p\}$ . It would be useful to be able to index each of these  $2^p$  possible models somehow, and we achieve this by the use of indicator variables  $\gamma \in \{0, 1\}^p$ . Let  $\gamma_j = 1$  if the variable  $X_j$  is selected, and  $\gamma_j = 0$  otherwise, for  $j = 1, \dots, p$ . As an example, the full model, where all the variables are included in the model, is denoted by  $\gamma = (1, \dots, 1)$ , while the intercept only model is denoted by  $\gamma = (0, \dots, 0)$ . Note that we do not consider the intercept to be selectable.

Following [Kuo1998](#), the linear model in (6.1) is expanded to include the indicator variables to form

$$y_i = \alpha + \sum_{k=1}^p x_{ik}\gamma_k\beta_k + \epsilon_i \quad (6.5)$$

$$\epsilon_i \stackrel{\text{iid}}{\sim} N_n(0, \sigma^2).$$

Hence, in addition to the usual model parameters  $(\beta, \sigma, \alpha)$ , we are also interested in conducting model inferences through the posterior distribution of the  $\gamma$ 's. The priors for the parameters are described below:

- **Model indicators**  $\gamma_j$ . An independent Bernoulli prior is specified for the model indicators

$$p(\gamma) = \prod_{j=1}^p \pi_j^{\gamma_j} (1 - \pi_j)^{1 - \gamma_j}. \quad (6.6)$$

We may choose to set all  $\pi_j = 0.5$  a priori to reflect equally likely probabilities that any model may be chosen. Alternatively, we might have some subjective beliefs about which predictor is more likely or unlikely to be included in the model. We may also choose to include  $\pi_j$  in the estimation procedure by assigning a hyperprior on  $\pi_j$  such as the Beta(1, 1) (uniform distribution), Beta(1/2, 1/2) (Jeffreys prior),

or any other suitable hyperprior. In any case, in this thesis we consider the simplest case of setting all  $\pi_j = 0.5$ .

- **Regression coefficients  $\beta$ .** The **Kuo1998** model is often known as the independent sampler due to the independence of model parameters and the indicator variables, i.e.,  $p(\beta, \gamma) = p(\beta)p(\gamma)$ . As such, prior choices for the regression coefficients can be any of the usual priors on  $\beta$ , including
  - the independent prior  $\beta \sim N_p(\mathbf{0}, c^2 \mathbf{I}_p)$  for some choice of  $c$  (e.g.  $c = 10$ );
  - the  $g$ -prior  $\beta|\sigma, g \sim N_p(\mathbf{0}, g\sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$  for some  $g$  either chosen a priori or estimated (Bayes or empirical Bayes); or
  - the I-prior  $\beta|\sigma, \kappa \sim N_p(\mathbf{0}, \kappa\sigma^2 \mathbf{X}^\top \mathbf{X})$ , which is the focus of this chapter.
- **Intercept  $\alpha$ .** A normal prior  $\alpha \sim N(0, \sigma^2 A)$ .
- **Scale  $\sigma$ .** An inverse gamma prior  $\sigma \sim \Gamma^{-1}(c, d)$ .

Priors for the intercept and scale parameters are chosen so as to maintain conjugacy to the normal regression model. Choices for the prior hyperparameters depend on the user’s prior beliefs, but it is reasonable to set vague and uninformative hyperparameters to let the data speak as much as it can, especially in the absence of prior information. With this in mind, we may choose large values of  $A$  (e.g. 100) and small values of the shape and scale parameters for the inverse gamma (e.g. 0.001). Note that as  $c, d \rightarrow 0$  in the inverse gamma distribution we get the Jeffreys prior<sup>2</sup> for scale parameters.

*Remark 6.1.* The BVS model (6.5) together with the choice of Bernoulli priors on  $\gamma$  and a normal prior  $N_p(\mathbf{0}, \mathbf{V}_\beta)$  for  $\beta$  can be seen a *spike-and-slab prior* prior for linear regression models, a mixture of a point mass at zero and a normal density (**mitchell1988bayesian**; **geweke1996variable**). Write  $\boldsymbol{\theta} = (\gamma_1\beta_1, \dots, \gamma_p\beta_p)^\top$ , which are interpreted as the “model-specific” regression coefficients. Then, the prior on  $\boldsymbol{\theta}$  is equivalently written

$$\boldsymbol{\theta}|\gamma \sim \begin{cases} N_p(\mathbf{0}, \mathbf{V}_\beta) & \text{w.p. } p(\gamma) \\ 0 & \text{w.p. } 1 - p(\gamma). \end{cases}$$

A subtle fact of these spike-and-slab priors is that the posterior distribution for  $\boldsymbol{\theta}$  will also be a combination of a point mass and a normal density (with appropriate posterior parameters). Looking at it from this perspective, regression coefficients are assigned zero values with positive probability, and it is this fact that allows covariates to be dropped from the model. As pointed out by **Kuo1998**, the form of the variable selection model allows the selection of important variables, while simultaneously shrinking the coefficients via prior information.

---

<sup>2</sup>The Jeffreys prior for a parameter  $\theta$  is defined as  $p(\theta) \propto |\mathcal{I}(\theta)|^{1/2}$  (**jeffreys1946invariant**).

64 '()  
 \_tl

### 6.3 Gibbs sampling for the I-prior BVS model

The Bayesian variable selection model can be estimated using Gibbs sampling, as demonstrated originally by [Kuo1998](#). In this section, we describe the Gibbs sampling procedure to obtain posterior samples of the parameters. For the I-prior specifically, the joint density of the responses and the priors is

$$p(\mathbf{y}, \gamma, \boldsymbol{\beta}, \alpha, \sigma^2, \kappa) = p(\mathbf{y}|\gamma, \boldsymbol{\beta}, \alpha, \sigma^2)p(\boldsymbol{\beta}|\sigma^2, \kappa)p(\alpha|\sigma^2)p(\gamma)p(\sigma^2)p(\kappa),$$

where the distribution of the model  $p(\mathbf{y}|\gamma, \boldsymbol{\beta}, \alpha, \sigma^2)$  and of the priors have been described in the previous section (except for  $\kappa$ , which we now assign an inverse gamma distribution). Let us denote  $\Theta = \{\alpha, \boldsymbol{\beta}, \gamma, \sigma^2, \kappa\}$  to be the full set of parameters that we wish to obtain posterior samples for. Starting with suitable initial values  $\Theta^{(0)}$ , we then proceed to obtain samples  $\Theta^{(1)}, \dots, \Theta^{(T)}$  by sampling each parameter from the conditional posterior density of that parameter given the rest of the parameters. A suggested set of initial values are the maximum likelihood (ML) estimates of  $\Theta$  or the posterior mean estimate of  $\Theta$  under the full model  $\gamma = (1, \dots, 1)$  after an initial MCMC run.

The Gibbs conditional densities are straightforward to obtain on account of model conjugacy (details of the derivation are given in [Appendix I, p. 297](#)). We start with  $\boldsymbol{\beta}$ . The conditional density of  $\boldsymbol{\beta}$  given  $\alpha, \gamma, \sigma^2, \kappa$  is multivariate normal with mean  $\tilde{\mathbf{B}}(\mathbf{y} - \alpha \mathbf{1}_n)$  and covariance matrix  $\sigma^2 \tilde{\mathbf{B}}$ , where  $\tilde{\mathbf{B}} = \mathbf{X}_\gamma^\top \mathbf{X}_\gamma + (\kappa \mathbf{X}^\top \mathbf{X})^{-1}$ , and  $\mathbf{X}_\gamma = (\gamma_1 X_1 \cdots \gamma_p X_p)$ . Interestingly, when  $X_j$  is dropped from the model ( $\gamma_j = 0$ ), the posterior mean and variance for  $j$ 'th component of  $\boldsymbol{\beta}$  is entirely informed by the prior ([Kuo1998](#)). The data-driven I-prior incorporates information from the covariates into the prior, which then informs the posterior. In a similar manner, the conditional density for the intercept  $\alpha$  is found to be  $N(\sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\theta})/\tilde{A}, \sigma^2 \tilde{A})$ , where  $\tilde{A} = n + A^{-1}$  and  $A$  is the prior variance for  $\alpha$ .

The (conditional) posterior samples of  $\gamma = (\gamma_1, \dots, \gamma_p)$  are obtained componentwise, and each conditional probability mass function for  $\gamma_j$  is Bernoulli with success probability  $\tilde{\pi}_j = u_j/(u_j + v_j)$ , where

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

and

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

Here, we have used the notation  $\boldsymbol{\theta}_j^{[1]}$  to indicate the vector  $\boldsymbol{\theta}$  with the  $j$ 'th component replaced by  $\beta_j$ , and  $\boldsymbol{\theta}_j^{[0]}$  to indicate the vector  $\boldsymbol{\theta}$  with the  $j$ 'th component replaced by 0. Values of 1 for  $\gamma$  are more likely to be sampled when the ratio  $u_j/v_j$  is greater than the prior odds  $\pi_j/(1 - \pi_j)$ . Specifically when the prior probabilities  $\pi_j$  are all set to be 0.5, then  $\gamma_j$  will be more likely to be sampled as '1' if  $u_j > v_j$ , i.e. if the residual sum of squares (RSS)  $\|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2$  is *smaller* when the  $j$ th component is non-zero, compared to the RSS when the  $j$ 'th component of  $\boldsymbol{\theta}$  is zero.

We can in fact draw parallels to a Bayesian hypothesis test, with the null hypothesis being  $H_0 : \beta_j = 0$  and the alternative being  $H_1 : \beta_j \neq 0$ , conditional on knowing all other values of the parameters. Under  $H_k$ ,  $\mathbf{y}|\Theta \sim N_n(\alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\theta}_j^{[k]}, \sigma^2\mathbf{I}_n)$ ,  $k = 0, 1$ . The conditional Bayes factor comparing the model in the alternative hypothesis  $M_1$  to the model in the null hypothesis  $M_0$  is therefore

$$BF(M_1, M_0) = \frac{u_j/\pi_j}{v_j/(1 - \pi_j)} = \frac{\tilde{\pi}_j}{1 - \tilde{\pi}_j} \Bigg/ \frac{\pi_j}{1 - \pi_j}.$$

Thus, it can be seen that the decision to include or exclude the  $j$ 'th variable from the model relates a hypothesis test using the Bayes factor rule, and this decision is embedded in the conditional posterior probabilities  $\tilde{\pi}_j$ . The Gibbs sampling procedure does something that can be described as "an automated stochastic F-test for subset selection" (**Kuo1998**).

Both scale parameters  $\sigma^2$  and  $\kappa$  follow the conditional inverse gamma distributions

$$\begin{aligned} \sigma^2 | \alpha, \beta, \gamma, \kappa &\sim \Gamma^{-1}(n/2 + c_\sigma + 1, \|\mathbf{y} - \alpha\mathbf{1}_n - \mathbf{X}\boldsymbol{\theta}\|^2/2 + d_\sigma) \\ \text{and} \\ \kappa | \alpha, \beta, \gamma, \sigma^2 &\sim \Gamma^{-1}(p/2 + c_\kappa + 1, \boldsymbol{\beta}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \boldsymbol{\beta} / \sigma^2 + d_\kappa). \end{aligned}$$

Note that the inverse gamma distribution that we specify here is defined by its shape and scale parameter, and has the density function described in [Appendix C.6](#). Here,  $\{c_\sigma, d_\sigma\}$  and  $\{c_\kappa, d_\kappa\}$  are the shape and scale hyperparameters of the inverse gamma priors on  $\sigma^2$  and  $\kappa$  respectively.

## 6.4 Posterior inferences

Having obtained posterior samples  $\Theta^{(t)} = \{\alpha^{(t)}, \boldsymbol{\beta}^{(t)}, \gamma^{(t)}, \sigma^{2(t)}, \kappa^{(t)}\}$ , there are two quantities of interest in relation to model inferences. The first is an estimate of posterior

model probabilities, given by

$$\hat{P}(\gamma = \gamma' | \mathbf{y}) = \frac{1}{T} \sum_{i=1}^T [\gamma^{(t)} = \gamma'], \quad (6.7)$$

where  $[\cdot]$  is the Iverson bracket. This gives an estimate of the probability of a model coded by  $\gamma'$  appearing in the posterior state space of models. The second is a quantification of the posterior inclusion for each of the  $p$  variables  $X_1, \dots, X_p$ , known as *posterior inclusion probabilities* (PIPs) for a variable being selected in any model. This is given by

$$\hat{P}(\gamma_j = 1 | \mathbf{y}) = \frac{1}{T} \sum_{i=1}^T [\gamma_j^{(t)} = 1], \quad j = 1, \dots, p. \quad (6.8)$$

Posterior inclusion probabilities are the marginals of the posterior model probabilities across each variable.

Table 6.1: Illustration of samples of  $\gamma$  from the Gibbs sampler for  $p = 3$ . As an example, to estimate the posterior model probability of  $\{X_1, X_3\}$ , we count the occurrences of the combination  $\gamma^{(t)} = (1, 0, 1)$  in the sample and divide by  $T$ . To estimate posterior inclusion probabilities for any of the three variables, we take the sample mean of the binary variates column-wise.

| $t$      | $\gamma_1^{(t)}$ | $\gamma_2^{(t)}$ | $\gamma_3^{(t)}$ |
|----------|------------------|------------------|------------------|
| 1        | 1                | 0                | 1                |
| 2        | 1                | 0                | 0                |
| 3        | 1                | 1                | 0                |
| $\vdots$ | $\vdots$         | $\vdots$         | $\vdots$         |
| $T$      | 1                | 0                | 1                |

Note, that the regression coefficient of interest is not  $\boldsymbol{\beta}$ , but rather the “model averaged” regression coefficients  $\boldsymbol{\theta} = (\gamma_1\beta_1, \dots, \gamma_p\beta_p)^\top$  ([madigan1994model](#)). Posterior variances for  $\boldsymbol{\theta}$  will typically be larger than variances for  $\boldsymbol{\beta}$ , because posterior estimates surrounding  $\boldsymbol{\theta}$  will have incorporated model uncertainty, but  $\boldsymbol{\beta}$  on the other hand, will not. Thus, any inferential procedure surrounding the regression coefficients avoids the risk of over-confidence. Note that, since  $\boldsymbol{\theta}$  will contain values of exactly zero when predictors are dropped out of the model, the posterior density for  $\boldsymbol{\theta}$  is a mixture of a point mass at zero and a normal density. In any case, the likelihood only provides sufficient information to identify the product of  $\boldsymbol{\beta}$  and  $\gamma$ , but not each of them separately ([Kuo1998](#)).

*Remark 6.2.* The intention of computing model-averaged regression coefficients  $\boldsymbol{\theta}$  is solely for the inclusion of model uncertainty. There is a strong agreement in the Bayesian

variable selection literature that that such coefficients are practically meaningless when it comes to explanatory inferences. **banner2017considerations** writes that “regression coefficients... may not hold equivalent interpretations across all of the models in which they appear”, and one reason for this might be “interpretation of partial regression coefficients can depend on other variables that have been included in the model”. The use of model-averaged effect sizes may result in misleading inferences (**cade2015model**).

Finally, any quantity of interest  $\Delta$  can be incorporated as part of the Gibbs sampling procedure. That is, at each Gibbs iteration  $t = 1, \dots, T$ , calculate  $\Delta^{(t)}$  as a function of the parameter values at iteration  $t$ . This can be done during the Gibbs sampling process, or even after the fact as part of a post-processing procedure. Any inference on the posterior of  $\Delta$  will then have incorporated the model uncertainty from a model averaging standpoint, as discussed earlier. As an example, suppose we are interested in the predicted value at a new covariate value  $\mathbf{x}_{\text{new}} \in \mathbb{R}^p$ . For each Gibbs sample, calculate

$$y_{\text{new}}^{(t)} = \alpha^{(t)} + \mathbf{x}_{\text{new}}^\top (\gamma_1 \beta_1, \dots, \gamma_p \beta_p),$$

and obtain a point estimate  $\hat{y}_{\text{new}}^{(t)}$  using the posterior mean or mode. The uncertainty for this estimate is contained in the standard deviation calculated from the sample  $y_{\text{new}}^{(1)}, \dots, y_{\text{new}}^{(T)}$ , from which a 95% credibility interval for this estimate can be obtained from the empirical upper and lower 0.025 cut off points.

## 6.5 Two stage procedure

The variable selection procedure can be improved by a “preselection” of variables to trim off unimportant variables which reduces the size of the model space being explored. Without appealing to other external preselection methods, there is actually information that we could use from Bayesian variable selection models in the form of posterior inclusion probabilities. The procedure would work as follows:

1. Run the Bayesian variable selection model and obtain posterior inclusion probabilities for each variable.
2. Discard variables with inclusion probabilities less than a certain threshold,  $\tau$ .
3. Re-run the Bayesian variable selection model on the set of reduced variables.

A natural choice for  $\tau$  would be 0.5, and therefore a two-stage approach to Bayesian variable selection can then be motivated as selecting the subset of variables which constitutes what is known as the *median probability model*. The median probability model is obtained by selecting all variables with a posterior inclusion probability of greater than or equal to a half. **Barbieri2004** show that the median probability model has

the property of being optimally predictive (minimises squared error loss for predictions) under certain strict conditions.

The notion of a two-stage approaches are not new, as many variable selection methods in the literature generally employ a preselection method of some kind before running their selection process proper. This can be based on subjective preconceptions about which variables to retain, substantive theory, or even an objective preselection criterion. Two-stage procedures for Bayesian variable selection models have been used in works by **Fouskakis2008** and **Ntzoufras2008**.

In the simulation studies conducted and observations from real-data examples, this two-stage approach does seem to provide a benefit. The complexity of estimating all model probabilities grows exponentially with  $p$ , therefore reducing this benefits the model selection procedure because the search of the model space is less cluttered. Of course, this idea works if the “correct” variables are deleted when proceeding to the second stage. We posit that the  $p$  posterior inclusion probabilities are easier to estimate than the  $2^p$  posterior model probabilities from the same amount of information coming from the MCMC samples. As a result, information summarised through the posterior inclusion probabilities are more precise than the posterior model probabilities.

## 6.6 Simulation study

In this section, we conduct a simulation study to compare the performance of different priors in the Bayesian variable selection framework described above. The priors on  $\beta$  that are compared are those mentioned in Section 6.2, i.e. the I-prior, the independent prior with large prior variance (flat/uninformative prior), and the  $g$ -prior with  $g = n$  (unit information prior, **Ntzoufras2008**). We also make a comparison the variable selection performance of the Lasso, which, from a Bayesian perspective, is similar to setting a double-exponential or Laplace priors on the regression coefficients (**park2008bayesian**). For clarity, the Lasso model employed in the simulations is of a frequentist regularisation framework as per **tibshirani1996regression**, and is neither a Bayesian variable selection model as described earlier, nor a fully Bayes implementation as per **park2008bayesian**. We felt it interesting to compare the Lasso as it is widely used for variable selection of linear models.

The experiment is to select from  $p = 100$  variables of a  $n = 150$  sample size artificial dataset which has pairwise correlations between the variables. This was inspired by the studies done by **George1993** and **Kuo1998** in their respective papers, albeit on a larger scale (in theirs,  $p = 30$ ). Five different scenarios were looked at. For each scenario, only  $s$  out of 100 variables were selected to form the “true” model and generate the responses according to the linear model  $\mathbf{y} \sim N_{100}(\mathbf{X}\beta, \sigma^2\mathbf{I}_{150})$ . The signal-to-noise ratio (SNR)

as a percentage is defined as  $s\%$ , and the five scenarios are made up of varying SNR from high to low: 90%, 75%, 50%, 25%, and 10%. Variables that were included in the model had true  $\beta$  coefficients equal to one. That is,  $\beta_{\text{true}} = (\mathbf{1}_s, \mathbf{0}_{100-s})^\top$ , where  $\mathbf{1}_s$  is a row-vector of  $s$  ones, and  $\mathbf{0}_{100-s}$  is a row-vector of  $100 - s$  zeroes. The data generation process is summarised as follows:

- Draw  $\mathbf{Z}_1, \dots, \mathbf{Z}_{100} \stackrel{\text{iid}}{\sim} N_{150}(\mathbf{0}, \mathbf{I}_{150})$ .
- Draw  $\mathbf{U} \sim N_{150}(\mathbf{0}, \mathbf{I}_{150})$ .
- Set  $\mathbf{X} = (\mathbf{Z}_1 + \mathbf{U}, \dots, \mathbf{Z}_{100} + \mathbf{U})$ . This induces pairwise correlations of about  $1/2$  between the columns of  $\mathbf{X}$ .<sup>3</sup>
- Draw  $\mathbf{y} \sim N_{150}(\mathbf{X}\beta_{\text{true}}, \sigma^2 \mathbf{I}_{150})$ , with  $\sigma = 2$ .

In each scenario, we are interested in obtaining the highest probability model and counting the number of false choices made in this model after a two-stage procedure of variable selection. False choices can either be selecting variables wrongly (false inclusion) or failing to select a variable (false exclusion). Each scenario was repeated a total of 100 times to account for variability in the data generation process, and the results averaged. A summary of the results is presented in Table 6.2. The overall results are also plotted in the form a frequency polygon (see Figure 6.1).

