Bayesian Model Selection under Noise - From Statistical to Practical Significance

Ph.D. Thesis

Andrade Silva Daniel Georg
SOKENDAI (The Graduate University for Advanced Studies)

Abstract

In this thesis, we consider the problem of model selection, and in particular the situation where modeling assumptions regarding exact zero partial correlations or exact zero regression coefficients are violated. For the purpose of getting scientific insights, one is often interested in model selection, i.e. choosing among two or more candidate models. In this thesis, we employ the marginal likelihood for Bayesian model selection. The marginal likelihood can identify the most plausible model, or a subset of plausible models, and this way can help the data analyst to gain new insights. One advantage of the marginal likelihood is that it incorporates a model complexity penalty that helps to prefer simpler models over complex models. In general, due to the ease of interpretation, models with low complexity (small dimensional parameter space) are preferred over complex ones. However, with small noise on the correlation between variables, or small noise on linear regression coefficients, more complex models tend to be selected independent of the effect size. This problem is especially pronounced by large sample sizes. In this thesis, we address this problem by carefully designing priors that absorb overly complex models with hyper-parameters that control the desirable effect size. In particular, we address the problem of clustering variables in the Gaussian graphical model (Chapter 2) and variable selection in linear regression (Chapter 3) under such small negligible noise.

In Chapter 2, we address the problem of clustering variables in the Gaussian graphical model. Variable clustering is important for explanatory analysis. However, only few dedicated methods for variable clustering with the Gaussian graphical model have been proposed. Even more severe, small insignificant partial correlations due to noise can dramatically change the clustering result when evaluating for example with the Bayesian information criterion (BIC). We address this issue by proposing a Bayesian model that accounts for negligible small, but not necessarily zero, partial correlations. To address the intractable calculation of the marginal likelihood, we propose two solutions: one based on a variational approximation, and another based on Markov Chain Monte Carlo (MCMC). Our variational approximation is based on a convex optimization problem for finding the maximum a-posterior (MAP) estimate and a low-dimensional non-convex optimization problem for identifying the variance around the MAP. Although, the former is a convex optimization problem, the high dimension and positive-definite constraint on the precision matrix are challenging for standard convex solvers. Therefore, we adapt a recently proposed 3-block alternating direction method of multipliers (ADMM) to our problem, which proves to be numerically stable and sufficiently precise. Experiments on simulated data shows that, in the no-noise setting, our proposed performs similar accurate to BIC in identifying the correct clusters, but is considerably more accurate when there are noisy partial correlations. Furthermore, on real data the proposed method provides clustering results that are intuitively sensible, which is not always the case when using BIC and its extensions. Experimentally, we also confirm that the variational approximation is considerably faster than MCMC while leading to similar accurate model selection.

In Chapter 3, we extend some of the ideas from Chapter 2 to variable selection under noise in linear regression. Sparseness of the regression coefficient vector is often a desirable property, since, among other benefits, sparseness improves interpretability. Therefore, in practice, we may want to trade in a small reduction in prediction accuracy for an increase in sparseness. The work in (Chipman et al., 2001) introduces two spike-and-slab priors that can potentially handle such a trade-off between prediction accuracy and sparseness. For that purpose, they introduce a threshold δ on the magnitude of each regression coefficient. Their first spike-and-slab model

couples the response variance with the variance on the regression coefficients leading to a closed-form analytic solution. However, as a result, their method is sensitive to the prior setting of the response variance and cannot guarantee anymore that the true model is selected. Their second spike-and-slab prior model solves the latter issue, but at the cost of losing conjugacy. Another subtle issue common to these spike-and-slab priors is that they lead to inconsistent Bayes factors. Due to the fact that their spike-and-slab priors have full support, the Bayes factors of any two models is bounded in probability for increasingly large sample sizes. This is an undesirable property for Bayesian hypotheses testing. Our proposed model decouples the response noise prior variance from the regression coefficients' prior variance, and thus makes the threshold parameter δ more meaningful than previous work. For example, δ can be set such that the Mean-Squared Error (MSE) of the prediction is only little influenced by ignoring covariates with coefficients' magnitude smaller than δ. In case where the specification of δ is difficult, we show that automatic selection of δ via the estimation of MSE can be a viable choice. Furthermore, by using disjunct support priors, our method guarantees consistent Bayes factors in the sense that the ratio of the true model's marginal likelihood to any other models' marginal likelihood converges to infinity for increasingly large sample sizes. Due to the non-conjugacy of the priors proposed by our method, estimating the marginal likelihood explicitly is computationally infeasible. Instead, we propose to estimate all model probabilities by introducing a latent variable indicator vector and sampling from its posterior distribution with an efficient Gibbs sampler. On several synthetic data sets, we evaluate our proposed method in terms of the ability to identify the true model. Here, we define the true model as the one correctly separating all variables into two sets S and C, where S contains all variables that have non-negligible regression coefficients, and C contains all remaining variables. We compare our method to the spike-and-slab priors as in EMVS (Ročková and George, 2014), (Bayarri et al., 2012), thresholding the mean regression coefficient vector of an horseshoe prior (Carvalho et al., 2010), and (penalized) maximum likelihood estimation combined with stability selection, AIC, BIC, and its extensions. In various settings: with/without noise and low/high dimensions the proposed method leads to consistently good model selection performance, which was not the case for any other baseline method. Finally, we evaluated our method also on three real data sets. Concerning the number of selected variables of our proposed method and all previous methods, we observe a similar behavior as for the synthetic data set. Furthermore, for $\delta = 0$, our proposed method seems to roughly agree with various previous methods, while the inspection of the results for $\delta = 0.5$, allows us to draw conclusions about the practical relevance of some of the selected variables.

Acknowledgment

First of all, I would like to thank Prof. Kenji Fukumizu for the many valuable discussions about several potential research topics and his advise on this work. I am also grateful to Prof. Akiko Takeda for her advise on convex optimization and in particular for her suggestion of a 3-block ADMM. I also grateful to the other examination committee members, Prof. Hideitsu Hino, Prof. Satoshi Kuriki, and Prof. Makoto Yamada for their thoughtful comments. Furthermore, I would like thank Prof. Daichi Mochihashi for his suggestion of comparing the variational approximation with MCMC methods. I also thankful for the administrative support by Ms. Akatsuka, Ms. Itsumi, and Ms. Matsukawa. I am also very grateful to NEC for their generous support of the PhD., in particular during the first two years as part of NEC's domestic study program. Finally, I am in debt to my wife Tomono and my son Yoshiya for their encouragement and time.

Contents

1 Introduction **3**
- 1.1 Frequentist hypothesis testing and model selection 4
- 1.2 Bayesian hypothesis testing and model selection 5
  - 1.2.1 Marginal likelihood 7
  - 1.2.2 Other Bayesian approaches to model selection 9

2 Robust Bayesian Model Selection for Variable Clustering with the Gaussian Graphical Model **11**
- 2.1 Introduction 11
- 2.2 Related work 12
- 2.3 The Bayesian Gaussian graphical model for clustering 13
- 2.4 Proposed method 14
  - 2.4.1 A Bayesian Gaussian graphical model for clustering under noisy conditions 14
  - 2.4.2 Estimation of the marginal likelihood 15
  - 2.4.3 Restricting the hypotheses space 24
- 2.5 Simulation study 26
  - 2.5.1 Evaluation of the restricted hypotheses space 27
  - 2.5.2 Evaluation of clustering selection criteria 27
  - 2.5.3 Comparison of variational and MCMC estimate 29
- 2.6 Real data experiments 29
  - 2.6.1 Mutual funds 29
  - 2.6.2 Gene regulations 30
  - 2.6.3 Aviation sensors 30
- 2.7 Discussion and conclusions 31

3 Disjunct Support Prior for Variable Selection in Regression **49**
- 3.1 Introduction 49
- 3.2 Related work 50
- 3.3 Proposed method 51
- 3.4 Asymptotic Bayes factors 54
- 3.5 Estimation of model probabilities 59
  - 3.5.1 Analytic solution for $p(z_j|\boldsymbol{\beta}_{-j}, \mathbf{z}_{-j}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X)$ 60
  - 3.5.2 Analytic solution for $p(\beta_j|\boldsymbol{\beta}_{-j}, \mathbf{z}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X)$ 61

3.5.3 Analytic solution for $p(\sigma_r^2|\boldsymbol{\beta}, \mathbf{z}, \mathbf{y}, X)$ . . . 62
3.5.4 Sampling from $p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{z}, \mathbf{y}, X)$ . . . 62
3.6 Specification of δ . . . 63
3.6.1 Bounding influence on Mean Squared Error . . . 64
3.6.2 Estimating expected increase of Mean Squared Error . . . 64
3.7 Evaluation on synthetic data . . . 65
3.8 Evaluation on real data . . . 68
3.9 Conclusions . . . 70

4 Conclusions and Discussion **95**
4.1 Methods for marginal likelihood estimation . . . 95
4.2 Robustness to small negligible noise . . . 96

Appendices

Appendix A Variable Clustering in the Gaussian Graphical Model **101**
A.1 Convergence of 3-block ADMM . . . 101
A.2 Derivation of variational approximation . . . 102
A.3 Spectral clustering for variable clustering . . . 105

Appendix B Disjunct Support Prior for Variable Selection in Regression **107**
B.1 Slice sampler . . . 107
B.2 Asymptotic approximation of $p(\mathbf{y}_n|X_n, S)$ . . . 108

Bibliography **111**

Chapter 1

Introduction

In this thesis, we consider the problem of model selection, and in particular the situation where modeling assumptions regarding exact zero partial correlations or exact zero regression coefficients are violated.

For the purpose of getting scientific insights, one is often interested in choosing among two or more models.[1] In most of the situations, the focus of model selection is one the following three:

- Choice of the likelihood function. (Frequentist and Bayesian modeling)
- Choice of the relevant covariates or interactions. (Frequentist and Bayesian modeling)
- Choice of the prior. (Only Bayesian modeling)

The choice of the likelihood function, for example, means to decide between Gaussian noise and Student-t noise in the linear regression model, or the choice between a poisson distribution and a negative binomial distribution to model count data. Though, the choice of the likelihood can impact the final conclusions[2], the choice of the likelihood function is often fixed due to computational convenience or strong prior beliefs about the data generation process.

The focus of this thesis in on the latter two. In Bayesian modeling, the choice of the relevant covariates and the choice of the prior is often intertwined. As a simple example, in linear regression, choosing as prior for a regression coefficient the Dirac measure with point mass at zero, is equivalent to excluding the corresponding covariate. The priors used in this thesis are of this type of nature: deciding the prior for the partial correlation (Chapter 2) or regression coefficient (Chapter 3) will correspond to the variable clustering (Chapter 2) and choice of relevant covariates (Chapter 3), respectively.[3] In frequentist modeling,

[1] If the focus is only on predictive performance, it is in general advisable to pursue model averaging rather than model selection, see e.g. (Piironen and Vehtari, 2017).

[2] For example if the data contains outlier, a student t-distribution can be more robust.

[3] Of course, the choice of prior can also involve more subtle decisions, like setting the scale parameter. However, these are often considered as nuisance parameters, that is parameters which are not of primary interest.

the choice of relevant covariates is restricted to the inclusion of covariates that are used for the likelihood function.

For reasons, which we discuss in Sections 1.1 and 1.2, we prefer the Bayesian paradigm to model selection over the frequentist one. In Bayesian model selection the marginal likelihood is the key quantity. The marginal likelihood can identify the most plausible model, or a subset of plausible models, and this way can help the data analyst to gain new insights. In general, due to the ease of interpretation, models with low complexity (small dimensional parameter space) are preferred over complex ones. The marginal likelihood incorporates a model complexity penalty that helps to prefer simpler models over complex models.

However, with small noise on the partial correlations between variables, or small noise on regression coefficients, more complex models tend to be selected independent of the effect size. This problem is especially pronounced by large sample sizes.

In this thesis, we address this problem by carefully designing priors that absorb overly complex models, and our hyper-parameters control the desirable effect size. In particular, we address the problem of clustering variables in the Gaussian graphical model (Chapter 2) and variable selection in linear regression (Chapter 3) under small negligible noise.

We emphasize that in this thesis, we define robustness to noise as robustness to the strict sparsity assumption. With strict sparsity assumption, we mean the assumption that many partial correlations (Chapter 2) or many regression coefficients (Chapter 3) are *exactly* zero. In that sense, robustness to noise is similar to robustness to model misspecification as in (Miller and Dunson, 2018), and the small negligible noise assumption is also sometimes called quasi-sparseness (Datta and Dunson, 2016).

In the remaining of this chapter, we discuss frequentists and Bayesian methods to model selection with its origins in hypothesis testing. In Section 1.1, we discuss major problems of classical and more recent methods for model selection that are based on the frequentist paradigm. In Section 1.2, we explain how the Bayesian paradigm to model selection can mitigate some of the problems, while also introducing new challenges: the choice of priors and the calculation of the marginal likelihood. The marginal likelihood is key to Bayesian model selection, and is discussed in more detail in Section 1.2.1. Some other methods for Bayesian model selection are discussed in Section 1.2.2.

1.1 Frequentist hypothesis testing and model selection

The problem of selecting between two models can be addressed with classical frequentist hypothesis testing. Traditionally, hypothesis testing defines the null hypothesis H_0 as the baseline, and the alternative hypothesis H_1 as the claim that one is hoping to prove. However, applying classical frequentist hypothesis testing to model selection has several shortcomings (Lopes and Polson, 2018;

Weakliem, 2016; Berger and Delampady, 1987):

(I) Asymmetry. The null hypothesis H_0 can only be rejected. Not rejection of H_0, does not mean acceptance of H_0. In other words, it is not possible to express evidence in favor of H_0. Therefore, the conclusions of hypothesis testing depend on which model is set as H_0 and which one is set as H_1.

(II) Lindley's paradox. Often H_0 is not rejected simply due to the sample size being too small. Conversely, with increasing sample size, H_0 tends to be rejected. In particular, testing for an exact value like $H_0 : \beta = 0$, is prone to be rejected for large sample sizes. In such situations Bayesian hypothesis testing, often leads to a different conclusion, which is sometimes referred to as Lindley's paradox (Tsao, 2006).

(III) Only partial order. Performing a hypothesis test for all pairs of models, does, in general, not lead to a full ranking of all models. In particular, it can happen that testing model A ($= H_0$) against model B ($= H_1$), and model B ($= H_0$) against model A ($= H_1$), are both rejected.

For theses reasons, penalized likelihood methods (Weakliem, 2016) are recently the preferred model selection methods. Penalized likelihood methods like AIC (Akaike, 1973), BIC (Schwarz, 1978) and EBIC (Chen and Chen, 2008) take the following form:

$$-2 \log p(\mathbf{y}|\hat{\boldsymbol{\theta}}) + \text{penalty}(n, d)\,,$$

where $p(\mathbf{y}|\hat{\boldsymbol{\theta}})$ is the probability of the observed data $\mathbf{y}$ given the maximum likelihood estimate $\hat{\boldsymbol{\theta}}$, and penalty$(n, d)$ is a penalty term that can depend on the sample size n and the number of parameter values d. Such a criterion can then be used to rank all methods to identify the best one. However, a major problem is that they require the maximum likelihood estimate which might not be defined, in particular in the setting $d \geq n$ (high-dimensional setting). Therefore, for the high-dimensional setting, a different methodology, like stability selection Meinshausen and Bühlmann (2010), is necessary.

1.2 Bayesian hypothesis testing and model selection

Bayesian methodology resolves some of the issues of frequentist hypothesis testing. In particular, instead of requiring different approaches for different problem settings[4], Bayesian testing and model selection can both be addressed with posterior model probabilities. Therefore, some consider the Bayesian methodology as more coherent (Robert, 2007).

[4]For example, hypothesis testing using p-values, model selection for $d < n$ using AIC, but stability selection for model selection with $d > p$.

The posterior probability for model M is given by the Bayes' rule

$$p(M|\mathbf{y}) = \frac{p(\mathbf{y}|M) \cdot p(M)}{\sum_{M'} p(\mathbf{y}|M') \cdot p(M')} \,. \tag{1.1}$$

For Bayesian hypothesis testing (Kass and Raftery, 1995), we are only interested in the odds of two models H_1 and H_0 after having observed data $\mathbf{y}$:

$$\frac{p(H_1|\mathbf{y})}{p(H_0|\mathbf{y})} = \frac{p(\mathbf{y}|H_1) \cdot p(H_1)}{p(\mathbf{y}|H_0) \cdot p(H_0)} \,. \tag{1.2}$$

Assuming that both models are a-priori equally likely, we see that the right-hand side reduces to just the posterior-odds between the marginal likelihood (see Section 1.2.1) under model H_1 and H_0, which is called the Bayes factor, denoted as B_{10}. The magnitude of the Bayes factor denotes the amount of evidence for H_1 when compared against H_0.

For model selection under $0/1$ loss, the actual magnitude is irrelevant, and we might just select

$$\arg\max_{M'} p(M'|\mathbf{y}) \,,$$

which of course is equivalent to $\arg\max_{M'} p(\mathbf{y}|M')$ for uniform model priors.

Contrasting to the problems of frequentist hypothesis testing, we have

(I) Symmetry. By definition of the Bayes factor, the evidence in favor of H_1 is just the reciprocal of the evidence in favor of H_0.

(II) Lindley's paradox. When testing for an exact value like $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$, the posterior probability now tends to be larger for H_0 than for H_1. In settings where $\beta = 0$ corresponds to a simpler model, this can be considered as a kind of Occam's razor. However, this can also be a problem. In general, for full support priors, the posterior odds for H_0 are still not much larger than 1, even if H_0 is true and the sample size n is large. In other words, even if H_0 is true and n is large, the evidence for H_0 might be small. This problem has been pointed out in (Johnson and Rossell, 2010), and might be resolved using non-local alternative priors. In this example, if H_1 uses a prior $p(\beta)$ such that $p(\beta) = 0$ in some non-empty interval around $\beta = 0$, the prior is called a non-local alternative prior (Johnson and Rossell, 2010). The idea is related to disjunct support priors, which we will introduce in Chapter 3.

(III) Total order. Every pair of model can be compared, and, in general, with probability 1, we have either $p(H_1|\mathbf{y}) > p(H_0|\mathbf{y})$ or $p(H_1|\mathbf{y}) < p(H_0|\mathbf{y})$.

Therefore, we see that some of the problems in particular (I) and (III) are resolved by the Bayesian paradigm to hypothesis testing (and model selection). However, (II) Lindley's paradox also exemplifies the problem of the dependence on the prior probabilities that need to be defined over all parameters of interest.

Compared to the frequentist paradigm, the $d \geq n$ setting does not require any new methodology, though the final results are more sensitive to the choice of the priors.

The Bayesian paradigm, also faces a computational problem. In Bayesian hypothesis testing and model selection, the key quantity that needs to be calculated is $p(\mathbf{y}|M)$ which is called the marginal likelihood. Except for over-simplified models, calculation of $p(\mathbf{y}|M)$ is, in general, analytically intractable, and computational methods are required.

1.2.1 Marginal likelihood

In this section, we review some of the properties of the marginal likelihood for model selection. For illustration purposes, let us denote by $\mathbf{y}$ all observed data and by $\boldsymbol{\theta}$ all model parameters. Then the marginal likelihood of a model M is defined as

$$p(\mathbf{y}|M) = \int p(\mathbf{y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}\,. \tag{1.3}$$

Furthermore, let us define

$$\boldsymbol{\theta}_{0,M} := \arg\max_{\boldsymbol{\theta}} p(\mathbf{y}|\boldsymbol{\theta}, M)\,.$$

From the definition in (1.3), we see that the marginal likelihood is high if the model explains the data well, i.e. high likelihood $p(\mathbf{y}|\boldsymbol{\theta}_{0,M}, M)$. Moreover, assuming a proper prior, the marginal likelihood also punishes model complexity, as we show in the following. We assume a proper prior, i.e.

$$1 = \int p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}\,. \tag{1.4}$$

Furthermore, let us assume two nested models M_1 and M_2, where M_2 contains model M_1, i.e. the parameter space of M_2 is larger than M_1, but $p(\mathbf{y}|\boldsymbol{\theta}_{0,M_1}, M_1) = p(\mathbf{y}|\boldsymbol{\theta}_{0,M_2}, M_2)$. Furthermore, let us assume factorized priors of the form

$$p(\boldsymbol{\theta}|M_1) = \prod_{i=1}^{d_1} p(\theta_i)\,,$$

and

$$p(\boldsymbol{\theta}|M_2) = \prod_{i=1}^{d_2} p(\theta_i)\,,$$

where $d_2 > d_1$. For illustration, let us also assume that the likelihood is roughly constant in the vicinity of some non-empty open set A_1 around $\boldsymbol{\theta}_{0,M_1}$ and A_2 around $\boldsymbol{\theta}_{0,M_2}$, i.e.

$$\forall\, \boldsymbol{\theta}_1 \in A_1,\, \boldsymbol{\theta}_2 \in A_2 :\; p(\mathbf{y}|\boldsymbol{\theta}_1, M_1) = p(\mathbf{y}|\boldsymbol{\theta}_2, M_2) \approx m\,,$$

and negligible in all other parts of the parameter space. Although, these are strong assumptions, they are actual guaranteed for large sample size n due to the Bayesian Central limit theorem (subject to some smoothness conditions on the likelihood and prior, see e.g. Ando (2010)). Then, the marginal likelihood for M_1 is given by

$$\begin{aligned} p(\mathbf{y}|M_1) &= \int p(\mathbf{y}|\boldsymbol{\theta}, M_1)p(\boldsymbol{\theta}|M_1)d\boldsymbol{\theta} \\ &\approx \int_{A_1} p(\mathbf{y}|\boldsymbol{\theta}, M_1)p(\boldsymbol{\theta}|M_1)d\boldsymbol{\theta} \\ &\approx m \cdot \int_{A_1} p(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1 \,. \end{aligned}$$

Analogously, we have

$$\begin{aligned} p(\mathbf{y}|M_2) &\approx m \cdot \int_{A_2} p(\boldsymbol{\theta})d\boldsymbol{\theta} \\ &= m \cdot \Big(\int_{A_1} p(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1 \Big) \cdot \Big(\int_{A_3} p(\boldsymbol{\theta}_3)d\boldsymbol{\theta}_3 \Big) \,, \end{aligned}$$

for some set A_3 such that $A_2 = A_1 \times A_3$. Due to Equation (1.4), we have, in general, that $\int_{A_3} p(\boldsymbol{\theta}_3)d\boldsymbol{\theta}_3 < 1$ and therefore this implies that $p(\mathbf{y}|M_2) < p(\mathbf{y}|M_1)$. In plain words, the overly complex model M_2 has lower prior probability around the optimal parameters than model M_1, and therefore the marginal likelihood is smaller for M_2 than for M_1. This property can also be expressed in more general terms, and is often referred to as the Bayesian Occam's razor (Barber, 2012).

Even if there is nothing such that a true model, the marginal likelihood can also indicate performance on held-out data (Kass and Raftery, 1995; van der Wilk et al., 2018):

$$p(\mathbf{y}|M) = p(y_n|y_1 \ldots, y_{n-1}, M) \cdot p(y_{n-1}|y_1 \ldots, y_{n-2}, M) \cdot \ldots \cdot p(y_2|y_1, M) \cdot p(y_1|M) \,,$$

showing that the marginal likelihood will be high, if the model trained on one part of the data predicts well other samples from the data.

Remark We remark that the above example of factorized priors is somehow simplified. However, in many situations where there is no knowledge a-priori, this is a reasonable and common choice. Indeed, for our proposed models in Chapter 2 and 3, we will employ such factorized priors. Finally, we remark that the fractional Bayes factor approach allows to generalize the concept of the marginal likelihood to *improper* priors, e.g. (O'Hagan, 1995). Though, in this thesis, all of our models described in Chapter 2 and 3 employ weakly informative *proper* priors.

1.2.2 Other Bayesian approaches to model selection

Cross-validation and related predictive evaluation criteria

Gelman et al. (2013) points out that the marginal likelihood can be very sensitive to the choice of the prior, and therefore recommends the use of predictive evaluation criteria. In particular, they recommend cross-validation (CV) and less computationally expensive alternatives like the Akaike information criterion (AIC) (Akaike, 1973), and the Watanabe-Akaike information criterion (WAIC) (Watanabe, 2013). If increasing prediction accuracy is the sole objective, we agree that cross-validation can be a better choice for model selection than the usage of the marginal likelihood. However, cross-validation tends to prefer overly complex models. For example, in linear regression, the horseshoe prior (Carvalho et al., 2010) provides excellent performance in terms of mean-squared error on held-out test data. However such modeling includes *all* variables and as such does not provide any decision on the set of relevant and not-relevant variables. Although our main focus is on model selection, we note that there is also some empirical evidence that the marginal likelihood approach performs better than CV and WAIC, in terms of selecting a sparse model with good predictive performance (Piironen and Vehtari, 2017).

C-posterior

Instead of using model posterior probabilities (or the marginal likelihood), Miller and Dunson (2018) proposes to use power posteriors for model selection. Power posteriors are similar to ordinary posterior distributions, but with the difference that the likelihood was raised to a power c. As such, power posteriors *cannot* be interpreted as probabilities (or probability density functions) anymore. Their motivation is similar to ours: making model selection robust to small deviations from the model assumptions. One advantage of their method is that it is model agnostic, in the sense that it can be used with any existing Bayesian model and no explicit noise model needs to be defined. On the contrary, the hyper-parameter c which provides a trade-off between model complexity and model fit, is difficult to interpret, and needs to be manually calibrated. Furthermore, their method does not allow anymore the computation of a marginal likelihood, but needs to resort to specialized MCMC methods.

We note that our proposed methods in Chapter 2 and Chapter 3, are very different from the approach taken by c-posteriors: we define explicit noise models with priors that control the degree of misspecification that we are willing to tolerate.

Chapter 2

Robust Bayesian Model Selection for Variable Clustering with the Gaussian Graphical Model

2.1 Introduction

The Gaussian graphical model (GGM) has become an invaluable tool for detecting partial correlations between variables. Assuming the variables are jointly drawn from a multivariate normal distribution, the sparsity pattern of the precision matrix reveals which pairs of variables are independent given all other variables (Anderson, 2004). In particular, we can find clusters of variables that are mutually independent, by grouping the variables according their entries in the precision matrix.

For example, in gene expression analysis, variable clustering is often considered to be helpful for data exploration (Palla et al., 2012; Tan et al., 2015).

However, in practice, it can be difficult to find a meaningful clustering due to the noise of the entries in the partial correlations. The noise can be due to the sampling, this is in particular the case when n the number of observations is small, or due to small non-zero partial correlations in the true precision matrix that might be considered as insignificant. Here in this work, we are particularly interested in the latter type of noise. In the extreme, small partial correlations might lead to a connected graph of variables, where no grouping of variables can be identified. For an exploratory analysis such a result might not be desirable.

As an alternative, we propose to cluster variables, such that the partial correlation between any two variables in different clusters is negligibly small, but not necessarily zero. The open question, which we try to address here, is

whether there is a principled model selection criteria for this scenario.

For example, the Bayesian Information Criteria (BIC) (Schwarz, 1978) is a popular model selection criteria for the Gaussian graphical model. However, in the noise setting it does not have any formal guarantees. As a solution, we propose here a Bayesian model that explicitly accounts for small partial correlations between variables in different clusters.

Under our proposed model, the marginal likelihood of the data can then be used to identify the correct (if there is a ground truth in theory), or at least a meaningful clustering (in practice) that helps analysis. Since the marginal likelihood of our model does not have an analytic solution, we provide two approximations. The first is a variational approximation, the second is based on MCMC.

Experiments on simulated data show that the proposed method is similarly accurate as BIC in the no noise setting, but considerably more accurate when there are noisy partial correlations. The proposed method also compares favorably to two previously proposed methods for variable clustering and model selection, namely the Clustered Graphical Lasso (CGL) (Tan et al., 2015) and the Dirichlet Process Variable Clustering (DPVC) (Palla et al., 2012) method.

Our paper is organized as follows. In Section 2.2, we discuss previous works related to variable clustering and model selection. In Section 2.3, we introduce a basic Bayesian model for evaluating variable clusterings, which we then extend in Section 2.4.1 to handle noise on the precision matrix. For the proposed model, the calculation of the marginal likelihood is infeasible and we describe two approximation strategies in Section 2.4.2. Furthermore, since enumerating all possible clusterings is also intractable, we describe in Section 2.4.3 an heuristic based on spectral clustering to limit the number of candidate clusterings. We evaluate the proposed method on synthetic and real data in Sections 2.5 and 2.6, respectively. Finally, we discuss our findings in Section 2.7.

2.2 Related work

Finding a clustering of variables is equivalent to finding an appropriate block structure of the covariance matrix. Recently, Tan et al. (2015) and Devijver and Gallopin (2018) suggested to detect block diagonal structure by thresholding the absolute values of the covariance matrix. Their methods perform model selection using the mean squared error of randomly left out elements of the covariance matrix (Tan et al., 2015), and a slope heuristic (Devijver and Gallopin, 2018).

Also several Bayesian latent variable models have been proposed for this task (Marlin and Murphy, 2009; Sun et al., 2014; Palla et al., 2012). Each clustering, including the number of clusters, is either evaluated using the variational lower bound (Marlin and Murphy, 2009), or by placing a Dirichlet Process prior over clusterings (Palla et al., 2012; Sun et al., 2014). However, all of the above methods assume that the partial correlations of variables across clusters are exactly zero.

An exception is the work in (Marlin et al., 2009) which proposes to regularize

the precision matrix such that partial correlations of variables that belong to the same cluster are penalized less than those belonging to different clusters. For that purpose they introduce three hyper-parameters, λ_1 (for within cluster penalty), λ_0 (for across clusters), with $\lambda_0 > \lambda_1$, and λ_D for a penalty of the diagonal elements. The clusters do not need to be known a-priori and are estimated by optimizing a lower bound on the marginal likelihood. As such their method can also find variable clusterings, even when the true partial correlation of variables in different clusters is not exactly zero. However, the clustering result is influenced by three hyperparameters λ_0, λ_1, and λ_D which have to be determined using cross-validation.

Recently, the work in (Sun et al., 2015; Hosseini and Lee, 2016) relaxes the assumption of a clean block structure by allowing some variables to correspond to two clusters. The model selection issue, in particular, determining the number of clusters, is either addressed with some heuristics (Sun et al., 2015) or cross-validation (Hosseini and Lee, 2016).

2.3 The Bayesian Gaussian graphical model for clustering

Our starting point for variable clustering is the following Bayesian Gaussian graphical model. Let us denote by d the number of variables, and n the number of observations. We assume that each observation $\mathbf{x} \in \mathbb{R}^d$ is generated i.i.d. from a multivariate normal distribution with zero mean and covariance matrix Σ. Assuming that there are k groups of variables that are mutually independent, we know that, after appropriate permutation of the variables, Σ has the following block structure

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_k \end{pmatrix},$$

where $\Sigma_j \in \mathbb{R}^{d_j \times d_j}$, and d_j is the number of variables in cluster j.

By placing an inverse Wishart prior over each block Σ_j, we arrive at the following Bayesian model

$$\begin{aligned} &p(\mathbf{x}_1, ..., \mathbf{x}_n, \Sigma | \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ &= \prod_{i=1}^{n} \text{Normal}(\mathbf{x}_i | \mathbf{0}, \Sigma) \prod_{j=1}^{k} \text{InvW}(\Sigma_j | \nu_j, \Sigma_{j,0}), \end{aligned} \tag{2.1}$$

where ν_j and $\Sigma_{j,0}$, are the degrees of freedom and the scale matrix, respectively. We set $\nu_j = d_j + 1, \Sigma_j = I_{d_j}$ leading to a non-informative prior on Σ_j. $\mathcal{C}$ denotes the variable clustering which imposes the block structure on Σ. We will refer to this model as the basic inverse Wishart prior model.

Assuming we are given a set of possible variable clusterings $\mathscr{C}$, we can then choose the clustering $\hat{\mathcal{C}}$ that maximizes the posterior probability of the clustering,

i.e.

$$\hat{\mathcal{C}} = \arg\max_{\mathcal{C}\in\mathscr{C}} p(\mathcal{C}|\mathscr{X}) = \arg\max_{\mathcal{C}\in\mathscr{C}} p(\mathscr{X}|\mathcal{C}) \cdot p(\mathcal{C}) \,, \tag{2.2}$$

where we denote by $\mathscr{X}$ the observations $\mathbf{x}_1, ..., \mathbf{x}_n$, and $p(\mathcal{C})$ is a prior over the clusterings which we assume to be uniform. Here, we refer to $p(\mathscr{X}|\mathcal{C})$ as the marginal likelihood (given the clustering). For the basic inverse Wishart prior model the marginal likelihood can be calculated analytically, see e.g. (Lenkoski and Dobra, 2011).

