LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER
APPLIED ECONOMETRICS AND FORECASTING

MASTER'S FINAL WORK
DISSERTATION

THE EFFECT OF HETEROSCEDASTICITY ON BAYESIAN VARIABLE SELECTION

HUGO FRANCISCO VICENTE MOREIRA

OCTOBER - 2019

LISBON
SCHOOL OF
ECONOMICS &
MANAGEMENT
UNIVERSIDADE DE LISBOA

MASTER
APPLIED ECONOMETRICS AND FORECASTING

MASTER'S FINAL WORK
DISSERTATION

THE EFFECT OF HETEROSCEDASTICITY ON BAYESIAN VARIABLE SELECTION

HUGO FRANCISCO VICENTE MOREIRA

SUPERVISION:
PROF. RUI PAULO

OCTOBER - 2019

Acronyms

AIC Akaike Information Criterion.

BF Bayes Factor.

BIC Bayesian Information Criterion.

BMA Bayesian Model Averaging.

BVS Bayesian Variable Selection.

DGP Data Generating Process.

GLS Generalised Least Squares.

HPD Highest Posterior Density.

HPM Highest Probability Model.

MPM Median Probability Model.

OLS Ordinary Least Squares.

RMSE Root Mean Squared Error.

RSE Relative Squared Error.

SD Standard Deviation.

SNR Signal-to-Noise Ratio.

SSE Sum of Squared Errors.

WLS Weighted Least Squares.

Abstract

This dissertation aims to study the effect of heteroscedasticity on Bayesian Variable Selection. It employs a simulation study, using two distinct datasets, to evaluate the effects of introducing heteroscedasticity in a linear regression, and whether transforming an heteroscedastic dataset into an homoscedastic one results in any considerable differences. We look at the variables selected, inclusion probabilities and performance measures. We find Bayesian Variable Selection to be robust to heteroscedasticity, although a better predictive performance may be attained if we take the error variance's structure explicitly into account.

KEYWORDS: Bayesian Variable Selection, Heteroscedasticity.

Resumo

Nesta dissertação estudamos o efeito da heterocedasticidade na seleção bayesiana de variáveis. Através de um estudo de simulação, e utilizando dois conjuntos de dados reais, avaliamos os efeitos de introduzir heteroscedasticidade numa regressão linear, bem como o efeito de transformar dados heterocedásticos em homocedásticos. Analisando as variáveis selecionadas, probabilidades de inclusão e medidas de performance preditiva, concluimos que a seleção bayesiana de variáveis é robusta à heterocedasticidade, mas é possível obter melhor perfomance preditiva se a estrutura de variância dos erros for tomada em conta.

PALAVRAS-CHAVE: Seleção bayesiana de Variáveis, Heteroscedasticidade.

Contents

Acronyms **i**

Abstract **ii**

Table of Contents **iv**

List of Figures **vi**

List of Tables **vii**

Acknowledgements **ix**

1 Introduction **1**

2 Literature Review **2**

2.1 Heteroscedasticity 2

2.2 Variable Selection and the Bayesian Paradigm 3

2.3 The frequentist approach to Variable Selection 6

2.4 Heteroscedasticity and BVS: intuition 7

3 Simulation Study **9**

3.1 Signal-to-Noise Ratio 12

3.2 Simple Linear Regression: Setup 13

3.2.1 Simple Linear Regression: original dataset 14

3.2.2 Simple Linear Regression: "corrected" datasets 19

3.3 Multiple Linear Regression: Setup 23

3.3.1 Multiple Linear Regression: original dataset 24

3.3.2 Multiple Linear Regression: "corrected" datasets 28

4 Conclusion **30**

References **32**

Appendices **36**

A Correlation tables 36

A.1 Correlation tables of regressors: Ozone dataset 36

A.2 Correlation tables of regressors: Crime dataset 37

B Heteroscedasticity tests ("corrected" datasets) 39

C R Code 40

List of Figures

1 Two instances of a simulated dependent variable (SNR=1) 17

2 Boxplot of inclusion probabilities: SNR = 1 18

List of Tables

I Percentage of rejections in heteroscedasticity tests: Ozone data . . . 14

II Selection frequencies (%) for Ozone data: SNR = 10 15

III Selection frequencies (%) for Ozone data: SNR=1 15

IV Predictive performance: Ozone 18

V Selection frequencies (%) for Ozone data ("corrected"): SNR = 10 . . 20

VI Selection frequencies (%) for Ozone data ("corrected"): SNR = 1 . . . 20

VII Predictive performance: Ozone ("corrected") 21

VIII Percentage of rejections in heteroscedasticity tests: Crime 24

IX Selection frequencies (%) for Crime data: SNR=10 25

X Selection frequencies (%) for Crime data: SNR=1 25

XI Predictive performance: Crime 26

XII Selection frequencies (%) for Crime data ("corrected"): SNR = 10 . . 28

XIII Selection frequencies (%) for Crime data ("corrected"): SNR = 1 . . . 28

XIV Predictive performance: Crime ("corrected") 29

A.I Correlation table, Ozone data: Homoscedasticity (original) 36

A.II Correlation table, Ozone data: Heteroscedasticity - X_7 ("corrected") . 36

A.III Correlation table, Ozone data: Heteroscedasticity - X_6 ("corrected") . 36

A.IV Correlation table, Crime data: Homoscedasticity (original) 37

A.V Correlation table, Crime data: Heteroscedasticity - exponential ("corrected") 37

A.VI Correlation table, Crime data: Heteroscedasticity - two values ("corrected") 38

A.VII Correlation table, Crime data: Heteroscedasticity - Pop ("corrected") 38

A.VIII Correlation table, Crime data: Heteroscedasticity - GDP ("corrected") 38

B.I Percentage of rejections in heteroscedasticity tests: Ozone ("corrected") 39

B.II Percentage of rejections in heteroscedasticity tests: Crime ("corrected") 39

Acknowledgements

I would like to thank my advisor, Professor Rui Paulo, for suggesting the topic of this thesis, and for all the advice and thoughtful remarks given throughout the development of this work.

I would also like to thank all my friends, especially Lídia, Patrícia and Francisco, for providing me with encouragement and interesting discussions.

Finally, I am very grateful to my family, especially my sister, for all the incredible support given every day.

1 Introduction

Heteroscedasticity is a concept that appears both in frequentist and Bayesian econometrics, usually identified as a form of model misspecification. In linear regressions, we observe a collection of random variables $\{y_1, \ldots, y_n\}$ and a collection of predictors $\{\boldsymbol{X_1}, \ldots, \boldsymbol{X_n}\}$. One of the assumptions of the model is that the conditional variances of y_i given $\boldsymbol{X_i}$ do not change with i, and this is usually referred to as the homoscedasticity assumption. If this assumption does not hold, we have heteroscedasticity. It is well-known, for instance, that in the frequentist approach heteroscedasticity leads to biased standard errors, invalidating inference on the parameters.

If heteroscedasticity is of known-form, there are simple solutions for the problem: Weighted Least Squares (WLS) for the frequentist case, and a transformation of the variables for the Bayesian approach. If the form of heteroscedasticity is unknown, the problem is more delicate. In particular, to the best of our knowledge, there is no study regarding the consequences on Bayesian Variable Selection (BVS) of assuming homoscedasticity, when in fact the error term is heteroscedastic. This work aims to shed some light on this question, by using real-data variables to evaluate the performance of BVS, when transitioning from an homoscedastic to a comparable heteroscedastic situation.

Our approach is twofold. First, we check direct impacts in terms of variables selected and predictive performance between homo and heteroscedasticity. Then, we use a transformation to correct for heteroscedasticity and check if any difference is present.

This thesis is divided in three main sections. A literature review is conducted in Section 2, where we define heteroscedasticity and introduce BVS, while mentioning some frequentist approaches to the problem of variable selection. We also provide an intuition behind the issue of heteroscedasticity within BVS.

We then proceed to a Simulation Study in Section 3, aiming to detail the behaviour of BVS when moving from homoscedasticity to its counterpart, and the consequences of trying to correct heteroscedasticity. Finally, we conclude our work in Section 4, and provide some discussion on future research.

A preliminary note should be given concerning the notation. Lowercase greek

letters are employed for parameters, and capital letters for matrices. Vectors are denoted in bold lowercase. The notation $N_n(\boldsymbol{\mu}, \Sigma)$ represents a multivariate normal distribution of a n-dimensional vector, with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ; $N_n(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma)$ denotes the corresponding density evaluated at a generic vector $\boldsymbol{y}$.

2 Literature Review

2.1 Heteroscedasticity

In a linear regression, the case we are concerned with in this work, heteroscedasticity arises when the error term, denoted by $\boldsymbol{\varepsilon}$, no longer has a covariance matrix proportional to I. In more rigorous terms, writing in matrix form, we are specially interested in the normal/gaussian linear regression, with n observations:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 I),$$

where $\boldsymbol{y}$ is a vector with n observations from the response variable, X is the $n \times p$ full column rank matrix of p independent variables, with $n > p$, and $\boldsymbol{\varepsilon}$ is the error term.

Under heteroscedasticity, the second assumption no longer holds, since now $Var(\boldsymbol{\varepsilon}) = \sigma^2\ \Omega$, where Ω is a diagonal matrix with elements ω_i, $i = 1, \ldots, n$, and σ^2 is a multiplicative constant.

Further (and common) assumptions are the independence between $\boldsymbol{\varepsilon}$ and X, and the strict exogeneity of the regressors.

In the Bayesian framework, the treatment for heteroscedasticity for estimation in the linear regression is part of textbook literature, whether its form is know or unknown (see, *inter alia*, Koop (2003), chapter 6). For the first scenario, when Ω is known up to a constant, a simple transformation of the variables, by premultiplication of all terms ($\boldsymbol{y}$, X and $\boldsymbol{\varepsilon}$) with a matrix P, that satisfies the condition $P\,\Omega\,P^T = I$, suffices. The simpler methods, which assume homoscedasticity, may then be employed.

When the form is unknown, an hierarchical prior in ω_i is used, and the problem is equivalent to a regression where each error term, ε_i, has a t-student distribution. The posterior is sampled via a Metropolis-Hastings algorithm.

In recent years, semiparametric and non-parametric Bayesian approaches for linear regression have been further developed, usually adapting frequentist procedures (among others, Crainiceanu et al. (2007), Pelenis (2014) and Norets (2015)).

In frequentist econometrics, the treatment follows a similar approach. When the form is known, the Generalised Least Squares (GLS) estimator (also called WLS if there is no correlation in the errors) may be used (Hayashi, 2000). On the other hand, when the form is unknown, the robust standard errors of White (1980) are widely available and employed, since they are robust to quite general forms of heteroscedasticity. Tests also exist to study its presence, the most common being the one proposed by White (1980) and the Breusch-Pagan test (Breusch and Pagan, 1979; Koenker, 1981).

2.2 Variable Selection and the Bayesian Paradigm

Following García-Donato and Martínez-Beneito (2013) and Forte et al. (2018), in a (normal) linear regression the variable selection problem may be stated as selecting the subset from $X = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p})$ that better represents the Data Generating Process (DGP) of the target variable $\boldsymbol{Y}$. Alternatively, consider the vector $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_p\}$. γ_j is a binary variable, assuming the value 1 if $\boldsymbol{x_j}$, $j = 1, \ldots, p$ is part of the model. This means we have to choose from 2^p competing models, which belong to the model space $\mathcal{M} = \{M_\gamma : \boldsymbol{\gamma} \in \{0,1\}^p\}$.

One advantage of the Bayesian approach to variable selection is the possibility of calculating the posterior probability for each proposed model, and make a decision based on those probabilities (Forte et al., 2015). BVS also works as an automatic Occam's Razor (Berger and Pericchi, 2001), preferring to select simpler models.