The simulation results seem to indicate that the I-prior performs consistently well across all five scenarios, making no more than five false choices out of 100 (i.e. a 95% correct selection rate) in at least 82% of the time in the worst scenario. We do not observe much difference between the  $g$ -prior and the independent prior, and while they behave poorly in high SNR scenarios, these two priors seem to perform extremely well in low SNR scenarios. A high propensity to drop variables in these scenarios is a likely explanation, which does not necessarily indicate good performance—they perform well by contentiously omitting of a large number of unnecessary variables, especially in a two-stage procedure. Finally, the Lasso is well known to yield poor selection performance under multicollinearity, so the results are expected. The Lasso procedure was not subject to a two-stage approach because the Lasso does not provide information regarding posterior inclusion probabilities for individual variables.

We also inspect the sensitivity of the hyperprior choice of  $\pi_j$  for the indicator variables on the number of false choices made. Figure 6.2 plots the mean number of false choices made in each of the five SNR scenarios with varying hyperprior setting for  $\pi_j$ . From the plot, it is seen that for high SNR scenarios, setting  $\pi_j$  too low increases the number of false exclusions. Conversely, for low SNR scenarios, setting  $\pi_j$  too high increases the number of false inclusions. This makes sense: when the true model size is small,

---

<sup>3</sup>For any row of  $\mathbf{X}$ ,  $\text{Cov}[X_j, X_k] = \text{Cov}[Z_j + U, Z_k + U] = \text{Var}[U] = 1$ , and  $\text{Var}[X_j] = \text{Var}[Z_j + U] = 2$ . Thus,  $\text{Corr}[X_j, X_k] = \text{Cov}[X_j, X_k]/(\text{Var}[X_j]\text{Var}[X_k])^{1/2} = 1/2$ .

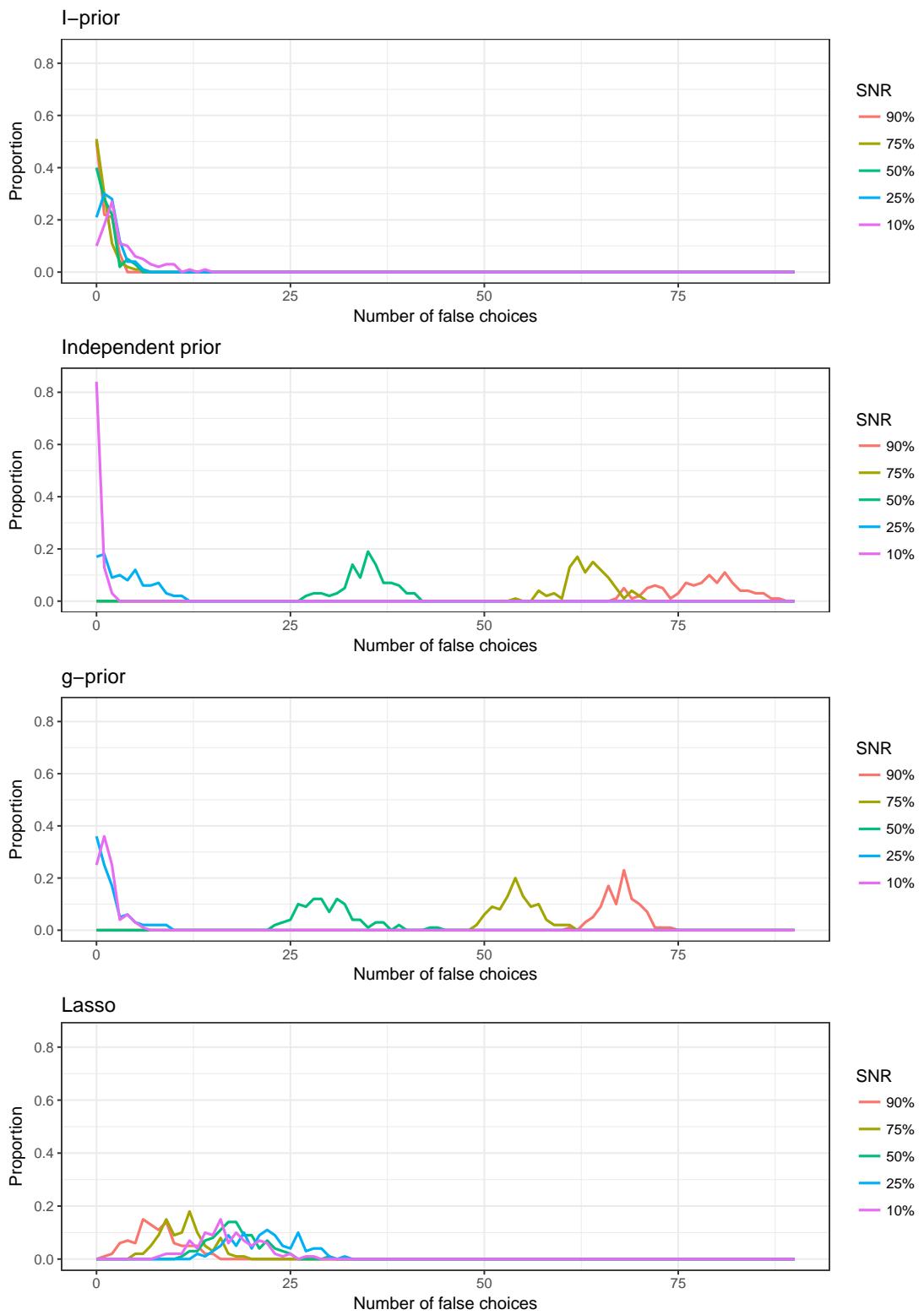


Figure 6.1: Frequency polygons for the number of false choices for each of the four priors. The I-prior performs robustly well across the five scenarios tested, mostly yielding five or fewer false inclusions or exclusions. Spurious exclusions led to the independent and *g*-prior simultaneously performing well in low SNR and badly in high SNR scenarios. The Lasso is known to be unreliable in the presence of collinearity.

Table 6.2: Simulation results (proportion of false choices) for the Bayesian variable selection experiment using the I-prior, an independent prior, the  $g$ -prior and the Lasso across varying SNR.

| False choices     | Signal-to-noise ratio |                    |                    |                    |                    |
|-------------------|-----------------------|--------------------|--------------------|--------------------|--------------------|
|                   | 90%                   | 75%                | 50%                | 25%                | 10%                |
| <i>I-prior</i>    |                       |                    |                    |                    |                    |
| 0-2               | <b>0.93</b> (0.03)    | <b>0.92</b> (0.03) | <b>0.90</b> (0.03) | <b>0.79</b> (0.04) | <b>0.55</b> (0.05) |
| 3-5               | 0.07 (0.03)           | 0.07 (0.03)        | 0.10 (0.03)        | 0.20 (0.04)        | 0.27 (0.04)        |
| >5                | 0.00 (0.00)           | 0.01 (0.01)        | 0.00 (0.00)        | 0.01 (0.01)        | 0.18 (0.04)        |
| <i>Ind. prior</i> |                       |                    |                    |                    |                    |
| 0-2               | 0.00 (0.00)           | 0.00 (0.00)        | 0.00 (0.00)        | <b>0.44</b> (0.05) | <b>1.00</b> (0.00) |
| 3-5               | 0.00 (0.00)           | 0.00 (0.00)        | 0.00 (0.00)        | 0.30 (0.05)        | 0.00 (0.00)        |
| >5                | <b>1.00</b> (0.00)    | <b>1.00</b> (0.00) | <b>1.00</b> (0.00) | 0.26 (0.04)        | 0.00 (0.00)        |
| <i>g-prior</i>    |                       |                    |                    |                    |                    |
| 0-2               | 0.00 (0.00)           | 0.00 (0.00)        | 0.00 (0.00)        | <b>0.78</b> (0.04) | <b>0.86</b> (0.03) |
| 3-5               | 0.00 (0.00)           | 0.00 (0.00)        | 0.00 (0.00)        | 0.14 (0.03)        | 0.13 (0.03)        |
| >5                | <b>1.00</b> (0.00)    | <b>1.00</b> (0.00) | <b>1.00</b> (0.00) | 0.08 (0.03)        | 0.01 (0.01)        |
| <i>Lasso</i>      |                       |                    |                    |                    |                    |
| 0-2               | 0.03 (0.02)           | 0.00 (0.00)        | 0.00 (0.00)        | 0.00 (0.00)        | 0.00 (0.00)        |
| 3-5               | 0.19 (0.04)           | 0.02 (0.01)        | 0.00 (0.00)        | 0.00 (0.00)        | 0.00 (0.00)        |
| >5                | <b>0.78</b> (0.04)    | <b>0.98</b> (0.01) | <b>1.00</b> (0.00) | <b>1.00</b> (0.00) | <b>1.00</b> (0.00) |

then setting  $\pi_j$  too high encourages variables to be retained in the model. While the optimal  $\pi_j$  corresponds directly to the true SNR (e.g. SNR = 10% performs best under  $\pi_j = 0.10$ ), Figure 6.2 makes a case for  $\pi_j = 0.5$  to be a “safe choice” in the face of prior ignorance on model size.

## 6.7 Examples

Now, we apply our I-prior Bayesian variable selection model to three real-world data sets that have all been previously analysed in the variable selection literature. All examples were analysed in R using our **ipriorBVS** package (**jamil2018ripriorBVS**) which contains a wrapper to JAGS (**plummer2003jags**). Reproducible code is available at <http://myphdcode.haziqj.ml>. In all analyses, a two-stage procedure was conducted for the I-prior model, where each stage consists of obtaining 15,000 MCMC samples (including 5,000 for burn-in).

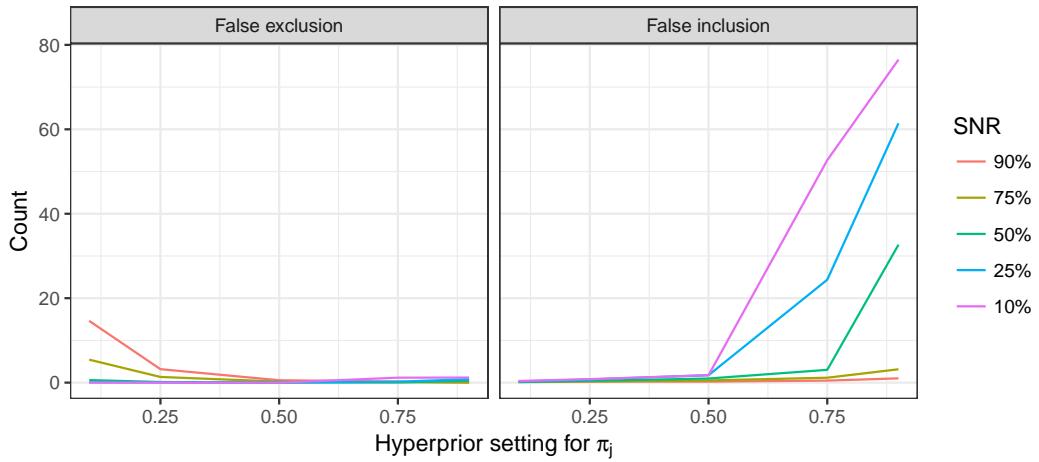


Figure 6.2: Average number of false choices (false inclusions or false exclusions) for the five different scenarios (SNR varied between 90%, 75%, 50%, 25% and 10%) with different hyperprior settings for  $\gamma_j \sim \text{Bern}(\pi_j)$ .

### 6.7.1 Aerobic data set

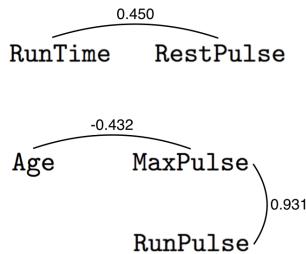


Figure 6.3: The sample correlations of interest in the aerobic fitness dataset. These show variables with correlations greater than 0.4 in magnitude.

This dataset appeared in the *SAS/STAT® User’s Guide* (**SAS2008**) and was also analysed by **Kuo1998**. It involves understanding the factors which affect aerobic fitness, which is measured by the ability to consume oxygen. A sample of  $n = 30$  male participants had their physical fitness measured by means of simple exercise tests. The response variable contains measurement of oxygen uptake rate in mL/kg body weight per minute. The six covariates were the participants’ age ( $X_1$ ), weight ( $X_2$ ), time taken to run one mile ( $X_3$ ), resting heart rate ( $X_4$ ), heart rate while running ( $X_5$ ), and maximum heart rate during the exercise ( $X_6$ ). This dataset, although small in size, is interesting to analyse because of the correlations between the variables, mainly due to the measurements being taken during the same exercise test. The sample correlations of interest are shown in Figure 6.3.

Notice that Table 6.3 reports only on four out of  $2^6 = 64$  possible models, but the sum of the posterior model probabilities add to one. Naturally, models which are deemed important by virtue of data evidence are sampled more often, and in fact, models which

Table 6.3: Results for variable selection of the Aerobic data set. Note that the Bayes factors reported are the Bayes factors comparing any of the models to Model 1 (base model).

|       | PIP   | $\theta$ est. (SD) | Model 1 | Model 2 | Model 3 | Model 4 |
|-------|-------|--------------------|---------|---------|---------|---------|
| $X_1$ | 0.685 | -0.169 (0.14)      | ✓       |         | ✓       |         |
| $X_2$ | 0.205 | -0.017 (0.05)      |         |         |         |         |
| $X_3$ | 1.000 | -0.745 (0.12)      | ✓       | ✓       | ✓       | ✓       |
| $X_4$ | 0.168 | -0.013 (0.05)      |         |         |         |         |
| $X_5$ | 0.663 | -0.163 (0.15)      | ✓       |         |         | ✓       |
| $X_6$ | 0.275 | 0.003 (0.10)       |         |         |         |         |
|       | PMP   |                    | 0.564   | 0.235   | 0.105   | 0.096   |
|       | BF    |                    | 1.000   | 0.418   | 0.187   | 0.170   |

are unpromising may not even get sampled. So, MCMC methods does not need to list out all possible models because models which are never visited in the posterior state space are assigned a probability of zero. The highest posterior model was found to be the model with the variables  $X_1$ ,  $X_3$  and  $X_5$  (PMP = 0.564). In Figure 6.4, we can see that the point mass at zero overwhelms the rest of the values in the density plots for  $X_2$ ,  $X_4$  and  $X_6$ , and hence these variables were dropped.

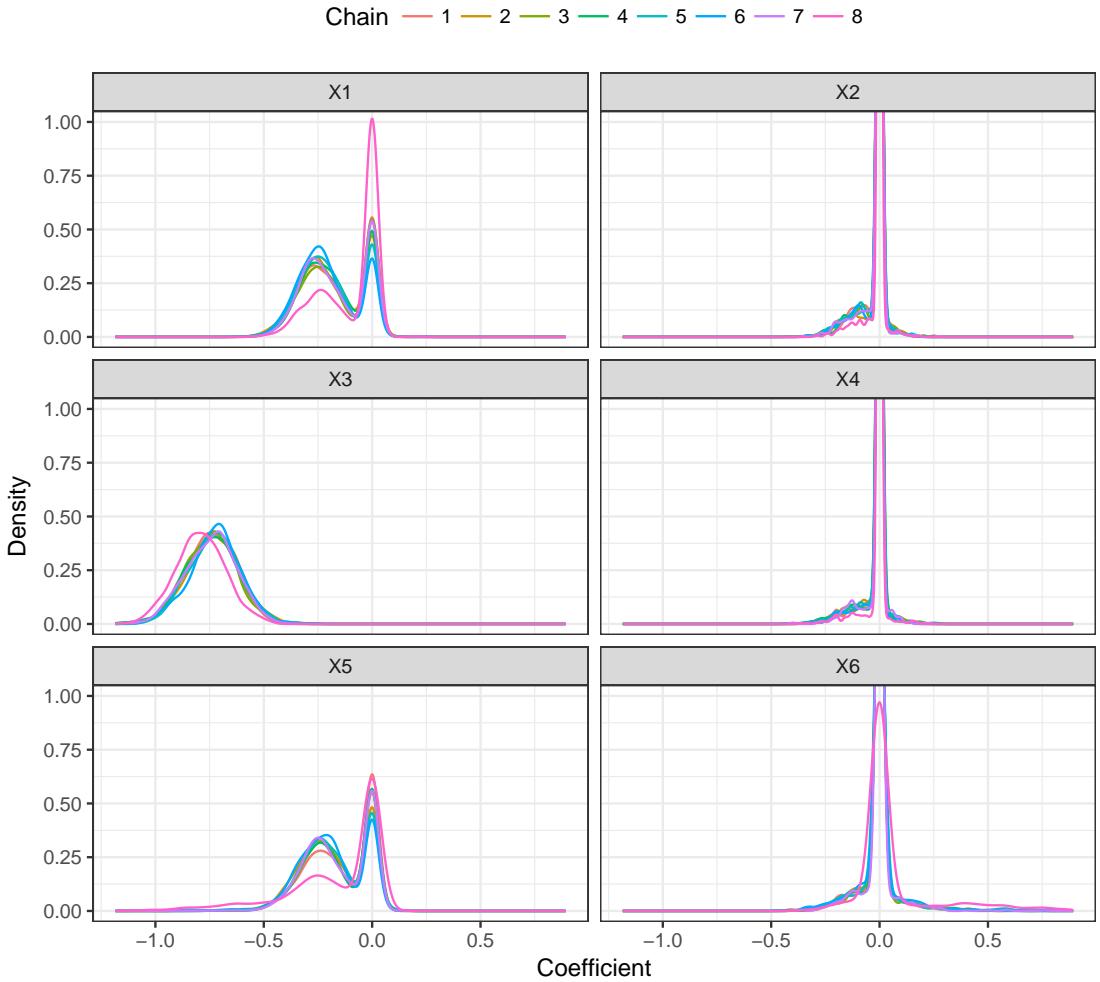


Figure 6.4: Posterior density plots of the regression coefficients  $\theta$  for the aerobic data set. The spike at zero observed in the density plots for  $X_2$ ,  $X_4$  and  $X_6$  is indicative of these variable being dropped often in the posterior samples.

### 6.7.2 Mortality and air pollution data

The next real world application comes from a paper by **McDonald1973**. In it, the effects of air pollution on mortality in a US metropolitan area ( $n = 60$  and  $p = 15$ ) were studied. The response variable is the total age adjusted mortality rate, and the main pollution effects of interest were that of hydrocarbons (HC), oxides of nitrogen ( $\text{NO}_x$ ) and sulphur dioxide ( $\text{SO}_2$ ). Several other environmental and socioeconomic considerations were taken into account, otherwise the model may include unexplained variation caused by factors other than pollution. For example, a metropolitan area with a high proportion of the elderly should expect to have a higher mortality rate than one with a low proportion. All of the variables can be considered as continuous and real; Table 6.4 provides a description of the variables.

Table 6.4: Description of the air pollution data set.

| Variable            | Description                                              |
|---------------------|----------------------------------------------------------|
| Mortality           | Total age adjusted mortality rate                        |
| Precipitation       | Mean annual precipitation (in)                           |
| Relative humidity   | Percent relative humidity, annual average at 1 p.m.      |
| January temperature | Mean January temperature ( $^{\circ}$ F)                 |
| July temperature    | Mean July temperature ( $^{\circ}$ F)                    |
| Population density  | Population per square mile in urbanised area             |
| Household size      | Population per household                                 |
| Education           | Median school years completed for those over 25          |
| Sound housing units | Percentage of sound housing units (no defects)           |
| Age >65 years       | Percent of population that is 65 years of age or over    |
| Non-white           | Percent of urbanised area population that is non-white   |
| White collar        | Percent employment in white-collar urbanised occupations |
| Income <\$3,000     | Percent of families with income under \$3,000            |
| HC                  | Relative population potential of hydrocarbons            |
| NO <sub>x</sub>     | Relative population potential of oxides of nitrogen      |
| SO <sub>2</sub>     | Relative population potential of sulphur dioxide         |

This dataset also contains several highly correlated variables which impedes a meaningful regression analysis. When the full model is fitted using ordinary least squares, none of the pollutant effects were found to be significant. Clearly, a variable selection method was required. **McDonald1973** used a ridge regression technique to determine which variables to select and eliminate “unstable” coefficients found from a trace analysis. In addition, the authors also looked at a variable elimination method based on total squared error via Mallow’s  $C_p$ . The results are summarised in Table 6.5.

In this case, the I-prior BVS model concurred with the overall finding of **McDonald1973**, in that SO<sub>2</sub> was found to be a significant contributing factor towards mortality rates, while the rest of the pollutants were not. the I-prior BVS model also obtained a model with the largest  $R^2$  and the smallest size. We note that the effect size for SO<sub>2</sub> is slightly larger under an I-prior, but generally, the rest of the I-prior coefficients are similar in magnitude and sign to the coefficients of the other two models.

