
To-do list

1. A lot of material here. Need to sort it out	1
2. Split into two: BVS by calculating explicit Bayes factors, and when p is large, need to use MCMC methods. Discussion on differences.	1
3. Investigate if letting the model estimate the common prior probability for the gamma variables will improve results.	8
4. The use of I-priors in Bayesian variable selection needs more convincing theoretical justification.	10
5. Under what conditions exactly are I-priors advantageous to be used for Bayesian variable selection? Design further simulation studies to gain insight.	11
6. How does the LASSO variable selection compare against I-priors?	11

Contents

8 I-prior BVS	1
8.1 Model selection (Empirical Bayes Factors)	1
8.1.1 Overview of Bayesian variable selection methods	5
8.1.2 The I-prior Bayesian variable selection model (I-prior)	9
8.1.3 Simulation study	11
8.1.4 Two-stage procedure	14
8.1.5 Real world applications	14
8.2 Special case: The canonical RKHS	20
8.3 Bayesian model selection	20
8.4 BVS using I-priors	20
8.5 Simulation study	20
8.6 Real-data examples	20
Bibliography	20
List of Figures	21
List of Tables	22

List of Theorems	23
------------------	----

List of Definitions	24
---------------------	----

Chapter 8

I-prior BVS

8.1 Model selection (Empirical Bayes Factors)

A lot of material here. Need to sort it out .

Split into two: BVS by calculating explicit Bayes factors, and when p is large, need to use MCMC methods. Discussion on differences.

Consider the linear regression model, where an $n \times 1$ response vector $\mathbf{y} = (y_1, \dots, y_n)$ relates to several predictors or covariates linearly through the following equation:

$$\begin{aligned}\mathbf{y} &= \boldsymbol{\alpha} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \psi^{-1}\mathbf{I}_n)\end{aligned}\tag{8.1}$$

where $\boldsymbol{\alpha}$ is the vector of intercepts ($\boldsymbol{\alpha} = \alpha\mathbf{1}_n$, with $\mathbf{1}_n$ being a vector of ones), \mathbf{X} is an $n \times p$ matrix containing (column-wise) the p observed explanatory variables, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ represents the errors. This linear model is undoubtedly familiar to any statistician, albeit written slightly differently. The constant term, or intercept, α , is segregated from the vector of coefficients $\boldsymbol{\beta}$, thereby allowing us to discard the column of ones typically reserved for the intercept in the design matrix \mathbf{X} . Also, we have chosen to work with the precision of the errors ψ , instead of the usual variance $\sigma^2 = 1/\psi$. These errors are assumed to be identically distributed as normal with mean zero and variance $1/\psi$, although one could of course choose to abandon this assumption by specifying $\boldsymbol{\Psi} = (\psi_{ij})$ as the variance-covariance matrix instead. All of these are chosen as a matter of convenience, especially on notation, as we will see later on.

The ordinary least squares (OLS) estimates for the regression coefficients are given as $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. This is obtained by maximising the normal likelihood of $\boldsymbol{\beta}$, but interestingly, the exact same solution is obtained by minimising the sum of squared errors - without having to set any distributional assumption on the errors. The form of the solution comes from only what is known to us: the data, \mathbf{X} and \mathbf{y} .

The Bayesian approach to estimating the linear model 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, or simply, priors. Inference about the parameters are then performed on the posterior

$$f(\boldsymbol{\Theta}|\mathbf{y}) \propto \overbrace{f(\mathbf{y}|\boldsymbol{\Theta})}^{\text{likelihood}} \times \overbrace{f(\boldsymbol{\Theta})}^{\text{prior}}$$

such as taking the mean, which is known as the Minimum Mean Squared Error estimate (MMSE), or the mode, which corresponds to the maximum a posteriori estimate (MAP). The Bayesian approach of MAP is similar to maximum likelihood, but differs only in the fact that the optimisation objective (the likelihood function) is augmented with a prior distribution about the parameters. It is critical then, that the prior chosen does not deter us in our cause of finding the correct estimates.

There are many ways of categorising different types of priors, but we like to think that priors can either be pure beliefs (subjective), or chosen according to some principle (objective). For instance, in estimating the chance of rain tomorrow, one might have their own personal feeling about this and elicit a certain probability based on no particular reason, but simply intuition. This is a subjective probability. However, one could also take into account historical data about the chances of rain on a particular day, somewhat more objectively.

In any case, we would also like to categorise priors as either being informative or uninformative, although one could always question the actual informative value in eliciting subjective priors. As the name implies, informative priors aim to help nudge the parameter estimation in the right direction, assuming the prior itself is correct. On the other hand, uninformative priors provide little or vague information about the parameters, and in these cases, the data take over and the prior has little influence on the outcome. One example is the transformation invariant **Jeffreys1946'** (**Jeffreys1946**) prior: $f(\theta) \propto \sqrt{I(\sigma)}$, where $I(\sigma)$ is the Fisher information for σ . For a scale parameter¹ $\sigma \in \mathbb{R}$, the **Jeffreys1946'** prior can be shown to be $f(\sigma) \propto 1/\sigma$, which isn't truly a distribution being a uniform distribution on the real line. Such distributions are known as improper priors. Regardless, these typically yield a proper posterior distribution which we can work with.

The type of prior that is of interest, at least for the purposes of Bayesian variable selection, is one which is objective and ideally informative. The I-prior fits this bill perfectly. A Gaussian I-prior on the regression coefficients $\boldsymbol{\beta}$ has some prior mean $\boldsymbol{\beta}_0$ and covariance matrix equal to the Fisher information for $\boldsymbol{\beta}$. This information theoretic prior for linear models has an intuitive appeal: when there is much Fisher information about the parameters, the covariance matrix for the prior will be large, and thus there will be little influence of the prior mean on the posterior estimate, and vice versa. We

¹A scale parameter σ for a family of probability distributions satisfies $F(x; \boldsymbol{\theta}, \sigma) = F(x/\sigma; \boldsymbol{\theta}, 1)$, where F is its cumulative distribution function.

typically set the prior mean to be zero for this intuition to work favourably.