2.4 Proposed method

In this section, we introduce our proposed method for finding variable clusters.

First, in Section 2.4.1, we extend the basic inverse Wishart prior model from Equation (2.1) in order to account for non-zero partial correlations between variables in different clusters. Given the proposed model, the marginal likelihood $p(\mathscr{X}|\mathcal{C})$ does not have a closed form solution anymore. Therefore, in Sections 2.4.2 and 2.4.2, we discuss two different methods for approximating the marginal likelihood. The first method is based on a variational approximation around the maximum a posteriori (MAP) solution. The second method is an MCMC method based on Chib's method (Chib, 1995; Chib and Jeliazkov, 2001). The latter has the advantage of being asymptotically correct for large number of posterior samples, but at considerably high computational costs. The former is considerably faster to evaluate and experimentally produces solutions similar to the MCMC method (see comparison in Section 2.5.3).

Finally, in Section 2.4.3, we propose to use a spectral clustering method to limit the clustering candidates to a set $\mathscr{C}^*$, where $\mathscr{C}^* \subseteq \mathscr{C}$. Based on this subset $\mathscr{C}^*$, we can then select the model maximizing the posterior probability (as in Equation (2.2)), or can also calculate approximate posterior distributions over clusterings. We restrict the hypotheses space to $\mathscr{C}^*$, since even for a moderate number of variables, say $d = 40$, the size of the hypotheses space $|\mathscr{C}|$ is $> 10^{36}$. Therefore, MCMC sampling over the hypotheses space could also only explore a small subset of the whole hypotheses space, but at higher computational costs (see also Hans et al. (2007); Scott and Carvalho (2008) for a discussion on related high-dimensional problems).

2.4.1 A Bayesian Gaussian graphical model for clustering under noisy conditions

In this section, we extend the Bayesian model from Equation (2.1) to account for non-zero partial correlations between variables in different clusters. For that purpose we introduce the matrix $\Sigma_\epsilon \in \mathbb{R}^{d\times d}$ that models the noise on the

precision matrix. The full joint probability of our model is given as follows:

$$\begin{aligned} & p(\mathbf{x}_1, ..., \mathbf{x}_n, \Sigma, \Sigma_\epsilon | \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ & = \prod_{i=1}^{n} \text{Normal}(\mathbf{x}_i | \mathbf{0}, \Xi) \\ & \quad \cdot \text{InvW}(\Sigma_\epsilon | \nu_\epsilon, \Sigma_{\epsilon,0}) \prod_{j=1}^{k} \text{InvW}(\Sigma_j | \nu_j, \Sigma_{j,0}) \,, \end{aligned} \tag{2.3}$$

where $\Xi := (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}$, and

$$\Sigma := \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Sigma_k \end{pmatrix} .$$

As before, the block structure of Σ is given by the clustering $\mathcal{C}$. The proposed model is the same model as in Equation (2.1), with the main difference that the noise term $\beta\Sigma_\epsilon^{-1}$ is added to the precision matrix of the normal distribution.

$1 \gg \beta > 0$ is a hyper-parameter that is fixed to a small positive value accounting for the degree of noise on the precision matrix. Furthermore, we assume non-informative priors on Σ_j and Σ_ϵ by setting $\nu_j = d_j + 1, \Sigma_j = I_{d_j}$ and $\nu_\epsilon = d + 1, \Sigma_{\epsilon,0} = I_d$.

Remark on the parameterization We note that as an alternative parameterization, we could have defined $\Xi := (\Sigma^{-1} + \Sigma_\epsilon^{-1})^{-1}$, and instead place a prior on Σ_ϵ that encourages Σ_ϵ^{-1} to be small in terms of some matrix norm. For example, we could have set $\Sigma_{\epsilon,0} = \frac{1}{\beta} I_d$. We chose the parameterization $\Xi := (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}$, since it allows us to set β to 0, which recovers the basic inverse Wishart prior model.

2.4.2 Estimation of the marginal likelihood

The marginal likelihood of the data given our proposed model can be expressed as follows:

$$\begin{aligned} & p(\mathbf{x}_1, ..., \mathbf{x}_n | \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ & = \int \text{Normal}(\mathbf{x}_1, ..., \mathbf{x}_n | \mathbf{0}, \Xi) \\ & \quad \cdot \prod_{j=1}^{k} \text{InvW}(\Sigma_j | \nu_j, \Sigma_{j,0}) d(\Sigma_j \succ 0) \\ & \quad \cdot \text{InvW}(\Sigma_\epsilon | \nu_\epsilon, \Sigma_{\epsilon,0}) d(\Sigma_\epsilon \succ 0) \,. \end{aligned}$$

where $\Xi := (\Sigma^{-1} + \beta\Sigma_\epsilon^{-1})^{-1}$.

Clearly, if $\beta = 0$, we recover the basic inverse Wishart prior model, as discussed in Section 2.3, and the marginal likelihood has a closed form solution due

to the conjugacy of the covariance matrix of the Gaussian and the inverse Wishart prior. However, if $\beta > 0$, there is no analytic solution anymore. Therefore, we propose to either use an estimate based on a variational approximation (Section 2.4.2) or on MCMC (Section 2.4.2). Both of our estimates require the calculation of the maximum a posterior (MAP) solution which we explain first in Section 2.4.2.

Remark on BIC type approximation of the marginal likelihood We note that for our proposed model an approximation of the marginal likelihood using BIC is not sensible. To see this, recall that BIC consists of two terms: the data log-likelihood under the model with the maximum likelihood estimate, and a penalty depending on the number of free parameters. The maximum likelihood estimate is

$$\hat{\Sigma}, \hat{\Sigma}_\epsilon = \arg\max_{\Sigma, \Sigma_\epsilon} \sum_{i=1}^{n} \log \text{Normal}(\mathbf{x}_i | \mathbf{0}, (\Sigma^{-1} + \beta \Sigma_\epsilon^{-1})^{-1}) \,,$$

where S is the sample covariance matrix. Note that without the specification of a prior, it is valid that $\hat{\Sigma}, \hat{\Sigma}_\epsilon$ are not positive definite as long as the matrix $\hat{\Sigma}^{-1} + \beta \hat{\Sigma}_\epsilon^{-1}$ is positive definite. Therefore $\hat{\Sigma}^{-1} + \beta \hat{\Sigma}_\epsilon^{-1} = S^{-1}$, and the data likelihood under the model with the maximum likelihood estimate is simply $\sum_{i=1}^{n} \log \text{Normal}(\mathbf{x}_i | \mathbf{0}, S)$, which is independent of the clustering. Furthermore, the number of *free* parameters is $(d^2 - d)/2$ which is also independent of the clustering. That means, for any clustering we end up with the same BIC.

Furthermore, a Laplacian approximation as used in the generalized Bayesian information criterion (Konishi et al., 2004) is also not suitable, since in our case the parameter space is over the positive definite matrices.

Calculation of maximum a posterior solution

Finding the exact MAP is crucial for the quality of the marginal likelihood approximation that we will describe later in Sections 2.4.2 and 2.4.2. In this section, we explain in detail how the corresponding optimization problem can be solved with a 3-block ADMM method, which is guaranteed to converge to the global optimum.

First note that

$$\begin{aligned} & p(\Sigma, \Sigma_\epsilon | \mathbf{x}_1, ..., \mathbf{x}_n, \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) \\ & \propto \text{Normal}(\mathbf{x}_1, ..., \mathbf{x}_n | \mathbf{0}, \Xi) \\ & \quad \cdot \prod_{j=1}^{k} \text{InvW}(\Sigma_j | \nu_j, \Sigma_{j,0}) \\ & \quad \cdot \text{InvW}(\Sigma_\epsilon | \nu_\epsilon, \Sigma_{\epsilon,0}) \end{aligned}$$

where $\Xi := (\Sigma^{-1} + \beta \Sigma_\epsilon^{-1})^{-1}$.

Therefore,

$$\begin{aligned}
&\log p(\Sigma, \Sigma_\epsilon | \mathbf{x}_1, ..., \mathbf{x}_n, \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C}) = \\
&- \frac{n}{2} \log |\Xi| - \frac{n}{2} \text{trace}(S\Xi^{-1}) \\
&- \frac{\nu_\epsilon + d + 1}{2} \log |\Sigma_\epsilon| - \frac{1}{2} \text{trace}(\Sigma_{\epsilon,0}\Sigma_\epsilon^{-1}) \\
&+ \sum_{j=1}^{k} \Big(- \frac{\nu_j + d_j + 1}{2} \log |\Sigma_j| - \frac{1}{2} \text{trace}(\Sigma_{j,0}\Sigma_j^{-1}) \Big) \\
&+ const \\
= &\frac{1}{2} \Big(n \cdot \log |\Xi^{-1}| - n \cdot \text{trace}(S\Xi^{-1}) \\
&+ (\nu_\epsilon + d + 1) \cdot \log |\Sigma_\epsilon^{-1}| - \text{trace}(\Sigma_{\epsilon,0}\Sigma_\epsilon^{-1}) \\
&+ \sum_{j=1}^{k} \Big((\nu_j + d_j + 1) \cdot \log |\Sigma_j^{-1}| - \text{trace}(\Sigma_{j,0}\Sigma_j^{-1}) \Big) \Big) \\
&+ const \,,
\end{aligned}$$

where the constant is with respect to $\Sigma_\epsilon, \Sigma_1, \dots \Sigma_k$, and d_j denotes the number of variables in cluster j.

Solution using a 3-block ADMM Finding the MAP can be formulated as a convex optimization problem by a change of parameterization: by defining $X := \Sigma^{-1}$, $X_j := \Sigma_j^{-1}$, and $X_\epsilon := \Sigma_\epsilon^{-1}$, we get the following convex optimization problem:

$$\begin{aligned}
\underset{X \succ 0, X_\epsilon \succ 0}{\text{minimize}} \; & n \cdot \text{trace}(S(X + \beta X_\epsilon)) - n \cdot \log |X + \beta X_\epsilon| \\
& + \text{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log |X_\epsilon| \\
& + \sum_{j=1}^{k} \Big(\text{trace}(A_j X_j) - a_j \cdot \log |X_j| \Big) \,,
\end{aligned} \tag{2.4}$$

where, for simplifying notation, we introduced the following constants:

$$\begin{aligned}
A_\epsilon &:= \Sigma_{\epsilon,0} \,, \\
a_\epsilon &:= \nu_\epsilon + d + 1 \,, \\
A_j &:= \Sigma_{j,0} \,, \\
a_j &:= \nu_j + d_j + 1 \,.
\end{aligned}$$

From this form, we see immediately that the problem is strictly convex jointly in X_ϵ and X.[1]

[1] Since $-log|X|$ is a strictly convex function and $\text{trace}(XS)$ is a linear function.

We further reformulate the problem by introducing an additional variable Z:

$$\begin{aligned}
&\text{minimize } f(X_\epsilon, X_1, \ldots, X_k, Z) \\
&\text{subject to} \\
&Z = X + \beta X_\epsilon \,, \\
&X_\epsilon, X_1, \ldots, X_k, Z \succeq 0 \,,
\end{aligned}$$

with

$$\begin{aligned}
f(X_\epsilon, X_1, \ldots, X_k, Z) := & \, n \cdot \text{trace}(SZ) - n \cdot \log |Z| \\
& + \text{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log |X_\epsilon| \\
& + \sum_{j=1}^{k} \Big(\text{trace}(A_j X_j) - a_j \cdot \log |X_j| \Big) \,.
\end{aligned}$$

It is tempting to use a 2-Block ADMM algorithm, like e.g. in (Boyd et al., 2011), which leads to two optimization problems: update of X, X_ϵ and update of Z. However, unfortunately, in our case the resulting optimization problem for updating X, X_ϵ does not have an analytic solution. Therefore, instead, we suggest the use of a 3-Block ADMM, which updates the following sequence:

$$\begin{aligned}
X^{t+1} := & \underset{X_1, \ldots, X_k \succ 0}{\arg\min} \sum_{j=1}^{k} \Big(\text{trace}(A_j X_j) - a_j \cdot \log |X_j| \Big) \\
& + \text{trace}(U^t (X + \beta X_\epsilon^t - Z^t)) \\
& + \frac{\rho}{2} ||X + \beta X_\epsilon^t - Z^t||_F^2 \,, \\
X_\epsilon^{t+1} := & \underset{X_\epsilon \succ 0}{\arg\min} \; \text{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log |X_\epsilon| \\
& + \text{trace}(U^t (X^{t+1} + \beta X_\epsilon - Z^t)) \\
& + \frac{\rho}{2} ||X^{t+1} + \beta X_\epsilon - Z^t||_F^2 \,, \\
Z^{t+1} := & \underset{Z \succ 0}{\arg\min} \; n \cdot \text{trace}(SZ) - n \cdot \log |Z| \\
& + \text{trace}(U^t (X^{t+1} + \beta X_\epsilon^{t+1} - Z)) \\
& + \frac{\rho}{2} ||X^{t+1} + \beta X_\epsilon^{t+1} - Z||_F^2 \,, \\
U^{t+1} := & \rho (X^{t+1} + \beta X_\epsilon^{t+1} - Z^{t+1}) + U^t \,,
\end{aligned}$$

where U is the Lagrange multiplier, and X^t, Z^t, U^t, denotes X, Z, U at iteration t; $\rho > 0$ is the learning rate.[2]

Each of the above sub-optimization problem can be solved efficiently via the following strategy. The zero gradient condition for the first optimization problem

[2]In our experiments, we set the learning rate ρ initially to 1.0, and increase it every 100 iterations by a factor of 1.1. We found experimentally that this speeds-up the convergence of ADMM.

with variable X is

$$-X_j^{-1} + \frac{\rho}{a_j} X_j = -\frac{1}{a_j}(A_j + U_j + \rho(\beta X_{\epsilon,j} - Z_j))\,.$$

The zero gradient condition for the second optimization problem with variable X_ϵ is

$$-X_\epsilon^{-1} + \frac{\rho\beta^2}{a_\epsilon} X_\epsilon = -\frac{1}{a_\epsilon}(A_\epsilon + \beta U + \rho\beta(X - Z))\,.$$

The zero gradient condition for the third optimization problem with variable Z is

$$-Z^{-1} + \frac{\rho}{n} Z = \frac{1}{n}(U - nS + \rho(X + \beta X_\epsilon))\,.$$

Each of the above three optimization problem can be solved via an eigenvalue decomposition as follows. We need to solve V such that it satisfies:

$$-V^{-1} + \lambda V = R \;\;\wedge\;\; V \succeq 0$$

Since R is a symmetric matrix (not necessarily positive or negative semi-definite), we have the eigenvalue decomposition:

$$QLQ^T = R\,,$$

where Q is an orthonormal matrix and L is a diagonal matrix with real values. Denoting $Y := Q^T V Q$, we have

$$-Y^{-1} + \lambda Y = L\,, \tag{2.5}$$

Since the solution Y must also be a diagonal matrix, we have $Y_{ij} = 0$, for $j \neq i$, and we must have that

$$-(Y_{ii})^{-1} + \lambda Y_{ii} = L_{ii}\,. \tag{2.6}$$

Then, Equation (2.6) is equivalent to

$$\lambda Y_{ii}^2 - L_{ii} Y_{ii} - 1 = 0\,,$$

and therefore one solution is

$$Y_{ii} = \frac{L_{ii} + \sqrt{L_{ii}^2 + 4\lambda}}{2\lambda}\,.$$

Note that for $\lambda > 0$, we have that $Y_{ii} > 0$. Therefore, we have that the resulting Y solves Equation (2.5) and moreover

$$V = QYQ^T \succ 0\,.$$

That means, we can solve the semi-definite problem with only one eigenvalue decomposition, and therefore is in $O(d^3)$.

Finally, we note that in contrast to the 2-block ADMM, a general 3-block ADMM does not have a convergence guarantee for any $\rho > 0$. However, using a recent result from (Lin et al., 2018), we can show in Appendix A.1 that in our case the conditions for convergence are met for any $\rho > 0$.

Variational approximation of the marginal likelihood

Here we explain our strategy for the calculation of a variational approximation of the marginal likelihood. For simplicity, let $\boldsymbol{\theta}$ denote the vector of all parameters, $\mathscr{X}$ the observed data, and $\boldsymbol{\eta}$ the vector of all hyper-parameters.

Let $\hat{\boldsymbol{\theta}}$ denote the posterior mode. Furthermore, let $g(\boldsymbol{\theta})$ be an approximation of the posterior distribution $p(\boldsymbol{\theta}|\mathscr{X}, \boldsymbol{\eta}, \mathcal{C})$ that is accurate around the mode $\hat{\boldsymbol{\theta}}$.

Then we have

$$\begin{aligned} p(\mathscr{X}|\boldsymbol{\eta}, \mathcal{C}) &= \frac{p(\boldsymbol{\theta}, \mathscr{X}|\boldsymbol{\eta}, \mathcal{C})}{p(\boldsymbol{\theta}|\mathscr{X}, \boldsymbol{\eta}, \mathcal{C})} \\ &= \frac{p(\hat{\boldsymbol{\theta}}, \mathscr{X}|\boldsymbol{\eta}, \mathcal{C})}{p(\hat{\boldsymbol{\theta}}|\mathscr{X}, \boldsymbol{\eta}, \mathcal{C})} \approx \frac{p(\hat{\boldsymbol{\theta}}, \mathscr{X}|\boldsymbol{\eta}, \mathcal{C})}{g(\hat{\boldsymbol{\theta}})} \,. \end{aligned} \tag{2.7}$$

Note that for the Laplace approximation we would use $g(\boldsymbol{\theta}) = N(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, V)$, where V is an appropriate covariance matrix. However, here the posterior $p(\boldsymbol{\theta}|\mathscr{X}, \boldsymbol{\eta}, \mathcal{C})$ is a probability measure over the positive definite matrices and not over $\mathbb{R}^d$, which makes the Laplace approximation inappropriate.

Instead, we suggest to approximate the posterior distribution $p(\Sigma_\epsilon, \Sigma_1, \dots \Sigma_k|\mathbf{x}_1, ..., \mathbf{x}_n, \nu_\epsilon, \Sigma_{\epsilon,0}, \{\nu_j\}_j, \{\Sigma_{j,0}\}_j, \mathcal{C})$ by the factorized distribution

$$g := g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j) \,.$$

We define $g_\epsilon(\Sigma_\epsilon)$ and $g_j(\Sigma_j)$ as follows:

$$g_\epsilon(\Sigma_\epsilon) := \mathrm{InvW}(\Sigma_\epsilon|\nu_{g,\epsilon}, \Sigma_{g,\epsilon}) \,,$$

with

$$\Sigma_{g,\epsilon} := (\nu_{g,\epsilon} + d + 1) \cdot \hat{\Sigma}_\epsilon \,,$$

where $\hat{\Sigma}_\epsilon$ is the mode of the posterior probability $p(\Sigma_\epsilon|\mathscr{X}, \boldsymbol{\eta}, \mathcal{C})$ (as calculated in the previous section). Note that this choice ensures that the mode of g_ϵ is the same as the mode of $p(\Sigma_\epsilon|\mathbf{x}_1, ..., \mathbf{x}_n, \boldsymbol{\eta}, \mathcal{C})$. Analogously, we set

$$g_j(\Sigma_j) := \mathrm{InvW}(\Sigma_j|\nu_{g,j}, \Sigma_{g,j}) \,,$$

with

$$\Sigma_{g,j} := (\nu_{g,j} + d_j + 1) \cdot \hat{\Sigma}_j \,,$$

where $\hat{\Sigma}_j$ is the mode of the posterior probability $p(\Sigma_j|\mathscr{X}, \boldsymbol{\eta}, \mathcal{C})$. The remaining parameters $\nu_{g,\epsilon} \in \mathbb{R}$ and $\nu_{g,j} \in \mathbb{R}$ are optimized by minimizing the KL-divergence between the the factorized distribution g and the posterior distribution $p(\Sigma_\epsilon, \Sigma_1, \dots \Sigma_k|\mathbf{x}_1, ..., \mathbf{x}_n, \boldsymbol{\eta}, \mathcal{C})$. The details of the following derivations are

given in Appendix A.2. For simplicity let us denote $g_J := \prod_{j=1}^{k} g_j$, then we have

$$\begin{aligned}
KL(g||p) = &- \int g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j) \\
&\log \frac{p(\Sigma_\epsilon, \Sigma_1, \dots \Sigma_k, \mathbf{x}_1, ..., \mathbf{x}_n | \boldsymbol{\eta}, \mathcal{C})}{g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j)} d\Sigma_\epsilon d\Sigma \\
&+ c \\
= &-\frac{1}{2} n \, \mathbb{E}_{g_J, g_\epsilon}[\log |\Sigma^{-1} + \beta \Sigma_\epsilon^{-1}|] \\
&+ \frac{1}{2}(\nu_\epsilon + d + 1) \, \mathbb{E}_{g_\epsilon}[\log |\Sigma_\epsilon|] \\
&+ \frac{1}{2} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S) \, \mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) \\
&- \mathrm{Entropy}[g_\epsilon] \\
&+ \frac{1}{2} \sum_{j=1}^{k} (\nu_j + d_j + 1) \, \mathbb{E}_{g_j}[\log |\Sigma_j|] \\
&+ \frac{1}{2} \sum_{j=1}^{k} \mathrm{trace}((\Sigma_{j,0} + n S_j) \, \mathbb{E}_{g_j}[\Sigma_j^{-1}]) \\
&- \sum_{j=1}^{k} \mathrm{Entropy}[g_j] + c \,,
\end{aligned}$$

where c is a constant with respect to g_ϵ and g_j. However, the term $E_{g_J, g_\epsilon}[\log |\Sigma^{-1} + \beta \Sigma_\epsilon^{-1}|]$ cannot be solved analytically, therefore we need to resort to some sort of approximation.

We assume that $E_{g_J, g_\epsilon}[\log |\Sigma^{-1} + \beta \Sigma_\epsilon^{-1}|]$
$\approx E_{g_J, g_\epsilon}[\log |\Sigma^{-1}|]$. This way, we get

$$\begin{aligned}
KL(g||p) \approx &\, KL(g_\epsilon \,||\, \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)) \\
&+ \sum_{j=1}^{k} KL(g_j \,||\, \mathrm{InvW}(\nu_j + n, \Sigma_{j,0} + n S_j)) \\
&+ c' \,,
\end{aligned}$$

where we used that

$$\mathbb{E}_{g_J, g_\epsilon}[\log |\Sigma^{-1}|] = -\sum_{j=1}^{k} \mathbb{E}_{g_j}[\log |\Sigma_j|] \,,$$

and c' is a constant with respect to g_ϵ and g_j.

From the above expression, we see that we can optimize the parameters of g_ϵ

and g_j independently from each other. The optimal parameter $\hat{\nu}_{g,\epsilon}$ for g_ϵ is

$$\begin{aligned}\hat{\nu}_{g,\epsilon} &= \arg\min_{\nu_{g,\epsilon}} KL(g_\epsilon \,||\, \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)) \\ &= \arg\min_{\nu_{g,\epsilon}} \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon} + d + 1} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\hat{\Sigma}_\epsilon^{-1}) \\ &\quad - 2\log\Gamma_d(\frac{\nu_{g,\epsilon}}{2}) - \nu_{g,\epsilon} d + d\nu_\epsilon \log(\nu_{g,\epsilon} + d + 1) \\ &\quad + (\nu_{g,\epsilon} - \nu_\epsilon) \sum_{i=1}^{d} \psi\Big(\frac{\nu_{g,\epsilon} - d + i}{2}\Big).\end{aligned}$$

And analogously, we have

$$\begin{aligned}\hat{\nu}_{g,j} = \arg\min_{\nu_{g,j}} & \frac{\nu_{g,j}}{\nu_{g,j} + d_j + 1} \mathrm{trace}((\Sigma_{j,0} + n S_j)\hat{\Sigma}_j^{-1}) \\ & - 2\log\Gamma_{d_j}(\frac{\nu_{g,j}}{2}) - \nu_{g,j} d_j \\ & + d_j(\nu_j + n)\log(\nu_{g,j} + d_j + 1) \\ & + (\nu_{g,j} - \nu_j - n) \sum_{i=1}^{d_j} \psi\Big(\frac{\nu_{g,j} - d_j + i}{2}\Big).\end{aligned}$$

Each is a one dimensional non-convex optimization problem that we solve with Brent's method (Brent, 1971).

Discussion: advantages over full variational approaches We described here an approximation to the marginal likelihood that can be considered as a blending of the ideas of the Laplace approximation (using the MAP) and a variational approximation where *all* parameters are learned by minimizing the Kullback-Leibler divergence between a variational distribution and the true posterior distribution. We refer to the latter as a *full* variational approximation. For simplicity, here, let us denote by Σ the positive definite matrix for which we seek the posterior distribution, and let Σ_g denote the parameter matrix of the variational distribution.

An obvious limitation of the full variational approach is that the expectation involving Σ cannot be calculated analytically anymore. As a solution, recent works on black-box variational inference propose to use a Monte Carlo estimate of the expectation of the gradient. In order to address high variance of the estimator, several techniques have been proposed (e.g. control variates and Rao-Blackwellization) among which the reparameterization trick appears to be the most promising (Ranganath et al., 2014; Kingma and Welling, 2013; Kucukelbir et al., 2017). In particular, Stan (Carpenter et al., 2017) provides a readily available implementation of the reparameterization trick (Kucukelbir et al., 2017) which is named automatic differentiation variational inference (ADVI). In ADVI, the transformation is $\Sigma_g := L^T L$ with L being a triangular matrix where each component is sampled from $N(0, 1)$. And the matrix L is the parameter of

the variational distribution that is optimized with stochastic gradient descent. However, note that this optimization problem is a stochastic non-convex problem. In contrast, finding the MAP is a non-stochastic convex optimization problem and the proposed solution has a guarantee of converging to the global minima. Apart from that, we note that a full variational approximation does not have any theoretic quality guarantees, including the case where $\beta \to 0$. In the general case, our approach also does not have such guarantees. However, in the special case where $\beta \to 0$, we know that the true posterior distribution is an inverse Wishart distribution and therefore matches our choice of the variational distribution.

MCMC estimation of marginal likelihood

As an alternative to the variational approximation, we investigate an MCMC estimation based on Chib's method (Chib, 1995; Chib and Jeliazkov, 2001).

To simplify the description, we introduction the following notations

$$\begin{aligned}\boldsymbol{\theta}_1 &:= \Sigma_\epsilon\,,\\ \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_{k+1} &:= \Sigma_1, \ldots, \Sigma_k\,.\end{aligned}$$

Furthermore, we define $\boldsymbol{\theta}_{<i} := \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}\}$ and $\boldsymbol{\theta}_{>i} := \{\boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_{k+1}\}$. For simplicity, we also suppress in the notation the explicit conditioning on the hyper-parameters $\boldsymbol{\eta}$ and the clustering $\mathcal{C}$, which are both fixed.

Following the strategy of Chib (1995), the marginal likelihood can be expressed as

$$\begin{aligned}p(\mathscr{X}) &= \frac{p(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{k+1}, \mathscr{X})}{p(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{k+1} | \mathscr{X})}\\ &= \frac{p(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_{k+1}, \mathscr{X})}{\prod_{i=1}^{k+1} p(\hat{\boldsymbol{\theta}}_i | \mathscr{X}, \hat{\boldsymbol{\theta}}_1 \ldots, \hat{\boldsymbol{\theta}}_{i-1})}\,.\end{aligned} \tag{2.8}$$

In order to approximate $p(\mathscr{X})$ with Equation (2.8), we need to estimate $p(\hat{\boldsymbol{\theta}}_i | \mathscr{X}, \hat{\boldsymbol{\theta}}_1, \ldots \hat{\boldsymbol{\theta}}_{i-1})$. First, note that we can express the value of the conditional posterior distribution at $\hat{\boldsymbol{\theta}}_i$, as follows (see Chib and Jeliazkov (2001), Section 2.3):

$$\begin{aligned}&p(\hat{\boldsymbol{\theta}}_i | \mathscr{X}, \hat{\boldsymbol{\theta}}_1, \ldots \hat{\boldsymbol{\theta}}_{i-1})\\ &= \frac{\mathbb{E}_{\boldsymbol{\theta}_{\geq i} \sim p(\boldsymbol{\theta}_{\geq i} | \mathscr{X}, \hat{\boldsymbol{\theta}}_{<i})}[\alpha(\boldsymbol{\theta}_i, \hat{\boldsymbol{\theta}}_i | \hat{\boldsymbol{\theta}}_{<i}, \boldsymbol{\theta}_{>i}) q_i(\hat{\boldsymbol{\theta}}_i)]}{\mathbb{E}_{\boldsymbol{\theta}_{\geq i} \sim p(\boldsymbol{\theta}_{>i} | \mathscr{X}, \hat{\boldsymbol{\theta}}_{\leq i}) q(\boldsymbol{\theta}_i)}[\alpha(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i | \hat{\boldsymbol{\theta}}_{<i}, \boldsymbol{\theta}_{>i})]}\,,\end{aligned} \tag{2.9}$$

where $q_i(\boldsymbol{\theta}_i)$ is a proposal distribution for $\boldsymbol{\theta}_i$, and the acceptance probability of moving from state $\boldsymbol{\theta}_i$ to state $\boldsymbol{\theta}'_i$, holding the other states fixed is defined as

$$\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}'_i | \boldsymbol{\theta}_{<i}, \boldsymbol{\theta}_{>i}) := \min\{1, \frac{p(\mathscr{X}, \boldsymbol{\theta}_{<i}, \boldsymbol{\theta}_{>i}, \boldsymbol{\theta}'_i) \cdot q_i(\boldsymbol{\theta}_i)}{p(\mathscr{X}, \boldsymbol{\theta}_{<i}, \boldsymbol{\theta}_{>i}, \boldsymbol{\theta}_i) \cdot q_i(\boldsymbol{\theta}'_i)}\}\,. \tag{2.10}$$

Next, using Equation (2.9), we can estimate $p(\hat{\boldsymbol{\theta}}_i|\mathscr{X}, \hat{\boldsymbol{\theta}}_1, \dots \hat{\boldsymbol{\theta}}_{i-1})$ with a Monte Carlo approximation with M samples:

$$\begin{aligned} & p(\hat{\boldsymbol{\theta}}_i|\mathscr{X}, \hat{\boldsymbol{\theta}}_1, \dots \hat{\boldsymbol{\theta}}_{i-1}) \\ & \approx \frac{\frac{1}{M}\sum_{m=1}^{M} \alpha(\boldsymbol{\theta}_i^{i,m}, \hat{\boldsymbol{\theta}}_i|\hat{\boldsymbol{\theta}}_{<i}, \boldsymbol{\theta}_{>i}^{i,m}) q_i(\hat{\boldsymbol{\theta}}_i)}{\frac{1}{M}\sum_{m=1}^{M} \alpha(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\theta}_i^{q,m}|\hat{\boldsymbol{\theta}}_{<i}, \boldsymbol{\theta}_{>i}^{i+1,m})} \end{aligned} \tag{2.11}$$

where $\boldsymbol{\theta}_i^{a,m} \sim p(\boldsymbol{\theta}_i|\mathscr{X}, \hat{\boldsymbol{\theta}}_{<a})$, $\boldsymbol{\theta}_{>i}^{a,m} \sim p(\boldsymbol{\theta}_{>i}|\mathscr{X}, \hat{\boldsymbol{\theta}}_{<a})$, and $\boldsymbol{\theta}_i^{q,m} \sim q(\boldsymbol{\theta}_i)$.

Finally, in order to sample from $p(\boldsymbol{\theta}_{\geq i}|\mathscr{X}, \hat{\boldsymbol{\theta}}_{<i})$, we propose to use the Metropolis-Hastings within Gibbs sampler as shown in Algorithm 1. $MH_j(\boldsymbol{\theta}_j^t, \boldsymbol{\psi})$ denotes the Metropolis-Hastings algorithm with current state $\boldsymbol{\theta}_j^t$, and acceptance probability $\alpha(\boldsymbol{\theta}_j, \boldsymbol{\theta}_j'|\boldsymbol{\psi})$, Equation (2.10), and $\boldsymbol{\theta}_{\geq i}^0$ is a sample after the burn-in. For the proposal distribution $q_i(\boldsymbol{\theta}_i)$, we use

$$q_i := \begin{cases} \text{InvW}(\nu, \hat{\Sigma}_\epsilon \cdot (\nu + d + 1)) & \\ \text{with } \nu = \beta\kappa \cdot n + \nu_\epsilon & \text{if } i = 1, \\ \text{InvW}(\nu, \hat{\Sigma}_{i-1} \cdot (\nu + d_{i-1} + 1)) & \\ \text{with } \nu = (1-\beta)\kappa \cdot n + \nu_{i-1} & \text{else.} \end{cases} \tag{2.12}$$

Here $\kappa > 0$ is a hyper-parameter of the MCMC algorithm that is chosen to control the acceptance probability. Note that if we choose $\kappa = 1$ and β is 0, then the proposal distribution $q_i(\boldsymbol{\theta}_i)$ equals the posterior distribution $p(\boldsymbol{\theta}_i|\mathscr{X}, \hat{\boldsymbol{\theta}}_1, \dots \hat{\boldsymbol{\theta}}_{i-1})$. However, in practice, we found that the acceptance probabilities can be too small, leading to unstable estimates and division by 0 in Equation (2.11). Therefore, for our experiments we chose $\kappa = 10$.