Table 6.5: A comparison of the coefficient values obtained using ordinary least squares (full model), **McDonald1973**'s minimum  $C_p$  and ridge analysis, and the I-prior. Standard errors/posterior standard deviations are given in parentheses. Values shaded grey indicate OLS regression coefficients not significant at the 10% level.

|                              | Full model    | Min. $C_p$    | Ridge         | I-prior       |
|------------------------------|---------------|---------------|---------------|---------------|
| <i>Environmental factors</i> |               |               |               |               |
| Precipitation                | 0.306 (0.14)  | 0.247 (0.07)  | 0.230 (0.07)  | 0.254 (0.12)  |
| Relative humidity            | 0.009 (0.10)  |               |               |               |
| January temperature          | -0.318 (0.18) | -0.164 (0.06) | -0.172 (0.06) | -0.195 (0.11) |
| July temperature             | -0.237 (0.15) | -0.073 (0.07) |               |               |
| <i>Demographic factors</i>   |               |               |               |               |
| Population density           | 0.084 (0.09)  |               | 0.091 (0.06)  |               |
| Household size               | -0.232 (0.15) |               |               |               |
| Education                    | -0.233 (0.16) | -0.190 (0.06) | -0.171 (0.07) | -0.151 (0.12) |
| Sound housing units          | -0.052 (0.15) |               |               |               |
| Age >65 years                | -0.213 (0.20) |               |               |               |
| Non-white                    | 0.640 (0.19)  | 0.481 (0.07)  | 0.462 (0.07)  | 0.517 (0.10)  |
| White collar                 | -0.014 (0.12) |               |               |               |
| Income <\$3,000              | -0.009 (0.22) |               |               |               |
| <i>Pollution potential</i>   |               |               |               |               |
| HC                           | -0.979 (0.72) |               |               |               |
| NO <sub>x</sub>              | 0.983 (0.75)  |               |               |               |
| SO <sub>2</sub>              | 0.090 (0.15)  | 0.255 (0.06)  | 0.232 (0.06)  | 0.302 (0.09)  |
| Size                         | 15            | 6             | 6             | 5             |
| $R^2$                        | 0.764         | 0.541         | 0.553         | 0.676         |

### 6.7.3 Ozone data set

In this section, we replicate the Bayesian variable selection analysis of the Ozone dataset done by **Casella2006** which appeared initially in **Breiman1985**, and also show how Bayesian variable selection can help select important interaction terms. The data consists of daily ozone readings and various meteorological quantities, and the aim was to see which of these quantities contributed to the ozone concentration. The variables considered are explained in Table 6.6.

The data contains 366 points, one for each day of the leap year 1976. There are 163 data points containing missing data on some of the predictors, so we did a complete case analysis on the remaining 203 samples. Out of these 203, we randomly set aside 25 to use for validation, thus the  $n$  used to train the model was  $n = 178$ . The training and test set were repeated multiple times and results averaged in order to make a comparison to the unknown training and test set used in the other studies. Out-of-sample prediction RMSE were obtained, as well as the coefficient of determination  $R^2$ .

Table 6.6: Description of the ozone data set used in this analysis. The data is available from the R package **mlbench** (**mlbench**).

| Variable | Description                                                            |
|----------|------------------------------------------------------------------------|
| $y$      | Daily maximum one-hour-average ozone reading (ppm) at Upland, CA       |
| $X_1$    | Month: 1 = January, ..., 12 = December                                 |
| $X_2$    | Day of month: 1, 2, ...                                                |
| $X_3$    | Day of week: 1 = Monday, ..., 7 = Sunday                               |
| $X_4$    | 500-millibar pressure height (m) measured at Vandenberg Air Force Base |
| $X_5$    | Wind speed (mph) at Los Angeles International Airport (LAX)            |
| $X_6$    | Humidity (%) at LAX                                                    |
| $X_7$    | Temperature ( $^{\circ}$ F) measured at Sandberg, CA                   |
| $X_8$    | Inversion base height (feet) at LAX                                    |
| $X_9$    | Pressure gradient (mmHg) from LAX to Daggett, CA                       |
| $X_{10}$ | Visibility (mi) measured at LAX                                        |
| $X_{11}$ | Temperature ( $^{\circ}$ F) measured at El Monte, CA                   |
| $X_{12}$ | Inversion base temperature (degrees Fahrenheit) at LAX                 |

C&M removed the variables  $X_{11}$  and  $X_{12}$  before running their selection model, citing multicollinearity causing ill-conditioned design matrices. Upon inspection, there are indeed correlations among the variables as high as 0.93 for some of them, but not enough to cause rank deficiency in the design matrix and a degenerate  $\mathbf{X}^T \mathbf{X}$  matrix. The sample correlations  $\widehat{\text{Corr}}(X_7, X_{11}) = 0.91$  and  $\widehat{\text{Corr}}(X_{11}, X_{12}) = 0.93$  seemed to drive the decision to drop the two variables, and while it is a valid concern, we still will conduct variable selection on the full set of 12 variables to see the performance of I-priors in the presence of multicollinearity in this real-world data set. On another note, the variables  $X_1$ ,  $X_2$  and  $X_3$  were presumably intended to be categorical as in modelling seasonality in a time series data, but these were treated as continuous, as did C&M. The results are compared in Table 6.7.

Table 6.7: Results for variable selection of the Ozone data set using only linear predictors.

| Method                   | Variables               | Size | $R^2$ | RMSE  |
|--------------------------|-------------------------|------|-------|-------|
| I-prior                  | $X_1, X_6, X_{11}$      | 3    | 0.708 | 0.554 |
| <b>Casella2006</b> (C&M) | $X_6, X_7, X_8$         | 3    | 0.686 | 0.992 |
| <b>Breiman1985</b> (B&F) | $X_7, X_8, X_9, X_{10}$ | 4    | 0.669 | 1.056 |

What we found was that the model selected using the I-prior does better in terms of  $R^2$  as well as RMSE compared to the methods used by C&M and B&F. The average posterior model probability for  $X_1, X_6, X_{11}$  as found by the I-prior was 0.722<sup>4</sup>. One thing to note is that the I-prior model selected the variable  $X_{11}$  instead of its highly correlated proxy  $X_7$ , which is what C&M selected. These two variables are temperature measurements at different locations in California. As C&M excluded  $X_{11}$  from the model search it was of course never considered in their model selection process, and because

we included it in ours, the variable selection model was able to consider both variables together and decide on the more appropriate one.

Interestingly, the distance as the crow flies between Sandberg, CA (location of temperature measurements for  $X_7$ ) and Upland, CA (location of ozone readings) is roughly 121 km, but El Monte, CA (location of temperature measurements for  $X_{11}$ ) is just 35 km away from Upland, CA. It stands to reason that  $X_{11}$  provides more geographical reliability than  $X_7$ . Unless there is a strong insistence on deleting variables beforehand, we might not know for sure whether the variable was rightfully removed from consideration, as this example seems to prove. Out of curiosity, running the variable selection model on the reduced variable space as C&M did, we arrive at the same results as theirs.

Figure 6.5: Locations<sup>5</sup> of the various points of interest in California, USA, related to the ozone measurements.

We then used the I-prior method to select between the squared terms and all level two interactions, in addition to all the variables, in an effort to improve model prediction. For 12 such variables, the number of variables to select becomes  $12 + 12 + 12(12 - 1)/2 = 90$ . By doing so, we were able to improve the model to give a slightly better predictive ability. The results are shown below in Table 6.8. The I-prior again selected a model which was superior in terms of  $R^2$  and RMSE compared to that obtained by C&M.

Table 6.8: Results for variable selection of the Ozone data set using linear, squared and two-way interaction terms.

| Method  | Variables                                                                     | Size | $R^2$ | RMSE  |
|---------|-------------------------------------------------------------------------------|------|-------|-------|
| I-prior | $X_1, X_5, X_6, X_{11}, X_{12}, X_1^2, X_9^2, X_6X_{11}, X_6X_{12}, X_7X_9$   | 10   | 0.812 | 0.503 |
| C&M     | $X_2, X_1^2, X_7^2, X_9^2, X_1X_5, X_2X_6, X_3X_7, X_4X_6, X_6X_8, X_6X_{10}$ | 10   | 0.758 | 0.873 |

<sup>5</sup>Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under CC BY-SA 3.0. Created using the ggmap package (ggmap) in R.

<sup>5</sup>Since the total model space used was different between our method, C&M and B&F, it does not make sense to compare posterior model probabilities which we obtained. C&M reported a model probability of 0.491 for their model, but this model was not selected at all using the I-prior.

## 6.8 Conclusion

The model selection problem is an important one in statistics, but highly contentious. **miller2002subset** writes that many statisticians view model selection as “unclean” or “distasteful”, and that “terms such as ‘fishing expeditions’, ‘torturing the data until they confess’, ‘data mining’, and others are used as descriptions of these practices”. The disagreement with the principle of model selection stems from the belief in the mantra that models should only be built by thoughtfully choosing variables which are expected to influence the response by appealing to substantive theory, and not by virtue of optimising some model selection criterion. However, variable selection as an exploratory study is certainly justified by many practical applications, especially when there is a genuine desire to know the most reasonable, parsimonious and interpretable model. Through variable selection exercises, we can learn which covariates are important, and which are negligible, in explaining the variation in the response.

The Bayesian variable selection method that we have seen has the appeal of reducing the problem of model search into one of estimation. At the outset, we aimed to seek a model which: 1) requires little tuning on the part of the user; 2) would work well in the presence of multicollinearity; and 3) is able to work well with little to no prior information. The I-prior on the regression coefficients in **Kuo1998**'s (**Kuo1998**) spike-and-slab stochastic search framework achieves this aim.

The attractive feature of a Bayesian approach to variable selection is the ability to simultaneously shrink and select predictors, thereby incorporating model uncertainty in the regressors. Sparsification is not “hard coded”, in the sense that regression coefficients are assigned a value of zero with some positive probability in the posterior. This is unlike the regularisation or penalised log-likelihood approach to variable selection using the Lasso, elastic net, and so on, whereby sparsity is induced at the mode, but not in the posterior distribution (**scott2014predicting**). This translates to being provided with a single variable selection decision, rather than information that is coded through a probability distribution.

We discuss three areas to concentrate on for future research and improvement:

1.  **$p > n$  cases.** Typically, when there is insufficient information in the data to inform the estimation, then additional information is sought from the priors. In our case, the I-prior covariance involves the inverse of a low rank matrix which is not invertible. A  $p$ -variate normal distribution with a singular covariance matrix will only have a probability distribution defined on a low dimensional subspace. The issue may however be computational—it might be worth exploring the generalised inverse, or study ways in which to avoid the inverse computation in the Gibbs sampler. As a matter of fact, we note that the posterior precision for  $\beta$  can be

written as

$$\begin{aligned}\tilde{\mathbf{B}}^{-1} &= (\mathbf{X}_\gamma^\top \mathbf{X}_\gamma + (\kappa \mathbf{X}^\top \mathbf{X})^{-1})^{-1} \\ &= \mathbf{X}_\gamma^\top \mathbf{X}_\gamma ((\mathbf{X}_\gamma^\top \mathbf{X}_\gamma)^2 + \kappa \mathbf{I}_p)^{-1}\end{aligned}$$

which avoids the need for inverting the low-rank matrix  $\mathbf{X}_\gamma^\top \mathbf{X}_\gamma$ .

2. **Improvement in computational time.** Although the model itself is not computationally intensive to run (roughly  $O(np^2)$  in time per Gibbs iteration), the main bottleneck is the reliance on a stochastic sampling algorithm. As in the previous chapter, variational inference is a promising area to look into, especially given that the Gibbs conditional distributions were straightforward to obtain, and these might be similar to a mean-field variational distribution. If this is successful, then it is expected to reduce computational time and avoid convergence issues that comes with traditional MCMCs. Variational inference with spike-and-slab priors on regression coefficients was studied by [ormerod2017variational](#).
3. **Extension to generalised linear models.** [Kuo1998](#) in their paper already provided a sketch of how the variable selection model would work. With the ideas in [Chapter 5](#), we can extend the I-prior variable selection to categorical responses when the continuous latent propensities are modelled using linear functions. Such an approach has been implemented in gene selection studies, for which the variables are gene expressions and the responses are presence of a particular disease ([lee2003gene](#)).

Finally, it should be mentioned that more complex variable selection models can be coded with the  $\gamma$  indicators. For instance, in selecting squared or interaction terms, we can insist on having the model select the main term if the squared or interaction term is selected:

$$y_i = \alpha + \gamma_1 \beta_1 x_{1i} + \gamma_2 \beta_2 x_{2i} + \gamma_1 \gamma_2 \gamma_3 \beta_3 x_{1i} x_{2i}.$$

Or perhaps, we could use a single  $\gamma$  indicator for the dummy variables which make up a single categorical covariate, which we would then infer on the selection of the single covariate rather than each individual category of the covariate.



# Chapter 7

## Summary

The work done in this thesis explores the concept of regression modelling using priors with Fisher information covariance kernels (I-priors, [bergsma2017](#)). It is best seen as a flexible regression technique which is able to fit both parametric and nonparametric models, and bears similarity to Gaussian process regression. For the regression model (1.1) subject to (1.2), stated again here for convenience,

$$y_i = \alpha + f(x_i) + \epsilon_i \quad (\text{from 1.1})$$

$$(\epsilon_1, \dots, \epsilon_n) \sim N_n(\mathbf{0}, \Psi^{-1}) \quad (\text{from 1.2})$$

$$i = 1, \dots, n,$$

and it is assumed that the regression function  $f$  lies in some reproducing kernel Hilbert or Krein space (RKHS/RKKS)  $\mathcal{F}$  with kernel  $h_\eta$  defined over the set of covariates  $\mathcal{X}$ . In [Chapter 2](#), we built a primer on basic functional analysis, and described various interesting RKHS/RKKS for regression modelling.

We then ascertained the form of the Fisher information for  $f$ , treated as a parameter of the model to be estimated, and from [Corollary 3.3.1](#) (p. 91), it is

$$\begin{aligned} \mathcal{I}(f(x), f(x')) &= \sum_{i,j=1}^n \psi_{ij} h_\eta(x, x_i) h_\eta(x', x_j) \\ &= \mathbf{h}_\eta(x)^\top \Psi \mathbf{h}_\eta(x'), \end{aligned}$$

for any two points  $x, x'$  in the domain of  $f$ , obtained using appropriate calculus for topological spaces detailed in [Chapter 3](#). An I-prior for  $f$  is defined as Gaussian with mean function  $f_0$  chosen a priori, and covariance function equal to the Fisher information.

The I-prior for  $f$  has the simple representation

$$\begin{aligned} f(x_i) &= f_0(x_i) + \sum_{k=1}^n h_\eta(x_i, x_k) w_k \\ (w_1, \dots, w_n)^\top &\sim N_n(\mathbf{0}, \boldsymbol{\Psi}) \\ i &= 1, \dots, n, \end{aligned}$$

and is written equivalently as the Gaussian process prior

$$(f(x_1), \dots, f(x_n))^\top \sim N_n(\mathbf{f}_0, \mathbf{H}_\eta \boldsymbol{\Psi} \mathbf{H}_\eta),$$

where  $\mathbf{H}_\eta = (h_\eta(x_i, x_j))_{i,j=1}^n$ .

In Chapter 4, we looked how the I-prior model has wide-ranging applications, from multilevel modelling, to longitudinal modelling, and modelling with functional covariates. Estimation was conducted mainly using a simple EM algorithm, although direct optimisation and Bayesian estimation using Markov chain Monte Carlo (MCMC) are also possible. In the case of polytomous responses, we used a latent variable framework in Chapter 5 to assign I-priors to latent propensities which drive the outcomes under a probit-transform scheme. An extension of the EM algorithm was considered, in which the E-step was replaced with variational inference, so as to overcome the intractability brought about by the conditional distributions. For both continuous and categorical response I-prior models, we find advantages of using I-priors, namely that model building and estimation is simple, inference straightforward, and predictions comparable, if not better, to similar state-of-the-art techniques.

Finally, in Chapter 6, we dealt with the problem of model selection, specifically for linear regression models. There, we used a fully Bayesian approach for estimating model probabilities in which regression coefficients are assigned an I-prior. We devised a model that requires minimal tuning on the part of the user, yet performs well in simulated and real-data examples, especially if multicollinearity exists among the covariates.

## 7.1 Summary of contributions

We give a summary of the novel contributions of this thesis.

- **Fisher information for infinite-dimensional parameters.** When the RKHS/RKKS  $\mathcal{F}$  is infinite dimensional (e.g. covariates are themselves functions), then the Fisher information involves derivatives with respect to an infinite-dimensional vector. Finite-dimensional results using componentwise/partial derivatives may fail in infinite dimensions. The technology of Fréchet and Gâteaux differentials

accommodate for the fact that  $f$  may be infinite dimensional, which, at minimum, requires  $\mathcal{F}$  to be a normed vector space. We foresee the work of Section 3.2 being applicable elsewhere, such as learning in (reproducing kernel) Banach spaces ([zhang2009reproducing](#); [zhang2012regularized](#)), or in the theory of parameter estimation for general exponential family type distributions of the form

$$p(X|\theta) = B(X) \exp(\langle \theta, T(X) \rangle_{\mathcal{H}} - A(\theta)),$$

in which  $\theta$  lies in some inner-product space  $\mathcal{H}$  which might be infinite dimensional ([sriperumbudur2013density](#)).

- **Efficient estimation methods for normal I-prior models.** The preferred estimation method for normal I-prior models for stability is the EM algorithm. Implementing the EM algorithm can be computationally costly, due to the squaring and inversion of the kernel matrices in the  $Q$  function in (4.18) on page 111. Unfortunately, not much can be done about the inversion, but we explored systematic ways in which to perform the squaring. Combining a “front-loading method” of the kernel matrices (Section 4.3.2, p. 117) and an exponential family ECM (expectation conditional maximisation) algorithm ([meng1993maximum](#)), the estimation procedure is streamlined. Our computational work culminated in the publicly available and well-documented R package **iprior** ([jamil2017iprior](#)) published on CRAN.
- **Methodological extension of I-priors to categorical responses.** An extension of the I-prior methodology to fit categorical responses was studied. We proposed a latent variable framework, in which there corresponds latent propensities for each category of the observations. Instead of modelling the responses directly, the latent propensities are modelled using an I-prior, and class probabilities obtained using a normal integral. We named this model the I-probit model. The challenge of estimation was overcoming said integral, and we used a variational EM algorithm in which the E-step uses a variational approximation to intractable conditional density. The variational EM algorithm was preferred over a fully Bayesian variational inference algorithm for two main reasons: 1) the work done in the normal I-prior EM algorithm applies directly; and 2) prior specification for hyperparameters can be dispensed with. Classification, meta-analysis and spatio-temporal modelling are specific examples of the applications of I-probit models.
- **Some distributional results for truncated normals.** In deriving the variational algorithm, some properties related to the conically truncated multivariate independent normal distribution (as defined in Appendix C.4, p. 267) were required. A small contribution of ours was to derive the closed-form expressions for

its first and second moments, and its entropy (Lemma C.5, p. 269). We have only seen closed-form expressions of the mean of such a distribution being used before ([girolami2006variational](#)) but not for the variance, nor an explicit derivation of these quantities.

- **Bayesian variable selection under collinearity.** Model comparison using likelihood ratio tests or Bayes factors is fine when the number of models under consideration is fairly small. Under a fully Bayesian scheme, we use MCMC to approximate posterior model probabilities of competing linear models. At the outset, we sought a model which required minimal intervention on the part of the user. The I-prior achieved this, with the added advantage of performing well under multicollinearity.

## 7.2 Open questions

In closing, we briefly discuss several questions which remain open during the course of completing this project.

- **Initialisation of EM or gradient-based methods.** Figure 4.1 (p. 110) indicates the impact that starting values can have on gradient-based optimisation. One can end up at a local optima on one of the two ridges. Usually, one of the ridges will have a higher maximum than the other, but it is not clear how to direct the algorithm in the direction of the “correct” ridge.

Importantly, the interpretation of a flat ridge in the likelihood is that there is insufficient information coming from the data to inform parameter estimation. In the EM algorithm, estimation is usually characterised by a fast increase in likelihood in the first few steps (as it climbs up the ridge), and then later iterations only improve the likelihood ever so slightly (as it moves along the ridge in search of the maximum). In some real-data cases (e.g. Tecator data set), we noticed that the EM sequence veers to the boundary of the parameter space, where the likelihood is infinite (e.g.  $L(\psi) \rightarrow \infty$  as  $\psi \rightarrow 0, \infty$ ).

Ill-posed problems similar to this are resolved by adding penalty terms to the log-likelihood. As to what penalty terms are appropriate remains an open question.

- **Standard errors for variational approximation.** Under a variational scheme, the log-likelihood function  $L(\theta)$  is replaced with the evidence lower bound (ELBO)  $\mathcal{L}_q(\theta)$  which serves as a conservative approximation to it. The question we have is whether the approximation degrades the asymptotic properties of the estimators obtained via variational inference? In particular, are the standard errors obtained

102 ‘()

\_tl

from the information matrix involving  $\mathcal{L}_q(\theta)$  reliable? This question has also been posed by [hall2011asymptotic](#); [bickel2013asymptotic](#); [chen2017use](#).

Variational methods for maximum likelihood learning can be seen as a deliberate misspecification of the model to achieve tractability. As such, the variational EM has been referred to as obtaining pseudo- or quasi-ML estimates. The quasi-likelihood literature has results relating to efficiency of parameter estimates (adjustments to the information matrix is needed), and we wonder if these are applicable for variational inference.