We realise there is an oddity in the classification of I-priors as informative. Previously, we alluded that a prior is said to be informative if it helps zone in on the “correct” estimate with the help of this prior, e.g. a normal prior assigned to a parameter with a small variance and prior mean close to the true value (assuming this is known somehow). Conversely, an uninformative prior would have a large variance. The I-prior is either informative or uninformative depending on the amount of Fisher information. Strictly speaking, since the informative-ness of the I-prior depends on the Fisher information, which in turn depends on the data, then technically the I-prior is considered to be uninformative as there really isn’t any new information that the prior brings². Any mention of the informativeness of I-priors is then just semantic - in fact, the ‘I’ in I-prior stands for information.

Circling back to the topic of interest: variable selection, or more generally, model selection. In an ideal world, model selection entails searching the entire model space to find the “best” model based on minimising a certain criterion. There are many such criteria, making model selection a huge topic to cover fully. These include criteria such as (adjusted) R^2 , Akaike’s information criteria (AIC) and other similar information criteria, Mallows’s C_p , (k -fold) cross-validation error, and many others. The obvious issue is that when the dimension of the full model is large, then a search of the entire model space may be computationally prohibitive or even downright unfeasible.

The Bayesian philosophy to model evaluation may be thought of as follows: it is believed that a dataset \mathbf{Y} had been generated from the pdf $f(\mathbf{y}|m_k, \boldsymbol{\Theta}_k)$, where m_k is one of a set of $M = \{m_1, \dots, m_K\}$ models³, and $\boldsymbol{\Theta}_k$ are the parameters associated with this model. The goal of model selection is then to infer which of the K models had generated the data. The Bayesian approach allows us to assign priors to the parameters and the model index, i.e. $f(\boldsymbol{\Theta}_k|m_k)$ and $f(m_k)$ respectively, and thereby computing the posterior model distribution as

$$\begin{aligned} f(m_k|\mathbf{y}) &\propto f(\mathbf{y}|m_k)f(m_k) \\ &\propto \int f(\mathbf{y}|m_k, \boldsymbol{\Theta}_k)f(\boldsymbol{\Theta}_k|m_k) d\boldsymbol{\Theta}_k f(m_k). \end{aligned}$$

The natural criteria for choosing m_k is the one which gives the highest posterior probability. We refer to this model as the maximum probability model.

If we are lucky, our problem may be simple enough that we are able to calculate all of the posterior probabilities, in which case the task is as simple as reading off the maximum probability model from a list of models with their corresponding probabilities.

²This is a similar argument as to why the **Jeffreys1946**’ prior is considered uninformative. **Liu2014** studied the Kullback-Leibler divergence between the prior and posterior, and noted that this divergence is maximised using **Jeffreys1946**’ prior. In other words, this is the prior for which the data brings the maximal amount of information.

³We refer to these as models not in the usual sense - more precisely, each m_k is a model class.

However, this is likely not the case, and we often have a large model set to consider. Even if the model set is small, we might find that the integral in the posterior is not analytically tractable. In either of these cases, Markov chain Monte Carlo (MCMC) methods is suitable to be used to overcome these issues of calculating the required posterior probabilities. In fact, MCMC methods can be quite efficient in the exploration of the model space because it will favour models which have great potential of being the true model, and will tend to ignore those that have little to no potential.

While the description we have just given for model selection is generic for most statistical models, variable selection is just a special case of model selection to where the model at hand is defined by the inclusion or exclusion of a finite number of variables. The linear regression model we were describing earlier is such an example. Much work has been done on Bayesian variable selection: **George1993** **Kuo1998** and **Dellaportas2002** to name a few. We will be reviewing these methods later on, comparing similarities and differences, strengths and weaknesses, for it is these methods that we intend to improve on by using I-priors. 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.

In this section, we put our Bayesian thinking caps on. Earlier, we introduced the concept of Bayesian model evaluation. Variable selection is just a special case of this whereby a model is defined by the inclusion or exclusion of variables. The linear model (8.1) defined at the beginning is an example of this, and for the remainder of this paper, we will only consider models of this type.

A model is defined as a subset of variables selected from the full set of variables $\{X_1, \dots, X_p\}$ and is linearly related to the response variables through the model equation in (8.1). As each of the p variables can either be selected or not selected, the size of the model space is 2^p . Even for moderate p we can see how the size of the model space can become exponentially large, such that a search of the entire space would be impractical. Note that we do not consider the intercept to be selectable. If this were the case, this would imply a model as having intercept equal to zero as being possible. For most practical modelling purposes, the intercept is almost always non-zero.

It would be useful to be able to index each of these 2^p possible models somehow. We do this by introducing the model identifier vector

$$\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p).$$

This vector of length p contains elements which indicate whether or not that particular variable was selected. In other words, $\gamma_j = 1$ if X_j was selected in the model, and $\gamma_j = 0$ otherwise for $j = 1, \dots, p$. The full model, where all the variables are included in the model, is denoted by $\boldsymbol{\gamma} = (1, \dots, 1)$. The intercept only model is denoted by $\boldsymbol{\gamma} = (0, \dots, 0)$.

With this in mind, we can then assign priors to the model $f(\boldsymbol{\gamma})$, and also to the

parameters of the model $f(\Theta|\gamma)$. Ultimately, we are interested in two things:

1. **Posterior inclusion probabilities** $P[\gamma_j = 1|\mathbf{y}]$ for variable X_j , for $j = 1, \dots, p$. This gives us an indication of how often each variable was selected in the posterior models.
2. **Posterior model probabilities** $P[\gamma = \gamma_k|\mathbf{y}]$. This gives us a sense of how likely a particular model would appear a posteriori.

The posterior inclusion probabilities can be thought of as the marginals of the posterior model probabilities across each variable. Also, as the distribution on the model probabilities are on a finite set, the posterior distribution is that of a probability distribution function, hence we speak of probabilities instead of densities.

These two types of quantities can be obtained by deriving the posterior distributions for the variable selection model if they are simple enough to be obtained. Sometimes, the relevant expressions are not available in closed form. Alternatively, MCMC methods such as Gibbs sampling can be employed to provide estimates of the quantities of interest. This is perhaps the preferred option, especially when p is large such that the computation all of the 2^p posterior model probabilities takes an unfeasible amount of time. MCMC usually does not list out all of the 2^p probabilities, but instead just the ones which are substantial enough to be deemed important. Models not visited in the MCMC posterior state space are assigned probability zero. Monte-Carlo errors are inevitably introduced into the estimates, but a large enough MCMC run can control these errors.

8.1.1 Overview of Bayesian variable selection methods

We start with an overview of the available methods, in chronological order of appearance in the literature. There are many good in-depth reviews to these methods and the reader may find **OHara2009** or **Chipman2008** useful.