Algorithm 1 Metropolis-Hastings within Gibbs sampler for sampling from $p(\boldsymbol{\theta}_{\geq i}|\mathscr{X}, \hat{\boldsymbol{\theta}}_{<i})$.

```
for t from 1 to M do
  for j from i to k + 1 do
    ψ := {θ̂_{<i}, θ_i^t, ..., θ_{j-1}^t, θ_{>j}^{t-1}}
    θ_j^t := MH_j(θ_j^{t-1}, ψ)
  end for
end for
```

2.4.3 Restricting the hypotheses space

The number of possible clusterings follow the Bell numbers, and therefore it is infeasible to enumerate all possible clusterings, even if the number of variables d is small. It is therefore crucial to restrict the hypotheses space to a subset of all clusterings that are likely to contain the true clustering. We denote this subset as $\mathscr{C}^*$.

We suggest to use spectral clustering on different estimates of the precision matrix to acquire the set of clusterings $\mathscr{C}^*$. A motivation for this heuristic is given in Appendix A.3.

First, for an appropriate λ, we estimate the precision matrix using

$$X^* := \underset{X \succeq 0}{\arg\min} - \log|X| + \operatorname{trace}(XS) + \lambda \sum_{i \neq j} |X_{ij}|^q \,. \tag{2.13}$$

In our experiments, we take $q = 1$, which is equivalent to the Graphical Lasso (Friedman et al., 2008) with an ℓ_1-penalty on all entries of X except the diagonal. In the next step, we then construct the Laplacian L as defined in the following.

$$\begin{aligned} L_{ii} &= \sum_{k \neq i} |X^*_{ik}|^q \,, \\ L_{ij} &= -|X^*_{ij}|^q \;\text{ for } i \neq j \,. \end{aligned} \tag{2.14}$$

Finally, we use k-means clustering on the eigenvectors of the Laplacian L. The details of acquiring the set of clusterings $\mathscr{C}^*$ using the spectral clustering method are summarized below:

Algorithm 2 Spectral Clustering for variable clustering with the Gaussian graphical model.

```
J := set of regularization parameter values.
K_max := maximum number of considered clusters.
C* := {}
for λ ∈ J do
  X* := solve optimization problem from Equation (2.13).
  (e_1, ..., e_{K_max}) := determine the eigenvectors corresponding to the K_max
  lowest eigenvalues of the Laplacian L as defined in Equations (2.14).
  for k ∈ {2, ..., K_max} do
    C_{λ,k} := cluster all variables into k partitions using k-means with
    (e_1, ..., e_k).
    C* := C* ∪ C_{λ,k}
  end for
end for
return restricted hypotheses space C*
```

In Section 2.5.1 we confirm experimentally that, even in the presence of noise, $\mathscr{C}^*$ often contains the true clustering, or clusterings that are close to the true clustering.

Posterior distribution over number of clusters

In principle, the posterior distribution for the number of clusters can be calculated using

$$p(k|\mathscr{X}) \propto \sum_{\mathcal{C} \in \mathscr{C}_k} p(\mathscr{X}|\mathcal{C}) \,,$$

where $\mathscr{C}_k$ denotes the set of all clusterings with number of clusters being equal to k. Since this is computationally infeasible, we use the following approximation

$$P(k|X) \propto \sum_{\mathcal{C} \in \mathscr{C}_k} p(X|\mathcal{C}) \approx \sum_{\mathcal{C} \in \mathscr{C}_k^*} p(X|\mathcal{C}) \,,$$

where $\mathscr{C}_k^*$ is the set of all clusterings with k clusters that are in the restricted hypotheses space $\mathscr{C}^*$.

2.5 Simulation study

In this section, we evaluate the proposed method on simulated data for which the ground truth is available. In sub-section 2.5.1, we evaluate the quality of the restricted hypotheses space $\mathscr{C}^*$, followed by sub-section 2.5.2, where we evaluated the proposed method's ability to select the best clustering in $\mathscr{C}^*$.

For the number of clusters we consider the range from 2 to 15. For the set of regularization parameters of the spectral clustering method we use $J := \{0.0001, 0.0005, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.01\}$ (see Algorithm 2).

In all experiments the number of variables is $d = 40$, and the ground truth is 4 clusters with 10 variables each.

For generating positive-definite covariance matrices, we consider the following two distributions: $\text{InvW}(d+1, I_d)$, and Uniform_d, with dimension d. We denote by $U \sim \text{Uniform}_d$ the positive-definite matrix generated in the following way

$$U = A + (0.001 - \lambda_{min}(A))I_d \,,$$

where $\lambda_{min}(A)$ is the smallest eigenvalue of A, and A is drawn as follows

$$\begin{aligned} A_{i,j} &= A_{j,i} \sim \text{Uniform}(-1, 1) \,, i \neq j \\ A_{i,i} &= 0 \,. \end{aligned}$$

For generating Σ, we either sample each block j from $\text{InvW}(d_j + 1, I_{d_j})$ or from Uniform_{d_j}.

For generating the noise matrix Σ_ϵ, we sample either from $\text{InvW}(d+1, I_d)$ or from Uniform_d. The final data is then sampled as follows

$$x \sim N(0, (\Sigma^{-1} + \eta\Sigma_\epsilon^{-1})^{-1}) \,,$$

where η defines the noise level.

For evaluation we use the adjusted normalized mutual information (ANMI), where 0.0 means that any correspondence with the true labels is at chance level, and 1.0 means that a perfect one-to-one correspondence exists (Vinh et al., 2010). We repeated all experiments 5 times and report the average ANMI score.

2.5.1 Evaluation of the restricted hypotheses space

First, independent of any model selection criteria, we check here the quality of the clusterings that are found with the spectral clustering algorithm from Section 2.4.3. We also compare to single and average linkage clustering as used in (Tan et al., 2015).

The set of all clusterings that are found is denoted by $\mathscr{C}^*$ (the restricted hypotheses space).

In order to evaluate the quality of the restricted hypotheses space $\mathscr{C}^*$, we report the oracle performance calculated by $\max_{\mathcal{C}\in\mathscr{C}^*} \text{ANMI}(\mathcal{C}, \mathcal{C}_T)$, where $\mathcal{C}_T$ denotes the true clustering, and $\text{ANMI}(\mathcal{C}, \mathcal{C}_T)$ denotes the ANMI score when comparing clustering $\mathcal{C}$ with the true clustering. In particular, a score of 1.0 means that the true clustering is contained in $\mathscr{C}^*$.

The results of all experiments with noise level $\eta \in \{0.0, 0.01, 0.1\}$ are shown in Tables 2.1, for balanced clusters, and Table 2.2, for unbalanced clusters.

From these results we see that the restricted hypotheses space of spectral clustering is around 100, considerably smaller than the number of all possible clusterings. More importantly, we also see that that $\mathscr{C}^*$ acquired by spectral clustering either contains the true clustering or a clustering that is close to the truth. In contrast, the hypotheses space restricted by single and average linkage is smaller, but more often misses the true clustering.

2.5.2 Evaluation of clustering selection criteria

Here, we evaluate the performance of our proposed method for selecting the correct clustering in the restricted hypotheses space $\mathscr{C}^*$. We compare our proposed method (variational) with several baselines and two previously proposed methods (Tan et al., 2015; Palla et al., 2012). Except for the two previously proposed methods, we created $\mathscr{C}^*$ with the spectral clustering algorithm from Section 2.4.3.

As a cluster selection criteria, we compare our method to the Extended Bayesian Information Criterion (EBIC) with $\gamma \in \{0, 0.5, 1\}$ (Chen and Chen, 2008; Foygel and Drton, 2010), Akaike Information Criteria (Akaike, 1973), and the Calinski-Harabasz Index (CHI) (Caliński and Harabasz, 1974). Note that EBIC and AIC are calculated based on the basic Gaussian graphical model (i.e. the model in Equation 2.1, but ignoring the prior specification).[3] Furthermore, we note that EBIC is model consistent, and therefore, assuming that the true precision matrix contains non-zero entries in each element, will choose asymptotically the clustering that has only one cluster with all variables in it. However, as an advantage for EBIC, we exclude that clustering. Furthermore, we note that in contrast to EBIC and AIC, the Calinski-Harabasz Index is not a model-based cluster evaluation criterion. The Calinski-Harabasz Index is an heuristic that uses as clustering criterion the ratio of the variance within and across clusters. As such it is expected to give reasonable clustering results if

[3] As discussed in Section 2.4.2, EBIC (and also AIC) cannot be used with our proposed model.

the noise is considerably smaller in magnitude than the within-cluster variable partial correlations.

We remark that EBIC and AIC is not well defined if the sample covariance matrix is singular, in particular if $n < d$ or $n \approx d$. As an ad-hoc remedy, which works well in practice[4], we always add 0.001 times the identity matrix to the covariance matrix (see also Ledoit and Wolf (2004)).

Finally, we also compare the proposed method to two previous approaches for variable clustering: the Clustered Graphical Lasso (CGL) as proposed in (Tan et al., 2015), and the Dirichlet process variable clustering (DPVC) model as proposed in (Palla et al., 2012), for which the implementation is available. DPVC models the number of clusters using a Dirichlet process. CGL uses for model selection the mean squared error for recovering randomly left-out elements of the covariance matrix. CGL uses for clustering either the single linkage clustering (SLC) or the average linkage clustering (ALC) method. For conciseness, we show only the results for ALC, since they tended to be better than SLC.

A summary of the experiments, with noise level $\eta \in \{0.0, 0.01, 0.1\}$, limited to the proposed method, basic inverse Wishart model, EBIC and Calinski-Harabasz Index, is shown in Figure 2.1 and Figure 2.2, for balanced and unbalanced clusters, respectively, and in Figure 2.3 for the high-dimensional setting ($d = 200$). Detailed results of all experiments are shown in Tables 2.3 and 2.4, for balanced clusters, and Tables 2.5 and 2.6, for unbalanced clusters, and Tables 2.7 and 2.8 for the high-dimensional setting. The tables also contain the performance of the proposed method for $\beta \in \{0, 0.01, 0.02, 0.03\}$. Note that $\beta = 0.0$ corresponds to the basic inverse Wishart prior model for which we can calculate the marginal likelihood analytically.

Comparing the proposed method with different β, we see that $\beta = 0.02$ offers good clustering performance in the no noise and noisy setting. In contrast, model selection with the basic inverse Wishart model, EBIC and AIC perform, as expected, well in the no noise scenario, however, in the noisy setting they tend to select incorrect clusterings. In particular, note that the basic inverse Wishart model and EBIC are model consistent, that means that for large enough n they are guaranteed to select the finest clustering such that there is no edge[5] between any two clusters. This is the reason why the basic inverse Wishart model and EBIC are guaranteed, at least asymptotically, to fail to ignore noisy edges. This is confirmed by our experiments in the noisy setting, showing that the basic inverse Wishart model and EBIC fail to identify the correct clusterings, when the number of samples n is large enough. On the other hand, if $n \leq d$, then the noise due to sampling, and the noise in the true precision matrix cannot be distinguished, and the basic inverse Wishart model performs similarly to the proposed method.

The Calinski-Harabasz Index performs well in the noisy settings, whereas in the no noise setting it performs unsatisfactory. Furthermore, note that the

[4] In particular for the mutual funds data in the next section, where the covariance matrix was bad conditioned.

[5] A non-zero entry in the precision matrix.

Calinski-Harabasz Index itself can only be used to rank different clustering results, but does not provide any probability estimates.

Finally, we show the posterior distribution of the number of clusters in Figures 2.4 and 2.5, with and without noise on the precision matrix, respectively.[6] In both cases, given that the sample size n is large enough, the proposed method is able to estimate correctly the number of clusters. In contrast, the basic inverse Wishart prior model underestimates the number of clusters for large n and existence of noise in the precision matrix.

2.5.3 Comparison of variational and MCMC estimate

Here, we compare our variational approximation with MCMC on a small scale simulated problem where it is computationally feasible to estimate the marginal likelihood with MCMC. We generated synthetic data as in the previous section, only with the difference that we set the number of variables d to 12.

The number of samples M for MCMC was set to 10000, where we used 10% as burn in. For two randomly picked clusterings for $n = 12$, and $n = 1200000$, we checked the acceptance rates and convergence using the multivariate extension of the Gelman-Rubin diagnostic (Brooks and Gelman, 1998). The average acceptance rates were around 80% and the potential scale reduction factor was 1.01.

The runtime of MCMC was around 40 minutes for evaluating one clustering, whereas for the variational approximation the runtime was around 2 seconds.[7] The results are shown in Table 2.9, suggesting that the quality of the selected clusterings using the variational approximation is similar to MCMC.

2.6 Real data experiments

In this section, we investigate the properties of the proposed model selection criterion on three real data sets. In all cases, we use the spectral clustering algorithm from Appendix A.3 to create cluster candidates. All variables were normalized to have mean 0 and variance 1. For all methods, except DPVC, the number of clusters is considered to be in $\{2, 3, 4, \ldots, \min(p-1, 15)\}$. DPVC automatically selects the number of clusters by assuming a Dirichlet process prior. We evaluated the proposed method with $\beta = 0.02$ using the variational approximation.

2.6.1 Mutual funds

Here we use the mutual funds data, which has been previously analyzed in (Scott and Carvalho, 2008; Marlin et al., 2009). The data contains 59 mutual funds (d =59) grouped into 4 clusters: U.S. bond funds, U.S. stock funds, balanced funds

[6] Same setting as before, $d = 40$, $\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$. Noise is $\Sigma_\epsilon \sim \text{InvW}(d+1, I_d), \eta = 0.01$. Proposed method $\beta = 0.02$.

[7] Runtime on one core of Intel(R) Xeon(R) CPU 2.30GHz.

(containing U.S. stocks and bonds), and international stock funds. The number of observations is 86.

The results of all methods are visualized in Table 2.10. It is difficult to interpret the results produced by EBIC ($\gamma = 1.0$), AIC and the Calinski-Harabasz Index. In contrast, the proposed method and EBIC ($\gamma = 0.0$) produce results that are easier to interpret. In particular, our results suggest that there is a considerable correlation between the balanced funds and the U.S. stock funds which was also observed in Marlin et al. (2009).

In Figure 2.6 we show a two dimensional representation of the data, that was found using Laplacian Eigenmaps (Belkin and Niyogi, 2003). The figure supports the claim that balanced funds and the U.S. stock funds have similar behavior.

2.6.2 Gene regulations

We tested our method also on the gene expression data that was analyzed in (Hirose et al., 2017). The data consists of 11 genes with 445 gene expressions. The true gene regularizations are known in this case and shown in Figure 2.7, adapted from (Hirose et al., 2017). The most important fact is that there are two independent groups of genes and any clustering that mixes these two can be considered as wrong.

We show the results of all methods in Figure 2.8, where we mark each cluster with a different color superimposed on the true regularization structure. Here only the clustering selected by the proposed method, EBIC ($\gamma = 1.0$) and Calinski-Harabasz correctly divide the two group of genes.

2.6.3 Aviation sensors

As a third data set, we use the flight aviation dataset from NASA[8]. The data set contains sensor information sampled from airplanes during operation. We extracted the information of 16 continuous-valued sensors that were recorded for different flights with a total of 25032364 samples.

The clustering results are shown in Table 2.11. The data set does not have any ground truth, but the clustering result of our proposed method is reasonable: Cluster 9 groups sensors that measure or affect altitude[9], Cluster 8 correctly clusters the left and right sensors for measuring the rotation around the axis pointing through the noise of the aircraft, in Cluster 2 all sensors that measure the angle between chord and flight direction are grouped together. It also appears reasonable that the yellow hydraulic system of the left part of the plane has little direct interaction with the green hydraulic system of the right part (Cluster 1 and Cluster 4). And the sensor for the rudder, influencing the direction of the plane, is mostly independent of the other sensors (Cluster 5).

[8] `https://c3.nasa.gov/dashlink/projects/85/` where we use all records from Tail 687.

[9] The elevator position of an airplane influences the altitude, and the static pressure system of an airplane measures the altitude.

In contrast, the clustering selected by the basic inverse Wishart prior, EBIC, and AIC is difficult to interpret. We note that we did not compare to DPVC, since the large number of samples made the MCMC algorithm of DPVC infeasible.

2.7 Discussion and conclusions

We have introduced a new method for evaluating variable clusterings based on the marginal likelihood of a Bayesian model that takes into account noise on the precision matrix. Since the calculation of the marginal likelihood is analytically intractable, we proposed two approximations: a variational approximation and an approximation based on MCMC. Experimentally, we found that the variational approximation is considerably faster than MCMC and also leads to accurate model selection.

We compared our proposed method to several standard model selection criteria. In particular, we compared to BIC and extended BIC (EBIC) which are often the method of choice for model selection in Gaussian graphical models. However, we emphasize that EBIC was designed to handle the situation where d is in the order of n, and has not been designed to handle noise. As a consequence, our experiments showed that in practice its performance depends highly on the choice of the γ parameter. In contrast, the proposed method, with fixed hyper-parameters, shows better performance on various simulated and real data.

We also compared our method to other two previously proposed methods, namely Cluster Graphical Lasso (CGL) (Tan et al., 2015), and Dirichlet Process Variable Clustering (DPVC) (Palla et al., 2012) that performs jointly clustering and model selection. However, it appears that in many situations the model selection algorithm of CGL is not able to detect the true model, even if there is no noise. On the other hand, the Dirichlet process assumption by DPVC appears to be very restrictive, leading again to many situations where the true model (clustering) is missed. Overall, our method performs better in terms of selecting the correct clustering on synthetic data with ground truth, and selects meaningful clusters on real data.

Apart from the differences in clustering performance, we note that our proposed method is a full probabilistic model allowing to quantify the uncertainty in all clustering results. This in contrast to using the Calinski-Harabasz Index for variable clustering. On the other hand, in contrast to DPVC, the sufficient statistic for our probabilistic model is the covariance matrix, whereas DPCV requires access to all samples which is prohibitive for large n.[10]

The python source code for variable clustering and model selection with the proposed method and all baselines is available at `https://github.com/andrade-stats/robustBayesClustering`.

[10] In detail, DPVC formulates a probabilistic model with number of latent variables increasing linearly in n.

Table 2.1: Evaluation of restricted hypotheses space for $d = 40$, $n \in \{20, 40, 400, 4000, 40000, 4000000\}$. Ground truth contains 4 balanced clusters. Shows the oracle performance measured by ANMI for spectral clustering, average linkage and single linkage. Note that that an ANMI score of 1.0 means that the true clustering is contained in the hypotheses space found by the clustering method. The size of the hypotheses space restricted by each clustering method is denoted by $|\mathscr{C}^*|$. Average results over 5 runs with standard deviation in brackets.

		20	40	400	4000	40000	4000000
		$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise					
spectral	ANMI	**0.77** (0.14)	**0.95** (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	140.8 (5.78)	139.0 (8.65)	112.8 (5.64)	99.8 (2.23)	101.4 (7.94)	98.4 (3.61)
average	ANMI	0.38 (0.09)	0.38 (0.06)	0.45 (0.05)	0.45 (0.03)	0.45 (0.07)	0.45 (0.03)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.32 (0.08)	0.34 (0.09)	0.39 (0.08)	0.39 (0.08)	0.42 (0.14)	0.41 (0.08)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d + 1, I_d), \eta = 0.01$					
spectral	ANMI	**0.49** (0.03)	**0.9** (0.03)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	143.2 (7.25)	144.4 (3.32)	108.6 (9.89)	105.4 (9.79)	103.6 (5.0)	97.0 (6.57)
average	ANMI	0.26 (0.05)	0.34 (0.04)	0.46 (0.07)	0.51 (0.08)	0.42 (0.09)	0.45 (0.06)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.16 (0.08)	0.25 (0.08)	0.37 (0.03)	0.4 (0.06)	0.3 (0.12)	0.32 (0.09)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d + 1, I_d), \eta = 0.1$					
spectral	ANMI	**0.34** (0.1)	**0.87** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	121.4 (7.34)	106.4 (18.51)	35.4 (5.12)	33.2 (11.48)	37.4 (5.54)	31.0 (8.65)
average	ANMI	0.1 (0.05)	0.15 (0.03)	0.34 (0.08)	0.37 (0.1)	0.26 (0.11)	0.28 (0.09)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.04 (0.03)	0.08 (0.04)	0.19 (0.11)	0.21 (0.06)	0.11 (0.03)	0.13 (0.02)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise					
spectral	ANMI	**0.34** (0.1)	**0.87** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	121.4 (7.34)	106.4 (18.51)	35.4 (5.12)	33.2 (11.48)	37.4 (5.54)	31.0 (8.65)
average	ANMI	0.1 (0.06)	0.26 (0.07)	0.92 (0.11)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.04 (0.02)	0.13 (0.08)	0.82 (0.25)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.01$					
spectral	ANMI	**0.28** (0.06)	**0.81** (0.1)	**0.94** (0.06)	**0.99** (0.03)	**0.99** (0.03)	**0.97** (0.03)
	$\lvert\mathscr{C}^*\rvert$	127.2 (3.6)	106.0 (5.29)	48.2 (9.77)	50.2 (5.95)	51.0 (8.94)	48.0 (5.69)
average	ANMI	0.14 (0.05)	0.22 (0.04)	0.81 (0.16)	0.89 (0.1)	0.87 (0.12)	0.94 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.04 (0.02)	0.1 (0.04)	0.78 (0.13)	0.71 (0.23)	0.78 (0.11)	0.79 (0.17)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.1$					
spectral	ANMI	**0.3** (0.03)	**0.72** (0.08)	**0.88** (0.07)	**0.9** (0.07)	**0.87** (0.11)	0.88 (0.04)
	$\lvert\mathscr{C}^*\rvert$	126.2 (2.23)	120.4 (9.35)	74.4 (19.41)	87.2 (7.93)	79.2 (13.61)	77.0 (14.25)
average	ANMI	0.08 (0.04)	0.26 (0.11)	0.83 (0.15)	0.88 (0.12)	**0.87** (0.11)	**0.94** (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.05 (0.03)	0.13 (0.07)	0.7 (0.14)	0.69 (0.15)	0.76 (0.12)	0.76 (0.14)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)

Table 2.2: Same setting as in Table 2.1 but with unbalanced clusters. Ground truth is 4 clusters with sizes 20, 10, 5, 5.

		$\Sigma_j \sim \text{InvW}(d_j+1, I_{d_j})$, no noise					
		20	40	400	4000	40000	4000000
spectral	ANMI	**0.52** (0.13)	**0.85** (0.11)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	141.2 (6.62)	133.2 (8.03)	80.8 (8.21)	73.4 (8.89)	62.0 (7.38)	62.6 (7.23)
average	ANMI	0.34 (0.06)	0.39 (0.05)	0.37 (0.04)	0.38 (0.07)	0.38 (0.06)	0.44 (0.09)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.33 (0.05)	0.35 (0.03)	0.32 (0.04)	0.32 (0.14)	0.27 (0.13)	0.39 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{InvW}(d_j+1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d+1, I_d), \eta = 0.01$					
spectral	ANMI	**0.55** (0.13)	**0.81** (0.07)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	148.8 (4.62)	136.0 (6.81)	80.4 (9.77)	68.8 (10.3)	67.0 (5.93)	63.0 (14.3)
average	ANMI	0.34 (0.06)	0.37 (0.08)	0.53 (0.12)	0.5 (0.1)	0.46 (0.1)	0.52 (0.1)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.29 (0.07)	0.29 (0.08)	0.41 (0.17)	0.4 (0.14)	0.37 (0.11)	0.32 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{InvW}(d_j+1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d+1, I_d), \eta = 0.1$					
spectral	ANMI	**0.26** (0.04)	**0.5** (0.06)	**0.93** (0.07)	**0.93** (0.07)	**0.99** (0.02)	**0.91** (0.08)
	$\lvert\mathscr{C}^*\rvert$	144.4 (5.54)	159.2 (1.83)	121.0 (10.43)	120.2 (6.62)	117.0 (3.41)	113.2 (11.91)
average	ANMI	0.2 (0.03)	0.22 (0.06)	0.37 (0.09)	0.36 (0.08)	0.41 (0.13)	0.44 (0.07)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.2 (0.08)	0.2 (0.07)	0.24 (0.04)	0.29 (0.05)	0.33 (0.07)	0.32 (0.05)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise					
spectral	ANMI	**0.36** (0.06)	**0.72** (0.13)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\lvert\mathscr{C}^*\rvert$	124.0 (7.29)	115.8 (9.89)	40.8 (12.5)	39.4 (5.2)	33.2 (4.79)	38.6 (5.24)
average	ANMI	0.09 (0.04)	0.05 (0.08)	0.12 (0.07)	0.29 (0.07)	0.37 (0.07)	0.34 (0.14)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.01 (0.04)	0.0 (0.0)	0.0 (0.01)	0.06 (0.1)	0.17 (0.19)	0.13 (0.12)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.01$					
spectral	ANMI	**0.39** (0.04)	**0.67** (0.11)	**0.85** (0.05)	**0.89** (0.07)	**0.87** (0.07)	**0.89** (0.06)
	$\lvert\mathscr{C}^*\rvert$	125.6 (8.06)	115.0 (12.85)	42.6 (7.09)	59.2 (11.55)	53.2 (9.2)	54.0 (6.69)
average	ANMI	0.04 (0.03)	0.06 (0.05)	0.12 (0.06)	0.21 (0.08)	0.18 (0.09)	0.21 (0.13)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.02)	0.01 (0.05)	0.02 (0.05)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
		$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.1$					
spectral	ANMI	**0.32** (0.06)	**0.68** (0.13)	**0.8** (0.09)	**0.81** (0.09)	**0.79** (0.07)	**0.78** (0.09)
	$\lvert\mathscr{C}^*\rvert$	124.2 (9.33)	109.6 (12.63)	66.6 (10.71)	74.2 (7.14)	62.8 (5.11)	65.2 (13.85)
average	ANMI	0.04 (0.03)	0.06 (0.05)	0.09 (0.05)	0.19 (0.05)	0.13 (0.06)	0.2 (0.13)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)
single	ANMI	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.02)
	$\lvert\mathscr{C}^*\rvert$	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)	14.0 (0.0)

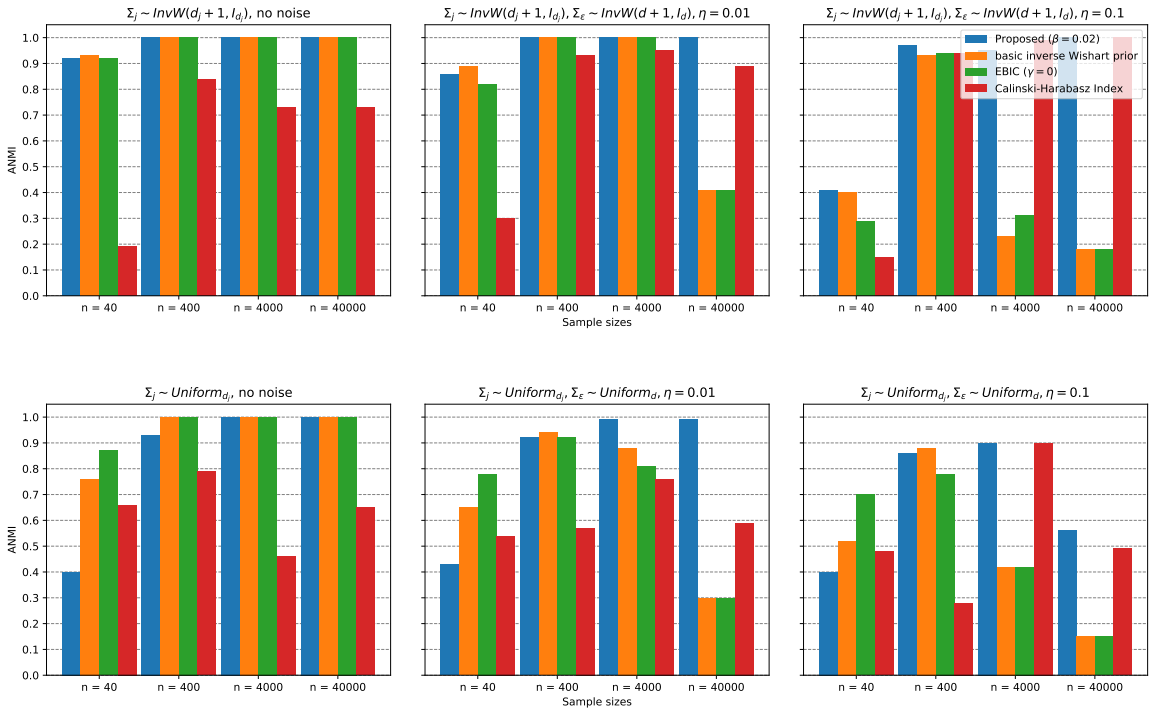

Figure 2.1: Shows the ANMI scores of the clustering selected by the proposed method (blue), basic inverse Wishart model (orange), EBIC (green), and Calinski-Harabasz Index (red) on synthetic data sets with $d = 40$ and ground truth being 4 *balanced* clusters. Upper row and lower row shows results where the true precision matrix was generated from an inverse Wishart distribution, and a uniform distribution, respectively. No noise setting (left column), small noise (middle column), large noise (right column). ANMI score of 0.0 means correspondence with true clustering at pure chance level and 1.0 means perfect correspondence. In both settings, with and without noise, the proposed method tends to be among the best. In contrast, EBIC and the basic inverse Wishart prior tend to suffer in the noise setting for large n, and Calinski-Harabasz Index performs sub-optimal in the no noise setting.

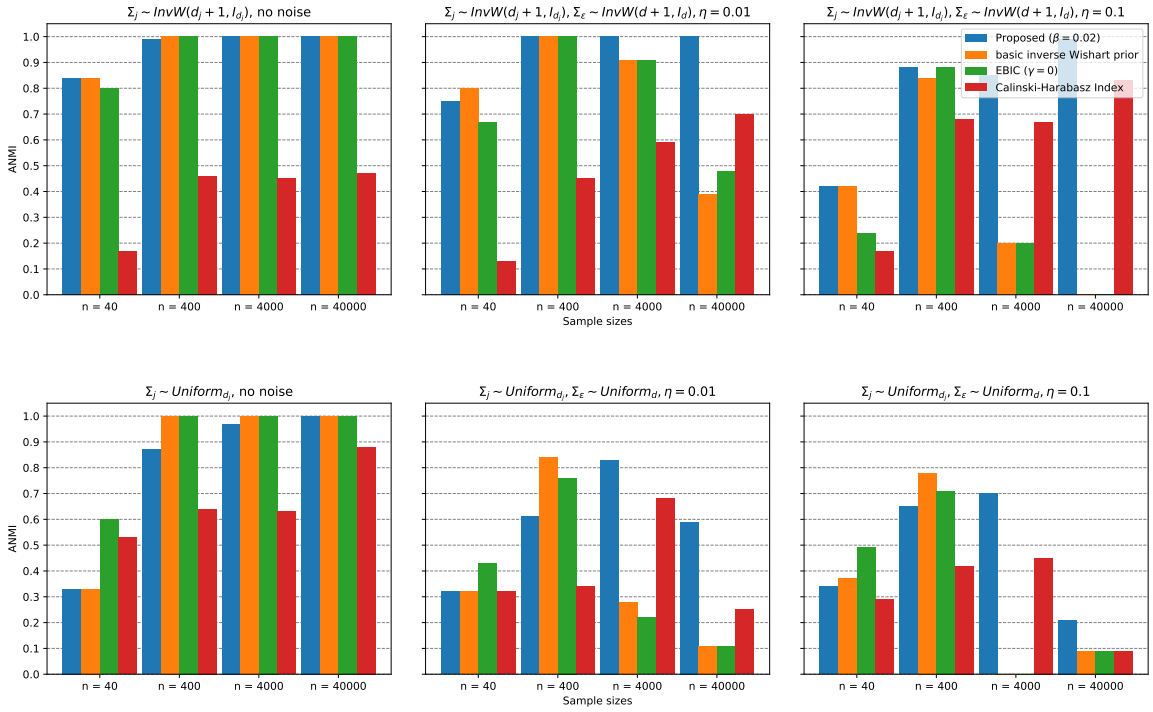

Figure 2.2: Same settings as in Figure 2.1, but ground truth being 4 *unbalanced* clusters.