Under each model M_γ, we can write:

$$M_\gamma : \mathbf{y} \sim N_n(\alpha\mathbf{1} + X_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}),$$

where X_γ contains the columns j of matrix X where $\gamma_j = 1$. The simplest and base model, M_0, contains the constant α. Applying Bayes' theorem, one can easily

write the posterior probability of M_γ:

$$f(M_\gamma|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|M_\gamma)f(M_\gamma)}{\sum_\gamma f(\boldsymbol{y}|M_\gamma)f(M_\gamma)},$$

where $f(M_\gamma)$ is the prior probability of model M_γ. The marginal likelihood, $f(\boldsymbol{y}|M_\gamma)$, is defined as:

$$f(\boldsymbol{y}|M_\gamma) = \int N_n(\boldsymbol{y}|\alpha\mathbf{1} + X_\gamma\boldsymbol{\beta}_\gamma,\ \sigma^2 I)\,\pi_\gamma(\alpha, \boldsymbol{\beta}_\gamma, \sigma)\ d\alpha\ d\boldsymbol{\beta}_\gamma\ d\sigma,$$

and $\pi_\gamma(\alpha, \boldsymbol{\beta}_\gamma, \sigma)$ is the prior for the parameters under model M_γ.

The posterior model probabilities may also be expressed in terms of Bayes Factors. The Bayes Factor (BF), denoted as $B_{\gamma\theta}$, is the ratio of the marginal likelihoods of M_γ and M_θ, and represents the evidence provided by the data (although it depends on the prior distributions) in favour of using model M_γ, against the alternative M_θ (Kass and Raftery, 1995). Rewriting the posterior probability, we get:

$$f(M_\gamma|\boldsymbol{y}) = \frac{B_{\gamma 0}\, f(M_\gamma)}{\sum_\gamma B_{\gamma 0}\, f(M_\gamma)}.$$

The issue of prior distribution choice elicits some discussion. Since there is usually no preexistent information available, the use of noninformative (or objective) priors is common.

This problematic is different whether we consider instances of Bayesian estimation or variable selection. For instance, in the first case, improper priors may be used, as long as the resulting posterior is proper. A fitting example is the well-known Jeffreys' prior (Jeffreys, 1946). In BVS, in general, the use of improper priors is not feasible. Consider the Bayes Factor comparing model M_γ to M_0:

$$B_{\gamma 0} = \frac{f(\mathbf{y}|M_\gamma)}{f(\mathbf{y}|M_0)} = \frac{\int N_n(\boldsymbol{y}|\alpha\mathbf{1} + X_\gamma\boldsymbol{\beta}_\gamma,\ \sigma^2\,\mathbf{I})\,\pi_\gamma(\alpha, \boldsymbol{\beta}_\gamma, \sigma)\ d\alpha\ d\boldsymbol{\beta}_\gamma\ d\sigma}{\int N_n(\boldsymbol{y}|\alpha\mathbf{1},\ \sigma^2\,\mathbf{I})\,\pi_0(\alpha, \sigma)\ d\alpha\ d\sigma}$$

If the prior were to be improper, the resulting Bayes Factor could take any value, as it would only be defined up to a multiplicative constant (Berger and Pericchi, 2001).

One of the approaches for developing objective priors mentioned in Berger and Pericchi (2001) is referred to as "conventional". More recently, Bayarri et al. (2012) found these priors satisfy a set of optimality criteria, based on what is

specified as Jeffreys' *desiderata*. This construct encompass priors of the form:

$$\pi_\gamma(\alpha, \boldsymbol{\beta}_\gamma, \sigma) = \pi_\gamma(\alpha, \sigma)\ \pi_\gamma(\boldsymbol{\beta}_\gamma) \propto \sigma^{-1}\, \pi_\gamma(\boldsymbol{\beta}_\gamma).$$

This prior is improper, since $\pi_\gamma(\alpha, \sigma)$ is not integrable, and hence only known up to a multiplicative constant. This constitutes, however, an exception to the above discussion: in this situation, there are reasons to consider α and σ as common parameters across models. Since they are common, they will have the same prior and when computing the BFs the unknown multiplicative constant in $\pi_\gamma(\alpha, \sigma)$ will cancel out, resulting in a well-defined posterior (Sansó et al., 1996).

Bayarri et al. (2012) also suggested a new "robust" prior of this form, which we will use in this work. The BF when applying this prior can be expressed in closed-form:

$$B_{i0} = \left[\frac{n+1}{p_i + p_0}\right]^{-p_i/2} \times \frac{Q_{i0}^{-(n-p_0)/2}}{p_i + 1}\, {}_2F_1\left[\frac{p_i+1}{2}; \frac{n-p_0}{2}; \frac{p_i+3}{2}; \frac{(1-Q_{i0}^{-1})(p_i+p_0)}{1+n}\right], \quad (1)$$

where p_0 is the number of independent variables of M_0 - in our case, $p_0 = 1$. p_i is the number of regressors in M_i except for the constant, and $Q_{i0} = \mathrm{SSE}_i/\mathrm{SSE}_0$, where SSE_i is the Sum of Squared Errors of M_i, after applying Ordinary Least Squares (OLS). ${}_2F_1$ is the hypergeometric function.

One also needs to decide on a prior probability for each of the models that we are entertaining. Scott and Berger (2010) recommend the use of a hierarchical prior, via a Bernoulli distribution with a uniform-distributed parameter, to control for multiplicity.

There are several quantities of interest when applying BVS. One can select the model which, after taking data into account, is the most probable - the Highest Probability Model (HPM). Or one can calculate the posterior probability of inclusion for each variable, x_l:

$$q_l = \sum_{\gamma:\, \gamma_l = 1} f(M_\gamma | \boldsymbol{y}),$$

which is an interesting measure of the importance of x_l. Using q_l, we can get the Median Probability Model (MPM), which includes the variables with probability of inclusion $> \frac{1}{2}$. This model is, under some conditions, the optimal choice among single models in terms of performance (Barbieri and Berger, 2004).

We can also use the posterior probabilities of each model as weights of a mix-

ture distribution, a procedure known as Bayesian Model Averaging (BMA) (developed in, for instance, Hoeting et al. (1999)). This constitutes another advantage of BVS, since one is able to explicitly consider uncertainty in the selection procedure. Generally, we use the distribution:

$$f(\Lambda|\boldsymbol{y}) = \sum_{\gamma} f(\Lambda|M_\gamma, \boldsymbol{y}) f(M_\gamma|\boldsymbol{y}),$$

where Λ is a quantity of interest. We may use it to calculate averaged coefficients for the regression. Alternatively, we may calculate the posterior predictive distribution for a new observation, y^*, for the values $\boldsymbol{x^*}$ of the independent variables:

$$f(y^*|\boldsymbol{y}, \boldsymbol{x^*}) = \sum_{\gamma} f(y^*|M_\gamma, \boldsymbol{y}, \boldsymbol{x^*}) f(M_\gamma|\boldsymbol{y}).$$

When there are many variables (and correspondly a very large number of models is being considered), it is computationally unfeasible to average over all possible models. However, we can keep only a set of most probable models, which contain a considerable total posterior probability, and perform averaging only among them, renormalising the posterior probabilities so that their sum equals 1. This approach is referred to by Madigan and Raftery (1994) as 'Occam's Window'.

Using a logarithmic scoring rule, Raftery et al. (1997) showed that using BMA gives better predictive performance than considering a single model. Piironen and Vehtari (2017), using a similar measure, found that better predictive performance also applies when comparing to other variable selection methods (including Bayesian approaches to cross-validation and information criteria).

2.3 The frequentist approach to Variable Selection

It is not clear how to classify frequentist approaches to variable selection into categories. We follow Jamil (2018), who separates between model comparison employing a criterion and shrinkage methods. In the former case, one compares the different models to get the highest (or lowest) criterion, an approach also called best subset selection. This approach may be computationally-intensive, since one needs to calculate the measure for 2^p models, but there are algorithms that make this computation feasible for a moderate p, such as the *leaps* from Furnival and Wilson (1974). Several measures are available, the most well known being R^2 (or its adjusted version), Akaike Information Criterion (AIC) and Bayesian In-

formation Criterion (BIC). AIC (Akaike, 1973) is defined as $n \log\left(\frac{\text{SSE}}{n}\right) + 2p$, where SSE is the Sum of Squared Errors. It should be noted that the AIC is prone to overfit when all models are enumerated (Burnham and Anderson, 2002, p. 436).

BIC (Schwarz, 1978) is similar to AIC, except the penalty term changes to the higher $p \times \log(n)$; BIC $= n \log\left(\frac{\text{SSE}}{n}\right) + p \log(n)$. The BIC is consistent (it selects the true model with probability 1 as n increases); it can be derived through a Laplace approximation, and the difference between the BIC of two models serves as an approximation of the logarithm of the Bayes Factor between them (Kass and Raftery, 1995).

Shrinkage methods consist in placing additional constraints in the optimisation procedure, in order to force (shrink) the less important parameters to be zero (or close to zero). The most popular method is the Lasso (Tibshirani, 1996), where the parameters $\boldsymbol{\beta}$ are obtained by minimising the expression $\sum_i^n (y_i - \sum_p \beta_j x_{ij})^2 + \lambda \sum_j |\beta_j|$ with respect to $\boldsymbol{\beta}$. The first term is the expression used to calculate the OLS estimate, while the second is the penalty term, which lowers the value of the estimated parameters. It can provide estimates which are exactly zero, thus providing guidance for variable selection decisions. This process may improve the predictive performance due to the bias-variance trade-off. It should be noted that, in order to work as desired, all (dependent and independent) variables used in the procedure should be standardised.

Lasso also benefits from a Bayesian interpretation: the results obtained are equivalent to the posterior mode of inducing an independent Laplace prior in each parameter. This fact motivated fully Bayesian approaches to the Lasso (Park and Casella, 2008), and Lasso is also the inspiration for penalised approaches to estimate-based Bayesian Variable Selection (spike-and-slab priors - Ročková and George (2018)), which are out of the scope of this work.

2.4 Heteroscedasticity and BVS: intuition

In Section 3.1, we will detail the circumstances under which we will compare variable selection in an homoscedastic situation with an heteroscedastic scenario. In essence, this will be controlled by the Signal-to-Noise Ratio (SNR). For now, start by noting that the BF comparing models i and j, using the prior defined in Bayarri et al. (2012) (Eq. 1), is only dependent on the models' fit through Q_{ij}, the

ratio of the Sum of Squared Errors (SSE) of the two considered models. Since heteroscedasticity only affects the standard errors (and not coefficient estimates) in a frequentist framework, the SSE will remain close in comparable situations. Therefore, BVS should work more or less the same when the error term goes from homo to heteroscedastic. For this same reason, we should also not expect any significant change from a method such as AIC.

We could not find any literature regarding how BVS behaves under heteroscedasticity, but we may look into studies which might indirectly lead us to some preliminary conclusions. One example is Pelenis (2014), who offers a comparison (in terms of predictive accuracy, coverage of credibility intervals and credible interval lengths) of a Bayesian linear regression with the usual priors (normal and inverse-gamma, the closest case to ours) between homoscedastic and heteroscedastic situations. Although performance, measured with Root-Mean-Square Error, does not change, the coverage and length of credible intervals both get shorter: while uncertainty is reduced (the interval length decreases), heteroscedasticity makes the normal linear regression lose accuracy (when measured by coverage), but this issue is not explored any further.

We may also take an analytical approach. We assume homoscedasticity of the error term in BVS, which means the error variance is proportional to the identity matrix ($\sigma^2 I$). However, in an heteroscedastic situation, the true variance may be written as $\sigma^2\,\Omega$, where σ^2 is no longer the variance of a single error observation, but an arbitrary constant, and Ω is a diagonal matrix.