George1993's Stochastic Search Variable Selection (SSVS)

$$\begin{aligned}
 y_i &= \alpha + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i \\
 \epsilon_i &\sim N(0, \psi^{-1}) \text{ iid} \\
 i &= 1, \dots, n
 \end{aligned}
 \tag{8.2}$$

Prior for β

$$\begin{aligned}
 \beta_j | \gamma_j &\sim \gamma_j N(0, c_j^2 t_j^2) + (1 - \gamma_j) N(0, t_j^2) \\
 j &= 1, \dots, p
 \end{aligned}$$

One of the early works on Bayesian variable selection for linear models come from the **George1993** paper by **George1993**. In it, they augmented the indicator variables γ into the prior for β , while the linear model itself remained the same. The prior for β_j is essentially one of two normal distributions, depending on whether or not variable X_j was selected.

The idea behind this type of prior is this: when variable X_j is not important, then γ_j should be equal to zero and the coefficient associated with it β_j should be small and close to zero as possible. Therefore, the prior on β_j should be normal with mean zero and have a small variance t_j^2 . Conversely, when the variable X_j is important, then γ_j is one and β_j should be non-zero, and thus the prior on β_j should have a large variance $c_j^2 t_j^2$. In essence, t_j and c_j are tuning parameters that the user must choose. The authors give some suggested values for these tuning parameters: $(\text{SE}(\hat{\beta}_j)/t_j, c_j) = (1, 5), (1, 10), (10, 100)$, or $(10, 500)$, where $\text{SE}(\hat{\beta}_j) = \sqrt{\hat{\psi}^{-1}(\mathbf{X}^\top \mathbf{X})_{jj}}$ under the full model.

The priors on β_j need not be independent of each other. Perhaps a more convenient notation for this prior is

$$\beta|\gamma \sim N(\mathbf{0}, \mathbf{R}_\gamma \mathbf{D} \mathbf{R}_\gamma),$$

where $\mathbf{D} = \mathbf{I}_p$, $\mathbf{R}_\gamma = \text{diag}[a_j t_j]$ and $a_j = \gamma_j c_j + (1 - \gamma_j)$, for $j = 1, \dots, p$. The matrix \mathbf{D} determines the independence of the β_j s. Setting this to be the identity matrix implies independence. On the other hand, **George1993** proposed setting this proportional to the inverse sample correlation matrix in order to capture the design correlation.

Kuo1998's sampler (KM)

$$\begin{aligned} y_i &= \alpha + \gamma_1 \beta_1 x_{i1} + \dots + \gamma_p \beta_p x_{ip} + \epsilon_i \\ \epsilon_i &\sim N(0, \psi^{-1}) \text{ iid} \\ i &= 1, \dots, n \end{aligned} \tag{8.3}$$

Prior for β

$$\begin{aligned} \beta_j &\sim N(b_j, d_j^2) \\ j &= 1, \dots, p \end{aligned}$$

Several years later in **Kuo1998** **Kuo1998** published their Bayesian variable selection model, commonly referred to as the independent sampler, so-called because of the independence of the β_j s and the γ_j s. Instead of having the γ_j s augmented into the prior, these are augmented into the model equation itself. Each term $\beta_j x_{ij}$ has its corresponding γ_j multiplied to it. Therefore, when $\gamma_j = 0$, the corresponding term drops out from the model.

The only hyperparameters one needs to choose for this model are the prior means and

variances for the normal distributions of the β_j s, similar to the Bayesian approach for estimating linear models. These choices reflect one's prior beliefs about the coefficients. In the absence of prior information, one can simply set $b_j = 0$, and choose $d_j = d$ such that $1/2 \leq d \leq 4$ after standardising the \mathbf{X} variables. Otherwise, the user must choose an appropriate value of d_j for each j that would reflect the uncertainty of the estimate β_j being zero.

The appeal of this method is its simplicity, and that also benefits the Gibbs sampling procedure, as the Gibbs conditional densities are easily worked out, and available in a recognisable closed form.

Dellaportas2002's Gibbs Variable Selection (GVS)

$$\begin{aligned}
 y_i &= \alpha + \gamma_1 \beta_1 x_{i1} + \cdots + \gamma_p \beta_p x_{ip} + \epsilon_i \\
 \epsilon_i &\sim N(0, \psi^{-1}) \text{ iid} \\
 i &= 1, \dots, n
 \end{aligned} \tag{8.4}$$

Prior for β

$$\begin{aligned}
 \beta_j | \gamma_j &\sim \gamma_j N(b_j, d_j^2) + (1 - \gamma_j) N(u_j, s_j^2) \\
 j &= 1, \dots, p
 \end{aligned}$$

The authors **Dellaportas2002** worked on an improvement to the current Bayesian variable selection methods, a method which they call the Gibbs Variable Selection (GVS). **Ntzoufras2008** provides an excellent reading about this method in his book, which also provides a good tutorial on using WinBUGS to estimate such models.

At first glance, their model looks like a cross between SSVS and KM, in that the γ indicators appear both in the model and in the prior. There are two priors for β : one is the actual prior, and one which they call the “pseudo prior”. This pseudo prior does not make its way into the posterior, and therefore does not influence the estimate at all. Instead, it is there just to make sampling more efficient, according to the authors. Why? When γ_j is one, then β_j is sampled from the posterior with the actual prior. This, coupled with the appropriate hyperparameters b_j and d_j , should encourage β_j to be non-zero. On the off-chance that γ_j is zero when the variable X_j is important, then β_j is sampled from the posterior with a pseudo prior which is designed such that good values for β_j are proposed. If the data (likelihood) also encourages β_j to be non-zero, then there is a high chance that γ_j will flip back to being one. In short, the pseudo prior helps flip the gamma in the right direction, if and when it needs to be flipped, and therefore spends less time being in the wrong state space.

With this model you do need to choose several tuning parameters. As before, we can

choose $b_j = 0$ and $d_j = d$ with large d (after standardising \mathbf{X}) if no prior information. As for the pseudo prior hyperparameter, **Dellaportas2002** suggests the following choices:

1. $u_j = \hat{\beta}_j$, the estimates of a full pilot MCMC run, and correspondingly $s_j^2 = \widehat{\text{Var}}(\hat{\beta}_j)$.
2. $u_j = 0$ and $s_j^2 \propto d_j^2$, but kept low.