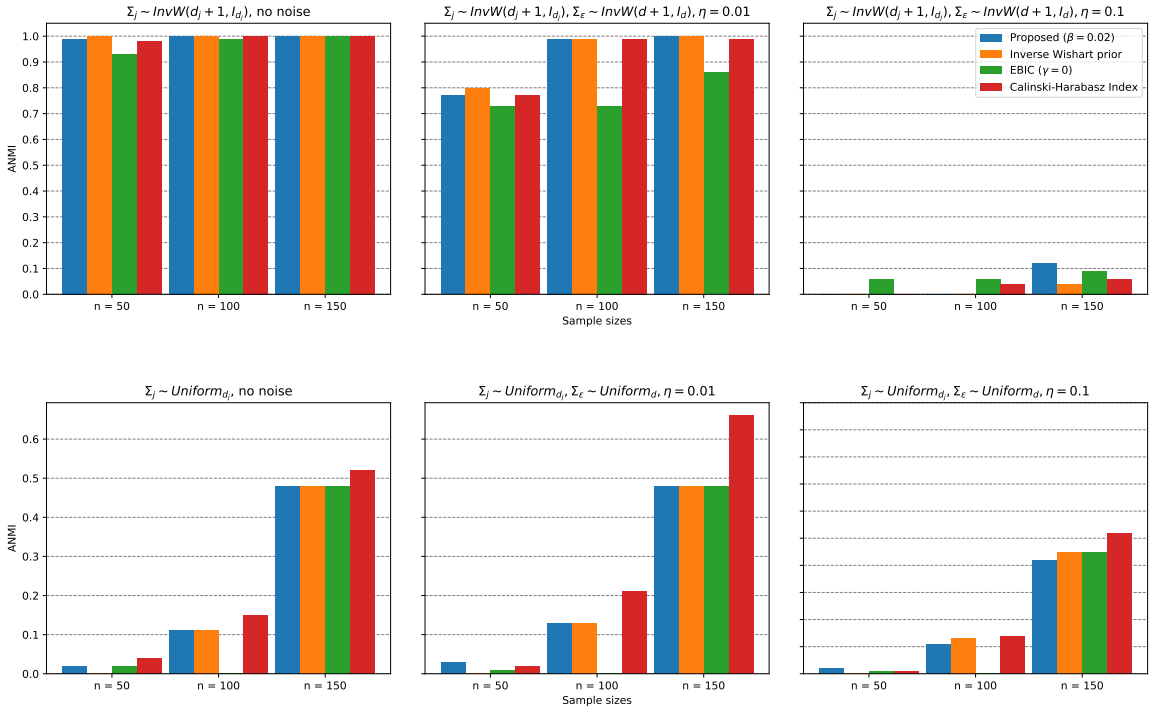

Figure 2.3: Same settings as in Figure 2.1, but number of variables $d = 200$. The proposed method, basic inverse Wishart prior, and the Calinski-Harabasz Index perform best, while the latter has some advantage when the ground truth is sampled from the uniform distribution.

Table 2.3: Evaluation of clustering results for $d = 40$, $n \in \{20, 40, 400, 4000, 40000, 4000000\}$. Ground truth is 4 balanced clusters. Shows the ANMI of the selected models (standard deviation in brackets). No noise is added.

	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise					
	20	40	400	4000	40000	4000000
Proposed ($\beta = 0.01$)	**0.76** (0.14)	**0.93** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.7 (0.2)	0.92 (0.08)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.03$)	0.67 (0.18)	0.88 (0.14)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
basic inverse Wishart prior	0.73 (0.17)	**0.93** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.12 (0.15)	0.92 (0.08)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.36 (0.03)	0.51 (0.04)	0.99 (0.03)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.35 (0.02)	0.39 (0.05)	0.96 (0.05)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.12 (0.15)	0.6 (0.49)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski-Harabasz Index	0.32 (0.03)	0.19 (0.16)	0.84 (0.13)	0.73 (0.0)	0.73 (0.0)	0.73 (0.0)
CGL (ALC)	0.06 (0.05)	0.03 (0.05)	0.11 (0.06)	0.04 (0.04)	0.06 (0.03)	0.06 (0.07)
DPVC	0.53 (0.07)	0.61 (0.17)	0.82 (0.06)	0.93 (0.09)	NA	NA
	$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise					
Proposed ($\beta = 0.01$)	0.12 (0.04)	0.48 (0.07)	0.94 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.12 (0.05)	0.4 (0.04)	0.93 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.03$)	0.12 (0.05)	0.39 (0.03)	0.93 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
basic inverse Wishart prior	0.14 (0.05)	0.76 (0.1)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.07 (0.04)	**0.87** (0.09)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.11 (0.05)	0.48 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.11 (0.05)	0.38 (0.05)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.07 (0.04)	0.66 (0.34)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski-Harabasz Index	**0.15** (0.05)	0.66 (0.16)	0.79 (0.11)	0.46 (0.14)	0.65 (0.23)	0.59 (0.17)
CGL (ALC)	0.03 (0.02)	0.02 (0.02)	0.37 (0.03)	0.39 (0.0)	0.39 (0.0)	0.51 (0.25)
DPVC	0.01 (0.02)	0.03 (0.03)	0.4 (0.2)	0.51 (0.22)	NA	NA

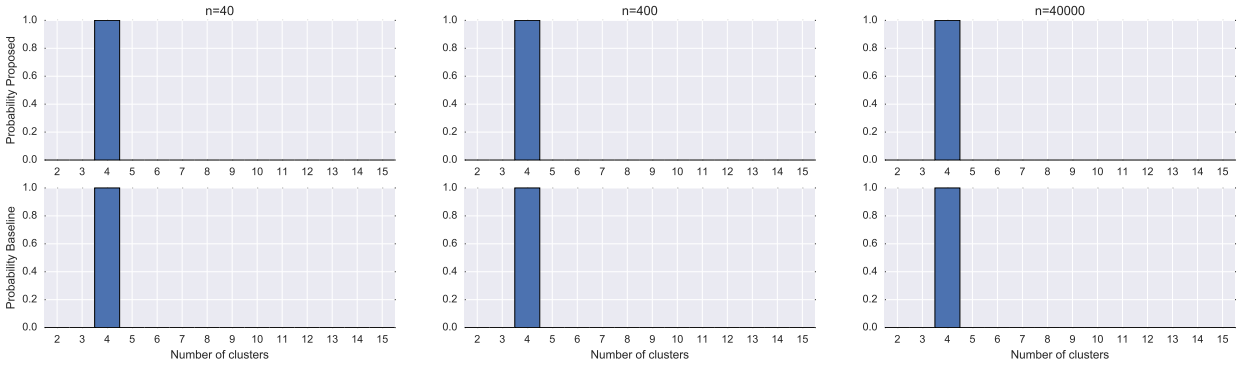

Figure 2.4: Posterior distribution of the number of clusters of the proposed method (top row) and the basic inverse Wishart prior model (bottom row). Ground truth is 4 clusters; there is no noise on the precision matrix.

Table 2.4: Evaluation of clustering results with $d = 40$, $n \in \{20, 40, 400, 4000, 40000, 4000000\}$. Ground truth is 4 balanced clusters. Shows the ANMI of the selected models (standard deviation in brackets). Noise is added to the precision matrix.

	$\Sigma_j \sim \text{InvW}(d_j+1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d+1, I_d), \eta = 0.01$					
	20	40	400	4000	40000	4000000
Proposed ($\beta = 0.01$)	0.44 (0.07)	0.86 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.41 (0.06)	0.86 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
Proposed ($\beta = 0.03$)	0.38 (0.06)	0.8 (0.06)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	0.99 (0.03)
basic inverse Wishart prior	**0.45** (0.07)	**0.89** (0.02)	**1.0** (0.0)	**1.0** (0.0)	0.41 (0.04)	0.39 (0.0)
EBIC ($\gamma = 0$)	0.02 (0.02)	0.82 (0.07)	**1.0** (0.0)	**1.0** (0.0)	0.41 (0.04)	0.39 (0.0)
EBIC ($\gamma = 0.5$)	0.25 (0.08)	0.32 (0.07)	0.98 (0.04)	**1.0** (0.0)	0.48 (0.13)	0.39 (0.0)
EBIC ($\gamma = 1.0$)	0.23 (0.07)	0.32 (0.07)	0.96 (0.06)	**1.0** (0.0)	0.66 (0.14)	0.39 (0.0)
AIC	0.0 (0.01)	0.54 (0.44)	**1.0** (0.0)	0.39 (0.0)	0.41 (0.04)	0.39 (0.0)
Calinski-Harabasz Index	0.26 (0.09)	0.3 (0.16)	0.93 (0.1)	0.95 (0.11)	0.89 (0.13)	0.84 (0.13)
CGL (ALC)	0.01 (0.02)	0.02 (0.05)	0.04 (0.05)	0.03 (0.02)	0.05 (0.06)	0.02 (0.02)
DPVC	0.33 (0.07)	0.42 (0.08)	0.59 (0.16)	0.21 (0.18)	NA	NA
	$\Sigma_j \sim \text{InvW}(d_j+1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d+1, I_d), \eta = 0.1$					
Proposed ($\beta = 0.01$)	0.1 (0.1)	0.4 (0.09)	0.93 (0.1)	0.39 (0.0)	0.33 (0.17)	0.29 (0.15)
Proposed ($\beta = 0.02$)	0.13 (0.09)	**0.41** (0.07)	**0.97** (0.04)	0.95 (0.11)	**1.0** (0.0)	**0.99** (0.03)
Proposed ($\beta = 0.03$)	0.13 (0.09)	0.4 (0.09)	0.95 (0.04)	**0.99** (0.03)	**1.0** (0.0)	**0.99** (0.03)
basic inverse Wishart prior	0.1 (0.1)	0.4 (0.09)	0.93 (0.1)	0.23 (0.19)	0.18 (0.21)	0.23 (0.19)
EBIC ($\gamma = 0$)	0.09 (0.09)	0.29 (0.06)	0.94 (0.05)	0.31 (0.15)	0.18 (0.21)	0.23 (0.19)
EBIC ($\gamma = 0.5$)	0.12 (0.05)	0.2 (0.02)	0.87 (0.02)	0.41 (0.04)	0.18 (0.21)	0.23 (0.19)
EBIC ($\gamma = 1.0$)	**0.14** (0.06)	0.2 (0.02)	0.54 (0.07)	0.86 (0.24)	0.18 (0.21)	0.23 (0.19)
AIC	0.0 (0.0)	0.0 (0.01)	0.09 (0.15)	0.23 (0.19)	0.18 (0.21)	0.23 (0.19)
Calinski-Harabasz Index	0.11 (0.05)	0.15 (0.13)	0.94 (0.05)	**0.99** (0.03)	**1.0** (0.0)	**0.99** (0.03)
CGL (ALC)	0.02 (0.03)	0.0 (0.01)	0.01 (0.01)	0.01 (0.02)	0.0 (0.0)	0.0 (0.0)
DPVC	0.11 (0.06)	0.16 (0.06)	0.27 (0.06)	0.04 (0.04)	NA	NA
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.01$					
Proposed ($\beta = 0.01$)	0.1 (0.04)	0.45 (0.05)	0.92 (0.06)	**0.99** (0.03)	**0.99** (0.03)	**0.93** (0.1)
Proposed ($\beta = 0.02$)	0.12 (0.03)	0.43 (0.06)	0.92 (0.06)	**0.99** (0.03)	**0.99** (0.03)	**0.93** (0.1)
Proposed ($\beta = 0.03$)	0.13 (0.02)	0.39 (0.03)	0.89 (0.07)	**0.99** (0.03)	**0.99** (0.03)	**0.93** (0.1)
basic inverse Wishart prior	0.11 (0.06)	0.65 (0.12)	**0.94** (0.06)	0.88 (0.12)	0.3 (0.28)	0.46 (0.14)
EBIC ($\gamma = 0$)	0.06 (0.04)	**0.78** (0.14)	0.92 (0.1)	0.81 (0.23)	0.3 (0.28)	0.46 (0.14)
EBIC ($\gamma = 0.5$)	0.1 (0.03)	0.44 (0.06)	**0.94** (0.06)	**0.99** (0.03)	0.3 (0.28)	0.46 (0.14)
EBIC ($\gamma = 1.0$)	0.1 (0.03)	0.39 (0.03)	**0.94** (0.06)	**0.99** (0.03)	0.3 (0.28)	0.46 (0.14)
AIC	0.06 (0.04)	0.24 (0.33)	0.35 (0.43)	0.44 (0.15)	0.3 (0.28)	0.46 (0.14)
Calinski-Harabasz Index	**0.14** (0.06)	0.54 (0.33)	0.57 (0.35)	0.76 (0.21)	0.59 (0.29)	0.66 (0.14)
CGL (ALC)	0.0 (0.01)	0.01 (0.01)	0.24 (0.18)	0.39 (0.0)	0.35 (0.08)	0.39 (0.0)
DPVC	0.0 (0.01)	0.06 (0.07)	0.29 (0.22)	0.44 (0.2)	NA	NA
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.1$					
Proposed ($\beta = 0.01$)	0.11 (0.02)	0.45 (0.05)	**0.88** (0.07)	0.79 (0.21)	**0.56** (0.34)	**0.64** (0.22)
Proposed ($\beta = 0.02$)	0.14 (0.04)	0.4 (0.02)	0.86 (0.07)	**0.9** (0.07)	**0.56** (0.34)	**0.64** (0.22)
Proposed ($\beta = 0.03$)	0.14 (0.04)	0.39 (0.03)	0.86 (0.07)	**0.9** (0.07)	**0.56** (0.34)	**0.64** (0.22)
basic inverse Wishart prior	0.13 (0.04)	0.52 (0.07)	**0.88** (0.07)	0.42 (0.33)	0.15 (0.19)	0.23 (0.19)
EBIC ($\gamma = 0$)	0.12 (0.06)	**0.7** (0.1)	0.78 (0.22)	0.42 (0.33)	0.15 (0.19)	0.16 (0.19)
EBIC ($\gamma = 0.5$)	0.13 (0.04)	0.44 (0.05)	**0.88** (0.07)	0.48 (0.26)	0.15 (0.19)	0.16 (0.19)
EBIC ($\gamma = 1.0$)	0.12 (0.05)	0.39 (0.03)	**0.88** (0.07)	0.6 (0.3)	0.15 (0.19)	0.16 (0.19)
AIC	0.12 (0.06)	0.2 (0.17)	0.06 (0.12)	0.42 (0.33)	0.15 (0.19)	0.16 (0.19)
Calinski-Harabasz Index	**0.17** (0.06)	0.48 (0.29)	0.28 (0.34)	**0.9** (0.07)	0.49 (0.27)	0.63 (0.22)
CGL (ALC)	0.01 (0.01)	0.07 (0.08)	0.31 (0.15)	0.39 (0.0)	0.33 (0.11)	0.38 (0.02)
DPVC	0.0 (0.0)	0.1 (0.09)	0.35 (0.12)	0.19 (0.18)	NA	NA

Table 2.5: Evaluation of clustering results for $d = 40$, $n \in \{20, 40, 400, 4000, 40000, 4000000\}$. Ground truth is 4 unbalanced clusters with sizes 20, 10, 5, 5. Shows the ANMI of the selected models (standard deviation in brackets). No noise is added.

	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise					
	20	40	400	4000	40000	4000000
Proposed ($\beta = 0.01$)	0.49 (0.15)	**0.84** (0.11)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.47 (0.17)	**0.84** (0.11)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.03$)	0.42 (0.19)	0.82 (0.13)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
basic inverse Wishart prior	0.5 (0.15)	**0.84** (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.2 (0.17)	0.8 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.24 (0.05)	0.37 (0.05)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.23 (0.06)	0.32 (0.04)	0.99 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.15 (0.19)	0.16 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski-Harabasz Index	0.17 (0.09)	0.17 (0.23)	0.46 (0.27)	0.45 (0.23)	0.47 (0.19)	0.4 (0.14)
CGL (ALC)	0.07 (0.11)	0.03 (0.04)	0.05 (0.07)	0.03 (0.03)	0.07 (0.07)	0.05 (0.06)
DPVC	**0.57** (0.13)	0.66 (0.07)	0.64 (0.14)	0.87 (0.17)	NA	NA
	$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise					
Proposed ($\beta = 0.01$)	0.15 (0.03)	0.33 (0.03)	0.87 (0.1)	0.98 (0.03)	**1.0** (0.0)	0.98 (0.03)
Proposed ($\beta = 0.02$)	0.15 (0.03)	0.33 (0.03)	0.87 (0.1)	0.97 (0.04)	**1.0** (0.0)	0.97 (0.04)
Proposed ($\beta = 0.03$)	0.16 (0.03)	0.31 (0.03)	0.67 (0.18)	0.97 (0.04)	0.98 (0.03)	0.97 (0.04)
basic inverse Wishart prior	**0.17** (0.05)	0.33 (0.02)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.08 (0.09)	**0.6** (0.23)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.16 (0.03)	0.33 (0.04)	0.98 (0.03)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 1.0$)	0.16 (0.03)	0.31 (0.03)	0.91 (0.12)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
AIC	0.08 (0.08)	0.52 (0.33)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Calinski-Harabasz Index	0.16 (0.06)	0.53 (0.3)	0.64 (0.15)	0.63 (0.28)	0.88 (0.17)	0.96 (0.08)
CGL (ALC)	0.0 (0.01)	0.0 (0.0)	0.0 (0.01)	0.15 (0.16)	0.15 (0.21)	0.12 (0.06)
DPVC	0.02 (0.01)	0.0 (0.04)	0.23 (0.14)	0.25 (0.13)	NA	NA

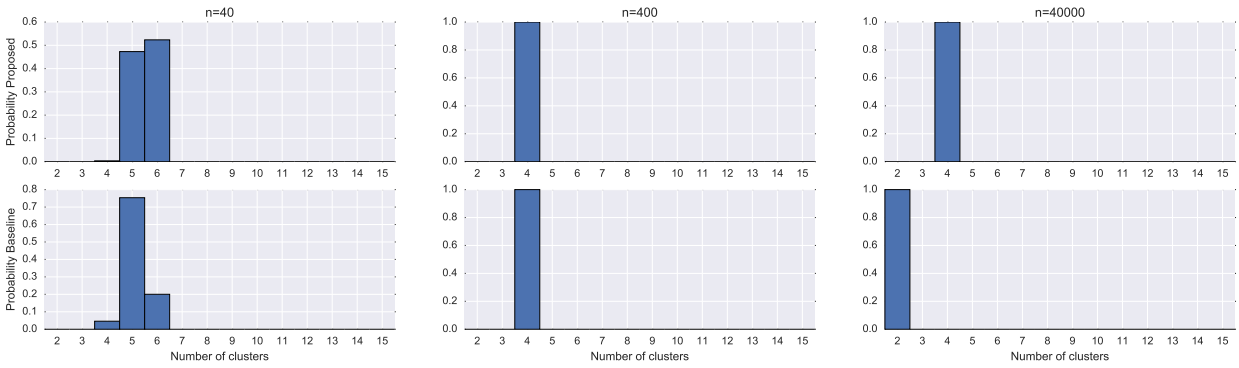

Figure 2.5: Posterior distribution of the number of clusters of the proposed method (top row) and the basic inverse Wishart prior model (bottom row). Ground truth is 4 clusters; noise was added to the precision matrix.

Table 2.6: Evaluation of clustering results with $d = 40$, $n \in \{20, 40, 400, 4000, 40000, 4000000\}$. Ground truth is 4 unbalanced clusters with sizes 20, 10, 5, 5. Shows the ANMI of the selected models (standard deviation in brackets). Noise is added to the precision matrix.

	$\Sigma_j \sim \text{InvW}(d_j+1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d+1, I_d), \eta = 0.01$					
	20	40	400	4000	40000	4000000
Proposed ($\beta = 0.01$)	0.45 (0.14)	0.75 (0.15)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed ($\beta = 0.02$)	0.39 (0.09)	0.75 (0.15)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)	0.98 (0.03)
Proposed ($\beta = 0.03$)	0.39 (0.09)	0.7 (0.18)	**1.0** (0.0)	0.97 (0.06)	**1.0** (0.0)	0.98 (0.03)
basic inverse Wishart prior	**0.48** (0.15)	**0.8** (0.09)	**1.0** (0.0)	0.91 (0.11)	0.39 (0.13)	0.42 (0.12)
EBIC ($\gamma = 0$)	0.12 (0.08)	0.67 (0.12)	**1.0** (0.0)	0.91 (0.11)	0.48 (0.17)	0.42 (0.12)
EBIC ($\gamma = 0.5$)	0.19 (0.08)	0.32 (0.04)	0.97 (0.03)	**1.0** (0.0)	0.54 (0.26)	0.42 (0.12)
EBIC ($\gamma = 1.0$)	0.17 (0.07)	0.28 (0.07)	0.96 (0.03)	**1.0** (0.0)	0.68 (0.24)	0.42 (0.12)
AIC	0.06 (0.09)	0.3 (0.34)	**1.0** (0.0)	0.4 (0.1)	0.39 (0.13)	0.42 (0.12)
Calinski-Harabasz Index	0.2 (0.06)	0.13 (0.2)	0.45 (0.27)	0.59 (0.17)	0.7 (0.21)	0.77 (0.03)
CGL (ALC)	0.08 (0.06)	0.05 (0.03)	0.04 (0.03)	0.03 (0.02)	0.03 (0.02)	0.04 (0.04)
DPVC	0.28 (0.04)	0.35 (0.07)	0.57 (0.08)	0.4 (0.12)	NA	NA
	$\Sigma_j \sim \text{InvW}(d_j+1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d+1, I_d), \eta = 0.1$					
Proposed ($\beta = 0.01$)	0.09 (0.11)	**0.42** (0.12)	0.84 (0.1)	0.42 (0.16)	0.18 (0.22)	0.24 (0.18)
Proposed ($\beta = 0.02$)	0.09 (0.11)	**0.42** (0.13)	0.88 (0.11)	0.85 (0.15)	**0.99** (0.02)	**0.9** (0.09)
Proposed ($\beta = 0.03$)	**0.15** (0.06)	**0.42** (0.13)	**0.89** (0.09)	**0.92** (0.07)	**0.99** (0.02)	**0.9** (0.09)
basic inverse Wishart prior	0.11 (0.14)	**0.42** (0.13)	0.84 (0.1)	0.2 (0.2)	0.0 (0.01)	0.1 (0.17)
EBIC ($\gamma = 0$)	0.04 (0.05)	0.24 (0.06)	0.88 (0.11)	0.2 (0.2)	0.0 (0.01)	0.1 (0.17)
EBIC ($\gamma = 0.5$)	0.05 (0.02)	0.19 (0.04)	0.74 (0.19)	0.44 (0.17)	0.0 (0.01)	0.1 (0.17)
EBIC ($\gamma = 1.0$)	0.05 (0.02)	0.19 (0.04)	0.41 (0.06)	0.78 (0.12)	0.0 (0.01)	0.1 (0.17)
AIC	0.0 (0.01)	0.15 (0.21)	0.19 (0.2)	0.2 (0.2)	0.0 (0.01)	0.1 (0.17)
Calinski-Harabasz Index	0.06 (0.03)	0.17 (0.11)	0.68 (0.25)	0.67 (0.2)	0.83 (0.17)	0.76 (0.04)
CGL (ALC)	0.04 (0.04)	0.03 (0.02)	0.05 (0.06)	0.1 (0.11)	0.05 (0.07)	0.08 (0.09)
DPVC	0.13 (0.05)	0.16 (0.05)	0.3 (0.13)	0.07 (0.03)	NA	NA
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.01$					
Proposed ($\beta = 0.01$)	0.11 (0.02)	0.32 (0.04)	0.74 (0.15)	**0.83** (0.1)	**0.59** (0.32)	0.5 (0.33)
Proposed ($\beta = 0.02$)	0.11 (0.02)	0.32 (0.04)	0.61 (0.17)	**0.83** (0.1)	**0.59** (0.32)	**0.59** (0.32)
Proposed ($\beta = 0.03$)	0.11 (0.02)	0.32 (0.04)	0.43 (0.06)	**0.83** (0.1)	**0.59** (0.32)	**0.59** (0.32)
basic inverse Wishart prior	0.11 (0.02)	0.32 (0.04)	**0.84** (0.05)	0.28 (0.0)	0.11 (0.14)	0.17 (0.23)
EBIC ($\gamma = 0$)	**0.18** (0.13)	**0.43** (0.05)	0.76 (0.13)	0.22 (0.12)	0.11 (0.14)	0.06 (0.11)
EBIC ($\gamma = 0.5$)	0.11 (0.02)	0.32 (0.04)	**0.84** (0.05)	0.51 (0.3)	0.11 (0.14)	0.06 (0.11)
EBIC ($\gamma = 1.0$)	0.11 (0.02)	0.32 (0.04)	0.79 (0.13)	0.67 (0.24)	0.11 (0.14)	0.06 (0.11)
AIC	0.14 (0.05)	0.16 (0.28)	0.17 (0.23)	0.22 (0.12)	0.09 (0.12)	0.06 (0.11)
Calinski-Harabasz Index	0.14 (0.08)	0.32 (0.3)	0.34 (0.33)	0.68 (0.22)	0.25 (0.27)	0.41 (0.32)
CGL (ALC)	0.0 (0.0)	0.0 (0.0)	0.01 (0.04)	0.0 (0.01)	0.02 (0.02)	0.01 (0.01)
DPVC	0.01 (0.01)	0.03 (0.06)	0.2 (0.05)	0.01 (0.02)	NA	NA
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.1$					
Proposed ($\beta = 0.01$)	0.1 (0.02)	0.34 (0.07)	0.68 (0.18)	0.6 (0.31)	0.09 (0.12)	0.06 (0.11)
Proposed ($\beta = 0.02$)	0.11 (0.02)	0.34 (0.07)	0.65 (0.21)	**0.7** (0.13)	0.21 (0.21)	0.28 (0.26)
Proposed ($\beta = 0.03$)	0.11 (0.02)	0.32 (0.06)	0.58 (0.2)	**0.7** (0.13)	**0.32** (0.22)	0.28 (0.26)
basic inverse Wishart prior	0.14 (0.03)	0.37 (0.08)	**0.78** (0.1)	0.0 (0.02)	0.09 (0.12)	0.06 (0.11)
EBIC ($\gamma = 0$)	**0.16** (0.05)	**0.49** (0.21)	0.71 (0.14)	0.0 (0.02)	0.09 (0.12)	0.06 (0.11)
EBIC ($\gamma = 0.5$)	0.11 (0.01)	0.36 (0.08)	0.77 (0.13)	0.06 (0.11)	0.09 (0.12)	0.06 (0.11)
EBIC ($\gamma = 1.0$)	0.11 (0.01)	0.31 (0.05)	0.7 (0.16)	0.12 (0.14)	0.09 (0.12)	0.06 (0.11)
AIC	0.15 (0.05)	0.05 (0.12)	0.06 (0.11)	0.0 (0.02)	0.09 (0.12)	0.06 (0.11)
Calinski-Harabasz Index	**0.16** (0.05)	0.29 (0.26)	0.42 (0.23)	0.45 (0.38)	0.09 (0.12)	**0.33** (0.31)
CGL (ALC)	0.0 (0.01)	0.0 (0.01)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.01)
DPVC	0.0 (0.04)	0.03 (0.05)	0.11 (0.13)	0.02 (0.03)	NA	NA

Table 2.7: Evaluation of clustering results for $d = 200$, $n \in \{50, 100, 150\}$. Ground truth is 4 balanced clusters. Shows the ANMI of the selected models (standard deviation in brackets).

	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise		
	50	100	150
Proposed ($\beta = 0.02$)	0.99 (0.01)	**1.0** (0.0)	**1.0** (0.0)
Inverse Wishart prior	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.93 (0.06)	0.99 (0.02)	**1.0** (0.0)
EBIC ($\gamma = 0.5$)	0.68 (0.06)	0.72 (0.02)	0.72 (0.03)
EBIC ($\gamma = 1.0$)	0.68 (0.06)	0.72 (0.02)	0.7 (0.02)
AIC	0.53 (0.09)	0.0 (0.0)	0.0 (0.0)
Calinski-Harabasz Index	0.98 (0.01)	**1.0** (0.0)	**1.0** (0.01)
CGL (ALC)	0.02 (0.02)	0.08 (0.04)	0.04 (0.06)
DPVC	0.61 (0.04)	0.63 (0.05)	0.7 (0.06)
	$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise		
Proposed ($\beta = 0.02$)	0.02 (0.01)	0.11 (0.06)	0.48 (0.01)
Inverse Wishart prior	0.0 (0.0)	0.11 (0.06)	0.48 (0.01)
EBIC ($\gamma = 0$)	0.02 (0.02)	0.0 (0.0)	0.48 (0.01)
EBIC ($\gamma = 0.5$)	0.02 (0.01)	0.11 (0.06)	0.48 (0.01)
EBIC ($\gamma = 1.0$)	0.02 (0.01)	0.11 (0.06)	0.48 (0.01)
AIC	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Calinski-Harabasz Index	**0.04** (0.02)	**0.15** (0.06)	**0.52** (0.27)
CGL (ALC)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
DPVC	0.0 (0.0)	0.01 (0.01)	0.01 (0.02)

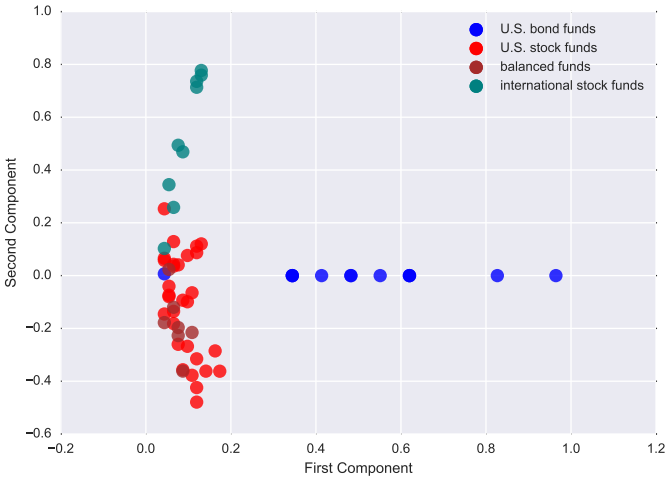

Figure 2.6: Two dimensional representation of the mutual funds data suggesting that balanced funds and U.S. stock funds are difficult to separate (one cluster), whereas U.S. bond funds and international stock funds appear to form mostly separate clusters.

Table 2.8: Evaluation of clustering results for $d = 200$, $n \in \{50, 100, 150\}$. Ground truth is 4 balanced clusters. Noise is added to the precision matrix. Shows the ANMI of the selected models (standard deviation in brackets).

	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d + 1, I_d), \eta = 0.01$		
	50	100	150
Proposed ($\beta = 0.02$)	0.77 (0.04)	**0.99** (0.01)	**1.0** (0.0)
Inverse Wishart prior	**0.8** (0.04)	**0.99** (0.01)	**1.0** (0.0)
EBIC ($\gamma = 0$)	0.73 (0.01)	0.73 (0.04)	0.86 (0.03)
EBIC ($\gamma = 0.5$)	0.6 (0.25)	0.72 (0.04)	0.72 (0.02)
EBIC ($\gamma = 1.0$)	0.6 (0.25)	0.72 (0.04)	0.72 (0.02)
AIC	0.54 (0.12)	0.0 (0.0)	0.0 (0.0)
Calinski-Harabasz Index	0.77 (0.05)	**0.99** (0.01)	0.99 (0.01)
CGL (ALC)	0.0 (0.0)	0.01 (0.02)	0.0 (0.0)
DPVC	0.22 (0.01)	0.35 (0.03)	0.38 (0.0)
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.01$		
Proposed ($\beta = 0.02$)	**0.03** (0.01)	0.13 (0.03)	0.48 (0.06)
Inverse Wishart prior	0.0 (0.0)	0.13 (0.03)	0.48 (0.06)
EBIC ($\gamma = 0$)	0.01 (0.01)	0.0 (0.0)	0.48 (0.07)
EBIC ($\gamma = 0.5$)	0.02 (0.02)	0.13 (0.03)	0.48 (0.06)
EBIC ($\gamma = 1.0$)	0.01 (0.02)	0.13 (0.03)	0.48 (0.06)
AIC	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Calinski-Harabasz Index	0.02 (0.02)	**0.21** (0.07)	**0.66** (0.09)
CGL (ALC)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
DPVC	0.0 (0.0)	0.0 (0.01)	0.03 (0.01)
	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d + 1, I_d), \eta = 0.1$		
Proposed ($\beta = 0.02$)	0.0 (0.0)	0.0 (0.0)	**0.12** (0.07)
Inverse Wishart prior	0.0 (0.0)	0.0 (0.0)	0.04 (0.06)
EBIC ($\gamma = 0$)	**0.06** (0.04)	0.06 (0.03)	0.09 (0.05)
EBIC ($\gamma = 0.5$)	**0.06** (0.04)	0.06 (0.03)	0.09 (0.05)
EBIC ($\gamma = 1.0$)	**0.06** (0.04)	0.06 (0.03)	0.09 (0.05)
AIC	0.03 (0.02)	0.01 (0.01)	0.0 (0.0)
Calinski-Harabasz Index	0.0 (0.0)	0.04 (0.04)	0.06 (0.06)
CGL (ALC)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
DPVC	0.03 (0.01)	**0.09** (0.03)	**0.12** (0.03)
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.1$		
Proposed ($\beta = 0.02$)	**0.02** (0.01)	0.11 (0.03)	0.42 (0.07)
Inverse Wishart prior	0.0 (0.0)	0.13 (0.03)	0.45 (0.07)
EBIC ($\gamma = 0$)	0.01 (0.01)	0.0 (0.0)	0.45 (0.07)
EBIC ($\gamma = 0.5$)	**0.02** (0.01)	0.13 (0.03)	0.45 (0.07)
EBIC ($\gamma = 1.0$)	**0.02** (0.01)	0.11 (0.03)	0.45 (0.07)
AIC	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Calinski-Harabasz Index	0.01 (0.01)	**0.14** (0.04)	**0.52** (0.27)
CGL (ALC)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
DPVC	0.0 (0.01)	0.0 (0.01)	0.01 (0.01)

Table 2.9: Comparison of variational and MCMC estimate. Evaluation of clustering results for $d = 12$, $n \in \{12, 120, 1200, 1200000\}$. Ground truth is 4 balanced clusters. $\beta = 0.02$. Shows the ANMI of the selected models (standard deviation in brackets).