Using the Cholesky decomposition, we can write:

$$\Omega^{-1} = LL^{\top},$$

where L is a lower triangular matrix. When treating heteroscedasticity of known-form, BVS is tantamount to the problem of parameter estimation in a linear regression. This means that we can pre-multiply the vector y and the matrix X by the matrix $L^{\top}$ to eliminate heteroscedasticity, as it will ensure that the error variance is constant: $Var(L^{\top}y) = L^{\top}\,Var(u)\,L = \sigma^2 L^{\top}\Omega L = \sigma^2 L^{\top}(LL^{\top})^{-1}L = \sigma^2 I$. The same is done when applying the GLS estimator.

In our case, since there is no serial correlation in the error term, the matrix Ω is diagonal. The procedure above may then be simplified, an analogous method to what is employed when applying Weighted Least Squares (WLS). It is sufficient

to divide each term by the square root of each variance term ω_i. The transformed variables will be $\frac{y_{ij}}{\sqrt{\omega_i}}$ and $\frac{x_{ij}}{\sqrt{\omega_i}}$, $i = 1, \ldots, n$; $j = 1, \ldots, p$.

When estimating the parameters $\boldsymbol{\beta}$ with OLS, we minimise the expression (Davidson and MacKinnon, 2004):

$$\text{SSE}_{\text{OLS}} \equiv \hat{u}_{\text{OLS}}^{\top}\hat{u}_{\text{OLS}} = (\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\text{OLS}})^{\top}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\text{OLS}}).$$

On the other hand, the GLS estimate is the one that minimises:

$$(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\text{GLS}})^{\top}\Omega^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}}_{\text{GLS}}).$$

Considering that the SSE for a general $\hat{\boldsymbol{\beta}}$ is $(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^{\top}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$, this makes it clear that $\text{SSE}_{\text{OLS}} \leq \text{SSE}_{\text{GLS}}$. Since these values will virtually never be the same, this may mean there exists a bias in the selection procedure, as we are always using OLS. However, it is not possible to quantify its proportion or direction. First of all, the posterior probability is not dependent on the SSE only. Secondly, and more importantly, the Bayes Factor (and posterior probability) is dependent on a ratio of SSE, rendering any comparisons to be very difficult. This leads us to believe that the relationship between the ratios is contingent on the intrinsic characteristics of each dataset. Furthermore, due to this nature of the ratio, it may happen that dismissing heteroscedasticity (and not applying any transformation) provides better results than taking it explicitly into account. It is the purpose of the next section to investigate this phenomenon in the context of real datasets and resorting to simulation.

3 Simulation Study

In this section, we perform a simulation study, employing datasets that are common in the BVS literature. We start with the simplest form of linear regression, with only one independent variable. We then move on to a multiple linear regression. In each, we simulate the dependent variable $\boldsymbol{y}$, keeping the DGP in terms of the parameters $\boldsymbol{\beta}$, only making changes to the variance(s) of the error term, simulated via a normal distribution. The balance between both situations (homo and heteroscedasticity) is maintained with the Signal-to-Noise Ratio, a measure detailed in the next subsection.

We compare the decisions made by the Highest Probability Model (HPM) and

Median Probability Model (MPM), calculating the average number of times a certain variable is selected. The same results are given for the frequentist methods mentioned in Section 2.3 (AIC, BIC and Lasso). We also analyse inclusion probabilities, looking for further insights.

Predictive performance is evaluated as well. We start by randomly subsetting the data between train and test datasets. BVS is applied on the training set only. We sample 10000 values from the predictive posterior distribution of y^*, getting $f(y^*|\boldsymbol{y}, \boldsymbol{x^*})$, where $\boldsymbol{x^*}$ is an observation of the test data, and compare the posterior mean $\hat{y}_i$ with the observed value y_i using the Relative Squared Error (RSE):

$$\text{RSE} = \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \bar{y})^2},$$

where $\bar{y}$ is the mean of all observations of the dependent variable in the test set.

This measure is utilised instead of other more popular ones, such as the Root Mean Squared Error (RMSE), because it allows comparisons for different scales of $\boldsymbol{y}$. This is especially important when comparing results between the original and the "corrected" datasets, detailed below.

We also calculate the interval length of each predictive distribution. Since the scale of this distribution is not constant, we calculate an "Interval Ratio Length", consisting on dividing the 95% Highest Posterior Density (HPD) interval length by the distance between the maximum and minimum values generated from $f(y^*|\boldsymbol{y})$.

Additionally, results on an estimated coverage probability are reported. It measures how often the credible interval contains the true value of $\boldsymbol{y}$. Both the interval length and coverage are employed by Pelenis (2014), who applies them to the posterior distribution of the parameter $\boldsymbol{\beta}$ in several univariate DGPs, but the interval length is modified to fit our purposes.

Since there are several repetitions, the final result is the average of the values obtained for every simulated $\boldsymbol{\varepsilon}$. We apply these measures to two different situations: after performing Bayesian Model Averaging (BMA), and using only the HPM. Results are expected to improve with BMA, but in an empirical setting a researcher might prefer to use a single model, so evaluating the performance under the HPM is clearly an useful exercise.

As mentioned in Section 2.4, heteroscedasticity can be easily accommodated into Bayesian Variable Selection by a process analogous to Weighted Least Squares. Since the DGP is fully known, we can divide each observation of every variable by the standard deviation of the respective error term (the weight), and repeat all the simulations above, verifying if any significant changes occur whether in the selected variables or the measures of performance. We refer to the resulting datasets as "corrected", and this process will provide an additional term of comparison, which comes from a different point of view to what is described above. If in the first case we will be comparing results when introducing heteroscedasticity (and it is not present before), here we evaluate if correcting for heteroscedasticity (comparing to not doing it) results in any considerable differences. Here, we expect the results to get closer to what happens under homoscedasticity, since the datasets will have, in theory, constant error variance.

Aside from the SNR, another decision one has to make is how to insert heteroscedasticity. It is usually assumed that it is dependent on variables that may or may not be part of the model (Wooldridge, 2015; Koop, 2003), whether through a squared or a exponential function (any function which returns positive values suffices). We focus mainly on variables which are also part of the selection procedure. Since $X = (\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p})$ is the vector of explanatory variables, we define our vector of variances, with size n, as an exponential function:

$$\boldsymbol{\omega} = \exp\{1 + \alpha_1 \boldsymbol{x_1} + \cdots + \alpha_p \boldsymbol{x_p}\}.$$

There is also a need to quantify "how much" heteroscedasticity is present. The idea presented by Gelfand (2015) may be of use here - the *SD Ratio*. To calculate this measure, one starts by selecting the 10% observations of the error term with highest Standard Deviations (SDs) and the 10% observations with lowest SD. Then, the average value of SD is calculated for both groups. Finally, we divide these averages (with the highest value on the numerator) to obtain the SD ratio. We decide to employ values around 3.5, slightly higher than the ones they find most proeminent in the studied datasets.

Our calculations are performed in R (R Core Team, 2017), and more particularly BVS is done with the R package *BayesVarSel* (García-Donato and Forte, 2017). For the calculations that follow, we use the objective prior for the parameters defined in Bayarri et al. (2012), as previously mentioned, and the model prior recommended in Scott and Berger (2010). These are the predefined choices in the

package. The HPD interval is calculated with the package *HDInterval* (Meredith and Kruschke, 2018).

For the multiple linear regression, AIC and BIC need to be applied with the algorithm defined in Furnival and Wilson (1974), with package *bestglm* (McLeod and Xu, 2018). Lasso is performed with the package *glmnet* (Friedman et al., 2010), and variables are standardised for the Lasso only.

3.1 Signal-to-Noise Ratio

In all the conducted simulations, the purpose is to compare an homoscedastic and an heteroscedastic dataset, and check if the procedure's performance is affected. Only by establishing a means of comparison one can make sure that they are fair, and that the introduction of heteroscedasticity does not create excessive noise, which may in itself deteriorate the results of the procedure.

To guarantee this fairness between datasets, we use the Signal-to-Noise Ratio (SNR). This concept suffers from the drawback of not having an universal definition: the meanings of signal and noise are not consensual. Besides, most definitions usually fail to consider the matrix of regressors X, rendering them inappropriate for non-simulated instances. However, this measure is the only one capable of comparing two different DGPs in relation to their coefficients and error terms. We turn to Friedman et al. (2009), who provide the definition:

$$\text{SNR} = \frac{\text{Var}(f(X))}{\text{Var}(\boldsymbol{\varepsilon})},$$

where Var(.) is the sample variance (the elements $f(X)$, which is equal to $X\boldsymbol{\beta}$ in the linear regression, and $\boldsymbol{\varepsilon}$ are known).

There is only an adaptation needed in order to accommodate heteroscedasticity in the error term. This is achieved by replacing the variance of $\boldsymbol{\varepsilon}$ with its average for all its observations, a reasoning that comes from the works of Dobriban and Su (2018) or Jia et al. (2013). Our final SNR definition is:

$$\text{SNR} = \frac{\text{Var}(X\boldsymbol{\beta})}{\frac{1}{n}\sum_i \text{Var}(\varepsilon_i)}.$$

After some simulations, and in order to get more flexibility in our results, we decide to utilise the values of 1 and 10 as references for low and high SNRs. It

should be mentioned that guaranteeing that the average variance of the error term remains constant is more important than the chosen values for the SNR.

3.2 Simple Linear Regression: Setup

For a starting point, we use the dataset *Ozone35*, available in the package *BayesVarSel*, and used by, for instance, Casella and Moreno (2006). We focus only in the original variables (no squared terms or cross products were considered), meaning there are 7 variables to be used as regressors, and one target variable - measures of ozone concentration - that we dismiss, since we need to have full control over the DGP. The number of observations is $n = 178$, and the candidate variables are named X_4 to X_{10}.

The correlations between the regressors are not substantial: something that may help in the procedure, as the prior includes the inverse of the matrix $X^\top X$. The correlation table may be found in Appendix A.1, table A.I.

Focusing on the simple linear regression, we set:

$$\boldsymbol{y} = 2 + 0.5X_7 + \boldsymbol{\varepsilon}$$

The variance of $\boldsymbol{\varepsilon}$ should adapt to conform to the values of SNR that we defined as low and high (1 and 10, respectively). The numerator for this measure is 48.31, meaning that we need an average variance of 48.31 and 4.831 to get our SNR as desired.

We will start by setting the error as homoscedastic, starting with the highest SNR:

$$\varepsilon_i \sim N(0, 4.831) \tag{2}$$

$$\varepsilon_i \sim N(0, 48.31). \tag{3}$$

For evaluation under heteroscedasticity, a function for the variance of the error term needs to be defined. As mentioned, an exponential is used as basis: we need then to adjust the parameters to the defined SD ratio, and finally adjusting the vector of variances to the averages used in the homoscedastic situation (a process that does not affect the SD ratio).

We start by setting the heteroscedasticity dependent only on X_7, the variable

that also enters the model:

$$\omega_i = \exp\{1 + 0.052\, X_{7,i}\} \tag{4}$$

The value 0.0052 is chosen so that the SD ratio is close to 3.5. We then renormalise the vector of variances so its average variance is 4.831 (SNR=10) and 48.31 (SNR=1).

We also make the heteroscedastic error vector dependent on another variable. X_6 is in the same order of magnitude of X_7, but their correlation is not significant (only 0.37). The parameter for an SD Ratio close to 3.5 is 0.038, meaning we use the expression:

$$\omega_i = \exp\{1 + 0.038\, X_{6,i}\},\ i = 1, \ldots, 178, \tag{5}$$

and then renormalise it so the average variances are the same as in the homoscedastic case.

3.2.1 Simple Linear Regression: original dataset

We now present the results of the procedures for this dataset, when working with the original variables.