Remark 1. In the long run, we expect the KM and GVS methods to give identical results if the same prior $N(b_j, d_j^2)$ for the β_j s are used. As mentioned, the pseudo prior in the GVS method merely improves efficiency of the Gibbs sampler.

Choices of priors

The main difference between the the three methods above, apart from the model structure itself, is the prior specified for β , but the priors for the rest of the common parameters are and can be similar. The priors for the intercept α and precision ψ are chosen as the conjugate normal-gamma prior, and the prior for each $\gamma_1, \dots, \gamma_p \in \{0, 1\}$ is of course chosen to be Bernoulli:

$$\begin{aligned} & \text{Priors for } \gamma, \alpha, \text{ and } \psi \\ & \gamma_j \sim \text{Bern}(p_j), \quad j = 1, \dots, p \\ & \alpha \sim N(a, b^2) \\ & \psi \sim \Gamma(c, d) \end{aligned}$$

Typically, in the absence of any prior knowledge, the hyperparameters are set to reflect an uninformative prior. For the normal-gamma, this implies the normal having mean $a = 0$ and large variance b^2 , while the gamma having both shape and scale parameters c and d small. Note that a $\Gamma(c, d)$ distribution becomes the Jeffrey's prior as c and d approaches zero. On the other hand one may actually have some prior knowledge about these and may set these hyperparameters accordingly. In any case, we are not too concerned about estimating the intercept and precision parameters.

For the Bernoulli prior on the indicator variables, we can appeal to the principle of indifference and set all $p_j = 1/2$, as each variable may either be selected or not selected. Another possibility is to let the model estimate this common probability $p_1 = \dots = p_p = p$ by assigning a hyperprior such as a beta distribution. The beta hyperprior can be chosen to be uninformative, such as Beta(1,1) (Uniform distribution) or Beta(1/2,1/2) (Jeffrey's prior). The user may also choose to code more complex relationship between the variables - e.g. if variable X_1 is included, then X_2 must be included - useful when performing variable selection on interaction effects. This way, the priors on $\gamma_1, \dots, \gamma_p$ will not be independent, and care must be taken when deriving the posteriors.

Estimate
common
 p

8.1.2 The I-prior Bayesian variable selection model (I-prior)

When we compare the three models side-by-side, there are obvious differences in the structure of the models. Things such as the structure of the parameter space, the amount of tuning parameters that need to be set, and how the two sets of parameters of interest γ and β behave in each model. Figure 8.1 summarises these differences.

However, in practice there isn't generally much to distinguish between these three models. It is more than likely that each of these methods will be optimal for a specific problem that the user faces, rather than one having an all-out advantage over the other in all situations. Having said that, these three methods have one thing in common, which is that they do not perform very well when faced with selection scenarios with correlated variables. This may be because of the use of independent β priors which lessens the posterior correlations, and thus smaller models tend to be selected (George1993).

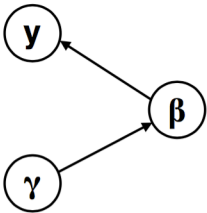
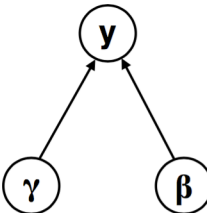
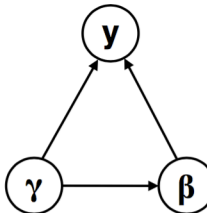
	 $f(y \beta)f(\beta \gamma)f(\gamma)$	 $f(y \gamma, \beta)f(\gamma)f(\beta)$	 $f(y \gamma, \beta)f(\beta \gamma)f(\gamma)$
	SSVS	KM	GVS
Parameter space	Retains original	Does not retain original	
Tuning parameters	Many	None	Some
Priors for β	$\beta \gamma \sim N(\mathbf{0}, \mathbf{R}_\gamma \mathbf{D} \mathbf{R}_\gamma)$ $\mathbf{D} = \mathbf{I}_p$ $\mathbf{R}_\gamma = \text{diag}(a_j t_j)$ $a_j = (1 - \gamma_j) + \gamma_j c_j$	$\beta \sim N(\mathbf{0}, \mathbf{D})$ $\mathbf{D} = d^2 \mathbf{I}_p$	$\beta \gamma \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu}_j = (1 - \gamma_j) u_j$ $\boldsymbol{\Sigma}_{jk} = \gamma_j \gamma_k (d^2 \mathbf{I}_p)_{jk}$ $+ (1 - \gamma_j \gamma_k) \mathbf{1}_{[j=k]} s_j^2$

Figure 8.1: A summarised comparison of the three Bayesian variable selection methods. Graphical models are also illustrated for each method.

We now see an opportunity to use I-priors in the Bayesian variable selection methods, by simply replacing the prior covariance matrix $\mathbf{D} = d^2 \mathbf{I}_p$, or in the case of SSVS $\mathbf{D} = \mathbf{I}_p$, by that of the I-prior covariance matrix $\psi \boldsymbol{\Lambda} \mathbf{X}^\top \mathbf{X} \boldsymbol{\Lambda}$ (see the end of Section ??). George1993 have already suggested to use $\mathbf{D} \propto (\mathbf{X}^\top \mathbf{X})^{-1}$ as a means of replicating the design correlation, and this turns out to be a generalisation of the g-prior for β , though it seems to have the opposite effect. This is discussed in Section ??.

The question is which of the three methods shall we peruse I-priors? The unappealing

feature of SSVS is the need to set the tuning parameters before running the model. **George1993** do give four possible suggestions in their paper, but this is thought to be non-exhaustive. In other words, the user must really know the optimal settings for their problem at hand before running the variable selection model. On this note, GVS also has some tuning parameters to set, but not as many in our opinion. As the prior for β comprises of a true prior and pseudo prior. The obvious choice for the true prior is the I-prior (if we want to employ I-priors, that is). As for the pseudo prior, **Ntzoufras2008** uses the estimates obtained from a full pilot MCMC run (see Section 8.1.1) in his examples, and this seems reasonable.

Out of all these methods, KM stands out as being the simplest. I-priors fit straight into the story by replacing the prior on β in the model. We think this simplicity outweighs the efficiency claimed to be brought about by introducing a pseudo prior in GVS. Further, the KM entails only specifying choices for the hyperparameters of the model as we would if we were estimate the linear model in a Bayesian manner. Since we are using the I-prior, there is no more hyperparameters to choose. This is a nice feature seeing it from a “hands-free plug-and-play” perspective.