Also, obtaining standard errors directly from an EM algorithm is of interest, especially under a variational EM setting. Though this is described in [mclachlan2007algorithm](#), we have not seen this implemented widely.

- **Comparison of logistic and probit links.** For general binary and multinomial models, the logistic link function sees more prevalent use than its probit counterpart. Of course, we chose the probit as it has distributional advantages which we can exploit for estimation using variational inference. However, is there a difference between the behaviour of the probit and logistic model? We know that there is a difference between the logistic and normal distribution, especially in scaling and behaviour in the tails, but do these affect the outcome of I-prior models?
- **Consistency of I-prior Bayesian variable selection.** We wondered about model selection consistency for I-priors in Bayesian variable selection. That is, assuming that model  $M_{\text{true}}$  is behind the true data generative process, do

$$\lim_{n \rightarrow \infty} P(M_{\text{true}} | \mathbf{y}) = 1 \quad \text{and} \quad \lim_{n \rightarrow \infty} P(M_k | \mathbf{y}) = 0, \forall M_k \neq M_{\text{true}}$$

hold for the I-prior Bayesian variable selection methodology? In machine learning, this property is referred to as the *oracle property*. For the  $g$ -prior specifically, model consistency results were obtained by [fernandez2001benchmark](#); [liang2008mixtures](#). [casella2009consistency](#) also looks at consistency of Bayesian procedures for a wide class of prior distributions, but we have yet to examine whether the I-prior falls under the remit of their work.



# Supplementary S1

## Basic estimation concepts

Statistics concerns what can be learned from data ([davison2003statistical](#)). A statistical model comprises of a probabilistic component which drives the data generative process, in addition to a systematic or deterministic component, which sets it apart from pure mathematical models. Real-valued observations  $\mathbf{y} := \{y_1, \dots, y_n\}$  are treated as realisations from an assumed probability distribution with parameters  $\theta = (\theta_1, \dots, \theta_p)^\top \in \Theta \subseteq \mathbb{R}^p$ . The crux of statistical inference is to estimate  $\theta$  given the observed values, so that this optimised value may be used in the model to make deductions. We describe the *frequentist* and *Bayesian* paradigms for parameter estimation.

### S1.1 Maximum likelihood estimation

In the frequentist setting, the *likelihood* function, or simply likelihood, is a function of the parameters  $\theta$  which measures the plausibility of the parameter value given the observed data to fit a statistical model. It is defined as the mapping  $\theta \mapsto p(\mathbf{y}|\theta)$ , where  $p(\mathbf{y}|\theta)$  is the probability density function (or in the case of discrete observations, the probability mass function) of the modelled distribution of the observations.

It is logical to consider the parameter which provides the largest likelihood value,

$$\hat{\theta}_{\text{ML}} = \arg \max_{\theta} p(\mathbf{y}|\theta). \quad (\text{S1.1})$$

The value  $\hat{\theta}_{\text{ML}}$  is referred to as the *maximum likelihood estimate* for  $\theta$ . For convenience, the *log-likelihood* function  $L(\theta) = \log p(\mathbf{y}|\theta)$  is maximised instead; as the logarithm is a monotonically increasing function, the maximiser of the log-likelihood function is exactly the maximiser of the likelihood function itself.

When ML estimates are unable to be found in closed-form, the maximisation problem of (S1.1) requires iterative, numerical methods to find the maximum. These methods are often *gradient based*, i.e. algorithms that make use of the gradient of the objective function to be optimised. Examples include Newton's method, Fisher's scoring, quasi-Newton methods, gradient descent, and conjugate gradient methods. As the name suggests, these methods require evaluation of gradients or approximate gradients, and in some cases, the Hessian. Depending on the situation, gradients or Hessians can be expensive or inconvenient to compute or approximate. In cases of multi-modality of the objective function, the algorithms can potentially converge to a local optima, as it is known that the algorithms are quite sensitive to starting locations.

Besides invariance, the ML estimate comes with the attractive limiting property  $\sqrt{n}(\hat{\theta}_{\text{ML}} - \theta_{\text{true}}) \xrightarrow{\text{dist.}} N_p(0, \mathcal{I}(\theta)^{-1})$  (casella2002statistical) as sample size  $n \rightarrow \infty$ , where  $\mathcal{I}(\theta)$  is the Fisher information for  $\theta$ . Other asymptotic properties of the ML estimate include consistency, i.e.  $P(\|\hat{\theta}_{\text{ML}} - \theta_{\text{true}}\| > \epsilon) \xrightarrow{\text{prob.}} 0$  for any  $\epsilon > 0$ , and efficiency, i.e. it achieves the Cramér-Rao lower bound  $\text{Var}(\theta_{\text{ML}}) \geq \mathcal{I}(\theta)^{-1}$ .

As the likelihood measures the plausibility of a parameter value given the data, it can be used to compare two competing models. Let  $\Theta_0 = \{\theta \mid \theta_{d+1} = \theta_{d+1,0}, \dots, \theta_p = \theta_{p,0}\}$  be the set of parameters with restrictions on the last  $d$  components of  $\theta$ . The *likelihood ratio test* statistic for testing the null hypothesis  $H_0 : \theta \in \Theta_0$  against the alternative  $H_1 : \theta \notin \Theta_0$  is

$$\lambda = -2 \log \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} = -2(\log L(\hat{\theta}_0) - \log L(\hat{\theta})), \quad (\text{S1.2})$$

where  $\hat{\theta}_0 = \arg \max_{\theta \in \Theta_0} \log p(\mathbf{y} | \theta)$ . Wilks' theorem states that  $\lambda$  has an asymptotic chi-squared distribution with degrees of freedom equal to the number of restrictions imposed (or rather, the difference in dimensionality of  $\Theta$  and  $\Theta_0$ ). This gives a convenient way of comparing nested models.

As a remark, models with more parameters will always have higher, or similar, log-likelihood, than models with fewer parameters, because the model has a better ability to fit the data with more free parameters. In a linear regression setting, this relates to overfitting: a linear model with as many explanatory variables as there are data points ( $n = p$ ) will extrapolate every point in the data set. Overfitting is an oft cited problem of maximum likelihood.

## S1.2 Bayesian estimation

The *Bayesian* approach to estimating  $\theta$  takes a different outlook, in that it supplements what is already known from the data with additional information in the form of prior

beliefs about the parameters. This usually means treating the parameters as random, following some distribution dictated by a *prior density*  $p(\theta)$ . There are many ways of categorising different types of priors. Broadly speaking, priors, and hence Bayesian analysis ([robert2007bayesian](#); [kadane2011principles](#)), can be either *subjective* or *objective*, with the demonym ‘subjectivists’ and ‘objectivists’ used to refer to those subscribing to each respective principle. Subjectivists assert that probabilities are merely opinions, while objectivists, in contrast, view probabilities as an extension of logic. In this regard, objective Bayes seek to minimise the statistician’s contribution to inference and “let data speak for itself”, while subjective Bayes does the opposite.

In either case, inference about the parameters are then performed using the *posterior density*

$$p(\theta|\mathbf{y}) \propto \underbrace{p(\mathbf{y}|\theta)}_{\text{likelihood}} \times \underbrace{p(\theta)}_{\text{prior}}, \quad (\text{S1.3})$$

rather than through a single point estimate such as the ML estimate in the frequentist case. The posterior density encapsulates the uncertainty surrounding the parameters  $\theta$  after observing the data  $\mathbf{y}$ . The *posterior mean*

$$\tilde{\theta} = \int \theta p(\theta|\mathbf{y}) d\theta \quad (\text{S1.4})$$

is normally taken to be the point estimate for  $\theta$ , with its uncertainty usually reported in the form of a *credible interval*: if  $\theta_k$  is the  $k$ ’th component of  $\theta$ , then a  $(1-\alpha) \times 100\%$  credible interval for  $\theta_k$  is  $(\theta_k^l, \theta_k^u)$ , where  $P(\theta_k^l \leq \theta_k \leq \theta_k^u) = (1-\alpha) \times 100\%$ . Under a quadratic loss function,  $\tilde{\theta}$  minimises the expected loss  $E[(\theta - \theta_{\text{true}})^2]$  ([berger2013statistical](#)), and is hence also viewed as the *minimum mean squared error* (MMSE) estimator.

On a practical note, integration over the parameter space may be intractable, for instance, the model consists of a large number of parameters for which we would like the posterior mean of, or the marginalising integral cannot be found in closed form. Markov chain Monte Carlo (MCMC) methods are the standard way of approximating such integrals, by way of random sampling from the posterior. The sample  $\{\theta^{(1)}, \dots, \theta^{(T)}\}$  is then manipulated in a way to derive its approximation. In the case of the posterior mean,

$$\hat{E}[\theta|\mathbf{y}] = \frac{1}{T} \sum_{i=1}^T \theta^{(t)} \quad (\text{S1.5})$$

gives an approximation, and its  $(1 - \alpha) \times 100\%$  credible interval can be approximated using the lower  $\alpha/2 \times 100\%$  and upper  $(1 - \alpha/2) \times 100\%$  quantile of the sample.

The normalising constant is the marginal likelihood over the distribution of the parameters,  $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta) d\theta$ . The quantity  $p(\mathbf{y})$  is also known as the *model evidence*, or simply, *evidence*. As its name suggests, model evidence is used as a measure of how

much support there is for a particular model. As such, it is used as a basis for model comparison. Let  $p(\mathbf{y}|M_0)$  and  $p(\mathbf{y}|M_1)$  be the model evidence for two competing models  $M_0$  and  $M_1$  respectively. Define the *Bayes factor* for comparing model  $M_0$  against an alternative model  $M_1$  as

$$\text{BF}(M_0, M_1) = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)}. \quad (\text{S1.6})$$

Values of  $\text{BF}(M_0, M_1) < 1$  would suggest that the data provides more evidence for model  $M_1$  over  $M_0$ .

Note that the model evidence is free of  $\theta$  because all of the parameters have been marginalised out, or put another way, considered in entirety and averaged over all possible values of  $\theta$  drawn from its prior density. Thus, model comparison using Bayes factors differs from the frequentist likelihood ratio comparison in that it does not depend on any one particular set of values for the parameters.

### S1.3 Maximum a posteriori estimation

One may also find the value of  $\theta$  which maximises the posterior,

$$\hat{\theta}_{\text{MAP}} = \arg \max_{\theta} p(\mathbf{y}|\theta)p(\theta), \quad (\text{S1.7})$$

which is the mode of the posterior distribution. This quantity is known as the *maximum a posteriori* (MAP) estimate. It is different from the ML estimate in that the maximisation objective is augmented with the prior density for  $\theta$ . In this sense, MAP estimation can be seen as regularisation of the ML estimation procedure, whereby a “penalty” term is added to avoid overfitting.

MAP estimation is often criticised for not being representative of Bayesian methods. That is, MAP estimation returns a point estimate with no apparent way of quantifying its uncertainty. Furthermore, unlike ML estimators, MAP estimators are not invariant under reparameterisation. If  $\theta$  is a random variable with density  $p(\theta)$ , then the pdf of  $\xi := g(\theta)$ , where  $g : \theta \mapsto g(\theta)$  is a one-to-one transformation, is

$$p_\xi(\xi) = p_\theta(g^{-1}(\xi)) \left| \frac{d}{d\xi} g^{-1}(\xi) \right|. \quad (\text{S1.8})$$

The second term in (S1.8) is called the *Jacobian (determinant)*. Therefore, a different parameterisation of  $\theta$  will impact the location of the maximum because of the introduction of the Jacobian into the optimisation objective (S1.7).

## S1.4 Empirical Bayes

The term *empirical Bayes* ([robbins1956empirical](#); [casella1985introduction](#)) refers to a procedure in which features of the prior is informed by the data. This is realised by parameterising the prior by a hyperparameter  $\eta$ , i.e.  $\theta \sim p(\theta|\eta)$ . Values for the hyperparameter are clearly important, because they appear in the posterior for  $\theta$ :

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta|\eta)}{p(\mathbf{y}|\eta)} \quad (\text{S1.9})$$

To avoid the subjectivist's approach of specifying values for  $\eta$  a priori, one instead turns to the data for guidance. Information concerning  $\eta$  is contained in the marginal likelihood  $p(\mathbf{y}|\eta) = \int p(\mathbf{y}|\theta)p(\theta|\eta) d\theta$ . This paves the way for using the *maximum marginal likelihood* estimate

$$\hat{\eta} = \arg \max_{\eta} p(\mathbf{y}|\eta) \quad (\text{S1.10})$$

in place of  $\eta$  in the equation of (S1.9). This procedure is also coined *maximum likelihood type-II* ([bishop2006pattern](#)), and is commonly referred to as such in the machine learning literature. It is also commonplace in statistics, especially in random-effects or latent variable models which employ a maximum likelihood procedure such as EM algorithm.

As a remark, estimation of  $\eta$  itself can be made to conform to Bayesian philosophy, i.e., by placing priors on it and inferring  $\eta$  through its posterior. Such a procedure is referred to as *Bayesian hierarchical modelling*. A motivation for doing this is because the ML estimate of  $\eta$  ignores any uncertainty in it. Of course, the hyperprior for  $\eta$  could be parameterised by a hyper-hyperparameter, and itself have a prior, and so on and so forth. Evidently the model is specified until such a point where there are parameters of the model which are left unoptimised and must be specified in subjective manner ([beal2003variational](#)).



## Supplementary S2

# The EM algorithm

Often times, there are unobserved, random variables  $\mathbf{w} = \{w_1, \dots, w_n\}$  that are assumed to make up the data generative process, prescribed in the statistical model through the *joint pdf*  $p(\mathbf{y}, \mathbf{w}|\theta)$ . Examples of models that include latent variables are plentiful: Gaussian mixture models, latent class analysis, factor models, random coefficient models, and so on. In order to obtain maximum likelihood (ML) estimates through a direct maximisation of the likelihood, it is necessary to first marginalise out the latent variables,

$$p(\mathbf{y}|\theta) = \int \overbrace{p(\mathbf{y}|\mathbf{w}, \theta)p(\mathbf{w}|\theta)}^{p(\mathbf{y}, \mathbf{w}|\theta)} d\mathbf{w}, \quad (\text{S2.1})$$

and obtain the *marginal likelihood*. Note that the integral is replaced by a summation over all possible values in the case of discrete latent variables  $\mathbf{w}$ .

Direct maximisation of the marginal (log-)likelihood might not be favourable due to intractability in obtaining ML solutions. The form of the marginal likelihood might not be conducive for closed-form estimates to be found, necessitating the use of numerical, gradient-based methods which is subject to its own undesirable quirks. Moreover, when the evaluation of the (log-)likelihood, gradient and/or Hessian are expensive to compute, then numerical methods are burdensome to execute.

It is usually the case that if the latent variables  $\mathbf{w}$  were somehow known, estimation would be made simpler. That is, the solution to  $\arg \max_{\theta} \log p(\mathbf{y}, \mathbf{w}|\theta)$  can be obtained in a simple manner. The expectation-maximisation algorithm ([dempster1977maximum](#)), commonly known as the EM algorithm, is an iterative procedure which exploits the fact that the so-called *complete data likelihood* is easier to work with. Correspondingly, in EM terminology, the marginal likelihood is referred to as the *incomplete data likelihood*.

We describe a derivation of both a general EM algorithm and an EM algorithm for models whose data generative pdf belongs to an exponential family of pdfs. Interestingly, the EM algorithm can be modified to obtain maximum a posteriori estimates or penalised log-likelihood solutions. As a note, the EM algorithm is not an algorithm per se, in that it does not provide exact instructions as to what the E- and M-steps should comprise of. Rather, it is a generic device to obtain parameter estimates ([mclachlan2007algorithm](#)).

## S2.1 Derivation of the EM algorithm

For want of an iterative procedure to obtain maximum likelihood estimates, we seek a solution to

$$\arg \max_{\theta} \{L(\theta|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) \geq 0\}, \quad (\text{S2.2})$$

where the solution to (S2.2) yields an improvement to the current  $t$ 'th iteration of the log-likelihood value  $L(\theta^{(t)}|\mathbf{y})$ . Note that the objective function in (S2.2) forms an upper bound for the quantity  $Q(\theta|\theta^{(t)})$ , as shown below:

$$\begin{aligned} L(\theta|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) &= \log \int p(\mathbf{y}|\mathbf{w}, \theta) p(\mathbf{w}|\theta) \frac{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})} d\mathbf{w} - \log p(\mathbf{y}|\theta^{(t)}) \\ &\geq \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w}, \theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})} d\mathbf{w} \quad (\text{Jensen's inequality}) \\ &\quad - \log p(\mathbf{y}|\theta^{(t)}) \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) d\mathbf{w} \\ &= \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w}, \theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})p(\mathbf{y}|\theta^{(t)})} d\mathbf{w} \\ &=: Q(\theta|\theta^{(t)}). \end{aligned}$$

Evidently, to maximise  $L(\theta|\mathbf{y})$ , we can't do any worse than maximising  $Q(\theta|\theta^{(t)})$  in  $\theta$ . Denote by  $\theta^{(t+1)}$  as the maximiser of  $Q(\theta|\theta^{(t)})$ . Then,

$$\begin{aligned} \theta^{(t+1)} &= \arg \max_{\theta} \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w}, \theta)p(\mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})p(\mathbf{y}|\theta^{(t)})} d\mathbf{w} \\ &= \arg \max_{\theta} \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log p(\mathbf{y}|\mathbf{w}, \theta)p(\mathbf{w}|\theta) d\mathbf{w} \\ &= \arg \max_{\theta} \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log p(\mathbf{y}, \mathbf{w}|\theta) d\mathbf{w} \\ &= \arg \max_{\theta} \mathbb{E}_{\mathbf{w}} [\log p(\mathbf{w}, \mathbf{y}|\theta)|\mathbf{y}, \theta^{(t)}] \end{aligned}$$

We arrive at an iterative procedure summarised succinctly as the following:

---

**Algorithm S1** EM algorithm

---

- 1: **initialise**  $\theta^{(0)}$  and  $t \leftarrow 0$
  - 2: **while** not converged **do**
  - 3:   E-step: compute  $Q(\theta|\theta^{(t)}) = \mathbb{E}_{\mathbf{w}} [\log p(\mathbf{w}, \mathbf{y}|\theta)|\mathbf{y}, \theta^{(t)}]$
  - 4:   M-step:  $\theta^{(t+1)} \leftarrow \arg \max_{\theta} Q(\theta|\theta^{(t)})$
  - 5:    $t \leftarrow t + 1$
  - 6: **end while**
- 

Notice that the log-likelihood function satisfies

$$L(\theta|\mathbf{y}) \geq L(\theta^{(t)}|\mathbf{y}) + Q(\theta|\theta^{(t)}), \quad (\text{S2.3})$$

for which equality is achieved when  $\theta = \theta^{(t)}$ , since

$$\begin{aligned} Q(\theta^{(t)}|\theta^{(t)}) &= \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log \frac{p(\mathbf{y}|\mathbf{w}, \theta^{(t)})p(\mathbf{w}|\theta^{(t)})}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})p(\mathbf{y}|\theta^{(t)})} d\mathbf{w} \\ &= \int p(\mathbf{w}|\mathbf{y}, \theta^{(t)}) \log \frac{p(\mathbf{y}, \mathbf{w}|\theta^{(t)})}{p(\mathbf{y}, \mathbf{w}|\theta^{(t)})} d\mathbf{w}^0 \\ &= 0. \end{aligned}$$

This implies that the EM algorithm improves the log-likelihood values at each iteration, since

$$L(\theta^{(t+1)}|\mathbf{y}) - L(\theta^{(t)}|\mathbf{y}) \geq Q(\theta^{(t+1)}|\theta^{(t)}) \geq 0$$

and  $Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) = 0$  since  $\theta^{(t+1)}$  maximises  $Q(\cdot|\theta^{(t)})$ .

The expectation in the E-step involves the conditional pdf  $p(\mathbf{w}|\mathbf{y}, \theta^{(t)})$ . Viewed through a Bayesian lens, this is the posterior density of the latent variables using the  $t$ 'th iteration parameter values. The success of the E-step is predicated on the availability of the conditional pdf for the expectation. If not, approximations to the E-step can be explored, for example using Monte Carlo methods ([wei1990monte](#)) or a variational approximation ([beal2003variational](#)).

The solution to the M-step usually, but not always, exists in closed form. Maximising the  $Q$  function over all possible values of  $\theta$  may not be feasible ([mclachlan2007algorithm](#)). In such situations, the generalised EM algorithm (as defined by [dempster1977maximum](#)) requires only that  $\theta^{(t+1)}$  be chosen in a way that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}).$$

That is,  $\theta^{(t+1)}$  is chosen so as to increase the value of the  $Q$  function at its current parameter value. As seen in the argument above, this requirement is sufficient for a guaranteed increase in the log-likelihood function at each iteration.