To evaluate whether heteroscedasticity is detectable, we employ frequentist heteroscedasticity tests (where the null hypothesis is homoscedasticity). Breusch-Pagan and White tests are carried, as defined by Wooldridge (2015): in the first one, the auxiliar regression of the squared residuals is done on the elements of X, whereas in the second one the independent variables are $\hat{y}$, the fitted values, and $\hat{y}^2$ (the "alternative" test). We report the percentage of rejections of the null, after 10000 repetitions:

TABLE I: Percentage of rejections in heteroscedasticity tests: Ozone data

	Breusch-Pagan	White
SNR = 10		
Homoscedasticity	4.48	4.62
Heteroscedasticity - X_7	98.52	99.99
Heteroscedasticity - X_6	97.65	38.17
SNR = 1		
Homoscedasticity	4.73	5.16
Heteroscedasticity - X_7	98.38	99.99
Heteroscedasticity - X_6	97.66	38.37

Results are very similar between the two values of SNR. For the case of homoscedasticity, we get values close to 5% (the test's dimension). Heteroscedasticity is correctly rejected (test's power) most of the time, getting only a less-than-optimal result for the White test when the heteroscedasticity is dependent on X_6, which we do not believe should constitute any worry, as there is still some power to the test (it is still detecting heteroscedasticity) and the Breusch-Pagan has no problem rejecting the null.

The following tables report, in percentage, the number of times a certain variable is selected by each method for each error variance, for values of SNR of 10 and 1. The last column refers to the number of times (in percentage again) where the selected model is the correct one (i.e. the method selects X_7 only). 5000 repetitions were used.

TABLE II: Selection frequencies (%) for Ozone data: SNR = 10

	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Model
Homoscedasticity								
HPM	0.1	0.1	0.2	100	0.3	0.1	0.1	99.1
MPM	0.2	0.2	0.1	100	0.3	0.1	0.1	99.0
AIC	16.7	17.0	16.3	100	17.2	16.2	16.6	37.5
BIC	2.1	2.6	2.5	100	2.9	2.2	2.1	87.1
Lasso	38.9	30.7	28.4	100	34.3	27.1	32.3	20.1
Heteroscedasticity - x7								
HPM	0.0	0.0	0.0	100	0.0	0.1	0.0	99.9
MPM	0.0	0.0	0.0	100	0.0	0.1	0.0	99.9
AIC	8.0	12.9	11.0	100	9.2	12.6	8.6	55.8
BIC	0.4	1.1	0.9	100	0.5	1.1	0.5	95.7
Lasso	36.6	32.6	29.0	100	31.1	28.4	30.8	20.8
Heteroscedasticity - x6								
HPM	0.1	0.0	0.0	100	0.1	0.1	0.0	99.7
MPM	0.1	0.1	0.0	100	0.1	0.1	0.1	99.6
AIC	14.4	14.1	16.1	100	13.6	19.2	12.1	44.3
BIC	1.2	1.1	1.3	100	1.4	2.2	1.2	92.7
Lasso	38.9	29.9	30.2	100	33.0	30.0	32.0	19.6

TABLE III: Selection frequencies (%) for Ozone data: SNR=1

	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Model
Homoscedasticity								
HPM	0.6	0.7	0.6	100	0.7	0.9	0.5	96.5
MPM	0.8	1.2	0.7	100	1.0	1.0	0.8	95.2

	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Model
AIC	16.4	16.5	17.9	100	16.1	17.0	16.3	37.4
BIC	2.2	2.8	2.4	100	2.5	2.3	2.3	86.9
Lasso	37.8	29.8	29.2	100	32.7	27.5	32.7	20.3
Heteroscedasticity - X_7								
HPM	0.1	0.2	0.3	100	0.2	0.2	0.1	99.1
MPM	0.1	0.3	0.3	100	0.3	0.3	0.1	98.7
AIC	7.5	12.6	12.2	100	8.7	12.4	8.2	56.7
BIC	0.2	1.0	1.0	100	0.7	1.1	0.5	95.9
Lasso	36.2	31.5	29.9	100	31.7	27.5	29.4	20.9
Heteroscedasticity - X_6								
HPM	0.3	0.3	0.5	100	0.3	0.6	0.4	98.2
MPM	0.5	0.6	0.6	100	0.5	0.9	0.5	97.1
AIC	13.9	14.6	16.0	100	12.2	17.7	12.7	45.6
BIC	1.2	1.3	1.5	100	1.3	1.8	1.4	92.8
Lasso	38.2	30.8	29.5	100	33.0	28.3	32.1	20.5

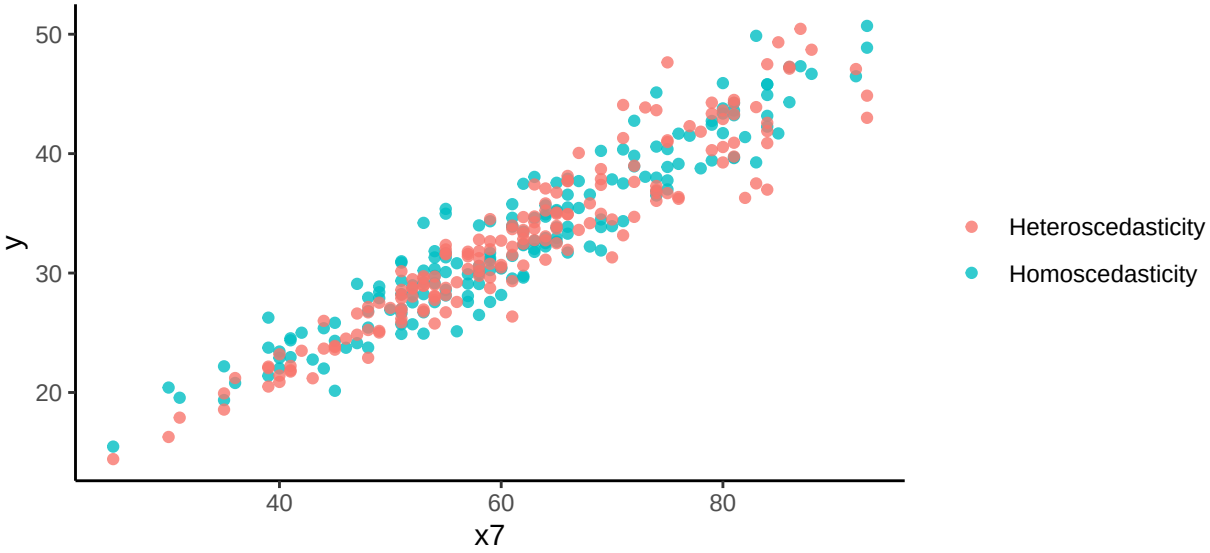
The results show that, on the one hand (and most importantly), BVS does not seem to deteriorate when heteroscedasticity is introduced. The proportion of wrong selections is low for the three considered scenarios and for both values of SNR, even though the percentage of right model's choice is higher when the average error variance is lower. These results are expected, considering our reasoning in Section 2.4.

On the other hand, BVS consistently outperforms the frequentist methods. Both AIC and Lasso are often selecting overfitted models. The BIC is the one that closest resembles the (good) selections of BVS.

Another interesting property is that BVS seems to behave better when the heteroscedasticity is dependent on X_7, the variable that also enters the model. One motive for this may be that observations with higher values will also be the ones with higher variances, diluting the impact of heteroscedasticity; moreover, since the average variance is constant, several observations in the dataset will contain a lower variance comparing to its homoscedastic counterpart, making it easier to capture the signal, and consequently easier for BVS to properly select the HPM and MPM (the same rationale extends to the AIC and BIC).

A simple graphical representation is illustrative of this question. The following figure plots two averages (10 simulations each) of homoscedastic and heteroscedastic instances of $\boldsymbol{y}$ against X_7 (SNR = 1).

FIGURE 1: Two instances of a simulated dependent variable (SNR=1)



We can easily verify that, when the data is heteroscedastic, the points become more "scattered" as X_7 increases, since heteroscedasticity is positively dependent on this variable.

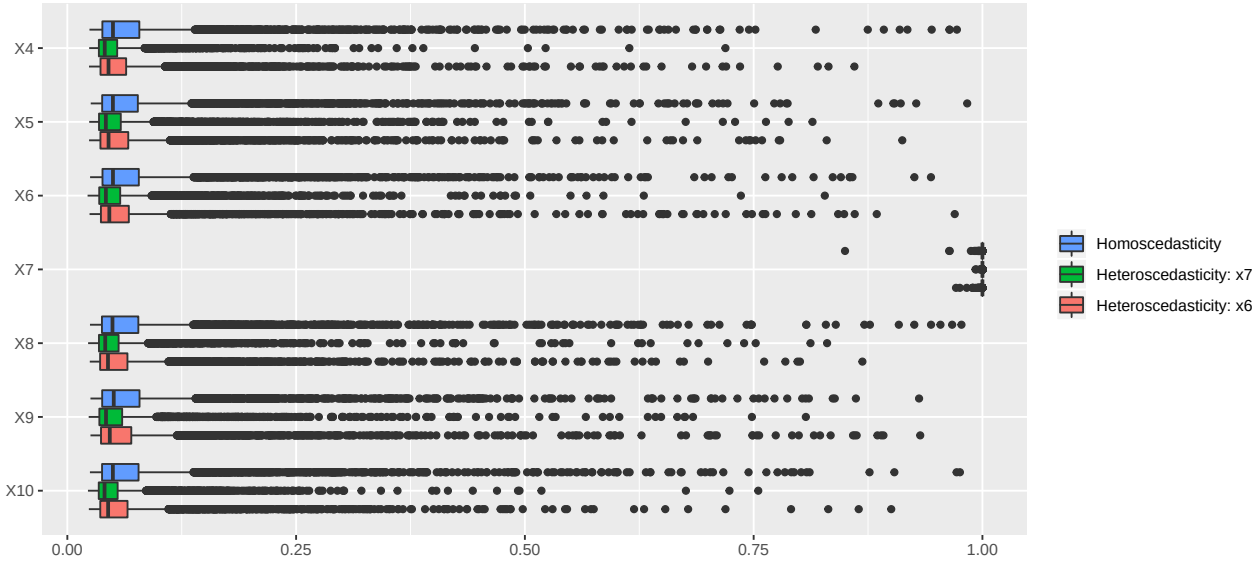
The same explanation may probably be given for the third case considered, where heteroscedasticity depends on X_6, and BVS seems to improve as well. Although a graphical representation is difficult in this case, the fact that selection via AIC and BIC shows less false positives may lead us to conclude that the generation of the dependent variable $\boldsymbol{y}$, and the induced change in the model's fit, is responsible for the majority of these differences (a reasoning that can be extended to dependence on X_7).

An analysis of the SSEs also yields similar conclusions, as there are differences in the ratios between M_γ and M_0 when the variance's specification is altered. For instance, when heteroscedasticity depends on X_7, the ratio with lowest value (corresponding to the model with all variables included) is higher.

For BVS, a further look into the inclusion probabilities is also meritory. To recapitalute, the MPM is the model which includes all variables with an inclusion probability larger than 0.5. From our simulation results, it is expected that X_7 has a more substantial inclusion probability, with the others rarely passing the 0.5 threshold. We can boxplot the results from all repetitions for the three different cases.

The inclusion probabilities for X_7 are concentrated in 1, confirming good per-

FIGURE 2: Boxplot of inclusion probabilities: SNR = 1



formance in selection. Moreover, the quartiles for the remaining variables are all close to zero, although some dispersion is present.

Once again, there seems to be a better performance when the heteroscedasticity depends on X_7 (lower quartiles for the variables not included in the model and less dispersion). We could also extend this conclusion to the other case of heteroscedasticity, although the difference is not so evident.