The I-prior Bayesian variable selection model is given below:

$$\begin{aligned} y_i &= \alpha + \gamma_1 \beta_1 x_{i1} + \cdots + \gamma_p \beta_p x_{ip} + \epsilon_i \\ \epsilon_i &\sim N(0, \psi^{-1}) \\ i &= 1, \dots, n \end{aligned}$$

Priors

(8.5)

$$\begin{aligned} \beta &\sim N(\mathbf{0}, \psi \mathbf{A} \mathbf{X}^\top \mathbf{X} \mathbf{A}), \text{ where } \mathbf{A} = \text{diag}[\lambda_1, \dots, \lambda_p] \\ \gamma_j &\sim \text{Bern}(p_j), \quad j = 1, \dots, p \\ \alpha &\sim N(a, b^2) \\ \psi, \lambda_1^{-2}, \dots, \lambda_p^{-2} &\sim \Gamma(c, d) \end{aligned}$$

By virtue of the I-prior being a maximum entropy prior, meaning that it is suitable to be used in the absence of prior information, we then complete the model specification above by also choosing uninformative hyperpriors (see Section 8.1.1).

The scale parameters $\lambda_1, \dots, \lambda_p$ originally came from I-prior modelling in a function space framework, whereby these scale parameters help resolve the arbitrary scale of the space of functions over the set of covariates. As these scale parameters make their way into the covariance matrix of the β prior, we can interpret them as follows: if no scale parameters are introduced, or equivalently all scale parameters are equal to one, then the covariance matrix is proportional to $\mathbf{X}^\top \mathbf{X}$. As the covariates are likely to be measured on differing scales, such as age in years, height in metres, weight in kilograms, etc., the entries of $\mathbf{X}^\top \mathbf{X}$ will be large for measurements on a large scale range (e.g. body weight in grams), and small for measurements on a small scale range (e.g. body height in metres).

More justification required

This in turn affects the precision of the prior and consequently the estimation of the β parameter. For instance, a high precision (small variance) supports the predictor not being selected. What is ideal for us is that important variables should have β_j s estimated as non-zero and vice-versa, but simply putting $\mathbf{X}^\top \mathbf{X}$ as the covariance matrix does not contribute towards this goal. Therefore, scaling the prior covariance matrix of β is necessary, and the I-prior method of scaling is a natural choice here.

An alternative solution, as is practiced by the three methods in the previous section, is to standardise both the \mathbf{X} and \mathbf{y} variables. This is indeed a good idea, but is slightly unsatisfactory - scaling the variables so that each has variance one feels ad-hoc in the face of it. Having scale parameters estimated through the model seems more elegant and conforms more to the original I-prior methodology. Having said that, standardising the variables while using a single estimable scale parameter λ is certainly an option, as all the variables would then have been scaled equally via standardisation. This has the advantage of taming extremely large entries of $\mathbf{X}^\top \mathbf{X}$ which may be problematic computationally when we require the inverse.

Remark 2. On another note, it might also be possible to treat the scale parameters $\lambda_1, \dots, \lambda_p$ as fixed, having been estimated from the full model using the original I-prior framework described in Section ???. This is an idea yet to be explored, and is not known whether this would yield good results. There is also a convergency and accuracy issue in obtaining reliable estimates of a large number of scale parameters through the EM procedure of maximising the likelihood of the I-prior model.

We can estimate this model by Gibbs sampling. Unlike the KM model however, one of the Gibbs conditional posterior was not found to be in closed form, which was the posterior for the precision ψ . So to estimate this model, one has to incorporate a Metropolis-Hastings step for the estimation of ψ . The conditional posterior densities are given in Appendix ???. We can also feed this model into WinBUGS or JAGS which is then able to estimate this model for us.

8.1.3 Simulation study

In this section, we compare the performance of the four methods of Bayesian variable selection: SSVS, KM, GVS, and I-prior by means of a simulation study. 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, $p = 30$).

The data was generated as follows:

- Draw $\mathbf{Z}_1, \dots, \mathbf{Z}_{100} \sim N(\mathbf{0}, \mathbf{I}_{150})$.
- Draw $\mathbf{U} \sim N(\mathbf{0}, \mathbf{I}_{150})$.
- Let $\mathbf{X}_j = \mathbf{Z}_j + \mathbf{U}$. This induces pairwise correlations of about 0.5.⁴

More simulations required

Compare LASSO

- Draw $\epsilon \sim N(\mathbf{0}, 2^2 \mathbf{I}_{150})$.

- Generate response variables $\mathbf{Y} = \mathbf{X}\beta_{\text{true}} + \epsilon$.

Let $\beta_{\text{true}} = (\beta_{-k}, \beta_k)$, where $\beta_{-k} = (\beta_1, \dots, \beta_k) = (0, \dots, 0)$ and $\beta_k = (\beta_{k+1}, \dots, \beta_{100}) = (1, \dots, 1)$. In other words, only variables X_{k+1} to X_{100} are used. The experiment involves varying the value of k between 10, 25, 50, 75 and 90 to create five scenarios, which we label as Scenarios A to E respectively. The two extremes, Scenarios A and E, are meant to simulate situations in which there are a lot of non-zero betas in the true model (Scenario A) and situations in which there are very few non-zero betas in the true model (Scenario E). The variable selection is conducted with many correlated variables.

10,000 MCMC were samples obtained for each scenario, and the metric of interest is the number false choices the models make, i.e. selecting variables which were not in the true model and failing to select variables which were in the true model. Each experiment was repeated 10 times and results averaged, as this ensures that a good result was not simply due to chance of a good random seed in the data generation step. This experiment was conducted in R using JAGS, a variation of WinBUGS, and the results presented in the form of histograms in Figure 8.2. Note that the same prior for β was used in the KM and GVS method, so we expect the results to be similar for these two methods.

The ideal picture would be a histogram with a lot of mass towards the left side of the graph, indicating models which produced little false choices. The histograms indicate similar behaviour for the SSVS, KM and GVS methods. These methods perform poorly in the presence of many non-zero β s, but perform slightly better in the presence of few non-zero β s. For I-priors however, it is the opposite situation. The I-prior method seems to work quite well given many non-zero β s, but performance worsens when there are actually few non-zero β s in the true model. However, in the defence of I-priors, when it does well in Scenarios A, B and C, it does much better (fewer false choices) than when the other three methods do well (Scenarios D and E). I-prior is also less worse than when the other methods do terribly (maximum number of false choice for I-prior is 30, compared to ≈ 50 for any of the other three methods).