	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j})$, no noise			
	12	120	1200	1200000
Proposed, variational	**0.39** (0.23)	**0.89** (0.09)	0.96 (0.07)	0.82 (0.11)
Proposed, MCMC	0.37 (0.23)	**0.89** (0.09)	0.96 (0.07)	0.9 (0.14)
basic inverse Wishart prior	**0.39** (0.23)	**0.89** (0.09)	**1.0** (0.0)	**1.0** (0.0)
	$\Sigma_j \sim \text{Uniform}_{d_j}$, no noise			
Proposed, variational	**0.76** (0.17)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
Proposed, MCMC	0.66 (0.1)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
basic inverse Wishart prior	**0.76** (0.17)	**1.0** (0.0)	**1.0** (0.0)	**1.0** (0.0)
	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d + 1, I_d), \eta = 0.01$			
Proposed, variational	**0.42** (0.27)	0.8 (0.16)	**1.0** (0.0)	**0.96** (0.07)
Proposed, MCMC	0.17 (0.24)	0.8 (0.16)	**1.0** (0.0)	**0.96** (0.07)
basic inverse Wishart prior	**0.42** (0.27)	**0.94** (0.12)	0.93 (0.13)	0.34 (0.04)
	$\Sigma_j \sim \text{InvW}(d_j + 1, I_{d_j}), \Sigma_\epsilon \sim \text{InvW}(d + 1, I_d), \eta = 0.1$			
Proposed, variational	0.11 (0.16)	0.57 (0.07)	0.55 (0.26)	**0.78** (0.2)
Proposed, MCMC	0.09 (0.06)	**0.61** (0.13)	**0.61** (0.23)	**0.78** (0.2)
basic inverse Wishart prior	**0.16** (0.15)	0.54 (0.1)	0.28 (0.15)	0.21 (0.18)
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.01$			
Proposed, variational	0.79 (0.12)	0.82 (0.26)	0.73 (0.33)	**0.96** (0.07)
Proposed, MCMC	**0.82** (0.11)	**0.96** (0.09)	**0.75** (0.31)	**0.96** (0.07)
basic inverse Wishart prior	0.79 (0.12)	0.48 (0.15)	0.28 (0.09)	0.28 (0.09)
	$\Sigma_j \sim \text{Uniform}_{d_j}, \Sigma_\epsilon \sim \text{Uniform}_d, \eta = 0.1$			
Proposed, variational	0.67 (0.22)	**0.24** (0.24)	0.32 (0.0)	**0.35** (0.18)
Proposed, MCMC	0.68 (0.17)	**0.24** (0.24)	**0.46** (0.27)	**0.35** (0.18)
basic inverse Wishart prior	**0.69** (0.21)	0.13 (0.11)	0.26 (0.13)	0.28 (0.09)

Table 2.10: Evaluation of selected clusterings of the mutual funds data. Colors highlight the type of fund. Numbers denote the cluster id assigned by the respective method. Here the size of the restricted hypotheses space $|\mathscr{C}^*|$ found by spectral clustering was 128.

Proposed and EBIC ($\gamma = 0.0$) [number of clusters = 6, ANMI = 0.48]	
U.S. bond funds	2 2 2 2 2 2 2 4 2 2 2 2 2
U.S. stock funds	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 5 1 4 6
balanced funds	1 1 1 1 1 1 1
international stock funds	1 3 1 1 3 1 3 3 1
basic inverse Wishart prior [number of clusters = 3, ANMI = 0.42]	
U.S. bond funds	2 2 2 2 2 2 2 2 2 2 2 2 2
U.S. stock funds	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1
balanced funds	1 1 1 1 1 1 1
international stock funds	1 1 1 1 1 1 1 1 1
EBIC ($\gamma = 0.5$) [number of clusters = 11, ANMI = 0.32]	
U.S. bond funds	2 9 2 9 2 2 2 1 10 9 2 2 2
U.S. stock funds	7 11 7 11 7 11 7 7 11 5 7 11 5 1 8 7 11 5 5 5 5 5 5 5 8 5 4 8 8 6
balanced funds	11 7 8 7 11 7 11
international stock funds	1 3 1 1 3 1 3 3 3
EBIC ($\gamma = 1.0$) [number of clusters = 14, ANMI = 0.25]	
U.S. bond funds	2 9 2 9 2 14 2 1 14 9 10 10 10
U.S. stock funds	12 8 12 6 12 8 12 12 8 6 12 8 6 3 11 6 8 5 7 5 5 5 5 6 11 5 11 15 4 11
balanced funds	8 12 1 12 8 6 7
international stock funds	3 13 3 3 13 3 13 13 13
AIC and Calinski-Harabasz Index [number of clusters = 2, ANMI = 0]	
U.S. bond funds	1 1 1 1 1 1 1 1 1 1 1 1 1
U.S. stock funds	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1
balanced funds	1 1 1 1 1 1 1
international stock funds	1 1 1 1 1 1 1 1 1
CGL (ALC) [number of clusters = 3, ANMI = 0.36]	
U.S. bond funds	1 1 1 1 1 1 1 3 1 1 1 1 1
U.S. stock funds	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3
balanced funds	2 2 2 2 3 2 2
international stock funds	2 2 2 2 2 2 2 3 2
DPVC [number of clusters = 2, ANMI = 0.35]	
U.S. bond funds	1 1 1 1 1 1 1 2 1 1 1 1 1
U.S. stock funds	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
balanced funds	2 2 2 2 2 2 2
international stock funds	2 2 2 2 2 2 2 2 2

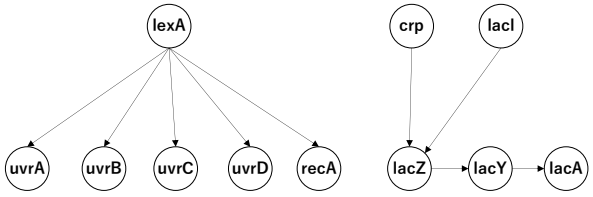

Figure 2.7: Gene regulations of E. coli. as given in (Hirose et al., 2017; Alberts et al., 2014) suggesting that the gene groups {lexA, uvrA, uvrB, uvrC, uvrD, recA} and {crp, lacl, lacZ, lacY, lacA} should be separated.

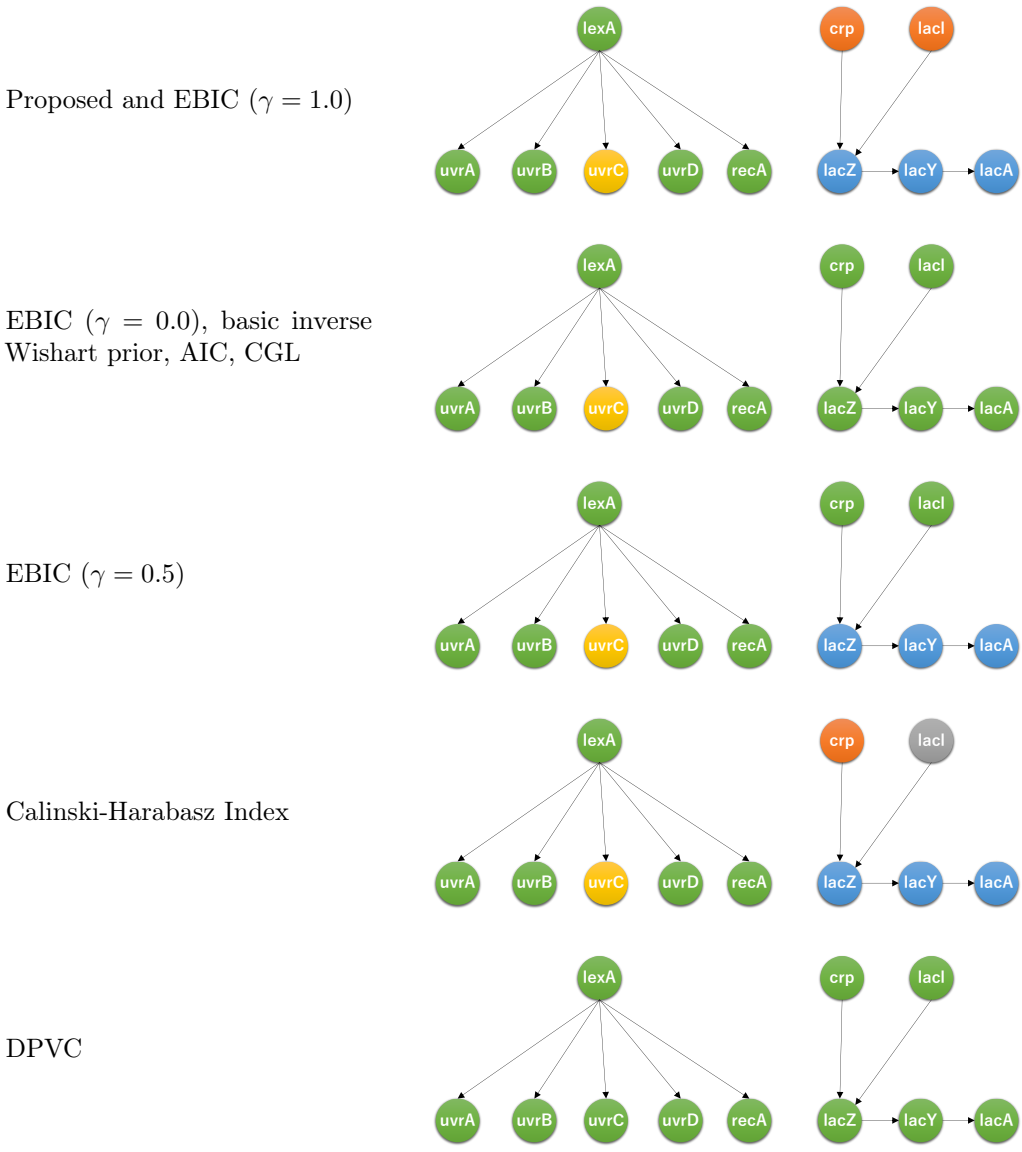

Figure 2.8: Clusterings of gene regulations network of E. coli. The clustering results are visualized by different colors. Here the size of the restricted hypotheses space $|\mathscr{C}^*|$ found by spectral clustering was 18. Only the proposed method, EBIC ($\gamma = 1.0$) and Calinski-Harabasz correctly divide the gene groups {lexA, uvrA, uvrB, uvrC, uvrD, recA} and {crp, lacI, lacZ, lacY, lacA}.

Table 2.11: Evaluation of selected clusterings of the Aviation Sensor Data with 16 variables. Here the size of the restricted hypotheses space $|\mathscr{C}^*|$ found by spectral clustering was 28.

	Proposed
Cluster 1	BRAKE PRESSURE LH YELLOW
Cluster 2	INDICATED ANGLE OF ATTACK, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1
Cluster 3	ROLL SPOILER RIGHT
Cluster 4	BRAKE PRESSURE RH GREEN
Cluster 5	RUDDER POSITION
Cluster 6	AILERON POSITION RH, AILERON POSITION LH
Cluster 7	ROLL SPOILER LEFT
Cluster 8	PITCH TRIM POSITION
Cluster 9	STATIC PRESSURE LSP, TOTAL PRESSURE LSP, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT,ELEVATOR POSITION RIGHT
	basic inverse Wishart prior, EBIC ($\gamma \in \{0.0, 0.5, 1.0\}$), AIC
Cluster 1	STATIC PRESSURE LSP, INDICATED ANGLE OF ATTACK, TOTAL PRESSURE LSP, RUDDER POSITION, AILERON POSITION RH, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT, ELEVATOR POSITION RIGHT, PITCH TRIM POSITION, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1, AILERON POSITION LH, ROLL SPOILER LEFT, BRAKE PRESSURE LH YELLOW, ROLL SPOILER RIGHT
Cluster 2	BRAKE PRESSURE RH GREEN
	Calinski-Harabasz Index
Cluster 1	STATIC PRESSURE LSP, TOTAL PRESSURE LSP, AILERON POSITION RH, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT, ELEVATOR POSITION RIGHT, BRAKE PRESSURE RH GREEN, AILERON POSITION LH, BRAKE PRESSURE LH YELLOW
Cluster 2	INDICATED ANGLE OF ATTACK, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1
Cluster 3	RUDDER POSITION, PITCH TRIM POSITION, ROLL SPOILER LEFT, ROLL SPOILER RIGHT
	CGL (ALC)
Cluster 1	STATIC PRESSURE LSP, TOTAL PRESSURE LSP, AVARAGE STATIC PRESSURE LSP, ELEVATOR POSITION LEFT, ELEVATOR POSITION RIGHT, BRAKE PRESSURE LH YELLOW
Cluster 2	INDICATED ANGLE OF ATTACK, RUDDER POSITION, AILERON POSITION RH, PITCH TRIM POSITION, BRAKE PRESSURE RH GREEN, ANGLE OF ATTACK 2, ANGLE OF ATTACK 1, AILERON POSITION LH, ROLL SPOILER LEFT, ROLL SPOILER RIGHT

Chapter 3

Disjunct Support Prior for Variable Selection in Regression

3.1 Introduction

In this chapter, we extend some of the ideas from the previous chapter to model selection under noise for variable selection in regression.

Sparseness of the regression coefficient vector is often a desirable property, since it (1) helps to improve interpretability, and (2) reduces the cost[1] of prediction. However, in practice, we may have to trade in a small reduction in prediction accuracy for an increase in sparseness. Spike-and-slab priors, as proposed by (Chipman et al., 2001), can potentially handle such a trade-off between prediction accuracy and sparseness. Though, manual setting of these priors is difficult, since they are either too restrictive, or depend on the unknown noise variance of the response variable. The limitations of these previous approaches are basically due to the desire for conjugate priors which results in closed-form solutions for the marginal likelihood.

Here, in this work, we propose a hierarchical spike-and-slab prior for the linear regression model that allows the user to explicitly specify the minimal magnitude δ of the regression coefficients that is considered practically significant. The proposed model decouples the response noise prior variance from the regression coefficients' prior variance, and thus making the threshold parameter δ more meaningful than previous work (Chipman et al., 2001). For example, δ can be set such that the Mean-Squared Error (MSE) of the prediction is only little influenced by ignoring covariates with coefficients' magnitude smaller than δ.

Our proposed method also resolves another subtle issue with previous spike-and-slab priors, namely inconsistent Bayes factors (BF). Due to the fact that the

[1] In case where acquiring the value of a covariate incurs a cost.

spike-and-slab priors of (Chipman et al., 2001) (and related work like (Ishwaran et al., 2005)) have full support, the Bayes factors of any two models is bounded in probability, for which we give a formal proof in Section 3.4. This is an undesirable property for Bayesian hypothesis testing, since we would like that the BF between the true and the wrong model grows with increasing sample size. In order to resolve this issue, our proposed method uses disjunct support priors, which allows us to guarantee consistent Bayes factors in the sense that the ratio of the true model's marginal likelihood to any other models' marginal likelihood converges to infinity for large sample sizes.

However, our choice of the prior does not enable the calculation of the marginal likelihood in closed-form anymore. Therefore, we introduce a latent variable indicator vector $\mathbf{z}$, and propose an efficient Gibbs sampler to sample from its posterior distribution. This allows us to estimate all model probabilities $p(S|\mathbf{y}, X, \delta)$, where S is a set of relevant variables.

The rest of this Chapter is organized as follows. In the next section, we summarize the properties of spike-and-slab priors from previous work. In Section 3.3, we introduce our model for variable selection based on disjunct support spike-and-slab priors. In Section 3.4, we prove that the disjunct support priors of our proposed method allows us to guarantee consistent Bayes factors. In Section 3.5, we explain our MCMC sampling strategy for estimating model probabilities. Since the elicitation of δ can be difficult, we discuss in Section 3.6 two strategies for determining δ: (1) bounding the increase in mean squared error (MSE) for prediction, and (2) estimating the expected MSE. We evaluate our proposed method on several synthetic data sets in Section 3.7, and real data sets in Section 3.8. Finally, we summarize our findings in Section 3.9.

3.2 Related work

To the best of our knowledge, the only Bayesian framework that allows to handle noise for variable selection are the spike-and-slab priors as proposed in (Chipman et al., 2001). The basic idea is to model the coefficients of the relevant and non-relevant variables by a normal distribution with variances σ_1^2 and σ_0^2, respectively, and $\sigma_1^2 \gg \sigma_0^2$. An example is shown in Figure 3.1.

The variance parameters σ_1^2 and σ_0^2 must be set manually. A difficulty of spike-and-slab priors is the correct setting of these parameters, and therefore (Ishwaran et al., 2005) proposed to place hyper-priors over these parameters in a such way that the resulting marginal prior $p(\boldsymbol{\beta})$ is little sensitive to the hyper-parameter choice. However, their prior choice does not allow for a closed-form marginal likelihood. Furthermore, their prior choice is only suitable for the situation where there is no noise, i.e. a variable j is considered to be relevant if and only if the true coefficient β_j is not zero.

In contrast, the spike and slab priors proposed in (Chipman et al., 2001) allow to specify practical significance (what we call here "relevance") by setting σ_1^2 to some large enough value (for example 100) and then set σ_0^2 such that the intersection points of the two priors occur at a pre-specified value δ (and $-\delta$),

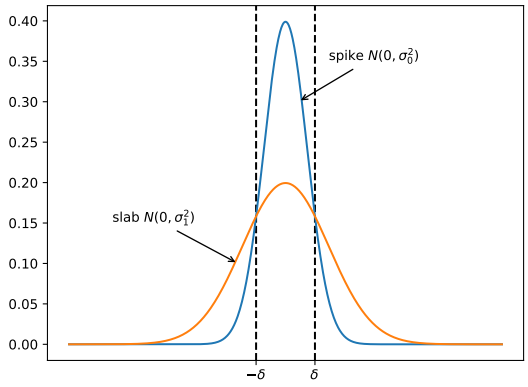

Figure 3.1: Example of spike and slab prior as proposed in (Chipman et al., 2001).

see Figure 3.1. However, their method has some drawbacks:

- Their conjugate prior formulation is sensitive to the prior for the response variance, whereas their non-conjugate formulation is not sensitive to the response variance, but has no closed-form solution anymore.
- For any $\delta > 0$, the Bayes factors are not consistent in the following sense. Let S be the true set of relevant variables and S' any other set, then we have

$$\frac{p(\mathbf{y}_n|X_n, S)}{p(\mathbf{y}_n|X_n, S')} \xrightarrow{P} O_p(1)\,,$$

  where $\mathbf{y}_n := (y_1, \dots y_n)$ and $X_n := (\mathbf{x}_1, \dots, \mathbf{x}_n)$, are the observed responses and covariates of n samples. This is due to the fact that the model dimension of spike-and-slab priors is the *same* for model S and S'. As a consequence, the influence of the prior can be ignored, in the sense that the influence of the prior is asymptotically the *same* for model S and S'. Since for both models $\boldsymbol{\beta}$ will concentrate around the true regression coefficient vector, the marginal likelihood cannot be distinguished any more. A formal proof, will be given in Section 3.4.
- It might be difficult to specify δ a-priori.

3.3 Proposed method

Let S be the indices of the selected covariates (i.e. the covariate that are considered to be relevant), and $\mathcal{C} := \{1, \dots, d\} \setminus \mathcal{S}$ the set of irrelevant covariates.

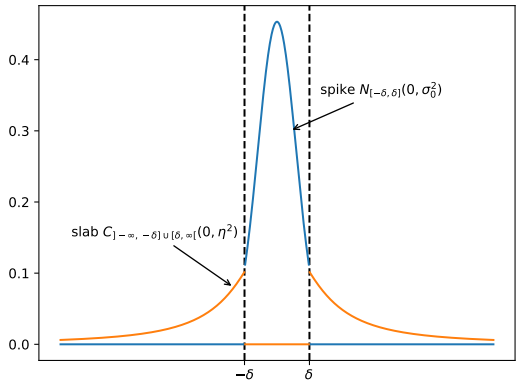

Figure 3.2: Illustration of the proposed spike and slab prior. $C_{]-\infty,-\delta]\cup[\delta,\infty[}(0,\eta^2)$ denotes the Cauchy distribution with mean 0 and scale η^2.

Furthermore, let $s := |\mathcal{S}|$ be the number of selected covariates. We consider the following linear model for $y \in \mathbb{R}$ regressed on $\mathbf{x} \in \mathbb{R}^d$:

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon\,,$$

where

$$\left.\begin{aligned} &\epsilon \sim N(0, \sigma_r^2)\,, \\ &\sigma_r^2 \sim \text{Inv-}\chi^2(\nu_r, \eta_r^2)\,, \end{aligned}\right\} \text{Prior for noise } \epsilon$$

$$\left.\begin{aligned} &s \sim \text{Multinomial}(p, \pi_{\text{rel}})\,, \\ &\pi_{\text{rel}} \sim \text{Beta}(1,1)\,, \end{aligned}\right\} \text{Prior for number of relevant covariates } s$$

$$\left.\begin{aligned} &\sigma_1^2 \sim \text{Inv-}\chi^2(\nu_1, \eta_1^2)\,, \\ &\text{for } j \in \{1, \ldots, d\}: \\ &\quad \text{if } j \in S, \text{ then} \\ &\quad\quad \beta_j \sim N_{]-\infty,-\delta]\cup[\delta,\infty[}(0, \sigma_1^2) \\ &\quad \text{else} \\ &\quad\quad \beta_j \sim N_{[-\delta,\delta]}(0, \sigma_0^2)\,. \end{aligned}\right\} \text{Prior for regression coefficients } \boldsymbol{\beta}$$

ν_r, η_r^2 are set such that Inv-$\chi^2(\nu_r, \eta_r^2)$ is a weakly informative prior. Inv-χ^2 denotes the scaled inverse chi-square distribution (see details below), where ν_r can be interpreted as the number of a-priori observations. For our experiments, we set ν_r and the prior variance σ_r^2 to 1.

$N_{[-\delta,\delta]}$ and $N_{]-\infty,-\delta]\cup[\delta,\infty[}$ denote the truncated normal distribution with support $[-\delta, \delta]$ and $]-\infty, -\delta] \cup [\delta, \infty[$ for the spike and slab prior, respectively. The specification of σ_0, and σ_1 determines the shape of the spike and slab prior,

respectively. For the slab prior, in order to allow for possibly large values of β_j, we place a diffuse hyper-prior on σ_1^2. In particular, we set $\nu_1 = 1$, and $\eta_1^2 = 100$ which corresponds to a truncated Cauchy distribution with mean zero and scale η_1^2 for $p(\beta_j | j \in S, \nu_1, \eta_1^2, \delta)$.

At the boundary $\beta_j = \delta$ (and, due to symmetry $\beta_j = -\delta$) we want to be indifferent about whether β_j was sampled from the spike or slab prior. Therefore, we set σ_0^2 such that

$$p(\beta_j = \delta | j \in S, \nu_1, \eta_1^2, \delta) = p(\beta_j = \delta | j \notin S, \sigma_0^2, \delta)\,. \tag{3.1}$$

The left hand side of Equation (3.1) does not have a closed-form solution. However, note that

$$p(\beta_j = \delta | j \in S, \nu_1, \eta_1^2, \delta) = \int N_{]-\infty,-\delta]\cup[\delta,\infty[}(\beta_j = \delta | 0, \sigma_1^2) \cdot \text{Inv-}\chi^2(\sigma_1^2 | \nu_1, \eta_1^2) d\sigma_1^2\,,$$

which we solve using numerical integration. Our proposed spike and slab prior is illustrated in Figure 3.2.

Therefore, the remaining critical hyper-parameter is only the specification of the threshold parameter δ. In Section 3.6, we discuss several methods for specifying δ.

Note that the prior on the number of relevant variables s ensures multiplicity control and has been extensively studied in (Scott et al., 2010; Scott and Berger, 2006). The probability of a variable being relevant π_{rel} can be integrated out leading to

$$p(s) = \frac{1}{d+1} \left(\begin{array}{c} d \\ s \end{array} \right)^{-1}\,. \tag{3.2}$$

Note that the scaled inverse chi-square distribution is defined as follows (see e.g. Gelman et al. (2013)):

$$\text{Inv-}\chi^2(\sigma^2 | \nu, \eta^2) = (\eta^2)^{\nu/2} \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} (\sigma^2)^{-(\frac{\nu}{2}+1)} e^{-\frac{1}{2\sigma^2}\nu\eta^2}\,.$$

Therefore, the joint probability density function is given by:

$$\begin{aligned}
p(\boldsymbol{\beta}, \sigma_r^2, \sigma_1^2, \mathbf{y}, S, |X) &= p(s) \cdot (2\pi)^{-\frac{n}{2}} \cdot (\sigma_r^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \\
&\quad \cdot (\eta_r^2)^{\nu_r/2} \frac{(\nu_r/2)^{\nu_r/2}}{\Gamma(\nu_r/2)} \cdot (\sigma_r^2)^{-(\frac{\nu_r}{2}+1)} e^{-\frac{1}{2\sigma_r^2}\nu_r\eta_r^2} \\
&\quad \cdot (\eta_1^2)^{\nu_1/2} \frac{(\nu_1/2)^{\nu_1/2}}{\Gamma(\nu_1/2)} \cdot (\sigma_1^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{1}{2\sigma_1^2}\nu_1\eta_1^2} \\
&\quad \cdot \Big(\prod_{j \in \mathcal{C}} \mathbb{1}_{\mathcal{N}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{N}, \sigma_0^2)} \cdot e^{-\frac{1}{2\sigma_0^2}\beta_j^2} \Big) \\
&\quad \cdot \Big(\prod_{j \in \mathcal{S}} \mathbb{1}_{\mathcal{R}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{R}, \sigma_1^2)} e^{-\frac{1}{2\sigma_1^2}\beta_j^2} \Big) \\
&= C_0 \cdot p(s) \cdot (\sigma_r^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \\
&\quad \cdot (\sigma_r^2)^{-(\frac{\nu_r}{2}+1)} e^{-\frac{1}{2\sigma_r^2}\nu_r\eta_r^2} \\
&\quad \cdot (\sigma_1^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{1}{2\sigma_1^2}\nu_1\eta_1^2} \\
&\quad \cdot \Big(\prod_{j \in \mathcal{C}} \mathbb{1}_{\mathcal{N}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{N}, \sigma_0^2)} \cdot e^{-\frac{1}{2\sigma_0^2}\beta_j^2} \Big) \\
&\quad \cdot \Big(\prod_{j \in \mathcal{S}} \mathbb{1}_{\mathcal{R}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{R}, \sigma_1^2)} e^{-\frac{1}{2\sigma_1^2}\beta_j^2} \Big),
\end{aligned}$$

where we defined $\mathcal{N} := [-\delta, \delta]$, and $\mathcal{R} :=]-\infty, -\delta] \cup [\delta, \infty[$, and

$$\iota(\mathcal{A}, \sigma^2) := \int \mathbb{1}_{\mathcal{A}}(x) e^{-\frac{1}{2\sigma^2}x^2} dx$$

and

$$C_0 := (2\pi)^{-\frac{n}{2}} \cdot (\eta_r^2)^{\nu_r/2} \frac{(\nu_r/2)^{\nu_r/2}}{\Gamma(\nu_r/2)} \cdot (\eta_1^2)^{\nu_1/2} \frac{(\nu_1/2)^{\nu_1/2}}{\Gamma(\nu_1/2)}.$$

3.4 Asymptotic Bayes factors

In this section, we formally prove the asymptotic behavior of the Bayes factors between the true model and any other model, first for our proposed method (Theorem 1), and then for previously proposed spike and slab priors (Theorem 2).

In the following, we define the true set of variables S as

$$S := \Big\{ j \in \{1, \dots, d\} \Big| \; | \beta_{j,t}| > \delta \Big\}.$$

Furthermore, we denote convergence in probability by $\xrightarrow{P}$.

Theorem 1. *Let S be the true set of relevant variables and S' any other set of variables. For the proposed method with disjunct support priors (as defined in Section 3.3), it holds that*

$$\frac{p(\mathbf{y}_n|X_n, S)}{p(\mathbf{y}_n|X_n, S')} \xrightarrow{P} \infty\,,$$

where $X_n := (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, are n samples drawn from a non-degenerated probability distribution $p(\mathbf{x})$ with finite covariance matrix, and $\mathbf{y}_n := (y_1, \ldots, y_n)$, where $y_i \sim p(y|\mathbf{x}_i, \sigma^2_{r,t}, \boldsymbol{\beta}_t)$, for some true parameters $\sigma^2_{r,t}$ and $\boldsymbol{\beta}_t$. We assume that $\boldsymbol{\beta}_t$ is not on the boundary of the support of the prior $p(\boldsymbol{\beta}|S)$.

We note that the convergence to infinity in Theorem 1 is exponentially fast in the number of samples n.

Proof. A general result in Bayesian hypothesis testing, as given in (Johnson and Rossell, 2010; Walker, 1969), states that the Bayes factor will converge exponentially fast favoring the alternative model, under the assumption that (1) the alternative model is true, (2) the support of the priors of the alternative and null model are disjunct, and (3) the models satisfy several regularity conditions. We show here that important regularity conditions are satisfied by our model, and complete the proof using some well-known asymptotic results.

First of all, let us do a change of variable using the one-to-one mapping $\tau := \sigma_r^{-2}$. For simplicity, let us denote $\boldsymbol{\theta} := (\tau, \boldsymbol{\beta})$, and the true parameter vector as $\boldsymbol{\theta}_t$.

Let us define the expected score function for a parameter vector $\boldsymbol{\theta}$ as

$$g(\boldsymbol{\theta}) := \mathbb{E}_{y\sim p(y|\boldsymbol{\theta},\mathbf{x})}\left[\log p(y|\boldsymbol{\theta}, \mathbf{x})\right].$$

First, we claim that the following function has a unique maximum

$$\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right],$$

where $\mathbf{x}$ is distributed according to some non-degenerated distribution with mean zero and positive definite covariance matrix C, and is such that we can exchange differentiation and integral.