## S2.2 Exponential family EM algorithm

Consider the density function  $p(\cdot|\boldsymbol{\theta})$  of the complete data  $\mathbf{z} = \{\mathbf{y}, \mathbf{w}\}$ , which depends on parameters  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_s)^\top \in \Theta \subseteq \mathbb{R}^s$ , belonging to an exponential family of distributions. This density takes the form  $p(\mathbf{z}|\boldsymbol{\theta}) = B(\mathbf{z}) \exp(\langle \boldsymbol{\eta}(\boldsymbol{\theta}), \mathbf{T}(\mathbf{z}) \rangle - A(\boldsymbol{\theta}))$ , where  $\boldsymbol{\eta} : \mathbb{R}^s \mapsto \mathbb{R}$  is a link function,  $\mathbf{T}(\mathbf{z}) = (T_1(\mathbf{z}), \dots, T_s(\mathbf{z}))^\top \in \mathbb{R}^s$  are the sufficient statistics of the distribution, and  $\langle \cdot, \cdot \rangle$  is the usual Euclidean dot product. It is often easier to work in the *natural parameterisation* of the exponential family distribution

$$p(\mathbf{z}|\boldsymbol{\eta}) = B(\mathbf{z}) \exp(\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z}) \rangle - A^*(\boldsymbol{\eta})) \quad (\text{S2.4})$$

by defining  $\boldsymbol{\eta} := (\eta_1(\boldsymbol{\theta}), \dots, \eta_r(\boldsymbol{\theta})) \in \mathcal{E}$ , and  $\exp A^*(\boldsymbol{\eta}) = \int B(\mathbf{z}) \exp \langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z}) \rangle d\mathbf{z}$  to ensure the density function normalises to one. As an aside, the set  $\mathcal{E} := \{\boldsymbol{\eta} = (\eta_1, \dots, \eta_s) \mid \int \exp A^*(\boldsymbol{\eta}) < \infty\}$  is called the *natural parameter space*. If  $\dim \mathcal{E} = r < s = \dim \Theta$ , then the pdf belongs to the *curved exponential family* of distributions. If  $\dim \mathcal{E} = r = s = \dim \Theta$ , then the family is a *full exponential family*.

Assuming the latent  $\mathbf{w}$  variables are observed and working with the natural parameterisation, then the complete maximum likelihood (ML) estimate for  $\boldsymbol{\eta}$  is obtained by solving

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{z}|\boldsymbol{\eta}) = \mathbf{T}(\mathbf{z}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0. \quad (\text{S2.5})$$

Of course, the variable  $\mathbf{w}$  are never observed, so the ML estimate for  $\boldsymbol{\eta}$  can only be informed from what is observed. Let  $p(\mathbf{y}|\boldsymbol{\eta}) = \int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) d\mathbf{w}$  represent the marginal density of the observations  $\mathbf{y}$ . Now, the ML estimate for  $\boldsymbol{\eta}$  is obtained by solving

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}|\boldsymbol{\eta}) &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}|\boldsymbol{\eta}) \\ &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \cdot \frac{\partial}{\partial \boldsymbol{\eta}} \left( \int p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) d\mathbf{w} \right) \\ &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \int \left( \frac{\partial}{\partial \boldsymbol{\eta}} p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) d\mathbf{w} \\ &= \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \int \left( p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \frac{\partial}{\partial \boldsymbol{\eta}} \log p(\mathbf{y}, \mathbf{w}|\boldsymbol{\eta}) \right) d\mathbf{w} \\ &= \int \left( \mathbf{T}(\mathbf{y}, \mathbf{w}) - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) \right) p(\mathbf{w}|\mathbf{y}, \boldsymbol{\eta}) d\mathbf{w} \\ &= \mathbb{E}_{\mathbf{w}} [\mathbf{T}(\mathbf{y}, \mathbf{w})|\mathbf{y}] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) \end{aligned} \quad (\text{S2.6})$$

equated to zero. Note that we are allowed to change the order of integration and differentiation provided the integrand is continuously differentiable. So the only difference

between the first order condition of (S2.5) and that of (S2.6) is that the sufficient statistics involving the unknown  $\mathbf{w}$  are replaced by their conditional or posterior expectations.

A useful identity to know is that  $\frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = \mathbb{E}_{\mathbf{z}}[\mathbf{T}(\mathbf{z})]$  (**casella2002statistical**), which can be expressed in terms of the original parameters  $\boldsymbol{\theta}$ . As a consequence, solving for the ML estimate for  $\boldsymbol{\theta}$  from the FOC equations (S2.6) is possible without having to deal with the derivative of  $A^*$  with respect to the natural parameters. Having said this, an analytical solution in  $\boldsymbol{\theta}$  may not exist, because the relationship of  $\boldsymbol{\theta}$  could be implicit in the set of equations  $\mathbb{E}_{\mathbf{w}} [\mathbf{T}(\mathbf{w}, \mathbf{y}) | \mathbf{y}, \boldsymbol{\theta}] = \mathbb{E}_{\mathbf{y}, \mathbf{w}} [\mathbf{T}(\mathbf{y}, \mathbf{w}) | \boldsymbol{\theta}]$ . One way around this is to employ an iterative procedure, as detailed in [Algorithm S2](#).

---

#### **Algorithm S2** Exponential family EM

---

- 1: **initialise**  $\boldsymbol{\theta}^{(0)}$  and  $t \leftarrow 0$
  - 2: **while** not converged **do**
  - 3:    E-step:  $\tilde{\mathbf{T}}^{(t+1)}(\mathbf{y}, \mathbf{w}) \leftarrow \mathbb{E}_{\mathbf{w}} [\mathbf{T}(\mathbf{w}, \mathbf{y}) | \mathbf{y}, \boldsymbol{\theta}^{(t)}]$
  - 4:    M-step:  $\boldsymbol{\theta}^{(t+1)} \leftarrow$  solution to  $\tilde{\mathbf{T}}^{(t+1)}(\mathbf{y}, \mathbf{w}) = \mathbb{E}_{\mathbf{y}, \mathbf{w}} [\mathbf{T}(\mathbf{y}, \mathbf{w}) | \boldsymbol{\theta}]$
  - 5:     $t \leftarrow t + 1$
  - 6: **end while**
- 

To see how [Algorithm S2](#) motivates the EM algorithm, consider the following argument. Recall that for the EM algorithm, the function  $Q_t(\boldsymbol{\eta}) = \mathbb{E}_{\mathbf{w}}[\log p(\mathbf{y}, \mathbf{w} | \boldsymbol{\eta}) | \mathbf{y}, \boldsymbol{\eta}^{(t)}]$  is maximised at each iteration  $t$ . For exponential families of the form (S2.4), the  $Q_t$  function turns out to be

$$Q_t(\boldsymbol{\eta}) = \mathbb{E}_{\mathbf{w}} [\langle \boldsymbol{\eta}, \mathbf{T}(\mathbf{z}) \rangle | \mathbf{y}, \boldsymbol{\eta}^{(t)}] - A^*(\boldsymbol{\eta}) + \log B(\mathbf{z}),$$

and this is maximised at the value of  $\boldsymbol{\eta}$  satisfying

$$\frac{\partial}{\partial \boldsymbol{\eta}} Q_t(\boldsymbol{\eta}) = \mathbb{E}_{\mathbf{w}} [\mathbf{T}(\mathbf{y}, \mathbf{w}) | \mathbf{y}, \boldsymbol{\eta}^{(t)}] - \frac{\partial}{\partial \boldsymbol{\eta}} A^*(\boldsymbol{\eta}) = 0,$$

a similar condition to (S2.6) when obtaining ML estimate of  $\boldsymbol{\eta}$ . Thus,  $Q_t$  is maximised by the solution to line 4 in [Algorithm S2](#).

### S2.3 Bayesian EM algorithm

A simple modification of the EM algorithm can be done to obtain maximum a posteriori estimates, or maximum penalised likelihood estimates. Under a Bayesian framework, a prior is assigned on the model parameters,  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ . Recall that the MAP estimate is obtained as the maximiser of the log-density  $\log p(\mathbf{y} | \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$ .

The EM algorithm works as before, but replaces the E-step with

$$E_{\mathbf{w}} \left[ \log p(\mathbf{w}, \mathbf{y} | \theta) + \log p(\theta) | \mathbf{y}, \theta^{(t)} \right] = Q(\theta | \theta^{(t)}) + \log p(\theta) \quad (\text{S2.7})$$

since  $\log p(\theta)$  has no terms involving the latent variables  $\mathbf{w}$ . The M-step now maximises (S2.7) with respect to  $\theta$ , which includes the log prior density (or a penalty term). It would seem that the regular EM algorithm maximises (S2.7) such that  $p(\theta) \propto \text{const.}$  is a diffuse prior for  $\theta$ . **beal2003** discuss a more Bayesian extension of EM, in which the output of the so-called *variational Bayes EM* algorithm are (approximate) posterior distributions of the parameters, rather than MAP estimates discussed here.

## Supplementary S3

# Hamiltonian Monte Carlo

Hamiltonian Monte Carlo had its beginnings in statistical physics, with the [duane1987hybrid](#) paper by [duane1987hybrid](#), using what they called ‘Hybrid Monte Carlo’ in lattice models of quantum theory. Their work merged the approaches of molecular dynamics and Markov chain Monte Carlo methods. As an interesting side note, their method abbreviates also to ‘HMC’, but throughout the statistical literature, it is more commonly referred to by its more descriptive name Hamiltonian Monte Carlo. Incidentally, the use of HMC started with applications to neural networks as early as 1996 (see [neal2011mcmc](#) for an excellent review of the subject matter). It was not until 2011 when active development of the method, and in particular, software for for statistical applications began. The [Stan](#) initiative ([carpenter2016stan](#)) began in response to difficulties faced when performing full Bayesian inference on multilevel generalised linear models. These difficulties mainly involved poor efficiency in usual MCMC samplers, particularly due to high autocorrelations in the posterior chains, which meant that many chains and many iterations were required to get an adequate sample. It was a case of exhausting all possible algorithmic remedies for existing samplers (Gibbs samplers, Metropolis samplers, etc.), and realising that fundamentally not much improvement can be had unless a novel sampling technique was discovered.

The basic idea behind HMC is to use Hamiltonian dynamics to propose new states in the posterior sampling, rather than relying on random walks. If one were to understand and use the geometry of the posterior density to one’s benefit, then it should be possible to generate new proposal states with high probabilities of acceptance and move far away from the current state. Hamiltonian dynamics, like classical Newtonian mechanics, provides a framework for modelling the motion of a body in space across time  $t$ . Additionally, Hamiltonian dynamics concatenates the position vector  $x$  with its momentum  $z$ , and the motion of  $x$  in  $d$ -dimensional space is then described through

Hamilton's equations

$$\frac{dx}{dt} = \frac{\partial H}{\partial z} \quad \text{and} \quad \frac{dz}{dt} = -\frac{\partial H}{\partial x}, \quad (\text{S3.1})$$

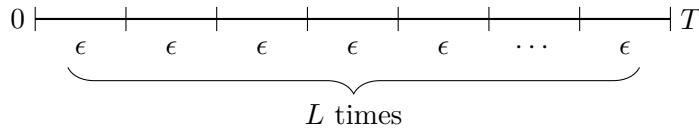
where  $H = H(x, z)$  is called the Hamiltonian of the system. The Hamiltonian is an operator which encapsulates the total energy of the system. In a closed system, one can express the sum of operators corresponding to the kinetic energy  $K(p)$  and the potential energy  $U(z)$  of the system

$$H(x, z) = K(z) + U(x). \quad (\text{S3.2})$$

Substituting (S3.2) into (S3.1), we get the system of partial differential equations (PDEs)

$$\frac{dx}{dt} = \frac{\partial}{\partial z} K(z) \quad \text{and} \quad \frac{dz}{dt} = -\frac{\partial}{\partial x} U(x). \quad (\text{S3.3})$$

To describe the evolution of  $(x(t), z(t))$  from time  $t$  to  $t+T$ , it is necessary to discretise time, and split  $T = L\epsilon$ . The quantity  $L$  is known as the number of *leapfrogs*, and  $\epsilon$  the *step size*.



The system of PDEs is solved using Euler's method, or the more commonly used leapfrog integration, which is a three-step process:

1. **Half-step momentum.**  $z(t + \epsilon/2) = z(t) - \frac{\epsilon}{2} \frac{\partial}{\partial x} U(x(t))$
2. **Full-step position.**  $x(t + \epsilon) = x(t) + \epsilon \frac{\partial}{\partial z} K(z(t + \epsilon/2))$
3. **Half-step momentum.**  $z(t + \epsilon) = z(t + \epsilon/2) = z(t) - \frac{\epsilon}{2} \frac{\partial}{\partial x} U(x(t))$

in which steps 1–3 are repeated  $L$  times.

Having knowing the formula for how particles move in space, we can use this information to treat random points drawn from some probability density as “particles”. Randomness of position and momentum are prescribed through probability densities on each. Given some energy function  $E(\theta)$  over states  $\theta$ , the *canonical distribution* of the states  $\theta$  (otherwise known as the *canonical ensemble*) is given by the probability density function

$$p(\theta) \propto \exp\left(-\frac{E(\theta)}{k\tau}\right),$$

where  $k$  is Boltzmann's constant,  $\tau$  is the absolute temperature of the system. The Hamiltonian is one such energy function over states  $(x, z)$ . By replacing  $E(\theta)$  by (S3.2) in the pdf above, we realise that the distribution for  $x$  and  $z$  are independent. The system can be manipulated such that  $k\tau = 1$ —in any case, these are constants which can be absorbed into one of the terms in the pdf anyway.

Using a *quadratic kinetic energy* function  $K(z) = z^\top M^{-1}z/2^1$ , we find that the probability density function for  $z$  is

$$p(z) \propto \exp\left(-\frac{1}{2}z^\top M^{-1}z\right),$$

implying  $z \sim N_d(0, M)$ . Here,  $M = \text{diag}(m_1, \dots, m_d)$  is called the *mass matrix*, which obviously serves as the variance for the randomly distributed  $z$ . As for the potential energy, choose a function such that  $U(x) = -\log p(x)$ , implying  $p(x) \propto \exp(-U(x))$ . Here,  $p(x)$  represents the target density from which we wish to sample, for instance, a posterior density of interest. Thus, to sample variables  $x$  from  $p(x)$ , one artificially introduces momentum variables  $z$  and sample jointly instead from  $p(x, z) = p(z)p(x)$ , and discarding  $z$  thereafter. The HMC algorithm is summarised in [Algorithm S3](#).

---

**Algorithm S3** Hamiltonian Monte Carlo

---

- 1: initialise  $x^{(0)}, z^{(0)}$  and choose values for  $L, \epsilon$  and  $M$
- 2: **for**  $t = 1, \dots, T$  **do**
- 3:     Draw  $z \sim N_d(0, M)$  ▷ Perturb momentum
- 4:     Move  $(x^{(t)}, z^{(t)}) \mapsto (x^*, z^*)$  using Hamiltonian dynamics ▷ Proposal state
- 5:     Accept/reject proposal state, i.e. ▷ Metropolis update

$$(x^{(t+1)}, z^{(t+1)}) \leftarrow \begin{cases} (x^*, z^*) & \text{w.p. } \min(1, A) \\ (x^{(t)}, z^{(t)}) & \text{otherwise} \end{cases}$$

where

$$A = \frac{p(x^*, z^*)}{p(x^{(t)}, z^{(t)})} = \exp\left(H(x, z) - H(x^{(t)}, z^{(t)})\right)$$

- 6: **end for**
  - 7: **return** Samples  $\{x^{(1)}, \dots, x^{(T)}\}$
- 

HMC is often times superior to standard Gibbs sampling, for a variety of reasons. For one, conjugacy does not play any role in the efficiency of the HMC sampler, thus freeing the modeller to choose more appropriate and more intuitive prior densities for the parameters of the model. For another, the HMC sampler is designed to incite little autocorrelations between samples, and thus increasing efficiency.

Several drawbacks do exist with the HMC sampler. Firstly, it is impossible to directly sample from discrete distributions  $p(x)$ . More concretely, HMC requires that the domain of  $p(x)$  is continuous and that  $\partial \log p(x)/\partial x$  is inexpensive to compute. To work around this, one must reformulate the model by marginalising out the discrete variables, and obtain them back later by separately sampling from their posteriors. Alternatively, a Gibbs sampler specifically for the discrete variables could be augmented with the HMC sampler. The other drawback of HMC is that there are many tuning parameters (leapfrog  $L$ , step-size  $\epsilon$ , mass matrix  $M$ , etc.) that is not immediately easy to perfect, at least not to the novice user.

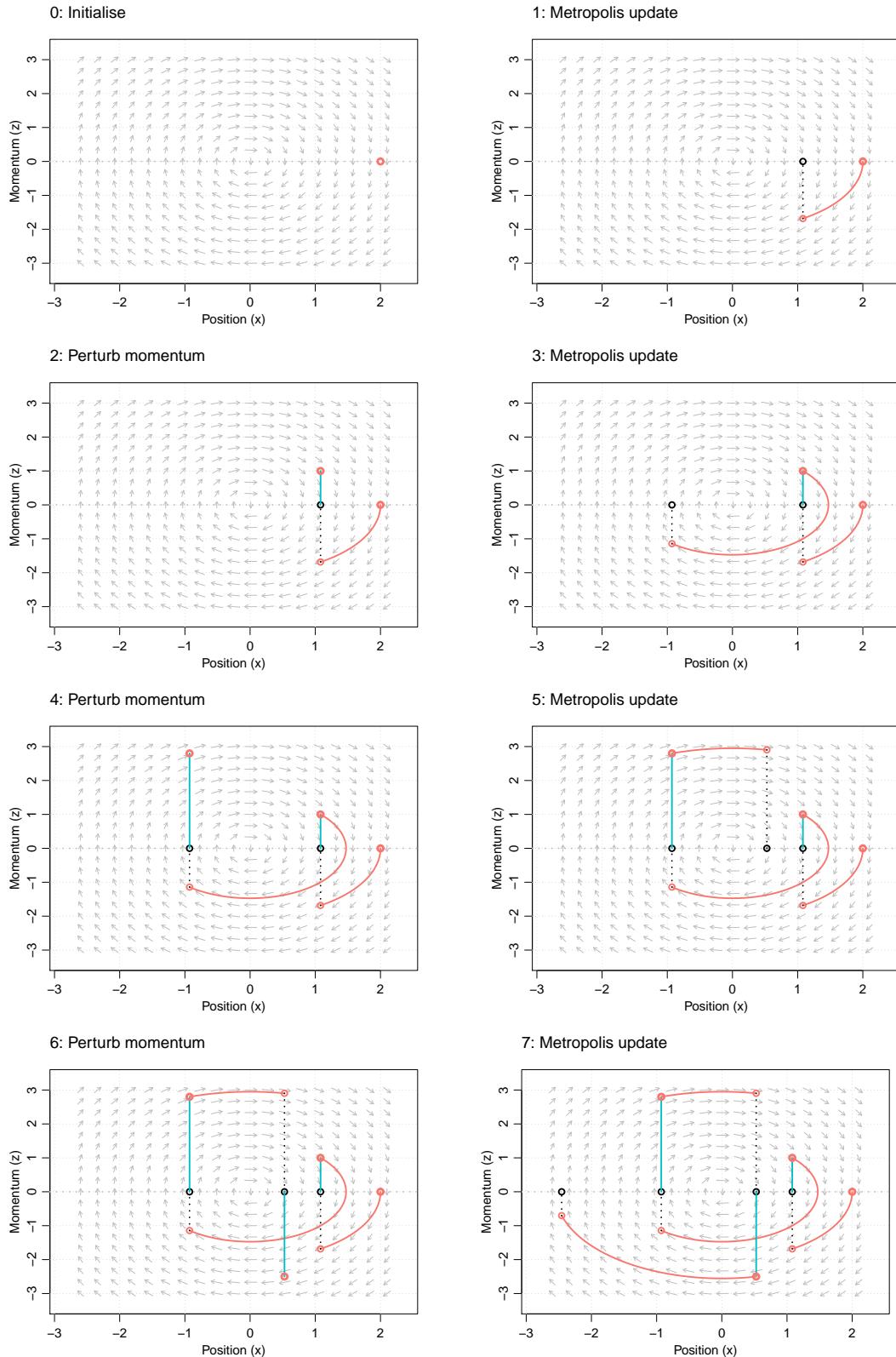


Figure S3.1: A phase diagram schematic of HMC sampling in one-dimension. At step 0, initialise values for momentum and position. At step 1, simulate movement using Hamiltonian dynamics, accept position and discard momentum. At step 2, perturb momentum using a normal density, then repeat.

The implementation of HMC by the programming language **Stan**, which interfaces many other programming languages including R, Python, MATLAB, Julia, Stata and **Mathematica**, is a huge step forward in computational Bayesian analysis. **Stan** takes the liberty of performing all the tuning necessary, and the practitioner is left with simply specifying the model. A vast library of differentiable probability functions are available, with the ability to bring your own code as well. Development is very active and many improvements and optimisations have been made since its inception.

---

<sup>1</sup>Thinking back to elementary mechanics, this is the familiar  $\frac{1}{2}mv^2$  formula for kinetic energy and substituting in the identity  $z = mv$ , where  $m$  is the mass of the object, and  $v$  is its velocity.