One possible explanation here is that when there are a lot of zero β s in the true model, the Fisher information in the covariance matrix of the prior only serves to confuse with all this evidently unnecessary information. Hence, we can't expect I-priors to benefit in situations like these. Scenarios D and E bode well for the other three methods because the lack of a correlation structure in the covariance matrix of the priors causes these methods to select fewer variables, and thus make fewer false choices.

⁴ $\text{Cov}(X_j, X_k) = \text{Cov}(Z_j + U, Z_k + U) = \text{Var } U = 1$, and $\text{Var}(X_j) = \text{Var}(Z_j + U) = 2$. Thus, $\text{Corr}(X_j, X_k) = \text{Cov}(X_j, X_k) / (\text{Var}(X_j)\text{Var}(X_k))^{1/2} = 1/2$.

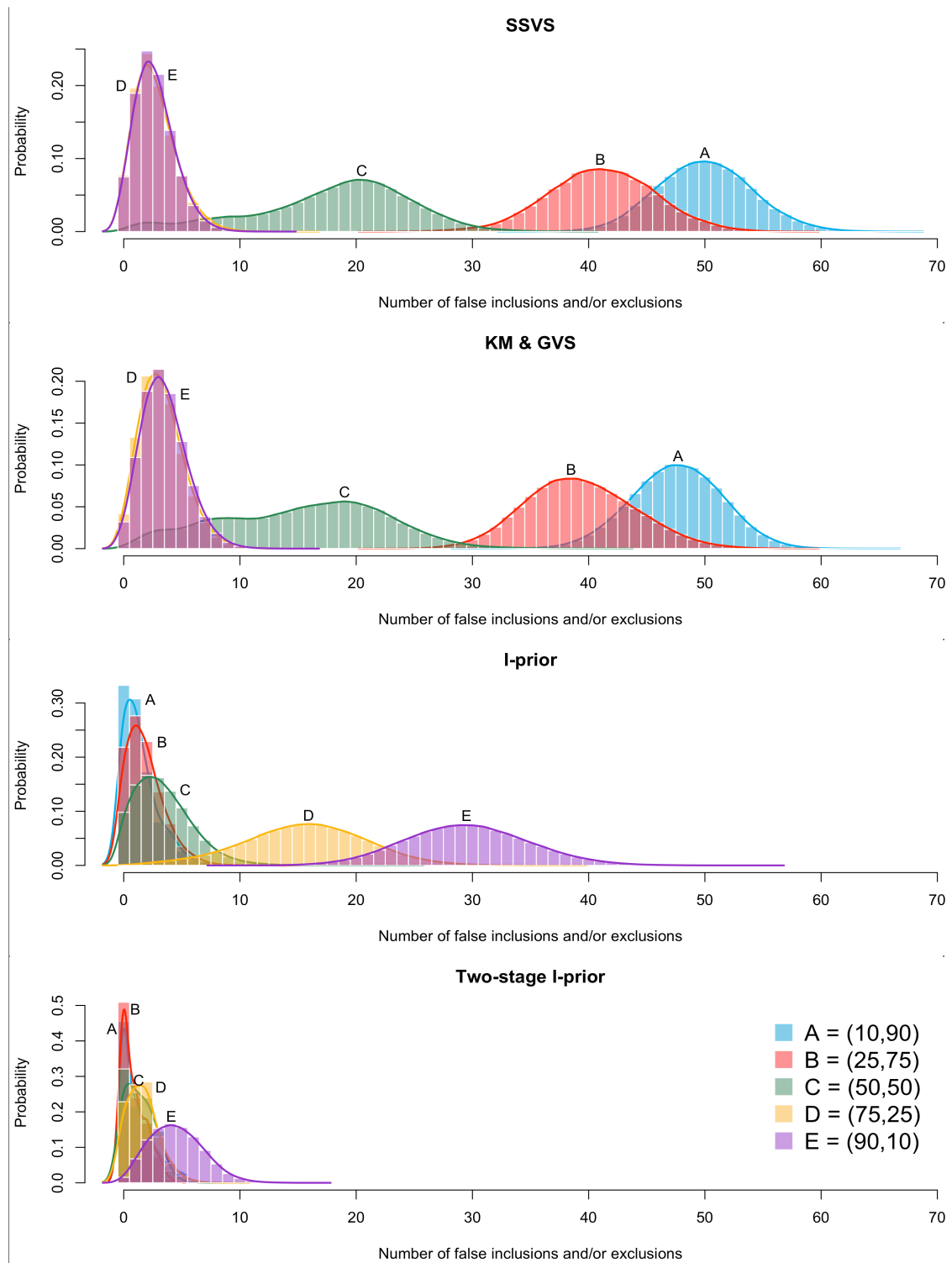


Figure 8.2: Histograms of false choices for SSVS, KM (equivalent to GVS), I-prior and two-stage I-prior compared across the five scenarios A to E. Each scenario is labelled as “($k, 100 - k$)”, where k denotes the number of zeros in the true value of β used.

8.1.4 Two-stage procedure

Since I-priors do quite well in Scenarios A-C, but not in D and E, why don't we try to make Scenarios D and E a bit more like Scenarios A-C? This requires some sort of pre-selection of the variables in order to trim off the unwanted variables before running the variable selection model. Without appealing to other pre-selection methods, there is some information from the Bayesian variable selection models that we can make use of - the posterior inclusion probabilities for each variable. As this gives an overall indication as to how valuable a particular variable is, we look for ways to incorporate this into our decision-making process of pre-selecting the variables. The obvious solution is to run the model twice:

1st Run the model. Discard variables with posterior inclusion probabilities less than τ , a threshold value.

2nd Re-run the model on the set of reduced variables.

A natural choice for τ would be 0.5. Setting it at 0.5 would mean that we only keep variables which have a better than equal chance of being selected. The model such that all posterior inclusion probabilities are greater than or equal to 0.5 is known as the median probability model. While this value of 0.5 may seem slightly arbitrary to some, **Barbieri2004** had done some work on median probability models, for which they had shown that under certain strict conditions, these models are also the most optimally predictive models that is able to be selected. The notion of two-stage procedures are not new, as many variable selection methods in the literature generally employ a pre-selection method before running their methods proper. For a two-stage procedure based on posterior inclusion probabilities, **Fouskakis2008** and **Ntzoufras2008** have employed this in their work.