We have

$$\log p(y|\boldsymbol{\theta}, \mathbf{x}) = \frac{1}{2}\log\tau - \frac{1}{2}\tau(y - \mathbf{x}^T\boldsymbol{\beta})^2 - \frac{1}{2}\log 2\pi\,,$$

and

$$\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right] = \mathbb{E}_{\mathbf{x},y}\left[\log p(y|\boldsymbol{\theta}, \mathbf{x})\right] \quad = \frac{1}{2}\log\tau - \frac{1}{2}\tau\,\mathbb{E}_{\mathbf{x},y}\left[(y - \mathbf{x}^T\boldsymbol{\beta})^2\right] - \frac{1}{2}\log 2\pi\,.$$

Since C is positive definite, we have that $\mathbb{E}_{\mathbf{x},y}\left[(y-\mathbf{x}^T\boldsymbol{\beta})^2\right]$ has a unique minimum at $\boldsymbol{\beta}=\boldsymbol{\beta}_t$. To see this note that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y}\left[(y-\mathbf{x}^T\boldsymbol{\beta})^2\right] &= \mathbb{E}_{\mathbf{x},\epsilon}\left[(\mathbf{x}^T\boldsymbol{\beta}_t+\epsilon-\mathbf{x}^T\boldsymbol{\beta})^2\right] \\
&= \mathbb{E}_{\mathbf{x},\epsilon}\left[\left(\mathbf{x}^T(\boldsymbol{\beta}_t-\boldsymbol{\beta})+\epsilon\right)^2\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\left(\mathbf{x}^T(\boldsymbol{\beta}_t-\boldsymbol{\beta})\right)^2\right] + 2\,\mathbb{E}_{\mathbf{x},\epsilon}\left[\epsilon\cdot\mathbf{x}^T(\boldsymbol{\beta}_t-\boldsymbol{\beta})\right] + E_\epsilon\left[\epsilon^2\right] \\
&= (\boldsymbol{\beta}_t-\boldsymbol{\beta})^T\,\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right](\boldsymbol{\beta}_t-\boldsymbol{\beta}) + 2\,\mathbb{E}_\epsilon\left[\epsilon\right]\cdot\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}^T\right](\boldsymbol{\beta}_t-\boldsymbol{\beta}) + E_\epsilon\left[\epsilon^2\right] \\
&= (\boldsymbol{\beta}_t-\boldsymbol{\beta})^T C(\boldsymbol{\beta}_t-\boldsymbol{\beta}) + \sigma_{r,t}^2\,.
\end{aligned}$$

where we used that $\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\right]=\mathbf{0}$, and $C=\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right]$. For $\boldsymbol{\beta}=\boldsymbol{\beta}_t$, we have $\mathbb{E}_{\mathbf{x},y}\left[(y-\mathbf{x}^T\boldsymbol{\beta})^2\right]=\frac{1}{\tau_t}$. Furthermore, since

$$\mathbb{E}_{\mathbf{x}}\left[g(\tau,\boldsymbol{\beta}_t)\right] = \frac{1}{2}\log\tau - \frac{1}{2}\tau\frac{1}{\tau_t} - \frac{1}{2}\log 2\pi$$

is strictly concave with respect to τ, with unique maximum at τ_r, we have that the unique maximum of $\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right]$ is given at $(\tau_t,\boldsymbol{\beta}_t)$. However, note that $\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right]$ is *not* jointly concave in τ and $\boldsymbol{\beta}$.

In detail, we have

$$\begin{aligned}
\frac{\partial}{\partial\tau}\log p(y|\boldsymbol{\theta},\mathbf{x}) &= \frac{1}{2}\frac{1}{\tau} - \frac{1}{2}(y-\mathbf{x}^T\boldsymbol{\beta})^2\,, \\
\frac{\partial^2}{\partial^2\tau}\log p(y|\boldsymbol{\theta},\mathbf{x}) &= -\frac{1}{2}\frac{1}{\tau^2}\,.
\end{aligned}$$

Furthermore, we have

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x}) &= \tau(y-\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}^T\,, \\
\frac{\partial^2}{\partial^2\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x}) &= -\tau\cdot\mathbf{x}\mathbf{x}^T\,, \\
\frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x}) &= (y-\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}^T\,.
\end{aligned}$$

And also note that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x})\right] &= \mathbb{E}_{\mathbf{x},y}\left[y\mathbf{x}^T-\boldsymbol{\beta}^T\mathbf{x}\mathbf{x}^T\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{y\sim p(y|\boldsymbol{\theta}_t,\mathbf{x})}\left[y\right]\mathbf{x}^T\right] - \boldsymbol{\beta}^T\,\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\boldsymbol{\beta}_t^T\mathbf{x}\mathbf{x}^T\right] - \boldsymbol{\beta}^T\,\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right] \\
&= (\boldsymbol{\beta}_t-\boldsymbol{\beta})^T C\,.
\end{aligned}$$

Therefore, the Schur complement of the Hessian of $\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right]$ is

$$
\begin{aligned}
w &:= \mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\tau}\log p(y|\boldsymbol{\theta},\mathbf{x})\right] \\
&\quad - \mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x})\right]\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x})\right]^{-1}\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x})\right]^T \\
&= -\frac{1}{2}\frac{1}{\tau^2} + \frac{1}{\tau}(\boldsymbol{\beta}_t-\boldsymbol{\beta})^T C C^{-1} C(\boldsymbol{\beta}_t-\boldsymbol{\beta}) \\
&= -\frac{1}{2}\frac{1}{\tau^2} + \frac{1}{\tau}(\boldsymbol{\beta}_t-\boldsymbol{\beta})^T C(\boldsymbol{\beta}_t-\boldsymbol{\beta})\,.
\end{aligned}
$$

Therefore, given that $\boldsymbol{\beta}$ is in a sufficiently small neighborhood around $\boldsymbol{\beta}_t$, we have that $w<0$. Combined with the fact that $\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\boldsymbol{\beta}}\log p(y|\boldsymbol{\theta},\mathbf{x})\right] = -\tau\cdot C \prec 0$ (definite negative), we have that $\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right]$ is locally concave around the true parameters $\boldsymbol{\theta}_t=(\tau_t,\boldsymbol{\beta}_t)$. In summary, we have[2]

$$
-\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\boldsymbol{\theta}}\log p(y|\boldsymbol{\theta}_t,\mathbf{x})\right] \succ 0\,.
$$

Asymptotic approximation of $p(\mathbf{y}_n|X_n,S)$ Let us define

$$
\hat{\boldsymbol{\theta}}_n := \underset{\boldsymbol{\theta}:p(\boldsymbol{\theta}|S)>0}{\arg\max}\; p(\mathbf{y}_n|X_n,\boldsymbol{\theta})\,.
$$

Then for large enough n, we have $\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_t$. Therefore, we have by the weak law of large numbers that

$$
J_n := -\frac{1}{n}\frac{\partial}{\partial^2\boldsymbol{\theta}}\log p(\mathbf{y}_n|X_n,\hat{\boldsymbol{\theta}}_n) \xrightarrow{P} -\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\boldsymbol{\theta}}\log p(y|\hat{\boldsymbol{\theta}}_n,\mathbf{x})\right],
$$

and by the continuous mapping theorem we have

$$
-\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\boldsymbol{\theta}}\log p(y|\hat{\boldsymbol{\theta}}_n,\mathbf{x})\right] \xrightarrow{P} -\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\boldsymbol{\theta}}\log p(y|\boldsymbol{\theta}_t,\mathbf{x})\right] \succ 0\,.
$$

That means, we have, in probability, for large enough n that $J_n \succ 0$. We can now follow the derivation of BIC (see e.g. (Ando, 2010), Chapter 8, pages 235, 236) to get

$$
\begin{aligned}
\log p(\mathbf{y}_n|X_n,S) &= \log p(\mathbf{y}_n|X_n,\hat{\boldsymbol{\theta}}_n) + \log p(\hat{\boldsymbol{\theta}}_n|S) - \frac{d+1}{2}\log n - \log|J_n| + O_p(1) \\
&= \log p(\mathbf{y}_n|X_n,\hat{\boldsymbol{\theta}}_n) - \frac{d+1}{2}\log n + O_p(1)\,. \qquad (3.3)
\end{aligned}
$$

A more detailed derivation of Equation (3.3) is given in Appendix B.2.

[2]The expression $\frac{\partial^2}{\partial^2\boldsymbol{\theta}}\log p(y|\boldsymbol{\theta}_t,\mathbf{x})$ denotes the second derivative of $\log p(y|\boldsymbol{\theta},\mathbf{x})$ evaluated at $\boldsymbol{\theta}_t$.

Upper bound for $p(\mathbf{y}_n|X_n, S')$ In the following, let us define

$$\hat{\boldsymbol{\theta}}_{S',n} := \underset{\boldsymbol{\theta}:p(\boldsymbol{\theta}|S')>0}{\arg\max}\, p(\mathbf{y}_n|X_n, \boldsymbol{\theta})$$

then we have

$$p(\mathbf{y}_n|X_n, S') = \int p(\mathbf{y}_n|X_n, \boldsymbol{\theta}) p(\boldsymbol{\theta}|S') d\boldsymbol{\theta} \leq p(\mathbf{y}_n|X_n, \hat{\boldsymbol{\theta}}_{S',n})\,. \tag{3.4}$$

Lower bound on $\log \frac{p(\mathbf{y}_n|X_n,S)}{p(\mathbf{y}_n|X_n,S')}$ Putting together the results from Equations (3.3) and (3.4), we get

$$\begin{aligned}
\log \frac{p(\mathbf{y}_n|X_n, S)}{p(\mathbf{y}_n|X_n, S')} &\geq \log p(\mathbf{y}_n|X_n, \hat{\boldsymbol{\theta}}_n) - \frac{d+1}{2}\log n + O_p(1) - \log p(\mathbf{y}_n|X_n, \hat{\boldsymbol{\theta}}_{S',n}) \\
&= n\Big(\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n) - \frac{1}{n}\sum_{i=1}^{n} \log p(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}_{S',n})\Big) - \frac{d+1}{2}\log n + O_p(1) \\
&\xrightarrow{P} n\Big(\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta}_t)\right] - \mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta}_{S'})\right]\Big) - \frac{d+1}{2}\log n + O_p(1)\,,
\end{aligned}$$

where $\boldsymbol{\theta}_{S'} := \arg\max_{\boldsymbol{\theta}:p(\boldsymbol{\theta}|S')>0} \mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right]$. Since $\boldsymbol{\theta}_t$ is the unique global maximizer of $\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right]$ and $p(\boldsymbol{\theta}_t|S') = 0$, we have that

$$c_\Delta := \mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta}_t)\right] - \mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta}_{S'})\right] > 0$$

and therefore

$$\log \frac{p(\mathbf{y}_n|X_n, S)}{p(\mathbf{y}_n|X_n, S')} \geq n \cdot c_\Delta - \frac{d+1}{2}\log n + O_p(1) \xrightarrow{P} \infty\,.$$

From the above line, we also see that the convergence of the Bayes factor $\frac{p(\mathbf{y}_n|X_n,S)}{p(\mathbf{y}_n|X_n,S')}$ is exponential in n. □

Next, let us investigate the Bayes factors for full support spike and slab priors, as for example in (Chipman et al., 2001; Ishwaran et al., 2005).

Theorem 2. *Under the same assumptions as in Theorem 1, but assuming full support spike and slab priors for the evaluation of the marginal likelihoods* $p(\mathbf{y}_n|X_n, S)$ *and* $p(\mathbf{y}_n|X_n, S')$*, we have the following result:*

$$\frac{p(\mathbf{y}_n|X_n, S)}{p(\mathbf{y}_n|X_n, S')} \xrightarrow{P} O_p(1)\,.$$

Proof. Since the priors have full support, the posterior distribution also has full support. Both posterior distributions contain the true regression coefficient vector $\boldsymbol{\beta}_t$, i.e.

$$p(\boldsymbol{\beta}_t|\mathbf{y}_n, X_n, S') > 0, \forall S'$$

Furthermore, since the likelihood function is the same as before in Theorem 1, we have, as was proven before, that $\mathbb{E}_{\mathbf{x}}\left[g(\boldsymbol{\theta})\right]$ has the unique maximizer $(\theta_t, \boldsymbol{\beta}_t)$ and is locally concave around this maximum. Therefore, the regularity conditions for the Bayesian central limit theorem are fulfilled for *all* models S', and we have:

$$\begin{aligned}\log p(\mathbf{y}_n|X_n, S') &\xrightarrow{P} \log p(\mathbf{y}_n|X_n, \hat{\boldsymbol{\theta}}_{S',n}) - \frac{d+1}{2}\log n + O_p(1) \\ &\xrightarrow{P} \log p(\mathbf{y}_n|X_n, \boldsymbol{\theta}_t) - \frac{d+1}{2}\log n + O_p(1)\,.\end{aligned}$$

And therefore

$$\log \frac{p(\mathbf{y}_n|X_n, S)}{p(\mathbf{y}_n|X_n, S')} \xrightarrow{P} O_p(1)\,.$$

□

3.5 Estimation of model probabilities

Calculating the marginal likelihood for each model explicitly is computationally challenging, due to the disjunct support priors on β:

- A Laplace approximation is not valid anymore, since the true parameter might not be contained in the support of the prior distribution.
- Chib's method (Chib, 1995; Chib and Jeliazkov, 2001) is computationally very expensive since, though we can sample, the normalization constants of each conditional probability is not available.

Instead, we estimate $p(S|\mathbf{y}, X)$, by introducing a model indicator vector $\mathbf{z} \in \{0,1\}^d$, where z_j indicates whether variable j should be included in S or not. We sample M samples from the posterior distribution of $\mathbf{z}$ by using the following MCMC algorithm:

Algorithm 3 Gibbs sampler for sampling from $p(\boldsymbol{z}|\sigma_0, \mathbf{y}, X)$.

for t from 1 to M **do**
  for j from 1 to d **do**
    $p(z_j) :=$ sample from $p(z_j|\boldsymbol{\beta}_{-j}, \mathbf{z}_{-j}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X)$
    $p(\beta_j) :=$ sample from $p(\beta_j|\boldsymbol{\beta}_{-j}, \mathbf{z}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X)$
  end for
  $\sigma_r^2 :=$ sample from $p(\sigma_r^2|\boldsymbol{\beta}, \mathbf{z}, \mathbf{y}, X)$
  $\sigma_1^2 :=$ sample from $p(\sigma_1^2|\boldsymbol{\beta}, \mathbf{z}, \mathbf{y}, X)$
end for

Sampling from each of the conditional distributions in Algorithm (3) is explained in the following. We note that all of the conditional distributions, except $p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{z}, \mathbf{y}, X)$, have an analytic solution that can be expressed by standard distributions. Therefore, we find that even for high-dimensional spaces, using Algorithm 3 is computationally feasible.

3.5.1 Analytic solution for $p(z_j|\boldsymbol{\beta}_{-j}, \mathbf{z}_{-j}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X)$

Let $\mathbf{x}_j$ denote the j-th column of X, and X_{-j} the matrix X where column j is removed. Then we have

$$||\mathbf{y} - X\boldsymbol{\beta}||_2^2 = ||\mathbf{y} - (\mathbf{x}_j\beta_j + X_{-j}\boldsymbol{\beta}_{-j})||_2^2 = ||\tilde{\mathbf{y}} - \mathbf{x}_j\beta_j||_2^2\,,$$

with $\tilde{\mathbf{y}} := \mathbf{y} - X_{-j}\boldsymbol{\beta}_{-j}$.

$$\begin{aligned}
p(z_j|\boldsymbol{\beta}_{-j}, \mathbf{z}_{-j}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X) &\propto \int p(\boldsymbol{\beta}, \mathbf{z}, \sigma_r, \sigma_1, \mathbf{y}|X, \sigma_0)d\beta_j \\
&= \int p(\mathbf{z}) \cdot C_0 \cdot (\sigma_r^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \\
&\quad \cdot (\sigma_r^2)^{-(\frac{\nu_r}{2}+1)} e^{-\frac{1}{2\sigma_r^2}\nu_r\eta_r^2} \\
&\quad \cdot \Big(\prod_{j\in\mathcal{C}} \mathbb{1}_{\mathcal{N}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{N}, \sigma_0^2)} \cdot e^{-\frac{1}{2\sigma_0^2}\beta_j^2}\Big) \\
&\quad \cdot \Big(\prod_{j\in\mathcal{S}} \mathbb{1}_{\mathcal{R}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{R}, \sigma_1^2)} e^{-\frac{1}{2\sigma_1^2}\beta_j^2}\Big) d\beta_j \\
&\propto \frac{p(\mathbf{z})}{\iota(\mathcal{A}_{z_j}, \sigma_{z_j}^2)} \cdot \int e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) \cdot e^{-\frac{1}{2\sigma_{z_j}^2}\beta_j^2} d\beta_j \\
&= \frac{p(\mathbf{z})}{\iota(\mathcal{A}_{z_j}, \sigma_{z_j}^2)} \cdot \int e^{-\frac{1}{2\sigma_r^2}(||\tilde{\mathbf{y}}||_2^2 - 2\tilde{\mathbf{y}}^T\mathbf{x}_j\beta_j + ||\mathbf{x}_j||_2^2\beta_j^2)} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) \cdot e^{-\frac{1}{2\sigma_{z_j}^2}\beta_j^2} d\beta_j \\
&\propto \frac{p(\mathbf{z})}{\iota(\mathcal{A}_{z_j}, \sigma_{z_j}^2)} \cdot \int e^{-\frac{1}{2\sigma_r^2}(-2\tilde{\mathbf{y}}^T\mathbf{x}_j\beta_j + (||\mathbf{x}_j||_2^2 + \frac{\sigma_r^2}{\sigma_{z_j}^2})\beta_j^2)} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) d\beta_j \\
&= \frac{p(\mathbf{z})}{\iota(\mathcal{A}_{z_j}, \sigma_{z_j}^2)} \cdot \int e^{-\frac{1}{2\tilde{\sigma}^2}(\beta_j - \tilde{\mu})^2} e^{\frac{\tilde{\mu}}{2\sigma_r^2}\tilde{\mathbf{y}}^T\mathbf{x}_j} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) d\beta_j \\
&= p(\mathbf{z}) \cdot e^{\frac{\tilde{\mu}}{2\sigma_r^2}\tilde{\mathbf{y}}^T\mathbf{x}_j} \cdot \frac{\iota(\mathcal{A}_{z_j}, \tilde{\mu}, \tilde{\sigma}^2)}{\iota(\mathcal{A}_{z_j}, \sigma_{z_j}^2)}\,,
\end{aligned}$$

where $\tilde{\mu} := \frac{\tilde{\mathbf{y}}^T\mathbf{x}_j}{||\mathbf{x}_j||_2^2 + \frac{\sigma_r^2}{\sigma_{z_j}^2}}$, and $\tilde{\sigma}^2 := (\frac{1}{\sigma_r^2}||\mathbf{x}_j||_2^2 + \frac{1}{\sigma_{z_j}^2})^{-1}$, and $\iota(\mathcal{A}_{z_j}, \tilde{\mu}, \tilde{\sigma}^2)$ is the normalization constant of a truncated normal distribution given by

$$\iota(\mathcal{A}_{z_j}, \tilde{\mu}, \tilde{\sigma}^2) := \int e^{-\frac{1}{2\tilde{\sigma}^2}(\beta_j - \tilde{\mu})^2} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) d\beta_j\,.$$

Case $\delta = 0$.

In the case, where $\delta = 0$, some care is needed. First, consider $z_j = 1$, then we can proceed as before

$$p(z_j = 1|\beta_{-j}, \sigma_r, \sigma_0, \sigma_1, \mathbf{y}, X, \mathbf{z}_{-j}) = c \cdot p(\mathbf{z}) \cdot e^{\frac{\tilde{\mu}}{2\sigma_r^2}\tilde{\mathbf{y}}^T\mathbf{x}_j} \cdot \frac{\iota(\mathbb{R}, \tilde{\mu}, \tilde{\sigma}^2)}{\iota(\mathbb{R}, \sigma_1^2)}\,,$$

where c is a normalization constant. Second, for $z_j = 0$, the prior $p(\beta_j)$ is a Dirac measure with 1 at position 0, and otherwise 0. Therefore, we can use the same calculation as before, but replacing β_j by 0. This way, we get

$$p(z_j = 0|\beta_{-j}, \sigma_r, \sigma_0, \sigma_1, \mathbf{y}, X, \mathbf{z}_{-j}) = c \cdot p(\mathbf{z})\,.$$

Note that in both cases, we can integrate over β_j, and therefore the reversible jump MCMC methodology (Green, 1995; Green and Hastie, 2009) is not necessary here.

3.5.2 Analytic solution for $p(\beta_j|\boldsymbol{\beta}_{-j}, \mathbf{z}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X)$

For $\delta > 0$, we have

$$\begin{aligned}
p(\beta_j|\boldsymbol{\beta}_{-j}, \mathbf{z}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X) &\propto p(\boldsymbol{\beta}, \mathbf{z}, \sigma_r, \sigma_1, \mathbf{y}|X, \sigma_0) \\
&= p(\mathbf{z}) \cdot C_0 \cdot (\sigma_r^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \\
&\quad \cdot (\sigma_r^2)^{-(\frac{\nu_r}{2}+1)} e^{-\frac{1}{2\sigma_r^2}\nu_r\eta_r^2} \\
&\quad \cdot \Big(\prod_{j\in\mathcal{C}} \mathbb{1}_{\mathcal{N}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{N}, \sigma_0^2)} \cdot e^{-\frac{1}{2\sigma_0^2}\beta_j^2}\Big) \\
&\quad \cdot \Big(\prod_{j\in\mathcal{S}} \mathbb{1}_{\mathcal{R}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{R}, \sigma_1^2)} e^{-\frac{1}{2\sigma_1^2}\beta_j^2}\Big) \\
&\propto e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) \cdot e^{-\frac{1}{2\sigma_{\tilde{z}_j}^2}\beta_j^2} \\
&= e^{-\frac{1}{2\sigma_r^2}||\tilde{\mathbf{y}}-\mathbf{x}_j\beta_j||_2^2} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) \cdot e^{-\frac{1}{2\sigma_{\tilde{z}_j}^2}\beta_j^2} \\
&= e^{-\frac{1}{2\sigma_r^2}(||\tilde{\mathbf{y}}||_2^2 - 2\tilde{\mathbf{y}}^T\mathbf{x}_j\beta_j + ||\mathbf{x}_j||_2^2\beta_j^2)} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) \cdot e^{-\frac{1}{2\sigma_{\tilde{z}_j}^2}\beta_j^2} \\
&\propto e^{-\frac{1}{2\sigma_r^2}(-2\tilde{\mathbf{y}}^T\mathbf{x}_j\beta_j + (||\mathbf{x}_j||_2^2 + \frac{\sigma_r^2}{\sigma_{\tilde{z}_j}^2})\beta_j^2)} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) \\
&= e^{-\frac{1}{2\tilde{\sigma}^2}(\beta_j-\tilde{\mu})^2} e^{\frac{\tilde{\mu}}{2\sigma_r^2}\tilde{\mathbf{y}}^T\mathbf{x}_j} \cdot \mathbb{1}_{\mathcal{A}_{z_j}}(\beta_j) \\
&\propto N_{\mathcal{A}_{z_j}}(\beta_j|\tilde{\mu}, \tilde{\sigma}^2)\,.
\end{aligned}$$

Note that if $\delta = 0$, then

$$p(\beta_j|\boldsymbol{\beta}_{-j}, \mathbf{z}, \sigma_r, \sigma_1, \sigma_0, \mathbf{y}, X) = \begin{cases} N(\beta_j|\tilde{\mu}, \tilde{\sigma}^2) & \text{if } z_j = 1\,, \\ 1_{\{0\}}(\beta_j) & \text{if } z_j = 0\,. \end{cases}$$

3.5.3 Analytic solution for $p(\sigma_r^2|\boldsymbol{\beta}, \mathbf{z}, \mathbf{y}, X)$

For the conditional posterior $p(\sigma_r^2|\boldsymbol{\beta}, \mathbf{z}, \mathbf{y}, X)$, we have a closed form solution given by

$$
\begin{aligned}
p(\sigma_r^2|\boldsymbol{\beta}, \mathbf{y}, \mathbf{z}, X) &\propto p(\boldsymbol{\beta}, \sigma_r, \mathbf{y}, \mathbf{z}|X) \\
&= p(\mathbf{z}) \cdot C_0 \cdot (\sigma_r^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \\
&\quad \cdot (\sigma_r^2)^{-(\frac{\nu_r}{2}+1)} e^{-\frac{1}{2\sigma_r^2}\nu_r\eta_r^2} \\
&\quad \cdot \Big(\prod_{j \in \mathcal{C}} \mathbb{1}_{\mathcal{N}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{N}, \sigma_0^2)} \cdot e^{-\frac{1}{2\sigma_0^2}\beta_j^2} \Big) \\
&\quad \cdot \Big(\prod_{j \in \mathcal{S}} \mathbb{1}_{\mathcal{R}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{R}, \sigma_1^2)} e^{-\frac{1}{2\sigma_1^2}\beta_j^2} \Big) \\
&\propto (\sigma_r^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \cdot (\sigma_r^2)^{-(\frac{\nu_r}{2}+1)} e^{-\frac{1}{2\sigma_r^2}\nu_r\eta_r^2} \\
&\propto (\sigma_r^2)^{-(\frac{\nu_r+n}{2}+1)} e^{-\frac{1}{2\sigma_r^2}(||\mathbf{y}-X\boldsymbol{\beta}||_2^2+\nu_r\eta_r^2)} \\
&\propto (\sigma_r^2)^{-(\frac{\nu_r+n}{2}+1)} e^{-\frac{1}{2\sigma_r^2}(\nu_r+n)\frac{||\mathbf{y}-X\boldsymbol{\beta}||_2^2+\nu_r\eta_r^2}{\nu_r+n}} \\
&\propto \text{Inv-}\chi^2(\sigma_r^2 \,|\, \nu_r + n, \frac{||\mathbf{y} - X\boldsymbol{\beta}||_2^2 + \nu_r\eta_r^2}{\nu_r + n}) \,.
\end{aligned}
$$

3.5.4 Sampling from $p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{z}, \mathbf{y}, X)$

For sampling from $p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{z}, \mathbf{y}, X)$, we employ a Slice sampler as described in the following. First note that

$$
\begin{aligned}
p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{y}, \mathbf{z}, X) &\propto p(\sigma_1^2, \boldsymbol{\beta}, \sigma_r^2, \mathbf{y}, \mathbf{z}|X) \\
&= p(\mathbf{z}) \cdot C_0 \cdot (\sigma_r^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma_r^2}||\mathbf{y}-X\boldsymbol{\beta}||_2^2} \\
&\quad \cdot (\sigma_r^2)^{-(\frac{\nu_r}{2}+1)} e^{-\frac{1}{2\sigma_r^2}\nu_r\eta_r^2} \\
&\quad \cdot \Big(\prod_{j \in \mathcal{C}} \mathbb{1}_{\mathcal{N}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{N}, \sigma_0^2)} \cdot e^{-\frac{1}{2\sigma_0^2}\beta_j^2} \Big) \\
&\quad \cdot \Big(\prod_{j \in \mathcal{S}} \mathbb{1}_{\mathcal{R}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{R}, \sigma_1^2)} e^{-\frac{1}{2\sigma_1^2}\beta_j^2} \Big) \\
&\quad \cdot (\sigma_1^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{1}{2\sigma_1^2}\nu_1\eta_1^2} \\
&\propto \Big(\prod_{j \in \mathcal{S}} \mathbb{1}_{\mathcal{R}}(\beta_j) \cdot \frac{1}{\iota(\mathcal{R}, \sigma_1^2)} \cdot e^{-\frac{1}{2\sigma_1^2}\beta_j^2} \Big) \\
&\quad \cdot (\sigma_1^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{1}{2\sigma_1^2}\nu_1\eta_1^2} \\
&\propto \frac{1}{\iota(\mathcal{R}, \sigma_1^2)^s} \cdot (\sigma_1^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{1}{2\sigma_1^2}(\nu_1\eta_1^2 + \sum_{j \in \mathcal{S}} \beta_j^2)} \,.
\end{aligned}
$$

If $\sigma_1^2 >> 1$, and $\delta << 1$, we have *approximately* that

$$\iota(\mathcal{R}, \sigma_1^2) \propto \iota(\mathbb{R}, \sigma_1^2) = (2\pi\sigma_1^2)^{\frac{1}{2}}\,, \tag{3.5}$$

and we have *exactly* (not approximately) that

$$\begin{aligned}
p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{y}, \mathbf{z}, X) &\propto \Big(\frac{(2\pi\sigma_1^2)^{\frac{s}{2}}}{\iota(\mathcal{R}, \sigma_1^2)^s}\Big) \cdot (2\pi\sigma_1^2)^{-\frac{s}{2}} \cdot \Big((\sigma_1^2)^{-(\frac{\nu_1}{2}+1)} e^{-\frac{1}{2\sigma_1^2}(\nu_1\eta_1^2 + \sum_{j\in\mathcal{S}}\beta_j^2)}\Big) \\
&\propto \Big(\frac{(2\pi\sigma_1^2)^{\frac{s}{2}}}{\iota(\mathcal{R}, \sigma_1^2)^s}\Big) \cdot \Big((\sigma_1^2)^{-(\frac{\nu_1+s}{2}+1)} e^{-\frac{1}{2\sigma_1^2}(\nu_1\eta_1^2 + \sum_{j\in\mathcal{S}}\beta_j^2)}\Big) \\
&\propto \Big(\frac{(2\pi\sigma_1^2)^{\frac{s}{2}}}{\iota(\mathcal{R}, \sigma_1^2)^s}\Big) \cdot \Big((\sigma_1^2)^{-(\frac{\nu_1+s}{2}+1)} e^{-\frac{1}{2\sigma_1^2}(\nu_1+s)\frac{(\nu_1\eta_1^2 + \sum_{j\in\mathcal{S}}\beta_j^2)}{\nu_1+s}}\Big) \\
&\propto \Big(\frac{(2\pi\sigma_1^2)^{\frac{s}{2}}}{\iota(\mathcal{R}, \sigma_1^2)^s}\Big) \cdot \text{Inv-}\chi^2(\nu_1 + s, \frac{(\nu_1\eta_1^2 + \sum_{j\in\mathcal{S}}\beta_j^2)}{\nu_1 + s}\Big)\,.
\end{aligned}$$

That means we have that

$$p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{y}, X, \mathcal{S}) \propto h(\sigma_1^2) \cdot \text{Inv-}\chi^2(\sigma_1^2 \,|\, \tilde{\nu}, \tilde{\eta}^2)\,,$$

for $\tilde{\nu} := \nu_1 + s$, $\tilde{\eta}^2 := \frac{(\nu_1\eta_1^2 + \sum_{j\in\mathcal{S}}\beta_j^2)}{\nu_1+s}$, and the function $h(\sigma_1^2) := \frac{(2\pi\sigma_1^2)^{\frac{s}{2}}}{\iota(\mathcal{R}, \sigma_1^2)^s}$ is changing slowly with σ_1^2. Therefore, we use a slice sampler (see e.g. Carlin and Louis (2008)) as follows. We start from the (approximate) mode given by $\sigma_1^2 := \frac{\tilde{\nu}\tilde{\eta}^2}{\tilde{\nu}+2}$, and then run the following two steps, until we retain a sample in the second step:[3]

1. Sample $U \sim \text{Uniform}([0, h(\sigma_1^2)])$.
2. Sample $\sigma_1^2 \sim \text{Inv-}\chi^2(\tilde{\nu}, \tilde{\eta}^2)$, and retain the sample if $U < h(\sigma_1^2)$.

Note that the sampling scheme is guaranteed to sample exactly from $p(\sigma_1^2|\boldsymbol{\beta}, \sigma_r^2, \mathbf{y}, X, \mathcal{S})$, independently of how well the approximation $h(\sigma_1^2) \propto 1$ holds. The correctness of the sampling scheme is shown in Appendix B.1. However, of course, the efficiency (whether we accept the sample in step 2) will depend on the closeness of the approximation in Equation (3.5). In practice, we observe that the sampling method is efficient if s is small. In detail, for several settings, for $s = 1$, and $s = 10$, the lowest acceptance rates were around 97% and 67%, respectively, where we tested $\sum_{j\in\mathcal{S}}\beta_j^2 \in \{0.1, 1.0, 10.0, 100.0\}$, and $\delta = \{0.8, 0.05, 0.001\}$.

3.6 Specification of δ

In some situations, where prior knowledge is given in the form of similar regression tasks from the past, it might possible to directly elicit a suitable threshold value δ. However, such prior knowledge might not be available, and therefore, we consider here two methods to specify δ.

[3] We assume that we started in a high probability region, and therefore use a burn-in of only 10.

3.6.1 Bounding influence on Mean Squared Error

The influence of ignoring small magnitude coefficients on the response variable can be bounded according to Theorem 3.

Theorem 3. *The increase in mean squared error (MSE) of the model selected by ignoring regression coefficients in $[-\delta, \delta]$ is upper bounded by*

$$d\delta^2 \lambda_{max} \,,$$

where λ_{max} is the largest eigenvalue of the covariance matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ (assuming x is centered).