The last measures of comparison are related to the predictive performance of BVS. For each of 10000 repetitions, 18 random observations are kept for testing, and BVS is applied with the remaining 160. Then, we sample 10000 values from the predictive distribution conditional on every vector of independent variables in the test set, meaning there are 18 different distributions for every data realisation. When applying BMA, all possible 128 models are used. The following table contains the values obtained for the 3 different measures discussed in the beginning of Section 3:

TABLE IV: Predictive performance: Ozone

		RSE	Interval Length	Coverage
SNR = 10				
Homoscedasticity	HPM	0.107	0.505	0.951
	BMA	0.106	0.505	0.952
Heteroscedasticity - X_7	HPM	0.105	0.505	0.943
	BMA	0.106	0.505	0.943

		RSE	Interval Length	Coverage
Heteroscedasticity - X_6	HPM	0.106	0.505	0.942
	BMA	0.107	0.505	0.940
SNR = 1				
Homoscedasticity	HPM	0.571	0.505	0.950
	BMA	0.574	0.505	0.952
Heteroscedasticity - X_7	HPM	0.579	0.505	0.942
	BMA	0.582	0.505	0.942
Heteroscedasticity - X_6	HPM	0.567	0.505	0.941
	BMA	0.574	0.505	0.940

The interval length shows virtually no change between SNR, number of models used or variance types. The Relative Squared Error, however, is inconsistent among SNRs: when it equals 10, this measure is lower for heteroscedastic situations under the HPM. When SNR=1, it gets higher for the first case of heteroscedasticity, and lower when it is dependent on X_6.

Coverage probability is the measure where differences between homoscedasticity and heteroscedasticity are more profound. When the error variance is equal, this measure reaches 0.95 (and even surpasses it), which is the expected result if we approach it from a frequentist standpoint of confidence intervals. When we introduce heteroscedasticity, coverage diminishes: the HPD intervals fail to contain the true value of $\boldsymbol{y}$ more often.

One additional interesting result is that the Relative Squared Error (RSE) is lower when using only the Highest Probability Model in most situations. This result is unexpected, considering the properties of BMA discussed in 2.2. The fact that the posterior probability of the HPM is very close to 1 may be a factor that justifies these values.

3.2.2 Simple Linear Regression: "corrected" datasets

In this section, we utilise the "corrected" datasets, which theoretically accommodate heteroscedasticity in the data. After dividing each observation by the corresponding standard error for the two considered heteroscedastic situations, we report the same measures (choice frequencies and predictive performance), but we restrict them only to fully Bayesian procedures (HPM and MPM) and

AIC[1] in the frequencies table. A comparison with a frequentist method is useful to analyse if the observed differences also apply to non-Bayesian approaches, a situation that, if happening, would corroborate some of our conclusions in the previous section. We also restrict our reported results to heteroscedastic cases only, since applying this approach to the homoscedastic (original) dataset would yield the same results as above.

To verify if heteroscedasticity is no longer detected, we ran the same frequentist heteroscedasticity tests (White and Breusch-Pagan) as in Section 3.2.1. As expected, percentages of rejection are close to 5% in every case, which indicate that heteroscedasticity is not present in the "corrected" datasets (table B.I, Appendix B).

Results concerning frequencies of choice for each variable are presented below:

TABLE V: Selection frequencies (%) for Ozone data ("corrected"): SNR = 10

	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Model
Heteroscedasticity - X_7								
HPM	0.2	0.3	0.2	100	0.2	0.1	0.2	98.8
MPM	0.2	0.3	0.2	100	0.1	0.1	0.2	98.8
AIC	50.3	17.8	17.0	100	18.1	17.0	16.3	19.6
Heteroscedasticity - X_6								
HPM	0.1	0.2	0.1	100	0.2	0.2	0.2	99.2
MPM	0.1	0.2	0.1	100	0.2	0.1	0.2	99.2
AIC	49.1	16.8	16.3	100	17.7	18.0	18.6	20.1

TABLE VI: Selection frequencies (%) for Ozone data ("corrected"): SNR = 1

	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Model
Heteroscedasticity - X_7								
HPM	0.6	0.4	0.3	100	0.7	0.4	0.4	97.2
MPM	0.8	0.7	0.5	100	1.0	0.7	0.7	96.1
AIC	25.3	18.7	17.7	100	18.8	16.5	17.7	28.2
Heteroscedasticity - X_6								
HPM	0.5	0.7	0.6	100	0.6	0.8	0.6	96.7
MPM	0.7	1.0	0.8	100	0.8	0.9	0.9	95.6

[1]Unfortunately, since the weights are also applied to the intercept, this term needs to be explicitly included in the design matrix X when performing the computations. Therefore, it is not possible to include it by default in every considered model when applying AIC. This fact may induce some bias in the procedure, but only for the frequentist method.

	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	Model
AIC	20.6	17.1	16.1	100	17.2	16.4	18.2	34.7

To recapitulate, when no correction is made, BVS works better if there is heteroscedasticity (especially when it depends on X_7), selecting spurious variables less often and, in consequence, increasing the number of times the right model is selected.

After applying weights to each observation, we see that the choices made by BVS become more resembling of the homoscedastic dataset. Comparing to the situations where homoscedasticity is assumed, we have more incorrect selections and less correct models overall, for both values of SNR. We might say that this correction eliminates, as intended, the differences in the model's fit that existed between homo and heteroscedasticity.

The frequentist method is, however, showing a different behaviour: the variable X_4 is being selected more often. We suspect that the high correlation between X_4 and X_7 leads to this issue, as it is close to 0.9 in both cases (Tables A.II and A.III, Appendix A.1).

We now analyse the three measures of predictive performance:

TABLE VII: Predictive performance: Ozone ("corrected")

		RSE	Interval Length	Coverage
SNR = 10				
Heteroscedasticity - X_7	HPM	0.182	0.505	0.952
	BMA	0.185	0.505	0.951
Heteroscedasticity - X_6	HPM	0.023	0.505	0.951
	BMA	0.023	0.505	0.951
SNR = 1				
Heteroscedasticity - X_7	HPM	0.733	0.505	0.951
	BMA	0.735	0.505	0.952
Heteroscedasticity - X_6	HPM	0.194	0.505	0.952
	BMA	0.195	0.505	0.952

The largest differences occur in the first performance measure, the RSE. When heteroscedasticity (and the weights) are dependent on X_7, this value is substantially higher than when considering homoscedasticity. This is reversed when we

move to the remaining heteroscedasticity function (X_6), where performance improves by this measure.

Looking more closely at the formula for the RSE, we see that the denominator is the sample variance of the dependent variable in the test set (missing only the division by the number of observations, which cancels out with the numerator). Calculating the numerator and denominator separately, the first component is more or less constant when comparing the two weights used, while the denominator is much lower (higher) using weights depending on X_7 (X_6), driving all the disparities encountered between each case.

Reasons for this are simple. $\boldsymbol{y}$ is a function of X_7. When the corrections use this same variable, and considering that the variance is a positive function of X_7, the values of this variable (and the values of $\boldsymbol{y}$) get more concentrated. When the weights depend on X_6, the opposite is verified.

This can be seen in the sampled predictive distributions. Scale almost does not change when heteroscedasticity depends on X_7 (by visual inspection, mass is usually in the 4-8 range), while in X_6 the support of each distribution shows more variability (the range length is similar, but mass is around the values 2-6 and 8-12). This fact does not interfere with the average variance of the distributions, which is virtually equal between these two situations, but the mean is higher in the second case.

The question to address is if these results reflect true predictive performance (in other words, if RSE is working as desired). The answer must be positive: the RSE is harmonising squared differences in different scales. This does not mean, however, that the use of weights should be automatically discouraged. This measure only utilises point estimates, while the remaining take full advantage of the predictive distribution.

As for the other two measures, the interval length, once again, does not show any difference. The coverage probability is where contrasts are more evident. Values close to 0.95 are always reached, which constitutes an improvement comparing to when heteroscedasticity is ignored.

3.3 Multiple Linear Regression: Setup

We now extend these results to the multiple linear regression, a case that, as more interesting, is also more nuanced. To illustrate this scenario, we use the Crime dataset, originally from Vandaele (1978), and available in *BayesVarSel*. It was used extensively by Raftery et al. (1997) to illustrate Bayesian Model Averaging, and also by Fernández et al. (2001) or Liang et al. (2008) to ellucidate problems of model uncertainty and selection.

This dataset is composed of 47 observations (one for each US state in this study from 1960) with 15 variables (we discard $\boldsymbol{y}$ once again), mostly economic and social indicators.

We arbitrarly decide on the following DGP:

$$\boldsymbol{y} = 0.1\,Ed + 0.5\,Po2 + 1.2\,Pop - 0.2\,Ineq + 100\,Prob - 0.1\,Time + \boldsymbol{\varepsilon}$$

In order to get a SNR with a value 1, the average variance of the error needs to be 3124.117. The considerable number of variables in the DGP may make this number a bit excessive, but having an equal average variance is more decisive in order to make realiable comparisons. We also expect that the existence of parameters with negative values will contribute to worsen the results, as the signal might be more difficult to capture.

For our heterocedastic datasets, we decide to start with a different approach. The first is setting the variance through an exponencial function, $\exp\{\boldsymbol{w}\}$, where $\boldsymbol{w}$ is the sequence of 47 equally spaced observations between 1 and 3.7 (this ensures an SD ratio of 3.5):

$$\omega_i = \exp\left\{1 + \frac{2.7}{46} i\right\},\ i = 1, \ldots, 47 \tag{6}$$

Secondly, we divide the dataset in two parts, the first one having a variance 12.25 times higher than the second (to get, once again, the desired SD ratio):

$$\omega_i = \begin{cases} 1, & i = 1, \ldots, 23 \\ 12.25, & i = 24, \ldots, 47 \end{cases} \tag{7}$$

Then, our variance function will depend on a variable included and one not in-

cluded in the DGP. We decide on:

$$\omega_i = \exp\{1 + 0.0186\, Pop_i\}\,,\ i = 1, \ldots, 47 \tag{8}$$

$$\omega_i = \exp\{1 + 0.0077\, GDP_i\}\,,\ i = 1, \ldots, 47 \tag{9}$$

Once again, we transform all these expressions to conform to the defined SNR.

3.3.1 Multiple Linear Regression: original dataset

Reported findings for the original dataset follow the sequence of Section 3.2.1. We begin by performing the heteroscedasticity tests described in that section, with 10000 generated datasets:

TABLE VIII: Percentage of rejections in heteroscedasticity tests: Crime

	Breusch-Pagan	White
SNR = 10		
Homoscedasticity	1.44	4.31
Heteroscedasticity - exponential	1.90	4.90
Heteroscedasticity - two values	1.23	4.72
Heteroscedasticity - Pop	23.81	53.91
Heteroscedasticity - GDP	5.71	17.5
SNR = 1		
Homoscedasticity	1.63	4.67
Heteroscedasticity - exponential	1.86	3.14
Heteroscedasticity - two values	1.25	3.37
Heteroscedasticity - Pop	23.96	63.36
Heteroscedasticity - GDP	5.94	21.29

The number of correct rejections can only be considered acceptable when heteroscedasticity is a function of the dataset variables. One factor that may be driving these results is that the variance is not ordered by values of $\boldsymbol{y}$. However, ordering the generated dependent variable does not affect the obtained power in any of the considered cases. Moreover, some simulations (like making $\boldsymbol{y}$ dependent on one variable and heteroscedasticity dependent on the same variable) did not improve the number of rejections, as did not changing the value of the SNR further. These are, however, tests based on an auxiliary regression which contains the regressors, or the fitted values of $\boldsymbol{y}$, which are a function of the regressors. It can, therefore, be expected that they perform poorer when the variance does not

depend on any independent variable. Further explanations are the low power these tests have for small samples (Long and Ervin, 2000), and the high number of degrees of freedom (15) in the Breusch-Pagan test.

The following tables report the selection frequencies for the variables in the dataset, along with the number of times the right model was selected. The variables belonging to the DGP are in bold. We use 5000 data realisations for this procedure.