The histogram at the bottom of Figure 8.2 shows that this two-stage procedure does indeed improve on the I-prior method. We see that a shift in the histogram towards the left-hand side of the graph for the second stage run of the model. Interestingly, not only does this improve the Scenarios D and E, we also seem improvements for Scenarios A-C. *Remark 3.* Among the reasons for a pre-selection of the variables are to remove highly correlated variables, removing variables which have no theoretical benefit, or simply to reduce the large number of variables for large especially when $p > n$. There is probably no justification why a two-stage procedure would work better than just a one-stage procedure other than for convenience. At the end of the day, it is the responsibility of the user to interpret the results of the variable selection methods carefully and not make inference blindly on the results of the model.

8.1.5 Real world applications

Here we look at three real-world applications on problems where there are some degree of multicollinearity in the datasets, and have been looked at before from a variable selection

standpoint so that we are able to compare results to I-priors.

Aerobic fitness data

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 **Oxygen** contains measurement of oxygen uptake rate in mL/kg body weight per minute. The six covariates were the participants' **Age**, **Weight**, time taken to run one mile (**RunTime**), resting heart rate (**RestPulse**), heart rate while running (**RunPulse**), and maximum heart rate during the exercise (**MaxPulse**). This dataset, although small in size, is interesting to analyze 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 8.3 below:

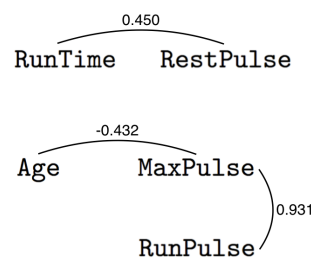


Figure 8.3: The sample correlations of interest in the aerobic fitness dataset. These show variables with correlations greater than 0.4 in magnitude.

The SAS analysis employed a forward selection and backwards elimination procedure and concluded that variables **Weight** and **RestPulse** were to be deleted. The KM procedure concurred with this finding. The I-prior method also did not choose the **Age** variable in addition to **Weight** and **RestPulse** in the maximum probability model. The variable **Age** only had a probability of 0.05 of being included in any posterior model, and also failed to appear in the top 92% of likely models. This can be explained by the correlation between **Age** and **MaxPulse** - supposedly the information encoded in **Age** has already been taken care of by **MaxPulse**, so the I-prior deemed this as surplus. However, the models that had **Age** selected performed better in terms of AIC, Mallows C_p , and 5-fold cross validation root mean squared error (RMSE). Despite this, the strength of the coefficients for the variables are comparable to that of the I-prior method, which is settling if one wishes to do inference on this.

	Full model	I-prior	Forward sel.	Back elim.
Intercept	104.2 (0.00)	80.8 (0.00)	103.3 (0.00)	98.6 (0.00)
Age	-0.24 (0.03)		-0.25 (0.02)	-0.21 (0.05)
Weight	-0.08 (0.15)		-0.08 (0.15)	
RunTime	-2.59 (0.00)	-2.97 (0.00)	-2.64 (0.00)	-2.75 (0.00)
RestPulse	-0.02 (0.72)			
RunPulse	-0.38 (0.00)	-0.38 (0.01)	-0.39 (0.00)	-0.36 (0.01)
MaxPulse	0.32 (0.03)	0.36 (0.02)	0.32 (0.03)	0.28 (0.05)
C_p	7.0	7.7	5.1	5.3
AIC	56.8	58.5	54.9	55.6
5f-CV RMSE	2.59	2.71	2.50	2.54

Table 8.1: The OLS estimates for each variable are given in the table above, along with the standard errors in parantheses. The table also shows the value for Mallows C_p , AIC and 5-fold cross validation RMSE given for each model.

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 **Mortality**, a total age adjusted mortality rate, and the main pollution effects of interest were that of hydrocarbons (HC), oxides of nitrogen (NO_x) and sulphur dioxide (SO₂). Several other environmental and socioeconomic considerations were taken into account, otherwise the model may include unexplained variation which may have been 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 lower proportion. All of the variables can be considered as continuous and real. A full description of the data can be found in Appendix ??.

This dataset also contains several highly correlated variables. When the full model is fitted, none of the pollutant effects were found to be significant. Clearly a variable selection method was required. **McDonald1973** used ridge regression analysis to determine which variables to select. We also have results from a backwards elimination procedure (using AIC as the selection criterion) for comparison. The results are summarised in Table 8.2.

		Full model	I-prior	Ridge	Back elim.
Environmental & demographic variables selected		All, but only Rain, JanTemp, NonW significant	Rain, JanTemp, JulTemp, Over65, Popn, Hous, NonW, Poor, Humid	Rain, JanTemp, Educ, Dens, NonW	JanTemp, Educ, NonW
Pollution effect	HC	✗	✗	✗	✓ $\beta = -0.98$
	NOx	✗	✗	✗	✓ $\beta = 1.99$
	SO2	✗	✓ $\beta = 0.33$	✓ $\beta = 0.24$	✗
C_p		16.0	13.4	5.6	8.7
AIC		439.8	439.2	431.3	435.0
BIC		49.5	36.5	20.3	21.2
5f-CV RMSE		50.6	41.7	39.3	38.6

Table 8.2: The results of the various variable selection methods compared. For each method, the variable selection procedure was conducted on the set of all variables, and then an OLS was fit on the resulting selected variables. The environmental and demographic variables selected are shown in the table for each model, but those in gray are the ones found to be not significant (at the 10% level).

It is noted that the I-prior method selected some variables which turned out to be insignificant, with only three significant variables selected in total. However, of importance is learning which of the three pollution factors has an effect on mortality rate. It is nice to see that the I-prior agrees with the ridge analysis done by **McDonald1973** on this, with only sulphur dioxide having a significant effect. The strength of this effect is also comparable (I-prior 0.33 c.f. ridge 0.24). The method of backwards elimination was found not only inconsistent with I-prior and ridge analysis, but also erroneous in that it seems to imply an increase in levels of hydrocarbons would bring about a reduction in mortality rate. Once again, we see that the I-prior is outperformed in terms of Mallows' C_p , AIC, BIC and 5-fold CV RMSE, but it is noted that the 5-fold CV RMSE is not too far off from its competitors.

Ozone data

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 are explained in Appendix ??.