Proof. Let $\boldsymbol{\beta} \in \mathbb{R}^d$ be the true regression coefficient vector and $\boldsymbol{\beta}^\delta \in \mathbb{R}^d$, be the thresholded true regression coefficient vector with $\boldsymbol{\beta}_i^\delta = 0$, if $\boldsymbol{\beta}_i \in [-\delta, \delta]$. The mean squared error when using $\boldsymbol{\beta}^\delta$ is given by

$$\begin{aligned}
\mathbb{E}_{y,\mathbf{x}}\left[(y - \mathbf{x}^T\boldsymbol{\beta}^\delta)^2\right] &= \mathbb{E}_{\epsilon,\mathbf{x}}\left[(\mathbf{x}^T\boldsymbol{\beta} + \epsilon - \mathbf{x}^T\boldsymbol{\beta}^\delta)^2\right] \\
&= \mathbb{E}_{\epsilon,\mathbf{x}}\left[(\mathbf{x}^T(\boldsymbol{\beta} - \boldsymbol{\beta}^\delta) + \epsilon)^2\right] \\
&= \mathbb{E}_{\epsilon,\mathbf{x}}\left[(\boldsymbol{\beta} - \boldsymbol{\beta}^\delta)^T\mathbf{x}\mathbf{x}^T(\boldsymbol{\beta} - \boldsymbol{\beta}^\delta) + 2\epsilon \cdot \mathbf{x}^T(\boldsymbol{\beta} - \boldsymbol{\beta}^\delta) + \epsilon^2\right] \\
&= (\boldsymbol{\beta} - \boldsymbol{\beta}^\delta)^T\,\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right](\boldsymbol{\beta} - \boldsymbol{\beta}^\delta) + \sigma_r^2 \,,
\end{aligned}$$

where we used that $\epsilon \sim N(0, \sigma_r^2)$. Next, note that

$$||\boldsymbol{\beta} - \boldsymbol{\beta}^\delta||_2^2 \leq d\delta^2 \,,$$

and

$$\begin{aligned}
\max_{||\mathbf{z}||_2^2 \leq d\delta^2} \mathbf{z}^T\,\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{z} &= \max_{||(d\delta^2)^{-\frac{1}{2}}\mathbf{z}||_2 \leq 1} \mathbf{z}^T\,\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{z} \\
&= d\delta^2 \max_{||\mathbf{z}||_2 \leq 1} \mathbf{z}^T\,\mathbb{E}_{\mathbf{x}}\left[\mathbf{x}\mathbf{x}^T\right]\mathbf{z} \\
&= d\delta^2 \lambda_{max} \,,
\end{aligned}$$

where λ_{max} is the largest eigenvalue of $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$. Therefore,

$$\mathbb{E}_{y,\mathbf{x}}\left[(y - \mathbf{x}^T\boldsymbol{\beta}^\delta)^2\right] \leq d\delta^2\lambda_{max} + \sigma_r^2 \,.$$

Since $\mathbb{E}_{y,\mathbf{x}}\left[(y - \mathbf{x}^T\boldsymbol{\beta})^2\right] = \sigma_r^2$, we have the desired result. □

3.6.2 Estimating expected increase of Mean Squared Error

The bound given in Theorem 3 can be too conservative. Furthermore, it only makes a statement in absolute terms of increase in MSE. However, often we are

interested in statements like "the selected model (with few variables) increases the mean squared error by no more than 5% when compared to the best model that can use all variables." (Piironen and Vehtari, 2017; Hahn and Carvalho, 2015).

For the "best model" we use the Bayesian model averaged (BMA) regression model. The BMA regression model is often considered the gold standard due to its good theoretic and practical performance (Fernandez et al., 2001; Piironen and Vehtari, 2017). The BMA model for the prediction of a new datapoint $(\tilde{y}, \tilde{\mathbf{x}})$ is defined as

$$p(\tilde{y}|\tilde{\mathbf{x}}) = \sum_{\mathbf{z}} \int p(\tilde{y}|\tilde{\mathbf{x}}, \mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y}, X) d\boldsymbol{\theta}\,,$$

where $\boldsymbol{\theta}$ denotes all parameters. The BMA model is a meta-model since it still requires the specification of the model for $p(\mathbf{z}, \boldsymbol{\theta}, y|X)$. Here, we use for $p(\mathbf{z}, \boldsymbol{\theta}, y|X)$, our proposed model with $\delta = 0$.

The expected mean squared error of BMA is therefore given by

$$\mathrm{MSE}_{\mathrm{bma}} := \mathbb{E}_{\mathbf{z}}[\mathbb{E}_{\sigma_r^2}[\sigma_r^2|\mathbf{z}, \mathbf{y}, X]|\mathbf{y}, X]\,,$$

which we estimate from the samples of our MCMC algorithm in Algorithm 3.

Given a threshold δ^*, and the best subset of variables specified by $\mathbf{z}^*$, we estimate the MSE as follows

$$\mathrm{MSE}_{\delta^*} := \mathbb{E}_{\sigma_r^2}[\sigma_r^2|\mathbf{z}^*, \mathbf{y}, X_{|\mathbf{z}^*}, \delta^*]\,,$$

where $X_{|\mathbf{z}^*}$ means that only the covariates index by $\mathbf{z}^*$ are used, where

$$\mathbf{z}^* := \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{y}, X, \delta^*)\,. \tag{3.6}$$

We can now estimate for each threshold δ the expected increase in MSE when compared to $\mathrm{MSE}_{\mathrm{bma}}$, i.e.:

$$\text{expected increase in MSE} = \frac{\mathrm{MSE}_{\delta^*}}{\mathrm{MSE}_{\mathrm{bma}}} - 1.0\,. \tag{3.7}$$

We then select the most parsimonious model that has an expected increase in MSE of less than 5%. We note that similar strategies for predictive model selection have been proposed in (Piironen and Vehtari, 2017; Hahn and Carvalho, 2015), though, their models are different from ours, and they do not make use of $p(\mathbf{z}|\mathbf{y}, X, \delta^*)$ as in Equation (3.6).

3.7 Evaluation on synthetic data

We study two settings, the low dimensional setting with $d < n$ and the high dimensional setting with $d \geq n$.

For the low dimensional experiments, we use the same regression setting as in (Tibshirani, 1996), namely the regression coefficient vector is set to

$$\boldsymbol{\beta}^T = (\mathbf{3}, \mathbf{1.5}, 0.0, 0.0, \mathbf{2.0}, 0.0, 0.0, 0.0)^T ,$$

and the response noise is set to $\sigma_r = 3.0$. For each sample, we draw a covariate vector $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = 0.5^{|i-j|}$. The number of samples is varied from $n = 10$ to $n = 100000$.

For the high dimensional experiments, we use the same setting as in (Ročková and George, 2014), with $d = 1000$ and $n \in \{100, 1000\}$, where the first three covariate are set to 3 ,2, and 1, and all others are set to zero. The covariate vector is drawn from $\mathbf{x} \sim N(\mathbf{0}, \Sigma)$, where $\Sigma_{ij} = 0.6^{|i-j|}$.

Furthermore, in the noise setting, we replace each zero entry of the original regression coefficient vector by a value sampled from Uniform($[-\eta, \eta]$), where $\eta \in \{0.2, 0.5\}$. In particular, when $\eta = 0.5$, the new regression coefficient vector for the low dimensional experiment becomes

$$\boldsymbol{\beta}^T = (\mathbf{3}, \mathbf{1.5}, -0.12, -0.35, \mathbf{2.0}, 0.16, 0.26, -0.01)^T ,$$

where the relevant variables are marked by bold font. The expected increase in mean squared error (MSE) for choosing the parsimonious model without the noise coefficients is about 0.4% and 2.8%, for $\eta = 0.2$, and $\eta = 0.5$, respectively.

In the high dimensional noise setting, we replace only 1% of the original zero entries (following the largest entries 3,2,1). This leads to an expected increase in mean squared error for choosing the parsimonious model of about 3.1% for $\eta = 0.2$.[4]

We show the results for $\delta \in \{0.8, 0.5, 0.05, 0.01, 0.001, 0.0\}$, and the results of the most parsimonious model that is estimated to lead to an increase in MSE of not more than 5%. For all methods based on MCMC we use 10000 samples, out of which 10% are used for burn in.

As our first baseline, we use the robust objective prior proposed in (Bayarri et al., 2012) together with a Gibbs sampler to explore the space of models, which we denote as "GibbsBvs".[5] Furthermore, we use the spike-and-slab prior and EM-algorithm as proposed in (Ročková and George, 2014) which we denote as "EMVS".[6]

The above two methods cannot account for negligible noise on the coefficient vectors. Therefore, we introduce another baseline using the horseshoe prior (Carvalho et al., 2010) as follows.[7] First, using the horseshoe prior, we estimate the mean coefficient vector $\boldsymbol{\beta}$ and the mean response variance $\sigma^2_{r,full}$ for the full model. Then, for each δ, we hard threshold $\boldsymbol{\beta}$, and this way get a model candidate $\mathbf{z}_\delta$. Finally, using again the horseshoe prior for the linear regression model but

[4] For the high-dimensional setting we do not consider $\eta = 0.5$, since this would correspond to an expected increase of in MSE of 19.0%.

[5] Implemented in the R package 'BayesVarSel'. As suggested by the authors, we use the g-Zellner prior (Zellner, 1986) in cases where the robust prior from (Bayarri et al., 2012) fails.

[6] Implemented in the R package 'EMVS'.

[7] Implemented in the R package 'horseshoe'.

reduced to the covariates $\mathbf{z}_\delta$, we estimate the mean response variance $\sigma^2_{r,\mathbf{z}_\delta}$, and then select the most parsimonious model that has lower expected increase in MSE than 5%. To estimate the expected increase in MSE, we use Formula (3.7), where we replace MSE_{δ^*} and MSE_{bma} by $\sigma^2_{r,\mathbf{z}_\delta}$ and $\sigma^2_{r,full}$, respectively.

Finally, we include also three frequentist methods for model search. As a first frequentist method, we use the popular Least Angle Regression (LARS) method (Efron et al., 2004) to get a set of candidate models. We then select the model using the Extended Bayesian information criterion (EBIC) with $\gamma \in \{0, 0.5, 1\}$ (Chen and Chen, 2008; Foygel and Drton, 2010), or the Akaike information criterion (AIC) (Akaike, 1973). Note that EBIC with $\gamma = 0$, is equal to the Bayesian information criterion (BIC) (Schwarz, 1978). As a third frequentist method, we use linear regression with Lasso (Tibshirani, 1996) combined with stability selection (Meinshausen and Bühlmann, 2010). Stability selection has two hyper-parameters that need to be specified: the "upper bound for the per-family error rate" (PFER) and "the number of (unique) selected variables" (denoted by q) as in the R package 'stabs'. For PFER we set always 1. However, we found that stability selection can be sensitive to the choice of q, and therefore show all results for three different values.

We evaluate all methods in terms of F1-Score. All experiments are repeated 5 times and we report average and standard deviations (shown in brackets). For large n, GibbsBvs did not execute correctly, which we mark as "-". For the high-dimensional setting GibbsBvs did not finish computation due to high memory requirements. For the proposed method, we select the threshold value δ, as described in Section 3.6.2, and for the horseshoe prior method as described in the previous paragraph. We refer to this as "automatic". If not reported otherwise, we use for all baselines the default settings.

Low dimensional setting The results for the low dimensional setting, with and without noise, are shown in Tables 3.1, 3.2 and 3.3. Overall, we see that the proposed method and the horseshoe prior method perform best.

GibbsBvs, EBIC and Stability selection (with $q \geq 4$) perform good for no noise or small noise. However, for $\eta = 0.5$, GibbsBvs, EBIC and Stability Selection start to select more irrelevant variables with increasing sample size n. Asymptotically, all three methods are expected to select all variables with coefficient regressions $\beta_j \neq 0$, no matter how small β_j is. However, if the sample size is small ($n \leq 100$), then all three methods are not influenced by the noise, i.e. they ignore the negligible small regression coefficients.

AIC performs similar to EBIC for $n \leq 100$, but for larger sample sizes it tends to select too many variables, even in the no-noise setting. This is not too surprising, since it is well known that AIC is not model selection consistent (see e.g. (Yang, 2005)).

Interestingly, in the noise setting ($\eta = 0.2$ and $\eta = 0.5$), even for large n, EMVS finds the correct relevant variables. However, for small sample sizes it tends to select too few variables. This suggests that EMVS has a strong inductive bias for sparse models, which can be helpful in the noise setting, but

is deteriorating performance for small to medium-sized n.

High dimensional setting The results for the high dimensional setting, with and without noise, are shown in Tables 3.7, and 3.8. Overall, we see that the proposed method, Stability selection (with $q \geq 50$) and EMVS perform best. In this setting the EMVS seems to profit from its inductive bias for sparse models. On the other hand, the horseshoe prior method performs somehow unsatisfactory, tending to select too many variables. AIC and EBIC performed very poorly in this setting, selecting too many variables. One reason seems to stem from the numerical instability of the maximum likelihood estimate for $d \leq n$. As an ad-hoc remedy we tried to combine it with a ridge estimate, but this did not seem to help.

Analysis of different δ In Tables 3.4, 3.5, and 3.6, we show the results for different fixed δ in the low-dimensional setting, and in Tables 3.9 and 3.10 for the high-dimensional setting. The proposed method is less sensitive to the choice of δ and tends to select sparse models even in the high-dimensional setting. However, as expected, the horseshoe prior method is highly sensitive to the choice of δ.

3.8 Evaluation on real data

In this section, we compare the results of our proposed and all baselines on three real data sets: crime data (Raftery et al., 1997; Liang et al., 2008), ozone data (Garcia-Donato and Martinez-Beneito, 2013), and GDP growth data (SDM) (Sala-i Martin et al., 2004). Details of the data sets are in Table 3.11; all variables are described in Tables 3.12, 3.13 and Tables 3.14 and 3.15 . In order to make the choice of all hyper-parameters invariant to the scale, we normalize the observations to have roughly the same scale as for the synthetic data set. In detail, we normalize the covariates to have mean 0 and variance 1, and the response variable to have mean 0 and variance 30. Furthermore, we log-transform the crime data as in (Liang et al., 2008).

For the experiments with the real data we use 100000 MCMC-samples for the proposed method, GibbsBvs, and the horseshoe prior.[8] Concerning the stability selection method, based on our findings from the simulated data, we set q to the values $\{0.1 \cdot d, 0.5 \cdot d, 0.8 \cdot d\}$.

The results for ozone, crime and SDM are shown in Tables 3.16, 3.17 and 3.18, respectively.

We see that the horseshoe method and EMVS perform similar as for the simulated data. The horseshoe prior with thresholding, finds models with relatively many variables, whereas EMVS tends to select models with only very few variables. In particular, EMVS suggests that in the ozone and SDM data, none of the variables are relevant, which is quite a strong statement that is contradicting the results from all other methods. The stability selection method

[8] Out of which 10% are used for burn in.

appears to select too few variables, independent of the setting of q. We note that for SDM, for $q = 0.8 \cdot d$, the stability selection method did not terminate correctly. The results for EBIC highlight the sensitivity to the hyper-parameter γ.

Our proposed method shows similar results to GibbsBvs, except for SDM. For SDM, our proposed model suggests that only EAST and MALFAL66 have relatively high regression coefficients, but our method also shows that the expected increase in mean-squared error is around 27% when compared to the Bayesian averaged model that uses all variables. For ozone, our model suggests that the model using x6.x7, x6.x8, x7.x7, and x6.x6, have relatively high regression coefficients, but not all of them are together in one model, possibly due to high correlation. For crime, our model suggests that all variables should be considered as relevant, whereas in particular M, Ed, Po1, Ineq have high regression coefficients.

To further analyze the results of our proposed method, we show the top 10 model probabilities and variable inclusion probabilities calculated for $\delta = 0$ and $\delta = 0.5$. The model probabilities for ozone, crime and SDM are shown in Tables 3.19, 3.21, and 3.23, respectively. Considering the low model probabilities, it is clear that there is no clearly winning model, and that care is needed when drawing conclusions from only the top model.

In order to investigate the importance of each individual variable, we also show the variable inclusion probabilities for ozone, crime and SDM in Tables 3.20, 3.22, and 3.24, respectively. In each of the Tables, we also show the results that were reported in previous studies. From the difference in the probabilities between previous studies, $\delta = 0$ and $\delta = 0.5$, we can draw some interesting conclusions.

Ozone data In Table 3.20, we show the inclusion probabilities of the proposed method together with the results reported in (Garcia-Donato and Martinez-Beneito, 2013). Comparing those results to the result of the proposed method, we find that the discrepancy between the results is not large, except in two cases: First, the importance of the variable x9, including its interaction terms, is much higher in (Garcia-Donato and Martinez-Beneito, 2013). Second, the squared term x7.x7 is considered as relevant by the proposed method, even when $\delta = 0.5$, which is in contrast to (Garcia-Donato and Martinez-Beneito, 2013), where an inclusion probability of only 45% is reported. Comparing the proposed method between $\delta = 0.0$ and $\delta = 0.5$, we see that the interaction variable x6.x8 is the most likely to be included for $\delta = 0.0$, with probability around 70%. However, looking at the result with $\delta = 0.5$, the effect size of x7.x7 is likely to be larger than x6.x8.

Crime data In Table 3.22, we show the inclusion probabilities of the proposed method together with the results reported in (Liang et al., 2008). For the proposed method with $\delta = 0$, we see good agreement with the results in (Liang et al., 2008). This is in particular true with respect to the median probability

model that includes all variables with probability larger than or equal to 0.5. However, inspecting the inclusion probabilities for $\delta = 0.5$, there is not enough evidence that the effect size of Po2 and U2 is high.

GDP growth data (SDM) In Table 3.24, we show the inclusion probabilities of the proposed method together with the results reported in (Sala-i Martin et al., 2004). We see that all the top 18 variables that have been considered as significant by (Sala-i Martin et al., 2004) are also listed in the top 18 of the proposed method ($\delta = 0$). Moreover, the results of the proposed method with $\delta = 0.5$, suggest, that among those 18 variables, only 7 variables have a probability of more than 20% of having a high effect size. In particular, it appears that DENS65C (density of costal population) seems to have only marginal influence on economic growth.

3.9 Conclusions

We proposed a new type of spike-and-slab prior that is particularly well suited for the situation where there are small negligible, but non-zero regression coefficients. These small negligible regression coefficients are considered as noise, since they can lead to the selection of overly complex models (i.e. models with many variables), although, only few variables should be considered as practically relevant. For that purpose, we introduced a disjunct support prior with a threshold parameter $\delta > 0$ in order to ignore small coefficients. We proved that for fixed δ, the proposed method leads to consistent Bayes factors, which is not the case for full support priors as proposed in (Chipman et al., 2001).

Due to the non-conjugacy of the priors proposed by our method, estimating the marginal likelihood explicitly is computationally infeasible. We therefore introduced a latent variable indicator vector $\mathbf{z}$, and proposed an efficient Gibbs sampler to sample from its posterior distribution. This allows us to estimate all model probabilities $p(S|\mathbf{y}, X, \delta)$, where S is a set of relevant variables and δ is a threshold parameter specifying practical relevance (effect size).

Since it can sometimes be difficult to specify δ explicitly, we showed how to estimate the mean squared error (MSE) of the final model selected for a specific δ. This way, for example, we can select the most parsimonious model that has MSE that is not worse more than 5% of a reference model. As a reference model, we suggest to use the Bayesian model averaged model that uses all variables.

For synthetic data with ground truth, we showed that the proposed method leads to good model selection performance in various settings: with/without noise and low/high dimensions. Furthermore, we compared our method to a commonly used spike-and-slab prior (Chipman et al., 2001; Ročková and George, 2014) (EMVS), Gibbs sampling for the objective prior as proposed in (Bayarri et al., 2012) (GibbsBvs), thresholding of the mean coefficient vector $\boldsymbol{\beta}$ estimated with the horseshoe prior (Carvalho et al., 2010), and three frequentist methods LARS + EBIC (Efron et al., 2004; Chen and Chen, 2008), LARS + AIC (Efron et al., 2004; Akaike, 1973) and Lasso + stability selection (Tibshirani, 1996;

Meinshausen and Bühlmann, 2010). The proposed method was always at par with the best previously proposed method which was varying between EMVS, GibbsBvs, stability selection and horseshoe prior with thresholding.

Finally, we evaluated our method also on three real data sets. Concerning the number of selected variables of our proposed method and the previous methods, we observed a similar behavior as for the synthetic data set. While for $\delta = 0$, our proposed method seems to roughly agree with various previous methods, the inspection of the results for $\delta = 0.5$, allowed us to draw conclusions about the practical relevance of some of the selected variables.

Table 3.1: Low-dimensional setting, $d = 8$ and $n \in \{10, 50, 100, 1000, 100000\}$. Evaluation results with no noise on regression coefficients.

F1-Scores					
	10	50	100	1000	100000
proposed (automatic)	0.51 (0.03)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
GibbsBvs	0.53 (0.02)	0.92 (0.1)	0.92 (0.1)	1.0 (0.0)	-
EMVS	0.5 (0.0)	0.16 (0.32)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)
horseshoe (automatic)	0.53 (0.07)	0.91 (0.18)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)
AIC	0.49 (0.07)	0.91 (0.07)	0.85 (0.13)	0.87 (0.11)	0.86 (0.08)
EBIC ($\gamma = 0.0$)	0.49 (0.07)	0.97 (0.06)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)
EBIC ($\gamma = 0.5$)	0.54 (0.04)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
EBIC ($\gamma = 1.0$)	0.53 (0.03)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
stability ($q = 1$)	0.3 (0.24)	0.2 (0.24)	0.3 (0.24)	0.5 (0.0)	0.5 (0.0)
stability ($q = 4$)	0.2 (0.24)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
stability ($q = 6$)	0.1 (0.2)	0.96 (0.08)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)
Average number of selected variables					
	10	50	100	1000	100000
proposed (automatic)	1.6 (1.2)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
GibbsBvs	5.2 (3.43)	2.6 (0.49)	3.6 (0.8)	3.0 (0.0)	-
EMVS	1.0 (0.0)	0.4 (0.8)	0.0 (0.0)	3.0 (0.0)	3.0 (0.0)
horseshoe (automatic)	5.8 (2.4)	4.0 (2.0)	3.6 (0.8)	3.0 (0.0)	3.0 (0.0)
AIC	6.6 (2.33)	3.6 (0.49)	4.2 (1.17)	4.0 (0.89)	4.0 (0.63)
EBIC ($\gamma = 0.0$)	6.6 (2.33)	3.2 (0.4)	3.6 (0.8)	3.0 (0.0)	3.0 (0.0)
EBIC ($\gamma = 0.5$)	5.0 (3.29)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
EBIC ($\gamma = 1.0$)	4.4 (3.14)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
stability ($q = 1$)	0.6 (0.49)	0.4 (0.49)	0.6 (0.49)	1.0 (0.0)	1.0 (0.0)
stability ($q = 4$)	0.4 (0.49)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
stability ($q = 6$)	0.2 (0.4)	2.8 (0.4)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)

Table 3.2: Low-dimensional setting, $d = 8$ and $n \in \{10, 50, 100, 1000, 100000\}$. Evaluation results with noise on regression coefficients $\eta = 0.2$.

F1-Scores					
	10	50	100	1000	100000
proposed (automatic)	0.5 (0.06)	0.97 (0.06)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
GibbsBvs	0.51 (0.06)	0.97 (0.06)	0.97 (0.06)	1.0 (0.0)	-
EMVS	0.42 (0.21)	0.1 (0.2)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)
horseshoe (automatic)	0.64 (0.19)	0.91 (0.07)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
AIC	0.6 (0.1)	0.91 (0.07)	0.91 (0.07)	0.88 (0.12)	0.58 (0.03)
EBIC ($\gamma = 0.0$)	0.6 (0.1)	0.94 (0.07)	0.97 (0.06)	1.0 (0.0)	0.63 (0.03)
EBIC ($\gamma = 0.5$)	0.61 (0.11)	0.9 (0.13)	1.0 (0.0)	1.0 (0.0)	0.63 (0.03)
EBIC ($\gamma = 1.0$)	0.61 (0.11)	0.9 (0.13)	0.96 (0.08)	1.0 (0.0)	0.65 (0.03)
stability ($q = 1$)	0.2 (0.24)	0.2 (0.24)	0.5 (0.0)	0.5 (0.0)	0.5 (0.0)
stability ($q = 4$)	0.2 (0.24)	0.8 (0.22)	0.96 (0.08)	1.0 (0.0)	0.91 (0.07)
stability ($q = 6$)	0.0 (0.0)	0.83 (0.18)	1.0 (0.0)	1.0 (0.0)	0.67 (0.0)
Average number of selected variables					
	10	50	100	1000	100000
proposed (automatic)	3.2 (2.4)	3.2 (0.4)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
GibbsBvs	5.4 (3.2)	3.2 (0.4)	3.2 (0.4)	3.0 (0.0)	-
EMVS	2.0 (2.53)	0.2 (0.4)	0.0 (0.0)	3.0 (0.0)	3.0 (0.0)
horseshoe (automatic)	4.6 (2.15)	3.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
AIC	6.8 (2.4)	3.6 (0.49)	3.6 (0.49)	4.0 (1.1)	7.4 (0.49)
EBIC ($\gamma = 0.0$)	6.8 (2.4)	3.4 (0.49)	3.2 (0.4)	3.0 (0.0)	6.6 (0.49)
EBIC ($\gamma = 0.5$)	4.4 (3.01)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)	6.6 (0.49)
EBIC ($\gamma = 1.0$)	4.4 (3.01)	3.2 (0.4)	2.8 (0.4)	3.0 (0.0)	6.2 (0.4)
stability ($q = 1$)	0.4 (0.49)	0.4 (0.49)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
stability ($q = 4$)	0.4 (0.49)	2.4 (0.49)	2.8 (0.4)	3.0 (0.0)	3.6 (0.49)
stability ($q = 6$)	0.0 (0.0)	2.6 (1.02)	3.0 (0.0)	3.0 (0.0)	6.0 (0.0)

Table 3.3: Low-dimensional setting, $d = 8$ and $n \in \{10, 50, 100, 1000, 100000\}$. Evaluation results with noise on regression coefficients $\eta = 0.5$.

F1-Scores					
	10	50	100	1000	100000
proposed (automatic)	0.55 (0.07)	0.97 (0.06)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
GibbsBvs	0.52 (0.07)	0.94 (0.07)	0.94 (0.07)	0.75 (0.0)	-
EMVS	0.42 (0.21)	0.1 (0.2)	0.0 (0.0)	1.0 (0.0)	1.0 (0.0)
horseshoe (automatic)	0.65 (0.17)	0.91 (0.07)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)
AIC	0.6 (0.1)	0.86 (0.0)	0.89 (0.06)	0.62 (0.05)	0.55 (0.0)
EBIC ($\gamma = 0.0$)	0.6 (0.1)	0.91 (0.07)	0.94 (0.07)	0.77 (0.14)	0.6 (0.0)
EBIC ($\gamma = 0.5$)	0.63 (0.1)	0.9 (0.13)	0.93 (0.09)	0.84 (0.15)	0.6 (0.0)
EBIC ($\gamma = 1.0$)	0.63 (0.1)	0.9 (0.13)	0.96 (0.08)	0.94 (0.07)	0.6 (0.0)
stability ($q = 1$)	0.2 (0.24)	0.2 (0.24)	0.5 (0.0)	0.5 (0.0)	0.5 (0.0)
stability ($q = 4$)	0.0 (0.0)	0.82 (0.18)	0.96 (0.08)	0.97 (0.06)	0.89 (0.06)
stability ($q = 6$)	0.0 (0.0)	0.83 (0.18)	1.0 (0.0)	0.91 (0.07)	0.67 (0.0)
Average number of selected variables					
	10	50	100	1000	100000
proposed (automatic)	3.4 (2.33)	3.2 (0.4)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
GibbsBvs	5.2 (3.06)	3.4 (0.49)	3.4 (0.49)	5.0 (0.0)	-
EMVS	2.0 (2.53)	0.2 (0.4)	0.0 (0.0)	3.0 (0.0)	3.0 (0.0)
horseshoe (automatic)	4.4 (2.33)	3.6 (0.49)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)
AIC	6.8 (2.4)	4.0 (0.0)	3.8 (0.4)	6.8 (0.75)	8.0 (0.0)
EBIC ($\gamma = 0.0$)	6.8 (2.4)	3.6 (0.49)	3.4 (0.49)	5.0 (1.26)	7.0 (0.0)
EBIC ($\gamma = 0.5$)	5.6 (2.58)	3.2 (0.4)	3.0 (0.63)	4.4 (1.36)	7.0 (0.0)
EBIC ($\gamma = 1.0$)	4.0 (2.53)	3.2 (0.4)	2.8 (0.4)	3.4 (0.49)	7.0 (0.0)
stability ($q = 1$)	0.4 (0.49)	0.4 (0.49)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
stability ($q = 4$)	0.0 (0.0)	2.2 (0.75)	2.8 (0.4)	3.2 (0.4)	3.8 (0.4)
stability ($q = 6$)	0.0 (0.0)	2.6 (1.02)	3.0 (0.0)	3.6 (0.49)	6.0 (0.0)

Table 3.4: Low-dimensional setting, $d = 8$ and $n \in \{10, 50, 100, 1000, 100000\}$. Evaluation results with no noise on regression coefficients. Comparison of the proposed method and horseshoe for different δ.

F1-Scores					
	10	50	100	1000	100000
proposed (delta = 0.8)	0.5 (0.0)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.5)	0.51 (0.03)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.05)	0.51 (0.02)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.01)	0.51 (0.02)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.001)	0.51 (0.02)	0.92 (0.1)	0.95 (0.1)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.0)	0.51 (0.02)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
proposed (automatic)	0.51 (0.03)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.8)	0.7 (0.16)	0.96 (0.08)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.5)	0.77 (0.15)	0.9 (0.08)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.05)	0.57 (0.03)	0.57 (0.03)	0.59 (0.04)	0.66 (0.06)	1.0 (0.0)
horseshoe (delta = 0.01)	0.55 (0.0)	0.56 (0.02)	0.56 (0.02)	0.57 (0.03)	0.86 (0.08)
horseshoe (delta = 0.001)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.56 (0.02)
horseshoe (delta = 0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)
horseshoe (automatic)	0.53 (0.07)	0.91 (0.18)	0.92 (0.1)	1.0 (0.0)	1.0 (0.0)
Average number of selected variables					
	10	50	100	1000	100000
proposed (delta = 0.8)	1.8 (1.6)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.5)	1.6 (1.2)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.05)	2.4 (2.8)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.01)	2.4 (2.8)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.001)	2.4 (2.8)	2.6 (0.49)	3.4 (0.8)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.0)	2.4 (2.8)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
proposed (automatic)	1.6 (1.2)	2.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.8)	2.6 (0.8)	2.8 (0.4)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.5)	3.8 (0.75)	3.2 (0.75)	3.6 (0.8)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.05)	7.6 (0.49)	7.6 (0.49)	7.2 (0.75)	6.2 (0.75)	3.0 (0.0)
horseshoe (delta = 0.01)	8.0 (0.0)	7.8 (0.4)	7.8 (0.4)	7.6 (0.49)	4.0 (0.63)
horseshoe (delta = 0.001)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	7.8 (0.4)
horseshoe (delta = 0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)
horseshoe (automatic)	5.8 (2.4)	4.0 (2.0)	3.6 (0.8)	3.0 (0.0)	3.0 (0.0)

Table 3.5: Low-dimensional setting, $d = 8$ and $n \in \{10, 50, 100, 1000, 100000\}$. Evaluation results with noise on regression coefficients $\eta = 0.2$. Comparison of the proposed method and horseshoe for different δ.

	F1-Scores				
	10	50	100	1000	100000
proposed (delta = 0.8)	0.5 (0.06)	0.87 (0.19)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.5)	0.5 (0.06)	0.97 (0.06)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.05)	0.51 (0.07)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)	0.68 (0.03)
proposed (delta = 0.01)	0.5 (0.05)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)	0.61 (0.03)
proposed (delta = 0.001)	0.5 (0.05)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)	0.61 (0.03)
proposed (delta = 0.0)	0.52 (0.09)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)	0.61 (0.03)
proposed (automatic)	0.5 (0.06)	0.97 (0.06)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.8)	0.73 (0.17)	0.93 (0.09)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.5)	0.7 (0.19)	0.91 (0.07)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.05)	0.63 (0.12)	0.62 (0.12)	0.65 (0.07)	0.64 (0.06)	0.68 (0.03)
horseshoe (delta = 0.01)	0.55 (0.0)	0.55 (0.0)	0.57 (0.03)	0.58 (0.03)	0.58 (0.03)
horseshoe (delta = 0.001)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)
horseshoe (delta = 0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)
horseshoe (automatic)	0.64 (0.19)	0.91 (0.07)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
	Average number of selected variables				
	10	50	100	1000	100000
proposed (delta = 0.8)	3.2 (2.4)	2.8 (0.98)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.5)	3.2 (2.4)	3.2 (0.4)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.05)	3.8 (3.06)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)	5.8 (0.4)
proposed (delta = 0.01)	4.0 (3.29)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)	6.8 (0.4)
proposed (delta = 0.001)	4.0 (3.29)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)	6.8 (0.4)
proposed (delta = 0.0)	3.6 (2.87)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)	6.8 (0.4)
proposed (automatic)	3.2 (2.4)	3.2 (0.4)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.8)	3.4 (1.2)	3.0 (0.63)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.5)	3.8 (1.17)	3.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.05)	6.8 (1.6)	7.0 (1.55)	6.4 (1.02)	6.4 (0.8)	5.8 (0.4)
horseshoe (delta = 0.01)	8.0 (0.0)	8.0 (0.0)	7.6 (0.49)	7.4 (0.49)	7.4 (0.49)
horseshoe (delta = 0.001)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)
horseshoe (delta = 0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)
horseshoe (automatic)	4.6 (2.15)	3.6 (0.49)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)

Table 3.6: Low-dimensional setting, $d = 8$ and $n \in \{10, 50, 100, 1000, 100000\}$. Evaluation results with noise on regression coefficients $\eta = 0.5$. Comparison of the proposed method and horseshoe for different δ.