TABLE IX: Selection frequencies (%) for Crime data: SNR=10

	M	So	**Ed**	Po1	**Po2**	LF	M.F	**Pop**	NW	U1	U2	GDP	**Ineq**	**Prob**	**Time**	Model
Homoscedasticity																
HPM	1.3	1.3	2.8	18.3	51.8	0.5	0.5	100	1.8	0.4	0.4	13.1	21.6	1.1	1.5	0.0
MPM	1.0	0.9	2.5	15.9	44.6	0.6	0.6	100	1.3	0.4	0.3	8.8	20.6	1.2	1.7	0.0
AIC	23.2	24.0	26.1	46.4	61.9	23.7	23.3	100	24.9	22.7	22.7	27.6	55.0	31.7	24.6	0.0
BIC	9.3	11.4	13.1	34.3	57.1	8.5	8.0	100	11.6	7.5	6.1	15.5	43.0	15.4	10.9	0.0
Lasso	41.9	27.3	35.3	47.7	73.9	31.3	32.1	100	23.0	29.2	29.3	43.8	72.1	32.7	32.9	0.1
Heteroscedasticity - Exponential																
HPM	2.0	1.3	2.9	20.7	51.9	1.2	0.4	100	1.5	0.4	0.7	11.4	23.4	1.2	2.6	0.0
MPM	1.8	0.8	2.5	18.3	46.2	1.2	0.4	100	0.8	0.2	0.6	6.7	22.0	1.5	2.9	0.0
AIC	24.6	24.5	28.1	55.0	64.1	31.2	25.7	100	24.0	23.2	24.5	22.5	56.0	29.5	30.4	0.0
BIC	11.0	9.8	13.9	41.1	57.9	13.2	9.6	100	10.4	8.1	8.5	12.9	43.1	13.3	14.6	0.0
Lasso	42.9	25.1	35.4	50.4	70.9	34.3	31.4	100	22.2	29.8	31.1	42.1	72.6	29.5	37.6	0.0
Heteroscedasticity - two values																
HPM	0.3	2.2	3.0	13.2	57.4	0.1	0.2	100	2.7	0.1	0.3	16.0	19.4	2.3	0.7	0.0
MPM	0.2	1.5	2.0	10.4	47.9	0.0	0.2	100	2.1	0.1	0.3	11.8	18.0	2.9	1.0	0.0
AIC	23.3	28.2	24.9	36.2	62.4	14.4	20.2	100	28.1	22.4	24.4	32.8	53.8	36.7	27.3	0.0
BIC	8.1	13.3	11.6	26.5	62.3	3.2	6.1	100	14.0	5.2	7.1	19.8	42.0	19.3	11.1	0.0
Lasso	40.1	31.0	36.7	45.5	81.1	24.4	30.8	100	26.6	27.5	30.3	45.7	72.4	37.4	34.1	0.0
Heteroscedasticity - Pop																
HPM	1.3	0.8	3.2	12.3	57.3	0.2	0.1	100	0.9	0.3	0.5	16.9	21.2	0.6	0.7	0.0
MPM	1.2	0.5	2.4	9.7	51.7	0.2	0.1	100	0.4	0.2	0.6	12.3	20.9	0.7	0.9	0.0
AIC	25.5	20.0	25.1	41.4	62.1	18.8	18.4	100	20.1	21.8	28.1	30.1	56.5	29.2	24.9	0.1
BIC	9.8	7.2	13.9	28.7	60.4	5.5	6.4	100	8.5	6.7	10.0	19.8	44.1	12.8	10.6	0.0
Lasso	43.5	25.7	36.9	45.8	76.3	28.0	31.4	100	20.4	26.6	34.0	45.1	74.7	30.5	33.5	0.1
Heteroscedasticity - GDP																
HPM	0.5	2.0	3.9	21.0	48.9	0.3	0.7	100	0.6	0.3	0.6	12.8	22.6	1.7	1.3	0.0
MPM	0.5	1.5	3.5	19.0	43.7	0.3	0.7	100	0.3	0.3	0.6	8.4	21.8	2.3	1.6	0.0
AIC	20.6	22.3	28.0	50.1	61.9	18.7	23.4	100	17.6	20.6	27.0	26.8	54.8	35.1	25.9	0.0
BIC	7.4	11.6	15.4	38.1	56.7	6.3	9.9	100	6.8	6.5	9.5	16.5	44.0	17.7	11.5	0.0
Lasso	41.8	28.7	37.6	47.4	73.0	28.8	36.1	100	17.7	27.6	32.6	44.6	74.5	35.5	33.4	0.0

TABLE X: Selection frequencies (%) for Crime data: SNR=1

	M	So	**Ed**	Po1	**Po2**	LF	M.F	**Pop**	NW	U1	U2	GDP	**Ineq**	**Prob**	**Time**	Model
Homoscedasticity																
HPM	3.8	3.7	5.6	9.3	11.7	2.7	3.3	95.8	3.3	2.3	2.5	8.4	8.4	3.0	3.2	0
MPM	4.0	3.6	4.8	7.7	8.0	3.5	4.4	98.1	3.5	2.9	3.2	6.1	6.2	4.0	4.2	0
AIC	22.9	23.3	24.8	31.1	33.5	23.2	23.3	99.3	24.3	23.5	23.6	27.4	29.3	24.1	23.2	0
BIC	9.2	9.3	11.8	17.2	20.5	8.1	8.4	98.4	8.7	7.5	6.8	14.9	16.7	8.7	8.4	0
Lasso	33.9	22.1	21.6	40.8	44.8	27.5	27.3	100.0	20.4	24.6	27.3	29.5	28.5	29.6	28.3	0
Heteroscedasticity - Exponential																
HPM	5.4	3.8	6.4	9.8	13.7	4.4	3.9	95.2	3.6	3.4	3.5	6.4	10.4	3.3	5.1	0
MPM	4.9	3.8	6.1	9.5	12.2	5.8	4.8	98.0	3.4	3.1	3.5	3.7	7.9	3.0	7.0	0
AIC	23.9	24.1	27.0	38.1	40.7	29.7	26.3	99.2	23.4	21.8	23.3	21.3	30.2	20.4	28.8	0
BIC	10.6	8.7	12.6	20.3	24.2	11.5	9.2	98.4	8.1	7.6	8.5	11.5	18.0	7.0	11.6	0
Lasso	35.5	21.7	23.1	40.6	46.8	29.1	28.3	100.0	18.9	25.7	27.9	25.5	29.7	26.5	33.4	0
Heteroscedasticity - two values																
HPM	2.2	3.8	5.0	6.3	11.0	1.6	2.4	97.4	3.6	1.6	1.7	11.3	8.0	3.9	2.3	0
MPM	3.4	4.2	4.1	4.0	5.7	1.1	2.8	98.9	5.3	2.1	2.7	8.7	6.8	5.9	3.8	0
AIC	23.4	25.7	22.5	22.8	25.7	12.5	20.0	99.6	28.0	22.2	23.8	32.2	28.0	31.3	24.1	0

	M	So	**Ed**	Po1	**Po2**	LF	M.F	**Pop**	NW	U1	U2	GDP	**Ineq**	**Prob**	**Time**	Model
BIC	7.3	10.1	10.8	11.9	18.4	3.0	5.6	98.8	11.5	5.8	6.7	19.5	15.7	13.3	8.1	0
Lasso	30.4	24.0	21.8	39.3	47.9	21.1	26.1	100.0	24.7	22.8	28.1	33.3	27.1	36.4	26.4	0
Heteroscedasticity - Pop																
HPM	3.2	2.3	5.6	6.5	17.5	1.5	2.1	88.2	1.9	1.8	1.9	7.4	7.6	1.6	2.1	0
MPM	6.8	1.9	4.5	4.4	11.9	1.6	2.2	90.3	1.6	1.9	3.2	5.6	5.4	1.5	2.7	0
AIC	24.2	20.0	24.3	27.0	37.9	18.0	18.3	94.1	19.2	21.1	27.2	29.7	27.9	18.5	20.0	0
BIC	9.6	6.0	12.3	12.9	27.9	4.8	6.3	92.3	5.1	6.1	8.5	17.1	16.3	5.2	6.2	0
Lasso	32.4	19.6	23.3	34.7	47.8	23.0	25.8	94.4	15.4	21.8	27.5	30.7	26.4	24.5	25.8	0
Heteroscedasticity - GDP																
HPM	2.8	3.3	4.0	14.9	14.3	2.1	4.2	91.2	2.0	2.2	2.4	6.6	8.0	3.7	3.2	0
MPM	3.2	3.2	3.9	13.7	11.6	1.9	5.2	94.6	1.3	2.2	3.4	5.7	6.5	4.8	3.9	0
AIC	19.5	21.3	25.1	38.6	38.0	16.8	22.7	98.2	15.7	20.6	26.2	26.4	28.2	27.1	23.0	0
BIC	7.0	8.1	10.6	23.9	24.5	5.0	9.8	95.9	3.9	5.8	8.2	14.1	15.7	10.8	8.2	0
Lasso	32.0	22.7	22.6	40.8	41.9	23.7	30.7	99.5	13.0	21.2	28.1	29.3	28.5	31.8	27.5	0

A mixture of the variance, which we believe is too high, and the chosen parameters result in selections that are not satisfactory. The correct model is virtually never selected. We see the same behaviour comparing to 3.2: the Lasso and AIC are selecting models containing more variables, which helps making some right decisions. HPM, MPM and BIC are selecting more parsimonious models.

The only variable selected most of the time is Pop. Since $\beta_{\text{Pop}} = 100$, this high value (even though the magnitude of the variable is small) might be capturing a considerable portion of the signal. For this and the remaining variables, while there are changes in all methods when heteroscedasticity is introduced, neither are they significant nor their signs unambiguous: the changes in percentages of inclusion for each variable do not show any distinct pattern. Moreover, more significant changes are usually related to selecting other variables which are highly correlated (e.g. Po1 instead of Po2).

For evaluation of predictive performance, 7 observations were randomly kept aside for testing, leaving the remaining 40 for training. For BMA, due to computational constraints, we keep only the 256 most probable models when simulating from the posterior predictive distribution: this allows a total posterior probability ranging from 0.5 to 0.9 when the SNR equals 1. 10000 instances of $\boldsymbol{y}$ were simulated.

TABLE XI: Predictive performance: Crime

		RSE	Interval Length	Coverage
SNR = 10				
Homoscedasticity	HPM	0.325	0.475	0.932
	BMA	0.301	0.472	0.951
Heteroscedasticity - exponential	HPM	0.324	0.475	0.925
	BMA	0.297	0.472	0.942

		RSE	Interval Length	Coverage
Heteroscedasticity - two values	HPM	0.315	0.475	0.918
	BMA	0.293	0.472	0.933
Heteroscedasticity - Pop	HPM	0.295	0.475	0.927
	BMA	0.280	0.472	0.942
Heteroscedasticity - GDP	HPM	0.316	0.475	0.921
	BMA	0.289	0.472	0.938
SNR = 1				
Homoscedasticity	HPM	1.040	0.475	0.930
	BMA	0.960	0.466	0.949
Heteroscedasticity - exponential	HPM	1.053	0.475	0.922
	BMA	0.953	0.467	0.938
Heteroscedasticity - two values	HPM	1.020	0.475	0.913
	BMA	0.934	0.466	0.930
Heteroscedasticity - Pop	HPM	1.403	0.475	0.923
	BMA	1.267	0.467	0.938
Heteroscedasticity - GDP	HPM	1.122	0.475	0.922
	BMA	1.031	0.467	0.935

For this dataset, conclusions from the RSE vary with the value of SNR considered: when the amount of noise is lower (higher SNR), the introduction of heteroscedasticity improves the predictive accuracy of BVS. However, when the level of noise increases, this is only true for the first two cases of heteroscedasticity, where the White and BP tests performed poorly in capturing heteroscedasticity. When heteroscedasticity is dependent on a dataset variable, it rises substantially, even after BMA. The high variance needs to be considered when looking at the values obtained from the lowest SNR.

Once again, differences in interval length are not substantial. As for the coverage probabilities, we get a confirmation of the results obtained in 3.2: the introduction of heteroscedasticity is detrimental to the coverage of the HPD interval. When BMA is applied, coverage under homoscedasticity is always around the desired 0.95.