The data contains 366 points, one for each day of the leap year 1976. For our analysis, we ignore the 163 missing data in the set, and use the remaining 203 datapoint

for our analysis. Out of these 203, we randomly set aside 25 to use for validation. So the n used for the Bayesian variable selection methods was $n = 178$. **Casella2006** removed the variables **TempElMon** and **ibtLAX** before running their selection model, citing multicollinearity. We won't do this, as we would like to see how well I-priors do in the presence of multicollinearity. On another note, the variables **Month**, **DayMonth** and **DayWeek** were presumably intended to be categorical as in modelling seasonality in a time series data, but these were treated as continuous, as did **Casella2006**. This is just as well, as our I-prior model is not able to handle categorical variables which have more than two levels. The results are compared below:

Method	Model	Post. prob.	R^2	RMSE
I-prior	Month HumLAX TempElMon	0.544	0.72	3.86
CM (MPM)	HumLAX TempSand ibhLAX	<0.001	0.69	4.47
CM (MSE)	Month HumLAX TempSand ibhLAX	<0.001	0.70	4.04
BF	TempSand ibhLAX PresGrad VisLAX	<0.001	0.66	4.27

Table 8.3: Table showing the comparison between the I-prior, Casella and Moreno (CM) analysis, and the ACE method by Breiman and Friedman (BF). MPM stands for maximum probability model, and MSE is the lowest RMSE model.

The maximum probability I-prior model was found to be much better in terms of RMSE compared to the maximum probability model of **Casella2006**. The variables selected using I-prior corresponded to the significant variables when the full OLS model was fitted. Our I-prior selected model also had a lower RMSE than **Casella2006**'s lowest RMSE model. The ACE method by **Breiman1985** was found to be the worst model for prediction. It is noted that neither **Casella2006**'s nor **Breiman1985**'s model were found in the posterior model space using the I-prior method. Out of interest, if we had removed the two variables **TempElMon** and **ibtLAX** at the beginning, then we arrive at the same results as **Casella2006**.

We now use the I-prior method to select between the squared terms and all level two interactions 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 8.4. The maximum probability model for I-prior method selected fewer variables as compared to **Casella2006**'s maximum probability model, yet was superior in terms of RMSE. For comparison, the backwards elimination resulted in a very complicated model which did not seem to improve on RMSE.

Remark 4. As the model fit involved randomly leaving out 25 data points which were later used for validation, the results between our analysis and **Casella2006**'s are bound to differ, as we did not use the same 25 data points for training and testing.

Remark 5. For this particular dataset, running the model without squared terms and linear predictors was straightforward. However, we ran into a numerical issue in the second part, whereby some entries of $\mathbf{X}^\top \mathbf{X}$ were found to be so large compared to others, that its inverse could not be computed. Thus, we standardised the \mathbf{X} and \mathbf{y}

Method	Model	Post. prob.	R^2	RMSE
I-prior	Month Month ² WindLAX HumLAX TempElMon			
	TempElMon ² ibtLAX PresGrad ²	0.103	0.83	3.74
	ibtLAX:HumLAX			
CM	DayMonth Month ² TempSand ² PresGrad ²			
	Month:WindLAX DayMonth:HumLAX			
	DayWeek:TempSand PresVand:HumLAX	<0.001	0.76	3.88
	HumLAX:ibhLAX HumLAX:VisLAX			
Back. elim.	Month Month ² DayWeek PresVand			
	HumLAX TempSand TempSand ² TempElMon			
	ibhLAX VisLAX PresVand ² WindLAX ²			
	PresGrad ² DayMonth:Month WindLAX:Month			
	HumLAX:Month ibhLAX:Month ibtLAX:Month			
	HumLAX:DayMonth TempSand:DayMonth			
	TempElMon:DayMonth VisLAX:DayMonth			
	WindLAX:PresVand HumLAX:PresVand	<0.001	0.87	4.29
	TempSand:PresVand TempSand:WindLAX			
	ibtLAX:WindLAX TempSand:HumLAX			
	TempElMon:HumLAX VisLAX:HumLAX			
	PresGrad:TempSand VisLAX:TempSand			
	PresGrad:TempElMon VisLAX:TempElMon			
	ibtLAX:PresGrad VisLAX:PresGrad			

Table 8.4: Results of variable selection for to look for squared and interaction terms in the ozone dataset.

variables and used a single scale parameter λ . See the second last paragraph of Section 8.1.2 on page 11.

Remark 6. The method that we employed was a naive I-prior variable selection method, whereby each of the 90 terms was considered independently. If one wishes a model such that its level one term is included when an interaction is present, then the variable selection needs to be adjusted accordingly. **Kuo1998** gives an example of this:

$$y_i = \alpha + \max(\gamma_1, \gamma_3)\beta_1 x_{i1} + \max(\gamma_2,$$

8.2 Special case: The canonical RKHS

8.3 Bayesian model selection

8.4 BVS using I-priors

8.5 Simulation study

8.6 Real-data examples

List of Figures

8.1	A summarised comparison of the three Bayesian variable selection methods. Graphical models are also illustrated for each method.	9
8.2	Histograms of false choices for SSVS, KM (equivalent to GVS), I-prior and two-stage I-prior compared across the five scenarios A to E. Each scenario is labelled as “ $(k, 100 - k)$ ”, where k denotes the number of zeros in the true value of β used.	13
8.3	The sample correlations of interest in the aerobic fitness dataset. These show variables with correlations greater than 0.4 in magnitude.	15

List of Tables

8.1	The OLS estimates for each variable are given in the table above, along with the standard errors in parantheses. The table also shows the value for Mallow's C_p , AIC and 5-fold cross validation RMSE given for each model.	16
8.2	The results of the various variable selection methods compared. For each method, the variable selection procedure was conducted on the set of all variables, and then an OLS was fit on the resulting selected variables. The environmental and demographic variables selected are shown in the table for each model, but those in gray are the ones found to be not significant (at the 10% level).	17
8.3	Table showing the comparison between the I-prior, Casella and Moreno (CM) analysis, and the ACE method by Breiman and Friedman (BF). MPM stands for maximum probability model, and MSE is the lowest RMSE model.	18
8.4	Results of variable selection for to look for squared and interaction terms in the ozone dataset.	19

List of Theorems

List of Definitions

Index

fractional Brownian motion, *see* fBm

reproducing kernel Hilbert space, *see*
RKHS