## Supplementary S4

# Variational inference

Consider a statistical model parameterised by  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$  for which we have observations  $\mathbf{y} := \{y_1, \dots, y_n\}$ , but also some latent variables  $\mathbf{w}$ . Typically, in such models, there is a want to evaluate the integral

$$I = \int p(\mathbf{y}|\mathbf{w})p(\mathbf{w}) d\mathbf{w}, \quad (\text{S4.1})$$

Marginalising out the latent variables in (S4.1) is usually a precursor to obtaining a log-likelihood function to be maximised in a frequentist setting, whereby there is an implicit dependence on the model parameters in the evaluation of  $I$ . In Bayesian analysis, priors are specified on the model parameters  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$ . By concatenating the latent variables and model parameters to form  $\mathbf{w}$ , the  $I$  corresponds to the marginal density for  $\mathbf{y}$ , on which the posterior depends.

In many instances, for one reason or another, evaluation of (S4.1) or is difficult, in which case inference is halted unless a way of overcoming the intractability is found. In this chapter, we discuss *variational inference* (VI) as a means of approximating the integral. The literature on variational inference is typically presented in a Bayesian light ([jordan1999introduction](#); [bishop2006pattern](#); [blei2017variational](#)), and as such, it is commonly known as *variational Bayes* method. The main attraction from a Bayesian point of view is that it provides a deterministic way of obtaining (approximate) posteriors, i.e. it does not involve sampling from posteriors.

Variational inference can be used in conjunction with an EM algorithm, in which the E-step is replaced with a variational E-step. This *variational EM algorithm* is used for maximum likelihood learning, but can modified to obtain maximum a posteriori estimates. In the works of ([beal2003variational](#); [beal2003](#)), the authors realised that the EM algorithm can be extended easily to obtain posterior densities of the latent variables and parameters if the statistical model is conjugate exponential family. They refer to

this as the *variational Bayes EM algorithm*, but in fact this is really just variational inference in which the algorithm resembles an EM algorithm with clear E- and M-steps.

We first briefly introduce variational methods for approximating the intractable integral, and this is usually considered a fully Bayesian treatment of the model. We then describe variational EM, and provide a comparison of the two methods.

## S4.1 A brief introduction to variational inference

The crux of variational inference is this: find a suitably close distribution function  $q(\mathbf{w})$  that approximates the true posterior  $p(\mathbf{w}|\mathbf{y})$ , where closeness here is defined in the Kullback-Leibler divergence sense,

$$D_{\text{KL}}(q||p) = \int \log \frac{q(\mathbf{w})}{p(\mathbf{w}|\mathbf{y})} q(\mathbf{w}) d\mathbf{w}.$$

Posterior inference is then conducted using  $q(\mathbf{w})$  in lieu of  $p(\mathbf{w}|\mathbf{y})$ . Advantages of this method are that 1) it is fast to implement computationally (compared to MCMC); 2) convergence is assessed simply by monitoring a single convergence criterion; and 3) it works well in practice, as attested to by the many studies implementing VI.

Briefly, we present the motivation behind variational inference and the minimisation of the KL divergence. Denote by  $q(\cdot)$  some density function of  $\mathbf{w}$ . One may show that log marginal density, i.e. the log of the intractable integral (S2.1), holds the following bound:

$$\begin{aligned} \log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{w}) - \log p(\mathbf{w}|\mathbf{y}) \quad (\text{Bayes' theorem}) \\ &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} - \log \frac{p(\mathbf{w}|\mathbf{y})}{q(\mathbf{w})} \right\} q(\mathbf{w}) d\mathbf{w} \quad (\text{expectation both sides}) \\ &= \mathcal{L}(q) + D_{\text{KL}}(q||p) \\ &\geq \mathcal{L}(q) \end{aligned} \tag{S4.2}$$

since the KL divergence is a non-negative quantity. The functional  $\mathcal{L}(q)$  given by

$$\begin{aligned} \mathcal{L}(q) &= \int \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} q(\mathbf{w}) d\mathbf{w} \\ &= \mathbb{E}_{\mathbf{w} \sim q} [\log p(\mathbf{y}, \mathbf{w})] + H(q), \end{aligned} \tag{S4.3}$$

where  $H$  is the entropy functional, is known as the *evidence lower bound* (ELBO). Evidently, the closer  $q$  is to the true  $p$ , the better, and this is achieved by maximising  $\mathcal{L}$ , or equivalently, minimising the KL divergence from  $p$  to  $q$ . Note that the bound (S4.2) achieves equality if and only if  $q(\mathbf{w}) \equiv p(\mathbf{w}|\mathbf{y})$ , but of course the true form of

the posterior is unknown to us—see Section S4.2 for a discussion. Maximising  $\mathcal{L}(q)$  or minimising  $D_{KL}(q\|p)$  with respect to the density  $q$  is a problem of calculus of variations, which incidentally, is where variational inference takes its name. The astute reader will realise that  $D_{KL}(q\|p)$  is impossible to compute, since one does not know the true distribution  $p(\mathbf{w}|\mathbf{y})$ . Efforts are concentrated on maximising the ELBO instead.

Maximising  $\mathcal{L}$  over all possible density functions  $q$  is not possible without considering certain constraints. Two such constraints are described. The first, is to make a distributional assumption regarding  $q$ , for which it is parameterised by  $\nu$ . For instance, we might choose the closest normal distribution to the posterior  $p(\mathbf{w}|\mathbf{y})$  in terms of KL divergence. In this case, the task is to find optimal mean and variance parameters of a normal distribution.

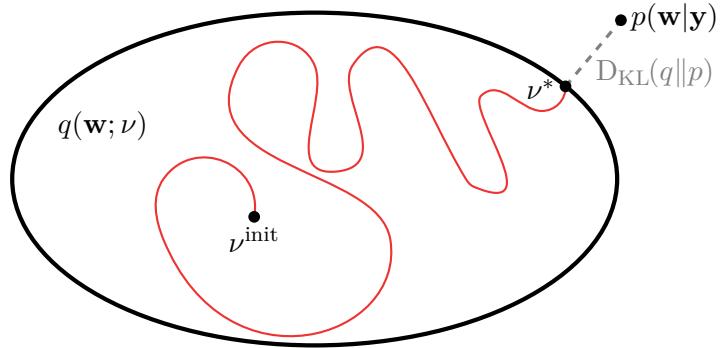


Figure S4.1: Schematic view of variational inference<sup>1</sup>. The aim is to find the closest distribution  $q$  (parameterised by a variational parameter  $\nu$ ) to  $p$  in terms of KL divergence within the set of variational distributions, represented by the ellipse.

The second type of constraint, and the one considered in this thesis, is simply an assumption that the approximate posterior  $q$  factorises into  $M$  disjoint factors. Partition  $\mathbf{w}$  into  $M$  disjoint groups  $\mathbf{w} = (w_{[1]}, \dots, w_{[M]})$ . Note that each factor  $w_{[k]}$  may be multidimensional. Then, the structure

$$q(\mathbf{w}) = \prod_{k=1}^M q_k(w_{[k]})$$

for  $q$  is considered. This factorised form of variational inference is known in the statistical physics literature as the *mean-field theory* ([itzykson1991statistical](#)).

*Remark S4.1.* The choice of factorisation is completely arbitrary, although forcing a factorisation also induces independence between the factors in the posterior, and this may or may not be suitable for the problem at hand. Landing the correct choice of factorisation is rather experimental, as the aim is to balance tractability and model

---

<sup>1</sup>Reproduced from the talk by David Blei entitled “Variational Inference: Foundations and Innovations”, 2017. URL: <https://simons.berkeley.edu/talks/david-blei-2017-5-1>.

misspecification. In a model with both latent variables and random parameters (in a Bayesian setting), then a good starting point would be to factorise the latent variables and parameters.

Let us denote the distributions which minimise the Kullback-Leibler divergence (maximise the variational lower bound) by the use of tildes. The impact of the mean-field factorisation on the ELBO is inspected:

$$\begin{aligned}\mathcal{L}(q) &= \int \cdots \int \log \frac{p(\mathbf{y}, \mathbf{w})}{\prod_{k=1}^M q_k(\mathbf{w})} \prod_{k=1}^m \{q_k(w_{[k]}) dw_{[k]}\} \\ &= \int \cdots \int \left( \log p(\mathbf{y}, \mathbf{w}) - \sum_{k=1}^M \log q_k(\mathbf{w}) \right) \prod_{k=1}^m \{q_k(w_{[k]}) dw_{[k]}\}\end{aligned}$$

and rearranging slightly for terms involving the  $j$ 'th component only, we get

$$\begin{aligned}\mathcal{L}(q) &= \int \cdots \int (\log p(\mathbf{y}, \mathbf{w}) - \log q_j(w_{[j]}) + \text{const.}) q_j(w_{[j]}) dw_{[j]} \prod_{k \neq j} \{q_k(w_{[k]}) dw_{[k]}\} \\ &= \int \left( \overbrace{\int \cdots \int \log p(\mathbf{y}, \mathbf{w}) \prod_{k \neq j} \{q_k(w_{[k]}) dw_{[k]}\}}^{\log \tilde{p}(\mathbf{y}, w_{[j]}) + \text{const.}} \right) q_j(w_{[j]}) dw_{[j]} \\ &\quad - \int \log q_j(w_{[j]}) q_j(w_{[j]}) dw_{[j]} + \text{const.} \\ &= -D_{\text{KL}}(q_{[j]} \| \tilde{p}) + \text{const.}\end{aligned}$$

The task of maximising  $\mathcal{L}$  is then equivalent to maximising  $-D_{\text{KL}}(q_{[j]} \| \tilde{p})$ , where  $\tilde{p}$  is defined in the overbrace of the second line in the equation above. Thus, for each  $w_{[k]}$ ,  $k = 1, \dots, M$ ,  $\tilde{q}_k$  satisfies

$$\log \tilde{q}_k(w_{[k]}) = E_{-k}[\log p(\mathbf{y}, \mathbf{w})] + \text{const.} \quad (\text{S4.4})$$

where expectation of the joint log density of  $\mathbf{y}$  and  $\mathbf{w}$  is taken with respect to all of the unknowns  $\mathbf{w}$ , except the one currently in consideration  $w_{[k]}$ , under their respective  $\tilde{q}_k$  densities. For further details, refer to **bishop2006pattern**.

In practice, rather than an explicit calculation of the normalising constant, one simply needs to inspect (S4.4) to recognise it as a known log-density function, which is the case when exponential family distributions are considered. That is, suppose that each complete conditional  $p(w_{[k]} | \mathbf{w}_{-k}, \mathbf{y})$ , where  $\mathbf{w}_{-k} = \{w_{[i]} | i \neq k\}$ , follows an exponential family distribution

$$p(w_{[k]} | \mathbf{w}_{-k}, \mathbf{y}) = B(w_{[k]}) \exp(\langle \zeta_k(\mathbf{w}_{-k}, \mathbf{y}), w_{[k]} \rangle - A(\zeta_k)).$$

Then, from (S4.4),

$$\begin{aligned}\tilde{q}(w_{[k]}) &\propto \exp\left(\mathbb{E}_{-k}[\log p(w_{[k]} | \mathbf{w}_{-k}, \mathbf{y})]\right) \\ &= \exp\left(\log B(w_{[k]}) + \mathbb{E}\langle \zeta_k(\mathbf{w}_{-k}, \mathbf{y}), w_{[k]} \rangle - \mathbb{E}[A(\zeta_k)]\right) \\ &\propto B(w_{[k]}) \exp \mathbb{E}\langle \zeta_\xi(\mathbf{w}_{-k}, \mathbf{y}), w_{[k]} \rangle\end{aligned}$$

is also in the same exponential family. In situations where there is no closed form expression for  $\tilde{q}$ , then one resorts to sampling methods such as a Metropolis random walk to estimate quantities of interest. This stochastic step within a deterministic algorithm has been explored before in the context of a Monte Carlo EM algorithm—see [meng1997algorithm](#) and references therein.

One notices that the optimal mean-field variational densities for each component are coupled with one another, in the sense that the distribution  $\tilde{q}_k$  depends on the moments of the rest of the components  $\mathbf{w}_{-k}$ . For very simple problems, an exact solution for each  $\tilde{q}_k$  can be found, but usually, the way around this is to employ an iterative procedure. The *coordinate ascent mean-field variational inference* (CAVI) algorithm cycles through each of the distributions in turn, updating them in sequence starting with arbitrary distributions as initial values.

---

#### Algorithm S4 The CAVI algorithm

---

```

1: initialise Variational factors  $q_k(w_{[k]})$ 
2: while ELBO  $\mathcal{L}(q)$  not converged do
3:   for  $k = 1, \dots, M$  do
4:      $\tilde{q}_k(w_{[k]}) \leftarrow \text{const.} \times \exp \mathbb{E}_{-k} [\log p(\mathbf{y}, \mathbf{w})]$             $\triangleright$  from (S4.4)
5:   end for
6:    $\mathcal{L}(q) \leftarrow \mathbb{E}_{\mathbf{w} \sim \prod_k \tilde{q}_k} \log p(\mathbf{y}, \mathbf{w}) + \sum_{k=1}^m H[q_k(w_{[k]})]$        $\triangleright$  Update ELBO
7: end while
8: return  $\tilde{q}(\mathbf{w}) = \prod_{k=1}^M \tilde{q}_k(w_{[k]})$ 

```

---

Each iteration of the CAVI brings about an improvement in the ELBO (hence the name coordinate ascent). The algorithm terminates when there is no more significant improvement in the ELBO, indicating a convergence of the CAVI. [blei2017variational](#) notes that the ELBO is typically a non-convex function, in which case convergence may be to (one of possibly many) local optima. A simple solution would be to restart the CAVI at multiple initial values, and the solution giving the highest ELBO is the distribution that is closest to the true posterior.

## S4.2 Variational EM algorithm

Consider again the latent variable setup described in [Supplementary Chapter S2](#), in which the goal is to maximise the (marginal) log-likelihood of the parameters  $\theta$  of the model, after integrating out the latent variables, as given by [\(S2.1\)](#). We will see how the EM algorithm relates to minimising the KL divergence between a density  $q(\mathbf{w})$  and the posterior of  $\mathbf{w}$ , and connect this idea to variational methods.

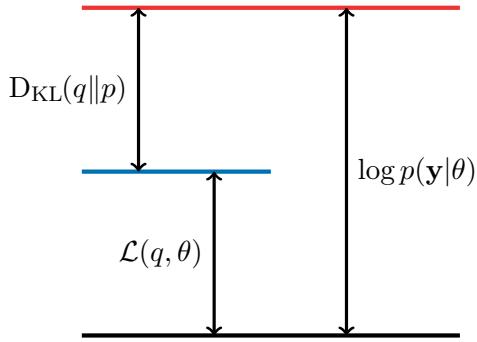


Figure S4.2: Illustration<sup>2</sup> of the decomposition of the log-likelihood into  $\mathcal{L}(q, \theta)$  and  $D_{KL}(q||p)$ . The quantity  $\mathcal{L}(q, \theta)$  is a lower bound for the log-likelihood.

As we did in deriving [\(S4.2\)](#), we decompose the (marginal) log-likelihood as

$$\begin{aligned} \log p(\mathbf{y}|\theta) &= \log p(\mathbf{y}, \mathbf{w}|\theta) - \log p(\mathbf{w}|\mathbf{y}, \theta) \\ &= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{w}|\theta)}{q(\mathbf{w})} - \log \frac{p(\mathbf{w}|\mathbf{y}, \theta)}{q(\mathbf{w})} \right\} q(\mathbf{w}) d\mathbf{w} \\ &= \underbrace{\mathbb{E}_{\mathbf{w} \sim q} \left[ \log \frac{p(\mathbf{y}, \mathbf{w}|\theta)}{q(\mathbf{w})} \right]}_{\mathcal{L}(q, \theta)} - \underbrace{\mathbb{E}_{\mathbf{w} \sim q} \left[ \log \frac{p(\mathbf{w}|\mathbf{y}, \theta)}{q(\mathbf{w})} \right]}_{-D_{KL}(q||p)}, \end{aligned}$$

where  $q(\mathbf{w})$  is any density function over the latent variables. This decomposition is shown in [Figure S4.2](#). The interest is then to have a density function  $q(\mathbf{w})$  which is as close as possible to the true posterior density  $p(\mathbf{y}|\mathbf{w}, \theta)$  in the KL divergence sense. Since the KL divergence is non-negative, minimising  $D_{KL}(q||p)$  is equivalent to maximising  $\mathcal{L}(q, \theta)$ .

As a remark, the above line of thought should be familiar as it is the exact same one made for variational inference. The twist here is that we will peruse a distribution which tightens the lower bound  $\mathcal{L}(q, \theta)$  to the marginal log-likelihood, and this happens when  $D_{KL}(q||p)$  is exactly zero, and this in turn happens when  $q$  is exactly the true posterior density. That is, for some parameter value,  $\theta = \theta^{(t)}$  say, the solution to

$$254 \arg \max_q \mathcal{L}(q, \theta^{(t)}) \tag{S4.5}$$

---

<sup>2</sup>Reproduced from [bishop2006pattern](#).

is  $q^{(t+1)}(\mathbf{w}) = p(\mathbf{w}|\mathbf{y}, \theta^{(t)})$ , because

$$D_{KL}(q||p) = E \left[ \log \frac{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})} \right] = 0.$$

At this stage, we have the equality

$$\log p(\mathbf{y}|\theta) = \mathcal{L}(q^{(t+1)}, \theta) \quad (\text{S4.6})$$

$$= E_{\mathbf{w} \sim q^{(t+1)}} \left[ \log \frac{p(\mathbf{y}, \mathbf{w}|\theta)}{p(\mathbf{w}|\mathbf{y}, \theta^{(t)})} \right] \quad (\text{S4.7})$$

$$= \underbrace{E_{\mathbf{w} \sim q^{(t+1)}} [\log p(\mathbf{y}, \mathbf{w}|\theta)]}_{Q(\theta|\theta^{(t)})} - \underbrace{E_{\mathbf{w} \sim q^{(t+1)}} [\log p(\mathbf{w}|\mathbf{y}, \theta^{(t)})]}_{-H(q^{(t+1)})}, \quad (\text{S4.8})$$

The term on the left is recognised as the  $Q$  function of the E-step

$$Q(\theta) = Q(\theta|\theta^{(t)}) = E_{\mathbf{w}} \left( \log p(\mathbf{y}, \mathbf{w}|\theta) \mid \mathbf{y}, \theta^{(t)} \right),$$

while the term on the left is an entropy term which does not depend on  $\theta$ . Thus, minimising the KL divergence, or maximising the lower bound  $\mathcal{L}$  with respect to  $q$ , corresponds to the E-step in the EM algorithm.

Furthermore, since equality between the log-likelihood and the lower bound is achieved after the E-step, increasing  $\mathcal{L}(q^{(t+1)}, \theta)$  with respect to  $\theta$  is sure to bring about an increase in the log-likelihood. That is, for any  $\theta$ , we find that

$$\begin{aligned} \log p(\mathbf{y}|\theta) - \log p(\mathbf{y}|\theta^{(t)}) &= Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}) + \Delta \text{entropy} \\ &\geq Q(\theta|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)}). \end{aligned}$$

because entropy differences are positive by Gibbs' inequality. We see that maximising  $Q$  with respect to  $\theta$  (the M-step) brings about an improvement to the log-likelihood value.

To summarise, given initial values  $q^{(0)}$  for the distribution and  $\theta^{(0)}$  for the parameters, the EM algorithm is seen as iterating between

- **E-step:**  $q^{(t+1)} \leftarrow \arg \max_q \mathcal{L}(q, \theta^{(t)})$ , i.e., maximise  $\mathcal{L}(q, \theta)$  with respect to  $q$ , keeping  $\theta$  fixed. This is equivalent to minimising the KL divergence  $D_{KL}(q||p)$ .
- **M-step.**  $\theta^{(t+1)} \leftarrow \arg \max_{\theta} \mathcal{L}(q^{(t+1)}, \theta)$ , i.e., maximise  $\mathcal{L}(q, \theta)$  with respect to  $\theta$ , keeping  $q(\mathbf{w})$  fixed.

When the true posterior distribution  $p(\mathbf{w}|\mathbf{y})$  is not tractable, then the E-step becomes intractable as well. By constraining the maximisation in the E-step to consider  $q$  belonging to a family of tractable densities, the E-step yields a variational approximation  $\tilde{q}$  to the true posterior. In [Section S4.1](#), we saw that constraining  $q$  to be of a

factorised form, then  $\tilde{q}$  is a mean-field density. After a variational E-step, the M-step proceeds as normal. This form of the EM is known as *variational EM algorithm* (VEM) ([blei2003variational](#)). The variational EM algorithm can also be modified to obtain MAP estimates by including the log prior density to the maximisation objective in the M-step.

Due to an approximation to the true posterior being used in the E-step, there is no guarantee that the log-likelihood value will increase at each iteration. This is seen pictorially in [Section S4.2](#): since the bound on the log-likelihood is not tight, increasing this bound will not necessarily cause an increase in log-likelihood value (Scenario C), and even if it did, it may not give as much an increase as it would under the true posterior density (Scenario B). Scenario A depicts an ideal case whereby the increase in log-likelihood is as much as it would be if the true posterior density was used.

On a practical note, if the posterior density is intractable, then so is the marginal likelihood, which means that we're unable to determine convergence of the EM using the log-likelihood. Instead, the lower bound  $\mathcal{L}(q, \theta)$  should be used, which monotonically increases to a local optima (as in the CAVI algorithm).