Lastly, and contrary to 3.2, Bayesian Model Averaging improves the results for every situation and measure considered.

3.3.2 Multiple Linear Regression: "corrected" datasets

We now repeat the approach used in Section 3.2.2, applying it to this dataset. We utilise 4 different sets of weights, corresponding to the 4 different heteroscedasticity functions defined in the previous section.

We also ran frequentist heteroscedasticity tests. While the low quality results for this dataset were already discussed, we get no evidence of heteroscedasticity (table B.II, Appendix B).

Choices made by BVS can be seen in the following tables:

TABLE XII: Selection frequencies (%) for Crime data ("corrected"): SNR = 10

	M	So	**Ed**	Po1	**Po2**	LF	M.F	**Pop**	NW	U1	U2	GDP	**Ineq**	**Prob**	**Time**	Model
Heteroscedasticity - Exponential																
HPM	1.3	1.1	2.5	19.2	65.1	0.5	0.4	100	2.1	0.3	0.4	7.4	30.9	1.5	1.9	0.0
MPM	0.9	0.5	1.8	16.7	61.2	0.4	0.4	100	1.2	0.2	0.3	5.2	30.2	1.5	1.9	0.0
AIC	27.2	25.6	29.1	47.9	67.0	26.1	28.1	100	25.7	23.7	24.3	28.0	68.9	36.7	27.4	0.0
Heteroscedasticity - two values																
HPM	1.9	5.4	7.5	18.5	79.3	0.5	0.5	100	2.6	0.4	0.3	6.1	64.2	5.4	3.9	0.0
MPM	1.2	3.4	5.7	17.6	80.0	0.4	0.3	100	1.6	0.2	0.2	5.0	60.9	4.4	3.0	0.0
AIC	26.9	26.4	32.2	39.7	74.7	26.1	27.4	100	24.2	24.2	23.3	27.3	84.3	42.9	31.0	0.2
Heteroscedasticity - Pop																
HPM	2.2	2.2	6.0	18.7	67.5	1.1	1.0	100	3.2	0.7	1.2	11.2	38.3	5.3	3.7	0.0
MPM	1.9	1.9	5.0	18.7	65.5	1.3	1.4	100	2.4	1.0	1.2	9.2	39.6	5.9	3.9	0.0
AIC	25.4	24.9	30.7	45.7	67.8	25.9	28.6	100	24.1	24.4	23.8	28.1	72.5	39.8	27.6	0.1
Heteroscedasticity - GDP																
HPM	1.4	2.5	5.9	18.8	45.3	0.7	0.6	100	2.1	0.5	0.6	8.1	39.2	2.6	2.0	0.0
MPM	1.0	1.8	4.7	16.0	41.7	0.6	0.6	100	1.4	0.4	0.6	5.8	36.9	2.9	2.0	0.0
AIC	26.4	25.1	28.8	47.4	61.2	25.6	28.0	100	26.9	24.6	24.2	27.3	68.8	35.9	26.9	0.0

TABLE XIII: Selection frequencies (%) for Crime data ("corrected"): SNR = 1

	M	So	**Ed**	Po1	**Po2**	LF	M.F	**Pop**	NW	U1	U2	GDP	**Ineq**	**Prob**	**Time**	Model
Heteroscedasticity - Exponential																
HPM	2.7	1.8	5.1	12.2	16.6	1.3	1.0	99.7	1.3	0.8	0.9	9.6	12.2	1.6	1.8	0.0
MPM	2.7	1.5	2.7	5.6	7.7	1.5	1.3	99.9	1.5	1.4	1.3	4.7	7.5	1.8	2.5	0.0
AIC	26.6	25.3	28.0	34.8	38.5	27.4	27.1	100	24.7	23.8	23.8	28.7	33.5	25.3	23.7	0.0
Heteroscedasticity - two values																
HPM	1.9	2.8	5.2	15.1	31.9	0.6	0.9	100	1.0	0.8	0.6	18.5	15.5	1.0	1.1	0.0
MPM	1.2	1.9	2.9	9.2	18.9	0.8	1.1	100	0.8	0.5	0.6	9.3	10.4	1.2	1.0	0.0
AIC	26.6	26.1	26.7	38.1	47.4	25.3	28.0	100	25.3	24.7	23.1	29.6	40.0	25.6	25.7	0.0
Heteroscedasticity - Pop																
HPM	13.3	12.8	15.5	21.7	28.1	12.1	12.9	75.1	12.7	12.1	12.1	20.0	21.3	12.3	12.3	0.0
MPM	14.7	9.8	10.7	20.1	24.7	8.7	10.5	88.5	9.5	9.0	8.9	14.3	16.7	10.1	9.2	0.0
AIC	26.8	24.3	27.9	35.0	39.4	25.9	27.0	96.7	23.8	24.5	24.1	30.0	34.2	26.3	25.1	0.0
Heteroscedasticity - GDP																
HPM	5.6	5.9	9.5	14.5	14.7	4.1	4.8	89.8	5.1	3.7	3.9	10.1	13.6	4.7	4.5	0.0
MPM	5.4	5.3	7.2	10.8	11.3	4.6	5.0	94.9	4.4	3.8	3.9	6.9	11.0	5.4	4.6	0.0
AIC	26.6	24.2	27.3	32.2	34.5	28.1	28.3	98.9	23.1	24.2	24.1	27.6	34.6	24.9	23.3	0.0

For some variables, the number of correct choices is improving: this is the case for Po2, Ineq and, more often than not, and when SNR=10, the remaining variables in the DGP. However, it is not possible to conclude that BVS is displaying a

better behaviour: the case of Po1 when the SNR equals 1 is an appropriate example. AIC is also showing these attributes, which leads us to believe that changes in the datasets' characteristics are the reason for the observed differences.

In fact, correlations are stronger when we correct for heteroscedasticity (tables may be found in Appendix A.2). Excluding the diagonal and the new constant, there are 105 different correlations between all independent variables. When no weights are applied, 30 of them are larger than 0.5 in absolute value; when weights are used, we get (in order) 70, 69, 51 and 75 values in the same situation.

Predictive performance is shown below:

TABLE XIV: Predictive performance: Crime ("corrected")

		RSE	Interval Length	Coverage
SNR = 10				
Heteroscedasticity - exponential	HPM	0.197	0.475	0.931
	BMA	0.178	0.471	0.951
Heteroscedasticity - two values	HPM	0.147	0.474	0.920
	BMA	0.129	0.468	0.951
Heteroscedasticity - Pop	HPM	0.606	0.474	0.921
	BMA	0.542	0.471	0.949
Heteroscedasticity - GDP	HPM	0.447	0.475	0.927
	BMA	0.398	0.472	0.950
SNR = 1				
Heteroscedasticity - exponential	HPM	0.803	0.475	0.931
	BMA	0.738	0.471	0.949
Heteroscedasticity - two values	HPM	0.625	0.475	0.928
	BMA	0.584	0.470	0.951
Heteroscedasticity - Pop	HPM	1.707	0.472	0.923
	BMA	1.451	0.456	0.949
Heteroscedasticity - GDP	HPM	1.322	0.474	0.926
	BMA	1.181	0.461	0.950

Compared to the results obtained under the original dataset, correcting for heteroscedasticity improves the Relative Squared Error (RSE) for the two first cases of heteroscedasticity, while for the last two this measure worsens when applying weights. These considerable differences are similar to the results verified in Section 3.2.2: once again, the numerators of the expression and the predictive distributions' scales are similar, but for the last two cases the denominator is con-

siderably lower, increasing the obtained RSE.

The remaining measures show more definite realities. The interval length only worsens for the first two situations, and when SNR=1; otherwise it remains constant or decreases, especially when applying BMA. The advantages of considering more than one model may be better demonstrated when looking at the coverage probabilities. Improvements are modest when looking at the HPM, but after BMA they reach the desired 0.95.

4 Conclusion

The objective of this thesis was to evaluate the consequences of heteroscedasticity on BVS. To this effect, we performed an extensive simulation study, comparing the selections made and predictive performance in two situations: when introducing heteroscedasticity, and when using a weighted dataset to correct for heteroscedasticity.

In the first situation, it is important to refer that special care was taken in guaranteeing that the introduction of heteroscedasticity did not bring excessive noise to the datasets. To this effect, we adapted the existing definitions of Signal-to-Noise Ratio to fit our needs.

We conclude that differences in the selected variables occur due to changes in the underlying datasets. This takes the form of better/worse fit when comparing to an homoscedastic dataset, and these changes in fit are not substantial. When applying weights to correct for heteroscedasticity in each heteroscedastic case, changes in the correlation structure are the driving factor behind modifications in the selected variables. BVS can be considered to be robust to the introduction of heteroscedasticity, although the quality of selection may not improve when taking the structure of the error's variance into account.

Even though the RSE shows mixed results, the interval length and coverage probability, especially this last measure, show better values when variance is explicitly considered. Therefore, using the standard deviations as weights may improve predictive performance. Since in a non-simulated environment this is not possible, fully Bayesian approaches may be developed to allow for heteroscedastic errors, which could in turn provide better predictions under BVS. The impor-

tance of considering Bayesian Model Averaging should also be highlighted.

In terms of limitations of our study, it is important to point out that a different approach could have been taken. Even though we looked at several different results (choice frequencies, predictive performance, and inclusion probabilities for the univariate linear regression) from BVS, one advantage of BVS is its richness, and other aspects of the analysis could have been explored in detail. For instance, it is possible to check the consequences in the posterior model probabilities, the estimates of the coefficients or the probabilities of selecting a certain number of variables, a task that could yield additional or even different conclusions.

For future work, the same strategy can be applied to slightly different situations. An example is the use of different priors, including the mentioned spike-and-slab distributions. It would also be interesting to study the case of serial correlation in the errors, which is more general than the one here considered, but where it is possible to employ the same measures, with some modifications. Although usually not considered, the same study may be done in a $p > n$ situation, where the family of conventional priors has already seen some developments (Berger et al., 2016).

References

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer New York, New York, NY.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Bayarri, M., Berger, J. O., Forte, A., and García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577.

Berger, J. O., García-Donato, G., Martínez-Beneito, M. A., and Peña, V. (2016). Bayesian variable selection in high dimensional problems without assumptions on prior model probabilities. *arXiv e-prints*, page arXiv:1807.00347.

Berger, J. O. and Pericchi, L. (2001). Objective bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, 38:135–207.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica*, 47(5):1287–1294.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag.

Casella, G. and Moreno, E. (2006). Objective bayesian variable selection. *Journal of the American Statistical Association*, 101.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A., and Goodner, B. (2007). Spatially adaptive bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16(2):265–288.

Davidson, R. and MacKinnon, J. (2004). *Econometric Theory and Methods*. Oxford University Press, New York, NY.

Dobriban, E. and Su, W. J. (2018). Robust Inference Under Heteroskedasticity via the Hadamard Estimator. *arXiv e-prints*, page arXiv:1807.00347.

Fernández, C., Ley, E., and Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

Forte, A., García-Donato, G., and Steel, M. (2018). Methods and Tools for Bayesian Variable Selection and Model Averaging in Normal Linear Regression. *International Statistical Review*, 86(2):237–258.

Forte, A., Peiró-Palomino, J., and Tortosa-Ausina, E. (2015). Does social capital matter for european regional growth? *European Economic Review*, 77:47–64.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag New York, 2nd edition.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Furnival, G. M. and Wilson, R. W. (1974). Regressions by leaps and bounds. *Technometrics*, 16(4):499–511.

García-Donato, G. and Martínez-Beneito, M. A. (2013). On Sampling Strategies in Bayesian Variable Selection Problems With Large Model Spaces. *Journal of the American Statistical Association*, 108(501):340–352.

García-Donato, G. and Forte, A. (2017). *BayesVarSel: Bayes Factors, Model Choice and Variable Selection in Linear Models*. R package version 1.8.0.