### S4.3 Comparing variational inference and variational EM

Variational inference is a fully Bayesian treatment of the model, for which the goal is to obtain approximate posterior densities for all latent variables and parameters. Variational EM algorithm on the other hand has the objective of obtaining ML or MAP estimates of the parameters using an EM algorithm in which the E-step is replaced with a variational E-step. In some cases, the CAVI algorithm can resemble an EM algorithm, especially when there is a distinction between latent variables and parameters, and a conjugate exponential family model is involved ([blei2017variational](#)).

Variational inference can yield exactly similar point estimates as variational EM if the approximate posterior is symmetric, e.g. a normal distribution. Under a normal posterior, its mean is used as a point estimate, which coincides with the mode, which is a MAP estimate, or in the case of diffuse priors, a ML estimate. However, since the output of variational inference are posterior densities instead of a single point estimate, one is able to obtain posterior standard deviations or credibility intervals about the parameters, something which is not so straightforward under a variational EM or even EM framework.

Derivation of the CAVI algorithm and ELBO for specific models is certainly more tedious than the derivation of the variational EM algorithm. Often, quantities that are required in the derivation include  $E(\theta)$ ,  $E(\theta^2)$ ,  $E(\theta^{-1})$ ,  $E(\log \theta)$  or any other moment of

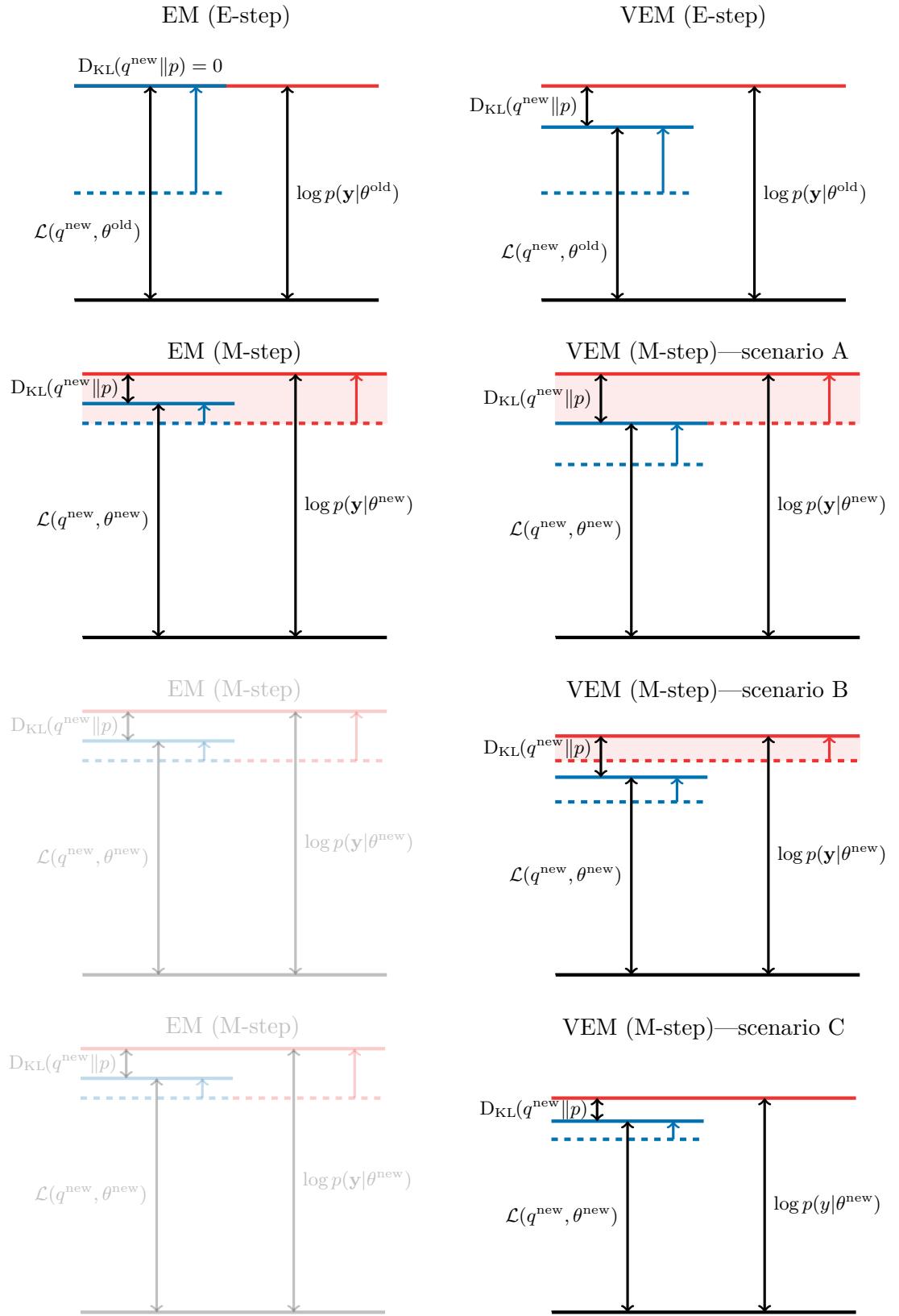


Figure S4.3: Illustration of EM vs Variational EM (VEM) algorithms. Whereas the EM guarantees an increase in log-likelihood value (red shaded region), the VEM does not.

Table S4.1: Comparison between variational inference and variational EM.

| Variational inference                                                                                                            | Variational EM                                                                                         |
|----------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| <b>GOAL:</b> Posterior densities for $(\mathbf{w}, \theta)$                                                                      | <b>GOAL:</b> ML/MAP estimates for $\theta$                                                             |
| Variational approximation for latent variables and parameters $q(\mathbf{w}, \theta) \approx p(\mathbf{w}, \theta   \mathbf{y})$ | Variational approximation for latent variables only $q(\mathbf{w}) \approx p(\mathbf{w}   \mathbf{y})$ |
| Priors required on $\theta$                                                                                                      | Priors not necessary for $\theta$                                                                      |
| Derivation can be tedious                                                                                                        | Derivation less tedious                                                                                |
| Inference on $\theta$ through posterior density $q(\theta)$                                                                      | Asymptotic distribution of $\theta$ not well studied; standard errors for $\theta$ not easily obtained |
| Suited to conjugate exponential family models: posteriors will be easily recognisable                                            | Suited to conjugate exponential family models, but not necessary                                       |

some function of  $\theta$ , where expectations are taken under the approximating  $q$  posterior density. For certain distributions  $q(\theta)$  these quantities can be awkward to compute, and may need approximating themselves.

The computational time and storage requirements of variational methods is virtually the same as EM algorithm ([beal2003variational](#); [blei2017variational](#)). Consider the mean-field variational approximation. In variational inference or variational EM, the updating step for the factors involve

$$\tilde{q}_k^{(t+1)}(w_{[k]}) \leftarrow \text{const.} \times \exp \left( \mathbb{E}_{\mathbf{w}_{-k} \sim \tilde{q}^{(t)}} [\log p(\mathbf{y}, \mathbf{w})] \right), \quad (\text{S4.9})$$

for each of the factors of the approximate posterior  $q(\mathbf{w}) = \prod_{k=1}^M q_k(w_{[k]})$ . In the EM algorithm E-step, one obtains the  $Q$  function

$$Q(\theta | \theta^{(t)}) = \mathbb{E}_{\mathbf{w}} (\log p(\mathbf{y}, \mathbf{w}) | \mathbf{y}, \theta^{(t)}). \quad (\text{S4.10})$$

We can see that in both equations (S4.9) and (S4.10), there is a need to compute the expectation of the joint log density, but the difference between the variational inference and EM or variational EM lies in the M-step. In variational inference one seeks a distribution, while in EM or variational EM one seeks a point estimate (posterior mode) of this distribution.

## Appendix A

# Functional derivative of the entropy

We present the functional derivative of the entropy  $H(p)$  in [Equation 3.6 \(p. 94\)](#). 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  $\rho$  with certain boundary conditions, and a functional  $F : M \rightarrow \mathbb{R}$ , the functional derivative of  $F(\rho)$  with respect to  $\rho$ , denoted  $\partial F / \partial \rho$ , is defined by

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

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

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

as the directional derivative at point  $\rho$  in the direction of  $\phi$ . 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  $\nu$  on  $\Omega$ , we can write

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

Using the definition of functional derivatives, we find that

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

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  $L^2(\Theta, \nu)$  with inner product  $\langle p, p' \rangle_{L^2(\Theta, \nu)} = \int pp' d\nu$ . 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:

$$\begin{aligned} \partial_q H(p) &= \frac{d}{dt} H(p + tq) \Big|_{t=0} \\ &= \frac{d}{dt} \left\{ - \int_{\Theta} (p(\theta) + tq(\theta)) \log (p(\theta) + tq(\theta)) d\nu(\theta) \right\} \Big|_{t=0} \\ &= - \int_{\Theta} \left\{ \frac{d}{dt} (p(\theta) + tq(\theta)) \log (p(\theta) + tq(\theta)) \Big|_{t=0} \right\} d\nu(\theta) \\ &= - \int_{\Theta} \left( \frac{p(\theta)q(\theta)}{p(\theta) + tq(\theta)} + \frac{tq(\theta)^2}{p(\theta) + tq(\theta)} + q(\theta) \log (p(\theta) + tq(\theta)) \right) \Big|_{t=0} d\nu(\theta) \\ &= - \int_{\Theta} q(\theta) (1 + \log p(\theta)) d\nu(\theta) \end{aligned}$$

$$\begin{aligned}
&= \langle -(1 + \log p), q \rangle_{\Theta} \\
&= dH(p)(q).
\end{aligned}$$

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.

**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 ([zhang2013kronecker](#)):

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

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

- **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  $\text{tr}(A \otimes B) = \text{tr}(A) \text{tr}(B)$ .
- **Rank.**  $\text{rank}(A \otimes B) = \text{rank}(A) \text{rank}(B)$ .
- **Matrix equations.**  $AXB = C \Leftrightarrow (B^\top \otimes A) \text{vec } X = \text{vec } C$ .

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

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

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

## Appendix C

# Statistical distributions and their properties

This appendix is intended as a reference relating to the multivariate normal, matrix normal, truncated univariate and multivariate normal, gamma and inverse gamma distributions, which are collated from various sources for convenience. Of interest are their probability density functions, first and second moments, and entropy (Definition 3.5, p. 94). Note that in this part of the appendix, boldface notation for matrix and vectors are not used.

### C.1 Multivariate normal distribution

**Definition C.1** (Multivariate normal distribution). 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  $\Sigma \in \mathbb{R}^{d \times d}$  (a square, symmetric, positive-definite matrix). We say that  $X \sim N_d(\mu, \Sigma)$ . Then,

- **Pdf.**  $p(X|\mu, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(X - \mu)^\top \Sigma^{-1} (X - \mu)\right).$
- **Moments.**  $E X = \mu$ ,  $E(XX^\top) = \Sigma + \mu\mu^\top$ .
- **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  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ , then the components of  $X = (X_1, \dots, X_d)^\top$  are independently distributed according to  $X_i \sim N(\mu_i, \sigma_i^2)$ .

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

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

Then,

- **Marginal distributions.**

$$X_a \sim N_{\dim X_a}(\mu_a, \Sigma_a) \quad \text{and} \quad X_b \sim N_{\dim X_b}(\mu_b, \Sigma_b).$$

- **Conditional distributions.**

$$X_a | X_b \sim N_{\dim X_a}(\tilde{\mu}_a, \tilde{\Sigma}_a) \quad \text{and} \quad X_b \sim N_{\dim X_b}(\tilde{\mu}_b, \tilde{\Sigma}_b),$$

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.**

$$AX + BY + C \sim N_d(A\mu + B\nu + C, A\Sigma A^\top + B\Psi B^\top)$$

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  $\mu \sim N(\nu, \Sigma + \Psi)$ .

*Proof.* Omitted—see **petersen2008matrix**. ■

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,

$$\begin{aligned}-\frac{1}{2}x^\top Ax + b^\top x &= -\frac{1}{2}(x - A^{-1}b)^\top A(x - A^{-1}b) + \frac{1}{2}b^\top A^{-1}b \\ -\frac{1}{2}\text{tr}(X^\top AX) + \text{tr}(B^\top X) &= -\frac{1}{2}\text{tr}((X - A^{-1}B)^\top A(X - A^{-1}B)) + \frac{1}{2}\text{tr}(B^\top A^{-1}B).\end{aligned}$$

*Proof.* Omitted—see [petersen2008matrix](#). ■

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

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

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

*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:

$$\frac{\partial}{\partial z} \text{tr} \Sigma = \text{tr} \frac{\partial \Sigma}{\partial z} \quad (\text{C.2})$$

$$\frac{\partial}{\partial z} \log |\Sigma| = \text{tr} \left( \Sigma^{-1} \frac{\partial \Sigma}{\partial z} \right) \quad (\text{C.3})$$

$$\frac{\partial \Sigma^{-1}}{\partial z} = -\Sigma^{-1} \frac{\partial \Sigma}{\partial z} \Sigma^{-1} \quad (\text{C.4})$$

A useful reference for these identities is [petersen2008matrix](#).

Differentiating the log-likelihood for  $\theta$  with respect to the  $i$ 'th component of  $\theta$  yields

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

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

- (A)

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

- (B)

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

- (C)

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

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

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

as required. ■

## C.2 Matrix normal distribution

**Definition C.2** (Matrix normal distribution). 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  $X \sim \text{MN}_{n,m}(\mu, \Sigma, \Psi)$ . Then,

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

The matrix normal distribution is simply an extension of the Gaussian distribution to matrices. A matrix normal random variable can be expressed as a multivariate normal random variable.

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

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

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

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.  $\blacksquare$

Some useful properties of the matrix normal distribution are listed:

- **Expected values.**

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

- **Transpose.**  $X^\top \sim \text{MN}_{m,n}(\mu^\top, \Psi, \Sigma)$ .
- **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  $AXB \sim \text{MN}_{a,b}(\mu^\top, A\Sigma A^\top, B^\top \Psi B)$ .
- **Iid.** If  $X_i \stackrel{\text{iid}}{\sim} N_m(\mu, \Psi)$  for  $i = 1, \dots, n$ , and we arranged these vectors row-wise into the matrix  $X = (X_1^\top, \dots, X_n^\top)^\top \in \mathbb{R}^{n \times m}$ , then  $X \sim \text{MN}(1_n \mu^\top, I_n, \Psi)$ .

### C.3 Truncated univariate normal distribution

**Definition C.3** (Truncated univariate normal distribution). Let  $X \sim N(\mu, \sigma^2)$  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  $X \sim {}^tN(\mu, \sigma^2, a, b)$ . 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.**

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

- **Entropy.**

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

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

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

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

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

with  $U \sim \text{Unif}(0, 1)$  will work too. Either of these methods will work reasonably well as long as the truncation region is not too far away from  $\mu$ , but neither is particularly efficient. Efficient algorithms have been explored which are along the lines of either accept/reject algorithms ([robert1995simulation](#)), Gibbs sampling ([damien2001sampling](#)), or pseudo-random number generation algorithms ([chopin2011fast](#)). The latter algorithm is inspired by the Ziggurat algorithm ([marsaglia2000ziggurat](#)) which is considered to be the fastest Gaussian random number generator.

## C.4 Truncated multivariate normal distribution

**Definition C.4** (Truncated multivariate normal distribution). Consider the restriction of  $X \sim N_d(\mu, \Sigma)$  to a convex subset<sup>1</sup>  $\mathcal{A} \subset \mathbb{R}^d$ . Call this distribution the truncated multivariate normal distribution, and denote it  $X \sim {}^tN_d(\mu, \Sigma, \mathcal{A})$ . The pdf is  $p(X|\mu, \Sigma, \mathcal{A}) = C^{-1}\phi(X|\mu, \Sigma)\mathbf{1}(X \in \mathcal{A})$ , where

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

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

Sampling from a truncated multivariate normal distribution is described by [robert1995simulation](#), who used a Gibbs-based approach, which we now describe. Assume that the one-dimensional slices of  $\mathcal{A}$

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

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

---

<sup>1</sup>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

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

According to **robert1995simulation**, 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  $\Sigma^{-1}$ . Therefore, the Gibbs sampler makes draws from truncated normal distributions in the following sequence, given initial values  $X^{(0)}$ :

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

In a later work, **damien2001sampling** introduce a latent variable  $Y \in \mathbb{R}$  such that the joint pdf of  $X$  and  $Y$  is

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

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

In the case where  $\Sigma$  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.

**Lemma C.5.** *Let  $X \sim {}^t\text{N}_d(\mu, \Sigma, \mathcal{C}_j)$ , with  $\mu = (\mu_1, \dots, \mu_d)^\top$  and  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ , and  $\mathcal{C}_j = \{X_j > X_k | k \neq j, \text{and } k = 1, \dots, m\}$  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  $\phi$  is the pdf of a standard normal distribution and

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

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

(ii) **Moments.** *The expectation  $\mathbb{E} X = (\mathbb{E} X_1, \dots, \mathbb{E} X_d)^\top$  is given by*

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

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

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

where we had defined

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

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

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

*Proof.* See Appendix D for the proof. ■

## C.5 Gamma distribution

**Definition C.5** (Gamma distribution). 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.**  $\mathbb{E} X = s/r$ ,  $\text{Var } X = s/r^2$ .
- **Entropy.**  $H(p) = s - \log r + \log \Gamma(s) + (1-s)\psi(s)$ .

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

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

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

**Definition C.6** (Inverse gamma distribution). For  $X \in \mathbb{R}_{\geq 0}$ , a random variable  $X$  distributed according to an inverse gamma distribution with parameters  $s$  (shape) and  $\sigma$  (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.**  $\mathbb{E} X = \sigma/(s-1)$ ,  $\text{Var } X = \sigma^2((s-1)^2(s-2))^{-1}$ .
- **Entropy.**  $H(p) = s + \log(\sigma \Gamma(s)) - (1+s)\psi(s)$ .

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

**Lemma C.6.** *If  $X \sim \Gamma(s, r)$  (shape and rate parameterisation), 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 conical truncations of multivariate normals

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

### D.1 Proof of Lemma C.5: Pdf

Using the fact that  $\int p(x) dx = 1$ , and that

$$\begin{aligned}
& \int \cdots \int [x_i < x_j, \forall i \neq j] \cdot \prod_{i=1}^d \phi(x_i | \mu_i, \sigma_i^2) dx_1 \cdots dx_d \\
&= \int \cdots \int \mathbb{1}[x_i < x_j, \forall i \neq j] \prod_{i=1}^d \left[ \frac{1}{\sigma_i} \phi\left(\frac{x_i - \mu_i}{\sigma_i}\right) \right] dx_1 \cdots dx_d \\
&= \int \cdots \int \mathbb{1}[x_i < x_j, \forall i \neq j] \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) \prod_{\substack{i=1 \\ i \neq j}}^d \left[ \frac{1}{\sigma_i} \phi\left(\frac{x_i - \mu_i}{\sigma_i}\right) \right] dx_1 \cdots dx_d \\
&= \int \prod_{\substack{i=1 \\ i \neq j}}^d \Phi\left(\frac{x_j - \mu_i}{\sigma_i}\right) \frac{1}{\sigma_j} \phi\left(\frac{x_j - \mu_j}{\sigma_j}\right) dx_j \\
&= \int \prod_{\substack{i=1 \\ i \neq j}}^d \Phi\left(\frac{\sigma_j z + \mu_j - \mu_i}{\sigma_i}\right) \phi(z) dz \\
&\quad (\text{by using the standardisation } z = (x_j - \mu_j)/\sigma_j)
\end{aligned}$$

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

the proof follows directly.

## D.2 Proof of Lemma C.5: Moments

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

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

and in particular, we have

$$\mathbb{E}(Y - \mu) = -\sigma \frac{\phi(\beta)}{\Phi(\beta)} \quad (\text{D.1})$$

$$\mathbb{E}(Y - \mu)^2 - \sigma^2 = -\sigma^2 \frac{\beta \phi(\beta)}{\Phi(\beta)} \quad (\text{D.2})$$

where  $\beta = (b - \mu)/\sigma$ . For the conically truncated multivariate normal distribution  $X \sim {}^t\text{N}_d(\mu, \Sigma, \mathcal{A}_j)$ , where  $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_d^2)$ , the independence structure of  $\Sigma$  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

$$\begin{aligned} \mathbb{E}[g(X_i)] &= C^{-1} \int \cdots \int [x_k < x_j, \forall k \neq j] \cdot g(x_i) \prod_{k=1}^d p(x_k) dx_1 \cdots dx_d \\ &= C^{-1} \frac{\Phi((x_j - \mu_j)/\sigma_j)}{\Phi((x_j - \mu_j)/\sigma_j)} \iint [x_i < x_j] \cdot g(x_i) p(x_i) p(x_j) \prod_{\substack{k=1 \\ k \neq i, j}}^d \Phi \left( \frac{x_j - \mu_k}{\sigma_k} \right) dx_i dx_j \\ &= C^{-1} \int \mathbb{E}_{X_i \sim {}^t\text{N}(\mu_i, \sigma_i^2, -\infty, x_j)} [g(X_i)] \prod_{\substack{k=1 \\ k \neq j}}^d \Phi \left( \frac{x_j - \mu_k}{\sigma_k} \right) p(x_j) dx_j \end{aligned} \quad (\text{D.3})$$

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

$$\begin{aligned} \mathbb{E}[g(X_j)] &= C^{-1} \int \cdots \int [x_k < x_j, \forall k \neq j] \cdot g(x_j) \prod_{k=1}^d p(x_k) dx_1 \cdots dx_d \\ &= C^{-1} \int g(x_j) \prod_{\substack{k=1 \\ k \neq j}}^d \Phi \left( \frac{x_j - \mu_k}{\sigma_k} \right) p(x_j) dx_d. \end{aligned} \quad (\text{D.4})$$

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

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

where  $Z$  is the distribution of  $N(0, 1)$ , 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  $dx_j = \sigma_j dz$ . For the  $j$ 'th component, substitute  $g(x_j) = x_j - \mu_j$  in (D.4) to get

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

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

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

And similarly, for the  $j$ 'th component

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

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

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

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

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

*Proof.* Use the fact that for any differentiable function  $g$ ,  $E[Zg(Z)] = E[g'(Z)]$ , 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:

$$\begin{aligned}\phi_i &= \phi(\sigma_i z + \mu_i) \\ \Phi_i &= \Phi(\sigma_i z + \mu_i)\end{aligned}$$

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  $m \in \{\mathbb{N} \mid m > 1\}$ . 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,

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

## Appendix E

# I-prior interpretation of the $g$ -prior

The I-prior for  $\beta$  in a standard linear model resembles the objective  $g$ -prior ([zellner1986assessing](#)) for regression coefficients,

$$\beta \sim N_p(\mathbf{0}, g(\mathbf{X}^\top \Psi \mathbf{X})^{-1}),$$

although they are quite different objects. The  $g$ -prior for  $\beta$  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  $\beta$  corresponds to a small prior variance, and hence less deviation away from the prior mean of zero in estimating  $\beta$ . 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  $\langle \mathbf{x}, \mathbf{x}' \rangle_{\mathcal{X}} = \mathbf{x}^\top (\mathbf{X}^\top \Psi \mathbf{X})^{-1} \mathbf{x}$ . With this inner product and from (3.3) (p. 88), the Fisher information on  $\beta$  is

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

and this, rather than the usual  $\mathbf{X}^\top \Psi \mathbf{X}$  as the prior covariance matrix for  $\beta$ , 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,  $k = 1, \dots, p$  for each unit  $i = 1, \dots, n$  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,  $\text{cm}^2$  and  $\text{kg}^2$  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. 100). These details relate to the standard linear multilevel model and the naïve classification model.

### F.1 The I-prior for standard multilevel models

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.,  $\Psi = \psi \mathbf{I}_n$ . The form of  $f \in \mathcal{F}$  is now  $f(\mathbf{x}_i^{(j)}, j) = \sum_{i'=1}^{n_j} \sum_{j'=1}^m h_\lambda((\mathbf{x}_i^{(j)}, j), (\mathbf{x}_{i'}^{(j')}, j')) w_{i'j'}$ , where each  $w_{i'j'} \sim N(0, \psi^{-1})$ .

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  $\beta_{0j}$  is normal with zero mean and variance

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

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

$$\begin{aligned}\text{Cov}(\beta_{0j}, \beta_{0j'}) &= \text{Cov} \left[ \lambda_2 \left( \frac{w_{+j}}{p_j} - w_{++} \right), \lambda_2 \left( \frac{w_{+j'}}{p_{j'}} - w_{++} \right) \right] \\ &= \frac{\lambda_2^2}{p_j p_{j'}} \underbrace{\text{Cov}(w_{+j}, w_{+j'})}_0 - \frac{\lambda_2^2}{p_j} \text{Cov}(w_{+j}, w_{++}) - \frac{\lambda_2^2}{p_{j'}} \text{Cov}(w_{++}, w_{+j'}) \\ &\quad + \lambda_2^2 \text{Cov}(w_{++}, w_{++}) \\ &= -\frac{\lambda_2^2}{n_j/n} n_j \psi - \frac{\lambda_2^2}{n_{j'}/n} n_{j'} \psi + \lambda_2^2 n \psi \\ &= -n\psi \lambda_2^2.\end{aligned}$$

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

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

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

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

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

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

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

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

and

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

## 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  $\mathcal{F}_M$  and  $\mathcal{F}_X$  need necessarily be zero-mean functions (as per the functional ANOVA definition in Definition 2.36 (p. 77), 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,

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

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  $\epsilon_{ij}$ ). 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  $\Psi$ . This means that  $\mathbf{w}$  follows a matrix normal distribution  $MN_{n,m}(\mathbf{0}, \mathbf{I}_n, \Psi)$ , which implies  $\text{vec } \mathbf{w} \sim N_{nm}(\mathbf{0}, \Psi \otimes \mathbf{I}_n)$ , and similarly,  $\boldsymbol{\epsilon} \sim N_{nm}(\mathbf{0}, \Psi^{-1} \otimes \mathbf{I}_n)$ . 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  $\text{vec } \mathbf{f} \sim N_{nm}(\mathbf{0}, \mathbf{A} \Psi \mathbf{A} \otimes \mathbf{B}_\eta^2)$ . 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

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

and

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

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

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

Suppose hypothetically, one uses the uncentered identity kernel  $a(j, j') = \delta_{jj'}$ , in which case centring of the intercepts  $\alpha_j$  must be handled separately. In conjunction with an assumption of iid errors ( $\Psi = \psi \mathbf{I}_n$ ), the above distributions simplify further.

Specifically, the variance in the marginal distribution becomes

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

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  $\text{Var}(\text{vec } \mathbf{f}) = \text{diag}(\psi \mathbf{B}_\eta^2, \dots, \psi \mathbf{B}_\eta^2)$ , which could be interpreted as having independent and identical I-priors on the regression functions for each class  $\mathbf{f}_{\cdot j} = (f(x_1, j), \dots, f(x_n, j))^\top$ .



## Appendix G

# Posterior distribution of the I-prior regression function

We derive the posterior distribution for the I-prior random effects  $\mathbf{w} = (w_1, \dots, w_n)^\top$ , 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,  $\mathbf{f} := (f(x_1), \dots, f(x_n))^\top = \mathbf{H}_\eta \mathbf{w}$ , and  $f \in \mathcal{F}$  an RKHS with kernel  $h_\eta$ . A closely related distribution of interest is the posterior predictive distribution of  $y_{\text{new}}$ , the prediction at a new data point  $x_{\text{new}}$ . We note the similarity of these results with the posterior distributions of Gaussian process regressions ([rasmussen2006gaussian](#)).

### G.1 Deriving the posterior distribution for $\mathbf{w}$

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

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

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

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

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

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

Thus,

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

and

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

as a direct consequence of the Woodbury matrix identity ([petersen2008matrix](#)).

## G.2 Deriving the posterior predictive distribution

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

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

where

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

The vector of covariances  $\boldsymbol{\sigma}_{\text{new}}$  between observations  $y_1, \dots, y_n$  and the predicted point  $y_{\text{new}}$  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

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

and variance

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



## Appendix H

# Variational EM algorithm for I-probit models

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  $\theta$  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

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

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

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.

- $p(\mathbf{y}|\mathbf{y}^*)$ . 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.

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

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

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

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

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

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  $\boldsymbol{\mu}_{i\cdot}$  and variance  $\boldsymbol{\Psi}^{-1}$ .

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

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

### 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

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

where  $\tilde{\boldsymbol{\mu}}_{i\cdot} = \boldsymbol{\alpha} + \tilde{\mathbf{w}}\mathbf{h}_\eta(x_i)$ ,  $\tilde{\mathbf{w}} = \mathbb{E}_{\mathbf{w} \sim \tilde{q}}[\mathbf{w}]$ . 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. 267).

*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,

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

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

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

$$\begin{aligned} \text{vec } \mathbf{y}^* | \boldsymbol{\alpha}, \mathbf{w}, \eta, \boldsymbol{\Psi} &\sim N_{nm} \left( \text{vec}(\mathbf{1}_n \boldsymbol{\alpha}^\top + \mathbf{H}_\eta \mathbf{w}), \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_n \right) \\ &\quad \text{and} \\ \text{vec } \mathbf{w} | \boldsymbol{\Psi} &\sim N_{nm}(\mathbf{0}, \boldsymbol{\Psi} \otimes \mathbf{I}_n) \end{aligned}$$

using properties of matrix normal distributions.

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

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

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

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

and

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

We will often refer to  $\tilde{\mathbf{w}}$  as the  $n \times m$  matrix constructed by filling in its entries with  $\text{vec} \tilde{\mathbf{w}}$  column-wise (akin to the opposite of vectorisation). This way, the  $\tilde{\mathbf{w}}$  contains posterior mean values arranged by class  $j = 1, \dots, m$  column-wise, and by observations  $i = 1, \dots, n$  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 (p. 175) for details.

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

$$\begin{aligned}\tilde{\mathbf{V}}_w &= (\text{diag}(\psi_1, \dots, \psi_m) \otimes \mathbf{H}_\eta^2 + \text{diag}(\psi_1, \dots, \psi_m) \otimes \mathbf{I}_n)^{-1} \\ &= \text{diag}\left((\psi_1 \mathbf{H}_\eta^2 + \psi_1^{-1} \mathbf{I}_n)^{-1}, \dots, (\psi_m \mathbf{H}_\eta^2 + \psi_m^{-1} \mathbf{I}_n)^{-1}\right) \\ &=: \text{diag}(\tilde{\mathbf{V}}_{w_1}, \dots, \tilde{\mathbf{V}}_{w_m}).\end{aligned}$$

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

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

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

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

A quantity that we will be requiring time and again will be  $\text{tr}(\mathbf{C} \mathbf{E}[\mathbf{w}^\top \mathbf{D} \mathbf{w}])$ , 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

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

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

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

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  $\delta$  is the Kronecker delta. Continuing on (H.1) leads us to

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

If  $\mathbf{C} = \text{diag}(c_1, \dots, c_m)$ , then

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

## H.2 Deriving the ELBO expression

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

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

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 $\mathbf{y}^*$

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

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

### H.2.2 Terms involving distributions of $\mathbf{w}$

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



## Appendix I

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

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

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

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  $\Theta$  the entire set of parameters, while  $\Theta_{-\xi}$  implies the set of parameters excluding the parameter  $\xi$ .

## I.1 Conditional posterior for $\beta$

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

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  $\gamma$ . This is of course recognised as the log density of a  $p$ -variate normal distribution with mean and variance

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

## I.2 Conditional posterior for $\gamma$

Consider each  $\gamma_j$  in turn. For  $j \in \{1, \dots, p\}$ ,

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

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

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

while for  $\gamma_j = 0$ , we have

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

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

$$\theta_j^{[\omega]} = \begin{cases} (\theta_1, \dots, \theta_{j-1}, \beta_j, \theta_{j+1}, \dots, \theta_p)^\top & \omega = 1 \\ (\theta_1, \dots, \theta_{j-1}, 0, \theta_{j+1}, \dots, \theta_p)^\top & \omega = 0. \end{cases}$$

Therefore, the conditional distribution for  $\gamma_j$  is Bernoulli with success probability

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

### I.3 Conditional posterior for $\alpha$

We can obtain the conditional posterior for  $\alpha$  in a similar fashion we obtained the conditional posterior for  $\beta$ . 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  $\alpha$  is normal with mean and variance which can be easily read off the final line above.

### I.4 Conditional posterior for $\sigma^2$

The conditional density for  $\sigma^2$  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 $\kappa$

Interestingly, since  $\kappa$  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  $\beta$ .

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

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

## 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 ([plummer2003jags](#)), 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  $\beta$ . Overall, if  $n \gg p$ , then time complexity is  $O(np^2)$ . Storage requirements are  $O(np)$ .

# Index

- additive model, 34, 137, 194
- aerobic data set, 215
- ANOVA
  - functional decomposition, 76
  - kernel/RKKS, 79, 103–190
  - multiway, 76
  - two-way, 75
- Bayes factor, 173, 205, 232
  - empirical, 127, 172
- Bayesian hypothesis test, 205, 208
- Bayesian model averaging, 203
- Bayesian model selection, 202
- Bayesian variable selection, 205
  - consistency, 227
- bilinear
  - form, 42, 46, 50, 89
- bootstrap, 172
- Borel space, 51
- bounded, 47
- bovine tuberculosis (BTB) data set, 190
- Brier score, 163
- Brownian bridge, 108
- Brownian motion, 64
- canonical kernel/RKHS, 62, 102, 129, 137, 179, 182, 186
  - centred, 63
- cardiac arrhythmia data set, 179
- categorical response, 147
- Cauchy sequence, 45
- Cauchy-Schwarz inequality, 45, 56, 65
- centring
  - RKHS, 107
- Cholesky decomposition, 111, 174
- classification
  - binary, 160, 179
  - multiclass, 186
- naïve, 283
- closed subspace, 46
- continuity
  - Hölder, 61, 65
  - Lipschitz, 47
- convergence, 45
- coordinate ascent variational inference (CAVI), 251
- covariance operator, 52
- cow growth data set, 134
- credibility interval, 128, 171
- curvature, 82
- degree, *see also* polynomial kernel, 73
- density, 51
- Deterding data set, *see* vowel recognition data set
- digamma function, 272
- Dirac functional, *see* evaluation functional
- distribution, 51
  - categorical, 147
  - folded-normal, 115
  - gamma, 272
    - inverse gamma, 272
    - matrix normal, 267
    - multivariate normal, 263
    - truncated multivariate normal, 269
- dual space
  - algebraic dual, 48
  - continuous dual, 48, 62
- ECM algorithm, 114, 167
- ELBO, 158, 250
- EM algorithm, 113, 235
  - Bayesian, 239
    - exponential family, 122, 238
- empirical Bayes factor, *see* Bayes factor

empirical distribution, 63, 107  
 entropy, 96, 257  
 evaluation functional, 53  
     continuous, 53  
 evidence, *see* model evidence  
 evidence lower bound, *see* ELBO  
 expectation propagation, 194  
 exponential family distribution, 238  
  
 false choice (inclusion/exclusion), 212  
 fBm kernel/RKHS, 64, 107, 115, 134,  
     142, 160, 179, 186, 190  
     centred, 64  
 feature map/space, 35, 55, 63, 64  
 Fisher information, 82  
     Gaussian, 113, 265  
     regression function, 223  
 fractional Brownian motion, *see* fBm  
 Fréchet derivative, 83–84, 258  
 functional derivative, 257  
 functional regression, 35, 109, 134, 137  
 fundamental projection, 59  
 fundamental symmetry, 59  
  
 $g$ -prior, 206  
 gamma function, 272  
 Gâteaux derivative, 84–86, 258  
 Gaussian process  
     classification, 194  
 Gaussian process, 64  
     classification, 179  
     regression, 37–38, 141, 287  
 Gaussian vector, 52  
 Geweke-Hajivassiliou-Keane (GHK)  
     simulator, 173  
 Gibbs sampler, 207  
 gradient  
     Hilbert space, 86  
 Gram matrix, *see* kernel matrix  
  
 Hadamard product, 72  
 Hamiltonian Monte Carlo, *see* HMC  
 Hessian, 157  
     Hilbert space, 87  
 highest probability model, 204  
 Hilbert space<sup>306</sup>  
     associated Hilbert space, 59  
     Fisher information, 83  
     pre-Hilbert space, 46  
  
 HMC, 114, 159, 241  
 Hurst, *see also* fBm kernel, 64  
 Hölder continuous, *see also* continuity,  
     65  
  
 I-prior, 33, 81, 99, 223  
     categorical, *see* I-probit  
     log-likelihood, 111  
     model, 37, 61, 109  
     posterior distribution, 34, 38, 50,  
         110, 287  
     posterior predictive distribution,  
         125, 288  
     theorem, 97  
 I-probit, 149  
     ELBO, 168  
     log-likelihood, 155  
     posterior distribution, 169  
     posterior predictive distribution,  
         171  
 identifiability, 116, 152  
 independent of irrelevant alternatives  
     (IIA), 153  
 inner product, 44  
     dot product, 92  
     indefinite, 58  
     negative-definite, 58  
     positive-definite, 44  
 insulin-like growth factor (IGF-I) data  
     set, 129  
 interaction, 79, 220  
 intercept, 110  
 isometry, 48  
 isomorphism, 48  
 Iverson bracket, 148  
  
 Jacobian, 232  
 Jeffreys prior, 206  
  
 $k$ -nearest neighbours, 179, 183  
 kernel, 35, 54  
     front loading, 119  
     matrix, 37, 55  
     method, 55  
     reproducing, 54  
     trick, 55  
     universality, 69  
 kernel smoothing, 137  
 KL divergence, 159, 164

Kronecker product, 177, 261  
 Kullback-Leibler, *see* KL divergence  
 L-BFGS algorithm, 112, 139, 168  
 Landau notation, 84  
 Laplace's method, 156  
 Lasso, 179, 183, 202, 212  
 latent propensity, 149  
 latent variable, 149  
 least squares, 36  
     partial, 137  
 lengthscale, *see also* SE kernel  
 linear  
     functional, 46  
 linear discriminant analysis, 179  
 linear kernel, *see* canonical kernel  
 linear regression, 102, 201  
 linear space, *see* vector space  
 log odds ratio, *see also* odds ratio, 184  
 log-likelihood  
     complete data, 123, 235  
     full data, *see* complete data  
     incomplete data, 235  
 logistic, 148, 197, 227  
 longitudinal model, 34, 134  
 Mallow's  $C_p$ , 217  
 MAP estimate, 124, 232  
 marginal likelihood, 235  
 Markov chain Monte Carlo, *see* MCMC  
 matrix normal distribution, 151  
 maximum a posteriori estimate, *see*  
     MAP estimate  
 maximum entropy, 81, 96  
 maximum likelihood, 82  
     estimate, 36, 111  
     penalised, 36, 124  
     Type-II, 39, 233  
 MCMC, 114, 159, 231  
 median probability model, 210  
 meta-analysis, 182  
 metric, 45  
 minimum mean squared error (MMSE)  
     estimate, 231  
 model evidence, 231  
 model indicator, 205  
 model probabilities, 204  
 model selection, 202  
 Monte Carlo integration, 173  
 Moore-Aronszajn theorem, 57, 60  
 mortality and air pollution data set, 217  
 multicollinearity, 211  
 multilevel model, 34, 103, 129, 281  
 negative definite, *see also* inner product  
 negative subspace, 59  
 neural network, 137  
 nonparametric, 35, 54, 223  
 norm, 44  
 normal regression model, 34  
 numerical issues, 114  
 Nyström method, 117, 142  
 odds, 183  
     ratio, 183  
 offset, *see also* polynomial kernel, 73  
 operator  
     averaging, 75  
     bounded, 47  
     covariance, *see* covariance operator  
     linear, *see* linear map  
     tensor product, 50, 89  
 orthogonal  
     complement, 49  
     decomposition, 49  
     projection, 49  
 outer product, *see also* tensor product  
 overfit, 39  
 ozone data set, 218  
 parallelogram law, 45  
 parametric, 31, 35, 223  
 Pearson kernel/RKHS, 69, 129, 134,  
     137, 142, 182, 190  
 polarisation identity, 49, 64  
 polynomial kernel/RKKS, 73, 137  
 positive definite, *see also* inner product,  
     55, 65  
 positive subspace, 59  
 posterior inclusion probability, 208  
 posterior model probability, 208  
 posterior predictive check, 128  
 power set, 75  
 preselection, 210  
 probit, 150, 153, 227  
 quadratic discriminant analysis, 179  
 quasi-Newton method, *see also* Newton  
     method, 111

random element, 51  
 random-effects model, *see* multilevel model  
 regression, 33, 90, 102  
 regularisation, 36, 202  
     Tikhonov, 36  
 residual sum of squares, 208  
 reverse triangle inequality, 45  
 ridge regression, 202  
 Riesz representation theorem, 48, 54, 62  
 Riesz-Fréchet theorem, *see* Riesz representation theorem  
 risk functional, 36  
 RKHS, 31, 35, 53  
     building block, 61  
     centring, 67  
     uniqueness, 56  
 RKKS, 31, 35, 58  
     uniqueness, 60  
 score function, 82  
 SE kernel/RKHS, 67, 137, 179, 186  
     centred, 69  
 sigma algebra, *see also* Borel space, 51  
 signal-to-noise ratio (SNR), 211  
 smoking cessation data set, 182  
 smoothing model, 35, 107, 115  
 spatio-temporal model, 190  
 spike-and-slab prior, 206  
 squared exponential, *see* SE  
 standard error, 113, 172  
 support vector machine, 175, 179, 183, 194  
 Tecator data set, 137, 226  
 tensor  
     product, 49  
     sum, 49  
 triangle inequality, 45  
 two stage procedure, 210  
 undersmooth, 39  
 variational Bayes, 197  
 variational calculus, 96  
 variational EM algorithm, 158, 163, 252  
     I-probit, 291  
     standard error, 172, 226  
 variational inference, 144, 247  
 vector space, 44  
     complete, 46  
 vectorisation, 261  
 vowel recognition data set, 186  
 Wiener process, *see* Brownian motion  
 Woodbury matrix identity, 117, 288