Gelfand, S. J. (2015). Understanding the Impact of Heteroscedasticity on the Predictive Ability of Modern Regression Methods. Master's thesis, Simon Fraser University.

Hayashi, F. (2000). *Econometrics*. Princeton University Press.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Jamil, H. (2018). *Regression modelling using priors depending on Fisher information covariance kernels (I-priors)*. PhD thesis, London School of Economics and Political Science.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461.

Jia, J., Rohe, K., and Yu, B. (2013). The lasso under poisson-like heteroscedasticity. *Statistica Sinica*, 23(1):99–118.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Koenker, R. (1981). A note on studentizing a test for heteroscedasticity. *Journal of Econometrics*, 17(1):107–112.

Koop, G. (2003). *Bayesian Econometrics*. J. Wiley, Hoboken, NJ.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

McLeod, A. and Xu, C. (2018). *bestglm: Best Subset GLM and Regression Utilities*. R package version 0.37.

Meredith, M. and Kruschke, J. (2018). *HDInterval: Highest (Posterior) Density Intervals*. R package version 0.2.0.

Norets, A. (2015). Bayesian regression with nonparametric heteroskedasticity. *Journal of Econometrics*, 185(2):409–419.

Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pelenis, J. (2014). Bayesian regression with heteroscedastic error density and parametric mean function. *Journal of Econometrics*, 178(PART 3):624–638.

Piironen, J. and Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Sansó, B., Pericchi, L. R., and Moreno, E. (1996). *On the robustness of the intrinsic Bayes factor for nested models*, volume Volume 29 of *Lecture Notes–Monograph Series*, pages 155–174. Institute of Mathematical Statistics, Hayward, CA.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Vandaele, W. (1978). Participation in illegitimate activities: Ehrlich revisited. In *Deterrence and Incapacitation*, pages 270–335. U.S. National Academy of Sciences, Washington, DC.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Wooldridge, J. M. (2015). *Introductory econometrics: A modern approach*. Nelson Education.

Appendices

A Correlation tables

A.1 Correlation tables of regressors: Ozone dataset

TABLE A.I: Correlation table, Ozone data: Homoscedasticity (original)

	x4	x5	x6	x7	x8	x9	x10
x4	1						
x5	-0.22	1					
x6	0.06	0.28	1				
x7	0.79	0.07	0.37	1			
x8	-0.49	0.13	-0.24	-0.48	1		
x9	-0.22	0.37	0.61	0.17	0.13	1	
x10	-0.32	0.02	-0.48	-0.4	0.41	-0.14	1

TABLE A.II: Correlation table, Ozone data: Heteroscedasticity - X_7 ("corrected")

	x4	x5	x6	x7	x8	x9	x10
x4	1						
x5	0.62	1					
x6	0.58	0.52	1				
x7	0.90	0.43	0.44	1			
x8	0.69	0.57	0.40	0.62	1		
x9	0.09	0.36	0.58	-0.03	0.27	1	
x10	0.64	0.49	0.17	0.61	0.6	-0.01	1

TABLE A.III: Correlation table, Ozone data: Heteroscedasticity - X_6 ("corrected")

	x4	x5	x6	x7	x8	x9	x10
x4	1						
x5	0.44	1					
x6	-0.74	-0.25	1				
x7	0.89	0.36	-0.67	1			
x8	0.65	0.42	-0.55	0.48	1		
x9	-0.76	-0.15	0.75	-0.68	-0.37	1	
x10	0.79	0.49	-0.61	0.65	0.65	-0.53	1

A.2 Correlation tables of regressors: Crime dataset

TABLE A.IV: Correlation table, Crime data: Homoscedasticity (original)

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time
M	1														
So	0.58	1													
Ed	-0.53	-0.7	1												
Po1	-0.51	-0.37	0.48	1											
Po2	-0.51	-0.38	0.50	0.99	1										
LF	-0.16	-0.51	0.56	0.12	0.11	1									
M.F	-0.03	-0.31	0.44	0.03	0.02	0.51	1								
Pop	-0.28	-0.05	-0.02	0.53	0.51	-0.12	-0.41	1							
NW	0.59	0.77	-0.66	-0.21	-0.22	-0.34	-0.33	0.10	1						
U1	-0.22	-0.17	0.02	-0.04	-0.05	-0.23	0.35	-0.04	-0.16	1					
U2	-0.24	0.07	-0.22	0.19	0.17	-0.42	-0.02	0.27	0.08	0.75	1				
GDP	-0.67	-0.64	0.74	0.79	0.79	0.29	0.18	0.31	-0.59	0.04	0.09	1			
Ineq	0.64	0.74	-0.77	-0.63	-0.65	-0.27	-0.17	-0.13	0.68	-0.06	0.02	-0.88	1		
Prob	0.36	0.53	-0.39	-0.47	-0.47	-0.25	-0.05	-0.35	0.43	-0.01	-0.06	-0.56	0.47	1	
Time	0.11	0.07	-0.25	0.1	0.08	-0.12	-0.43	0.46	0.23	-0.17	0.1	0	0.1	-0.44	1

TABLE A.V: Correlation table, Crime data: Heteroscedasticity - exponential ("corrected")

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time
M	1														
So	0.46	1													
Ed	0.91	0.16	1												
Po1	0.67	0.01	0.83	1											
Po2	0.67	0.01	0.83	1	1										
LF	0.95	0.25	0.98	0.76	0.76	1									
M.F	0.97	0.34	0.97	0.76	0.76	0.99	1								
Pop	0.31	-0.01	0.39	0.63	0.63	0.36	0.35	1							
NW	0.61	0.83	0.32	0.2	0.2	0.42	0.49	0.19	1						
U1	0.89	0.32	0.87	0.67	0.67	0.88	0.92	0.34	0.49	1					
U2	0.82	0.42	0.77	0.69	0.69	0.79	0.84	0.46	0.57	0.92	1				
GDP	0.78	0.03	0.94	0.92	0.92	0.89	0.88	0.49	0.19	0.79	0.74	1			
Ineq	0.93	0.64	0.74	0.45	0.45	0.83	0.87	0.25	0.75	0.81	0.78	0.56	1		
Prob	0.69	0.70	0.47	0.16	0.17	0.56	0.61	-0.06	0.69	0.59	0.52	0.29	0.77	1	
Time	0.81	0.27	0.77	0.66	0.66	0.82	0.81	0.57	0.51	0.7	0.73	0.73	0.76	0.36	1

TABLE A.VI: Correlation table, Crime data: Heteroscedasticity - two values ("corrected")

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time
M	1														
So	0.38	1													
Ed	0.94	0.11	1												
Po1	0.72	-0.04	0.83	1											
Po2	0.72	-0.04	0.84	1	1										
LF	0.97	0.2	0.98	0.8	0.8	1									
M.F	0.98	0.23	0.98	0.79	0.8	0.99	1								
Pop	0.3	0.16	0.32	0.55	0.54	0.32	0.32	1							
NW	0.46	0.77	0.2	0.17	0.16	0.32	0.31	0.30	1						
U1	0.87	0.12	0.91	0.73	0.73	0.9	0.93	0.31	0.16	1					
U2	0.83	0.18	0.85	0.76	0.75	0.84	0.88	0.42	0.23	0.96	1				
GDP	0.89	0.04	0.98	0.9	0.91	0.95	0.95	0.41	0.15	0.89	0.86	1			
Ineq	0.95	0.54	0.84	0.59	0.59	0.9	0.92	0.3	0.54	0.83	0.79	0.77	1		
Prob	0.76	0.41	0.70	0.38	0.39	0.73	0.75	0.08	0.38	0.66	0.60	0.60	0.80	1	
Time	0.87	0.33	0.81	0.75	0.74	0.83	0.84	0.54	0.43	0.76	0.80	0.83	0.80	0.44	1

TABLE A.VII: Correlation table, Crime data: Heteroscedasticity - Pop ("corrected")

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time
M	1														
So	0.19	1													
Ed	0.82	-0.22	1												
Po1	0.37	-0.29	0.64	1											
Po2	0.38	-0.29	0.65	0.99	1										
LF	0.9	-0.1	0.95	0.54	0.55	1									
M.F	0.94	-0.01	0.94	0.53	0.54	0.97	1								
Pop	-0.87	0.1	-0.91	-0.5	-0.51	-0.91	-0.94	1							
NW	0.22	0.77	-0.20	-0.19	-0.20	-0.05	0	0.09	1						
U1	0.76	-0.04	0.77	0.39	0.4	0.76	0.85	-0.84	-0.05	1					
U2	0.61	0.16	0.56	0.4	0.41	0.56	0.69	-0.65	0.09	0.87	1				
GDP	0.66	-0.31	0.92	0.8	0.82	0.83	0.83	-0.8	-0.33	0.68	0.54	1			
Ineq	0.84	0.48	0.52	0.02	0.01	0.68	0.73	-0.62	0.47	0.63	0.57	0.27	1		
Prob	0.60	0.43	0.42	0.01	0.02	0.48	0.55	-0.49	0.38	0.48	0.43	0.22	0.66	1	
Time	0.61	0.15	0.41	0.20	0.18	0.49	0.50	-0.43	0.22	0.34	0.34	0.35	0.51	0.05	1

TABLE A.VIII: Correlation table, Crime data: Heteroscedasticity - GDP ("corrected")

	M	So	Ed	Po1	Po2	LF	M.F	Pop	NW	U1	U2	GDP	Ineq	Prob	Time
M	1														
So	0.86	1													
Ed	0.94	0.74	1												
Po1	0.50	0.59	0.45	1											
Po2	0.51	0.59	0.45	0.99	1										
LF	0.97	0.78	0.97	0.46	0.46	1									
M.F	0.98	0.82	0.97	0.48	0.48	0.98	1								
Pop	0.18	0.36	0.07	0.49	0.46	0.12	0.14	1							
NW	0.85	0.85	0.74	0.63	0.64	0.78	0.80	0.34	1						
U1	0.85	0.69	0.86	0.37	0.37	0.84	0.90	0.07	0.68	1					
U2	0.82	0.77	0.77	0.48	0.47	0.77	0.84	0.23	0.71	0.94	1				
GDP	0.95	0.77	0.95	0.45	0.45	0.95	0.96	0.13	0.69	0.86	0.81	1			
Ineq	0.98	0.88	0.94	0.52	0.52	0.96	0.98	0.23	0.85	0.88	0.86	0.94	1		
Prob	0.89	0.79	0.89	0.39	0.40	0.86	0.90	0.10	0.79	0.82	0.78	0.83	0.89	1	
Time	0.88	0.80	0.79	0.51	0.49	0.83	0.84	0.35	0.86	0.73	0.75	0.78	0.88	0.73	1

B Heteroscedasticity tests ("corrected" datasets)

TABLE B.I: Percentage of rejections in heteroscedasticity tests: Ozone ("corrected")

	Breusch-Pagan	White
SNR = 10		
Homoscedasticity	4.48	4.62
Heteroscedasticity - x7	3.78	5.33
Heteroscedasticity - x6	4.56	4.56
SNR = 1		
Homoscedasticity	4.73	5.16
Heteroscedasticity - x7	4.10	5.18
Heteroscedasticity - x6	4.38	4.48

TABLE B.II: Percentage of rejections in heteroscedasticity tests: Crime ("corrected")

	Breusch-Pagan	White
SNR = 10		
Homoscedasticity	1.63	4.67
Heteroscedasticity - exponential	0.96	2.09
Heteroscedasticity - two values	0.72	1.67
Heteroscedasticity - Pop	1.73	3.81
Heteroscedasticity - GDP	1.39	4.72
SNR = 1		
Homoscedasticity	1.44	4.31
Heteroscedasticity - exponential	1.10	2.19
Heteroscedasticity - two values	0.54	1.66
Heteroscedasticity - Pop	1.93	3.32
Heteroscedasticity - GDP	1.31	3.86

C R Code

The R code used for all calculations can be found at `https://github.com/hugofvm/MFW_R_Code`.