F1-Scores					
	10	50	100	1000	100000
proposed (delta = 0.8)	0.55 (0.07)	0.83 (0.18)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.5)	0.55 (0.07)	0.97 (0.06)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
proposed (delta = 0.05)	0.49 (0.05)	0.97 (0.06)	0.93 (0.09)	0.8 (0.1)	0.6 (0.0)
proposed (delta = 0.01)	0.54 (0.06)	0.97 (0.06)	0.93 (0.09)	0.8 (0.1)	0.59 (0.02)
proposed (delta = 0.001)	0.54 (0.06)	0.97 (0.06)	0.93 (0.09)	0.8 (0.1)	0.59 (0.02)
proposed (delta = 0.0)	0.54 (0.06)	0.97 (0.06)	0.93 (0.09)	0.8 (0.1)	0.59 (0.02)
proposed (automatic)	0.55 (0.07)	0.97 (0.06)	0.96 (0.08)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.8)	0.67 (0.19)	0.93 (0.09)	1.0 (0.0)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.5)	0.7 (0.19)	0.91 (0.07)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)
horseshoe (delta = 0.05)	0.63 (0.12)	0.6 (0.08)	0.63 (0.07)	0.59 (0.04)	0.6 (0.0)
horseshoe (delta = 0.01)	0.57 (0.05)	0.57 (0.03)	0.57 (0.03)	0.57 (0.03)	0.55 (0.0)
horseshoe (delta = 0.001)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)
horseshoe (delta = 0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)	0.55 (0.0)
horseshoe (automatic)	0.65 (0.17)	0.91 (0.07)	0.97 (0.06)	1.0 (0.0)	1.0 (0.0)
Average number of selected variables					
	10	50	100	1000	100000
proposed (delta = 0.8)	3.4 (2.33)	2.6 (1.02)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.5)	3.4 (2.33)	3.2 (0.4)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
proposed (delta = 0.05)	3.4 (2.73)	3.2 (0.4)	3.0 (0.63)	4.6 (0.8)	7.0 (0.0)
proposed (delta = 0.01)	3.6 (2.65)	3.2 (0.4)	3.0 (0.63)	4.6 (0.8)	7.2 (0.4)
proposed (delta = 0.001)	3.6 (2.65)	3.2 (0.4)	3.0 (0.63)	4.6 (0.8)	7.2 (0.4)
proposed (delta = 0.0)	3.6 (2.65)	3.2 (0.4)	3.0 (0.63)	4.6 (0.8)	7.2 (0.4)
proposed (automatic)	3.4 (2.33)	3.2 (0.4)	2.8 (0.4)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.8)	3.2 (1.47)	3.0 (0.63)	3.0 (0.0)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.5)	3.8 (1.17)	3.6 (0.49)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)
horseshoe (delta = 0.05)	6.8 (1.6)	7.2 (1.17)	6.6 (1.02)	7.2 (0.75)	7.0 (0.0)
horseshoe (delta = 0.01)	7.6 (0.8)	7.6 (0.49)	7.6 (0.49)	7.6 (0.49)	8.0 (0.0)
horseshoe (delta = 0.001)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)
horseshoe (delta = 0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)	8.0 (0.0)
horseshoe (automatic)	4.4 (2.33)	3.6 (0.49)	3.2 (0.4)	3.0 (0.0)	3.0 (0.0)

Table 3.7: High-dimensional setting, $d = 1000$ and $n \in \{100, 1000\}$. Evaluation results with no noise on regression coefficients.

F1-Scores		
	100	1000
proposed (automatic)	0.96 (0.08)	1.0 (0.0)
EMVS	0.96 (0.08)	1.0 (0.0)
horseshoe (automatic)	0.46 (0.24)	1.0 (0.0)
AIC	0.06 (0.0)	0.01 (0.0)
EBIC ($\gamma = 0.0$)	0.06 (0.0)	0.01 (0.0)
EBIC ($\gamma = 0.5$)	0.06 (0.0)	0.01 (0.0)
EBIC ($\gamma = 1.0$)	0.06 (0.0)	0.01 (0.0)
stability ($q = 1$)	0.2 (0.24)	0.5 (0.0)
stability ($q = 50$)	1.0 (0.0)	1.0 (0.0)
stability ($q = 100$)	0.84 (0.08)	1.0 (0.0)
Average number of selected variables		
	100	1000
proposed (automatic)	2.8 (0.4)	3.0 (0.0)
EMVS	2.8 (0.4)	3.0 (0.0)
horseshoe (automatic)	14.6 (9.16)	3.0 (0.0)
AIC	99.0 (0.0)	999.0 (0.0)
EBIC ($\gamma = 0.0$)	99.0 (0.0)	999.0 (0.0)
EBIC ($\gamma = 0.5$)	99.0 (0.0)	999.0 (0.0)
EBIC ($\gamma = 1.0$)	99.0 (0.0)	999.0 (0.0)
stability ($q = 1$)	0.4 (0.49)	1.0 (0.0)
stability ($q = 50$)	3.0 (0.0)	3.0 (0.0)
stability ($q = 100$)	2.2 (0.4)	3.0 (0.0)

Table 3.8: High-dimensional setting, $d = 1000$ and $n \in \{100, 1000\}$. Evaluation results with noise on regression coefficients $\eta = 0.2$.

F1-Scores		
	100	1000
proposed (automatic)	0.84 (0.08)	1.0 (0.0)
EMVS	0.88 (0.1)	0.96 (0.08)
horseshoe (automatic)	0.52 (0.17)	0.66 (0.2)
AIC	0.06 (0.0)	0.01 (0.0)
EBIC ($\gamma = 0.0$)	0.06 (0.0)	0.01 (0.0)
EBIC ($\gamma = 0.5$)	0.06 (0.0)	0.01 (0.0)
EBIC ($\gamma = 1.0$)	0.06 (0.0)	0.01 (0.0)
stability ($q = 1$)	0.4 (0.2)	0.5 (0.0)
stability ($q = 50$)	0.92 (0.1)	0.97 (0.06)
stability ($q = 100$)	0.88 (0.1)	1.0 (0.0)
Average number of selected variables		
	100	1000
proposed (automatic)	2.2 (0.4)	3.0 (0.0)
EMVS	2.4 (0.49)	2.8 (0.4)
horseshoe (automatic)	10.4 (8.45)	7.6 (5.24)
AIC	99.0 (0.0)	999.0 (0.0)
EBIC ($\gamma = 0.0$)	99.0 (0.0)	999.0 (0.0)
EBIC ($\gamma = 0.5$)	99.0 (0.0)	999.0 (0.0)
EBIC ($\gamma = 1.0$)	99.0 (0.0)	999.0 (0.0)
stability ($q = 1$)	0.8 (0.4)	1.0 (0.0)
stability ($q = 50$)	2.6 (0.49)	3.2 (0.4)
stability ($q = 100$)	2.4 (0.49)	3.0 (0.0)

Table 3.9: High-dimensional setting, $d = 1000$ and $n \in \{100, 1000\}$. Evaluation results with no noise on regression coefficients. Comparison of the proposed method and horseshoe for different δ.

F1-Scores		
	100	1000
proposed (delta = 0.8)	0.5 (0.0)	0.8 (0.0)
proposed (delta = 0.5)	0.56 (0.12)	1.0 (0.0)
proposed (delta = 0.05)	0.96 (0.08)	1.0 (0.0)
proposed (delta = 0.01)	0.96 (0.08)	1.0 (0.0)
proposed (delta = 0.001)	0.96 (0.08)	1.0 (0.0)
proposed (delta = 0.0)	0.96 (0.08)	1.0 (0.0)
proposed (automatic)	0.96 (0.08)	1.0 (0.0)
horseshoe (delta = 0.8)	0.96 (0.08)	1.0 (0.0)
horseshoe (delta = 0.5)	0.96 (0.08)	1.0 (0.0)
horseshoe (delta = 0.05)	0.69 (0.16)	0.97 (0.06)
horseshoe (delta = 0.01)	0.2 (0.02)	0.34 (0.03)
horseshoe (delta = 0.001)	0.01 (0.0)	0.02 (0.0)
horseshoe (delta = 0.0)	0.01 (0.0)	0.01 (0.0)
horseshoe (automatic)	0.46 (0.24)	1.0 (0.0)
Average number of selected variables		
	100	1000
proposed (delta = 0.8)	1.0 (0.0)	2.0 (0.0)
proposed (delta = 0.5)	1.2 (0.4)	3.0 (0.0)
proposed (delta = 0.05)	2.8 (0.4)	3.0 (0.0)
proposed (delta = 0.01)	2.8 (0.4)	3.0 (0.0)
proposed (delta = 0.001)	2.8 (0.4)	3.0 (0.0)
proposed (delta = 0.0)	2.8 (0.4)	3.0 (0.0)
proposed (automatic)	2.8 (0.4)	3.0 (0.0)
horseshoe (delta = 0.8)	2.8 (0.4)	3.0 (0.0)
horseshoe (delta = 0.5)	2.8 (0.4)	3.0 (0.0)
horseshoe (delta = 0.05)	6.2 (2.4)	3.2 (0.4)
horseshoe (delta = 0.01)	27.2 (2.93)	15.0 (1.55)
horseshoe (delta = 0.001)	463.8 (23.79)	393.2 (8.84)
horseshoe (delta = 0.0)	1000.0 (0.0)	1000.0 (0.0)
horseshoe (automatic)	14.6 (9.16)	3.0 (0.0)

Table 3.10: High-dimensional setting, $d = 1000$ and $n \in \{100, 1000\}$. Evaluation results with noise on regression coefficients $\eta = 0.2$. Comparison of the proposed method and horseshoe for different δ.

F1-Scores		
	100	1000
proposed (delta = 0.8)	0.5 (0.0)	0.8 (0.0)
proposed (delta = 0.5)	0.5 (0.0)	0.96 (0.08)
proposed (delta = 0.05)	0.84 (0.08)	0.97 (0.06)
proposed (delta = 0.01)	0.84 (0.08)	0.97 (0.06)
proposed (delta = 0.001)	0.84 (0.08)	0.97 (0.06)
proposed (delta = 0.0)	0.84 (0.08)	0.97 (0.06)
proposed (automatic)	0.84 (0.08)	1.0 (0.0)
horseshoe (delta = 0.8)	0.88 (0.1)	1.0 (0.0)
horseshoe (delta = 0.5)	0.92 (0.1)	1.0 (0.0)
horseshoe (delta = 0.05)	0.63 (0.09)	0.8 (0.11)
horseshoe (delta = 0.01)	0.18 (0.04)	0.3 (0.05)
horseshoe (delta = 0.001)	0.01 (0.0)	0.01 (0.0)
horseshoe (delta = 0.0)	0.01 (0.0)	0.01 (0.0)
horseshoe (automatic)	0.52 (0.17)	0.66 (0.2)
Average number of selected variables		
	100	1000
proposed (delta = 0.8)	1.0 (0.0)	2.0 (0.0)
proposed (delta = 0.5)	1.0 (0.0)	2.8 (0.4)
proposed (delta = 0.05)	2.2 (0.4)	3.2 (0.4)
proposed (delta = 0.01)	2.2 (0.4)	3.2 (0.4)
proposed (delta = 0.001)	2.2 (0.4)	3.2 (0.4)
proposed (delta = 0.0)	2.2 (0.4)	3.2 (0.4)
proposed (automatic)	2.2 (0.4)	3.0 (0.0)
horseshoe (delta = 0.8)	2.4 (0.49)	3.0 (0.0)
horseshoe (delta = 0.5)	2.6 (0.49)	3.0 (0.0)
horseshoe (delta = 0.05)	6.0 (1.67)	4.6 (1.02)
horseshoe (delta = 0.01)	32.8 (9.97)	17.8 (3.76)
horseshoe (delta = 0.001)	525.4 (54.5)	426.2 (10.4)
horseshoe (delta = 0.0)	1000.0 (0.0)	1000.0 (0.0)
horseshoe (automatic)	10.4 (8.45)	7.6 (5.24)

Table 3.11: Statistics of real data sets

	ozone	crime	SDM
n	178	47	88
d	35	15	67

Table 3.12: Response y and covariates of ozone data. The data set contains all of the variables below, including all second-order terms and interactions. This table is partly copied from Table 5 in the supplement material of (Garcia-Donato and Martinez-Beneito, 2013).

y	Daily maximum 1-hour-average ozone reading (ppm) at Upland, CA
x4	500-millibar pressure height (m) measured at Vandenberg AFB
x5	Wind speed (mph) at Los Angeles International Airport (LAX)
x6	Humidity (%) at LAX
x7	Temperature (Fahrenheit degrees) measured at Sandburg, CA
x8	Inversion base height (feet) at LAX
x9	Pressure gradient (mm Hg) from LAX to Daggett, CA
x10	Visibility (miles) measured at LAX

Table 3.13: Response y and covariates of crime data. This table is partly copied from Table 4 in (Raftery et al., 1997).

y	crime rate
M	Percentage of males age 14-24
So	Indicator variable for southern state
Ed	Mean years of schooling
Po1	Police expenditure in 1960
Po2	Police expenditure in 1959
LF	Labor force participation rate
M.F	Number of males per 1,000 females
Pop	State population
NW	Number of nonwhites per 1,000 people
U1	Unemployment rate of urban males age 14-24
U2	Unemployment rate of urban males, age 35-39
GDP	Wealth
Ineq	Income inequality
Prob	Probability of imprisonment
Time	Average time served in state prisons

Table 3.14: Response y and covariates (first part) of SDM data. This table is copied from the description of R package 'BayesVarSel'.

y	Growth of GDP per capita at purchasing power parities between 1960 and 1996.
ABSLATIT	Absolute latitude.
AIRDIST	Logarithm of minimal distance (in km) from New York, Rotterdam, or Tokyo.
AVELF	Average of five different indices of ethnolinguistic fractionalization
BRIT	Dummy for former British colony after 1776.
BUDDHA	Fraction of population Buddhist in 1960.
CATH00	Fraction of population Catholic in 1960.
CIV72	Index of civil liberties index in 1972.
COLONY	Dummy for former colony.
CONFUC	Fraction of population Confucian.
DENS60	Population per area in 1960.
DENS65C	Coastal (within 100 km of coastline) population per coastal area in 1965.
DENS65I	Interior (more than 100 km from coastline) population per interior area in 1965.
DPOP6090	Average growth rate of population between 1960 and 1990.
EAST	Dummy for East Asian countries.
ECORG	Degree Capitalism index.
ENGFRAC	Fraction of population speaking English.
EUROPE	Dummy for European economies.
FERTLDC1	Fertility in 1960's.
GDE1	Average share public expenditures on defense as fraction of GDP between 1960 and 1965.
GDPCH60L	Logarithm of GDP per capita in 1960.
GEEREC1	Average share public expenditures on education as fraction of GDP between 1960 and 1965.
GGCFD3	Average share of expenditures on public investment as fraction of GDP between 1960 and 1965.
GOVNOM1	Average share of nominal government spending to nominal GDP between 1960 and 1964.
GOVSH61	Average share government spending to GDP between 1960 and 1964.
GVR61	Share of expenditures on government consumption to GDP in 1961.
H60	Enrollment rates in higher education.
HERF00	Religion measure.
HINDU00	Fraction of the population Hindu in 1960.
IPRICE1	Average investment price level between 1960 and 1964 on purchasing power parity basis.
LAAM	Dummy for Latin American countries.
LANDAREA	Area in km.
LANDLOCK	Dummy for landlocked countries.

Table 3.15: Covariates (second part) of SDM data. This table is copied from the description of R package 'BayesVarSel'.

LHCPC	Log of hydrocarbon deposits in 1993.
LIFE060	Life expectancy in 1960.
LT100CR	Proportion of country's land area within 100 km of ocean or ocean-navigable river.
MALFAL66	Index of malaria prevalence in 1966.
MINING	Fraction of GDP in mining.
MUSLIM00	Fraction of population Muslim in 1960.
NEWSTATE	National independence.
OIL	Dummy for oil-producing country.
OPENDEC1	Ratio of exports plus imports to GDP, averaged over 1965 to 1974.
ORTH00	Fraction of population Orthodox in 1960.
OTHFRAC	Fraction of population speaking foreign language.
P60	Enrollment rate in primary education in 1960.
PI6090	Average inflation rate between 1960 and 1990.
SQPI6090	Square of average inflation rate between 1960 and 1990.
PRIGHTS	Political rights index.
POP1560	Fraction of population younger than 15 years in 1960.
POP60	Population in 1960
POP6560	Fraction of population older than 65 years in 1960.
PRIEXP70	Fraction of primary exports in total exports in 1970.
PROT00	Fraction of population Protestant in 1960.
RERD	Real exchange rate distortions.
REVCOUP	Number of revolutions and military coups.
SAFRICA	Dummy for Sub-Saharan African countries.
SCOUT	Measure of outward orientation.
SIZE60	Logarithm of aggregate GDP in 1960.
SOCIALIST	Dummy for countries under Socialist rule for considerable time during 1950 to 1995.
SPAIN	Dummy variable for former Spanish colonies.
TOT1DEC1	Growth of terms of trade in the 1960's.
TOTIND	Terms of trade ranking
TROPICAR	Proportion of country's land area within geographical tropics.
TROPPOP	Proportion of country's population living in geographical tropics.
WARTIME	Fraction of time spent in war between 1960 and 1990.
WARTORN	Indicator for countries that participated in external war between 1960 and 1990.
YRSOPEN	Number of years economy has been open between 1950 and 1994.
ZTROPICS	Fraction tropical climate zone.

Table 3.16: Selected variables for the ozone data. For proposed method and horseshoe method we denote by "MSE inc" expected increase in mean squared error compared to choosing the full model.

method	selected variables
proposed (delta = 0.8, MSE inc = 37.31%)	x7.x7
proposed (delta = 0.5, MSE inc = 19.5%)	x6.x6, x6.x7
proposed (delta = 0.05, MSE inc = 5.43%)	x6.x7, x6.x8, x7.x7
proposed (delta = 0.01, MSE inc = 4.91%)	x6.x6, x6.x7, x6.x8
proposed (delta = 0.001, MSE inc = 4.94%)	x6.x6, x6.x7, x6.x8
proposed (delta = 0.0, MSE inc = 5.44%)	x6.x7, x6.x8, x7.x7
horseshoe (delta = 0.8, MSE inc = 15.47%)	x6.x7, x7.x7, x7.x10
horseshoe (delta = 0.5, MSE inc = 5.11%)	x6.x7, x6.x8, x7.x7, x7.x8, x7.x10
horseshoe (delta = 0.05, MSE inc = 0.0%)	all except x5, x4.x5, x4.x8, x5.x8, x6.x9, x6.x10, x8.x8, x8.x9, x9.x10
horseshoe (delta = 0.01, MSE inc = 0.0%)	all except x5.x8, x6.x9
horseshoe (delta = 0.001, MSE inc = 0.0%)	all except x6.x9
horseshoe (delta = 0.0, MSE inc = 0.0%)	all
GibbsBvs	x6.x6, x6.x7, x6.x8
EMVS	none
AIC	x9, x4.x4, x6.x7, x6.x8, x7.x7, x7.x8, x7.x10, x8.x10, x9.x9
EBIC ($\gamma = 0$)	x6.x7, x6.x8, x7.x7, x7.x8, x7.x10, x8.x10, x9.x9
EBIC ($\gamma = 0.5$)	x6.x7, x7.x7, x7.x8, x7.x10, x9.x9
EBIC ($\gamma = 1.0$)	x4.x8, x6.x7, x7.x7
stability ($q = 0.1 \cdot d$)	x7.x7
stability ($q = 0.5 \cdot d$)	x7.x10
stability ($q = 0.8 \cdot d$)	none

Table 3.17: Selected variables for the crime data. For proposed method and horseshoe method we denote by "MSE inc" expected increase in mean squared error compared to choosing the full model.

method	selected variables
proposed (delta = 0.8, MSE inc = 65.62%)	Po1, Ineq
proposed (delta = 0.5, MSE inc = 21.97%)	M, Ed, Po1, Ineq
proposed (delta = 0.05, MSE inc = 0.0%)	all
proposed (delta = 0.01, MSE inc = 0.0%)	all
proposed (delta = 0.001, MSE inc = 0.0%)	all
proposed (delta = 0.0, MSE inc = 0.0%)	all
horseshoe (delta = 0.8, MSE inc = 17.23%)	M, Ed, Po1, Po2, NW, Ineq, Prob
horseshoe (delta = 0.5, MSE inc = 3.07%)	all except So, LF, M.F, Pop, U1, Time
horseshoe (delta = 0.05, MSE inc = 0.0%)	all except M.F
horseshoe (delta = 0.01, MSE inc = 0.0%)	all except M.F
horseshoe (delta = 0.001, MSE inc = 0.0%)	all
horseshoe (delta = 0.0, MSE inc = 0.0%)	all
GibbsBvs	all
EMVS	Po1, Ineq
AIC	all except So, Po2, M.F, U1
EBIC ($\gamma = 0$)	all except So, Po2, LF, Pop, U1, GDP, Time
EBIC ($\gamma = 0.5$)	M, Ed, Po1, M.F, NW, Ineq, Prob
EBIC ($\gamma = 1.0$)	Po1, NW
stability ($q = 0.1 \cdot d$)	Po1
stability ($q = 0.5 \cdot d$)	NW
stability ($q = 0.8 \cdot d$)	none

Table 3.18: Selected variables for the SDM data. For proposed method and horseshoe method we denote by "MSE inc" expected increase in mean squared error compared to choosing the full model.

method	selected variables
proposed (delta = 0.8, MSE inc = 110.29%)	EAST
proposed (delta = 0.5, MSE inc = 27.07%)	EAST, MALFAL66
proposed (delta = 0.05, MSE inc = 27.25%)	EAST, MALFAL66
proposed (delta = 0.01, MSE inc = 27.2%)	EAST, MALFAL66
proposed (delta = 0.001, MSE inc = 27.14%)	EAST, MALFAL66
proposed (delta = 0.0, MSE inc = 27.22%)	EAST, MALFAL66
horseshoe (delta = 0.8, MSE inc = 69.31%)	EAST, GDPCH60L, IPRICE1, P60
horseshoe (delta = 0.5, MSE inc = 20.16%)	CONFUC, EAST, GDPCH60L, IPRICE1, LIFE060, P60, TROPICAR
horseshoe (delta = 0.05, MSE inc = 0.0%)	all except DENS65I, DPOP6090, ECORG, EUROPE, HERF00, LANDAREA, LANDLOCK, OIL, ORTH00, PI6090, SQPI6090, POP6560, SIZE60, TOT1DEC1, TOTIND, WARTIME, WARTORN
horseshoe (delta = 0.01, MSE inc = 0.0%)	all except DENS65I, ECORG, LANDAREA, SQPI6090, WARTIME
horseshoe (delta = 0.001, MSE inc = 0.0%)	all
horseshoe (delta = 0.0, MSE inc = 0.0%)	all
GibbsBvs	DENS65C, EAST, GDPCH60L, IPRICE1, P60, TROPICAR
EMVS	none
AIC	AVELF, BUDDHA, CIV72, CONFUC, DENS65C, EAST, GDPCH60L, GGCFD3, GOVNOM1, GVR61, HINDU00, IPRICE1, MALFAL66, MINING, MUSLIM00, OPENDEC1, OTHFRAC, P60, POP60, RERD, REVCOUP, SAFRICA, SPAIN, TROPICAR, TROPPOP, YRSOPEN
EBIC ($\gamma = 0$)	CONFUC, EAST, MALFAL66, P60, TROPPOP, YRSOPEN
EBIC ($\gamma = 0.5$)	EAST, TROPPOP, YRSOPEN
EBIC ($\gamma = 1.0$)	EAST, TROPPOP, YRSOPEN
stability ($q = 0.1 \cdot d$)	EAST, YRSOPEN
stability ($q = 0.5 \cdot d$)	none
stability ($q = 0.8 \cdot d$)	- (did not terminate)

Table 3.19: Top 10 selected models using the proposed method with $\delta = 0.0$ and $\delta = 0.5$ for the ozone data. Last column also shows the highest posterior probability model reported in (Garcia-Donato and Martinez-Beneito, 2013) using a g-prior where inclusion probabilities are calculated exactly (i.e. no MCMC).

model	probability
$\delta = 0.5$	
x6.x6, x6.x7	0.066
x6.x7, x7.x7	0.034
x4.x10, x7.x7, x7.x10	0.027
x7.x7	0.026
x10, x4.x7, x7.x10	0.02
x4.x7, x4.x10, x7.x10	0.019
x10, x7.x7, x7.x10	0.019
x7, x6.x7, x7.x7	0.016
x7.x7, x7.x10	0.015
x6.x6, x6.x7, x7.x8	0.013
$\delta = 0.0$	
x6.x7, x6.x8, x7.x7	0.031
x6.x6, x6.x7, x6.x8	0.029
x10, x6.x7, x6.x8, x7.x7, x7.x10	0.018
x4.x10, x6.x7, x6.x8, x7.x7, x7.x10	0.018
x4.x6, x4.x10, x6.x8, x7.x7, x7.x10	0.016
x10, x4.x6, x6.x8, x7.x7, x7.x10	0.013
x6.x7, x7.x7, x7.x8	0.01
x6, x4.x10, x6.x8, x7.x7, x7.x10	0.01
x4.x6, x6.x8, x7.x7	0.009
x6, x6.x8, x7.x7	0.009
Gracia-Donato	
x10, x4.x6, x6.x8, x7.x7, x7.x10	0.0009

Table 3.20: All inclusion probabilities using the proposed method with $\delta = 0.0$ and $\delta = 0.5$ for the ozone data. Last column also shows the results reported in (Garcia-Donato and Martinez-Beneito, 2013) using a g-prior where inclusion probabilities are calculated exactly (i.e. no MCMC).

variable	$\delta = 0.5$	$\delta = 0.0$	Gracia-Donato
x7.x7	0.58	0.67	0.450
x6.x7	0.568	0.603	0.636
x7.x10	0.5	0.649	0.743
x6.x6	0.313	0.245	0.532
x4.x10	0.233	0.334	0.361
x6.x8	0.226	0.702	0.560
x10	0.226	0.291	0.368
x4.x7	0.212	0.234	0.252
x7.x8	0.179	0.279	0.349
x4.x6	0.164	0.295	0.325
x6	0.139	0.246	0.297
x7	0.133	0.16	0.195
x7.x9	0.09	0.072	0.431
x8	0.076	0.139	0.200
x4.x9	0.064	0.059	0.301
x4.x8	0.064	0.132	0.208
x9.x9	0.059	0.156	0.434
x9	0.053	0.056	0.291
x8.x10	0.037	0.112	0.236
x10.x10	0.028	0.07	0.117
x8.x8	0.028	0.067	0.142
x8.x9	0.019	0.034	0.263
x5.x10	0.019	0.036	0.124
x6.x10	0.017	0.052	0.115
x6.x9	0.012	0.036	0.126
x4.x4	0.011	0.032	0.164
x5.x6	0.011	0.027	0.107
x4	0.011	0.031	0.164
x5.x8	0.009	0.031	0.098
x5.x5	0.008	0.024	0.124
x5.x7	0.008	0.025	0.094
x9.x10	0.007	0.024	0.103
x5	0.006	0.019	0.096
x4.x5	0.006	0.02	0.095
x5.x9	0.005	0.022	0.088

Table 3.21: Top 10 selected models using the proposed method with $\delta = 0.0$ and $\delta = 0.5$ for the crime data.

model	probability
$\delta = 0.5$	
M, Ed, Po1, Ineq	0.021
M, Ed, Po1, NW, Ineq, Prob	0.017
Po1, Ineq	0.017
Ed, Po1, Ineq	0.016
M, Ed, Po1, NW, U2, Ineq, Prob	0.015
M, Ed, Po1, Ineq, Prob	0.015
Ed, Po1, NW, Ineq, Prob	0.013
M, Ed, Po1, U2, Ineq	0.011
M, Ed, Po1, U2, Ineq, Prob	0.011
M, Ed, Po1, NW, Ineq, Prob, Time	0.011
$\delta = 0.0$	
all	0.02
M, Ed, Po1, Ineq	0.01
M, Ed, Po1, NW, U2, Ineq, Prob	0.01
Ed, Po1, Ineq	0.008
M, Ed, Po1, NW, Ineq, Prob	0.008
all except So, Po2, LF, M.F, Pop, U1, GDP	0.008
Po1, Ineq	0.007
all except So, LF, M.F, Pop, U1, GDP, Time	0.007
M, Ed, Po1, NW, Ineq, Prob, Time	0.007
M, Ed, Po1, U2, Ineq, Prob	0.007

Table 3.22: All inclusion probabilities using the proposed method with $\delta = 0.5$ and $\delta = 0.0$ for the crime data. Last column also shows the results reported in (Liang et al., 2008) with the Zellner-Siow Prior with the null-model as the reference model.

variable	$\delta = 0.5$	$\delta = 0.0$	Liang
Ineq	0.993	0.995	1.0
Ed	0.906	0.943	0.97
Prob	0.758	0.833	0.90
Po1	0.742	0.792	0.67
M	0.731	0.808	0.85
NW	0.604	0.711	0.69
Po2	0.52	0.591	0.45
U2	0.425	0.557	0.61
GDP	0.381	0.481	0.36
Time	0.256	0.395	0.37
Pop	0.244	0.368	0.37
So	0.207	0.32	0.27
U1	0.134	0.269	0.25
M.F	0.12	0.238	0.20
LF	0.115	0.233	0.20

Table 3.23: Top 10 selected models using the proposed method with $\delta = 0.0$ and $\delta = 0.5$ for the SDM data.

model	probability
$\delta = 0.5$	
EAST, MALFAL66	0.12
EAST, P60, TROPICAR	0.035
EAST, P60	0.034
EAST, MALFAL66, P60	0.026
EAST, TROPICAR	0.019
EAST, LIFE060	0.016
EAST, GDPCH60L, LIFE060, MALFAL66	0.012
EAST, IPRICE1, P60, TROPICAR	0.011
EAST, IPRICE1, P60	0.011
EAST, GDPCH60L, IPRICE1, LIFE060	0.01
$\delta = 0.0$	
EAST, MALFAL66	0.055
DENS65C, EAST, GDPCH60L, IPRICE1, P60, TROPICAR	0.02
EAST, MALFAL66, P60	0.015
EAST, P60, TROPICAR	0.012
EAST, MALFAL66, P60, SPAIN	0.007
EAST, MALFAL66, SPAIN	0.007
EAST, GVR61, MALFAL66	0.006
EAST, GDPCH60L, LIFE060, MALFAL66	0.006
EAST, LIFE060, MALFAL66	0.006
EAST, MALFAL66, YRSOPEN	0.006

Table 3.24: Top 30 inclusion probabilities using the proposed method with $\delta = 0.0$ and $\delta = 0.5$ for the SDM data. For reference, we also show the results that were reported in (Sala-i Martin et al., 2004) using the BACE method. Variables in bold mark the 18 variables that were considered as significant in (Sala-i Martin et al., 2004).

variable	$\delta = 0.5$	variable	$\delta = 0.0$	variable	BACE
EAST	0.892	**EAST**	0.855	**EAST**	0.823
P60	0.485	**P60**	0.623	**P60**	0.774
MALFAL66	0.341	**IPRICE1**	0.523	**IPRICE1**	0.774
GDPCH60L	0.34	**GDPCH60L**	0.475	**GDPCH60L**	0.685
IPRICE1	0.328	**MALFAL66**	0.434	**TROPICAR**	0.563
TROPICAR	0.285	**TROPICAR**	0.411	**DENS65C**	0.428
LIFE060	0.253	**DENS65C**	0.248	**MALFAL66**	0.252
CONFUC	0.113	**LIFE060**	0.231	**LIFE060**	0.209
YRSOPEN	0.085	**CONFUC**	0.189	**CUNFUC**	0.206
SAFRICA	0.079	**YRSOPEN**	0.14	**SAFRICA**	0.154
RERD	0.068	**SAFRICA**	0.136	**LAAM**	0.149
DENS65C	0.063	**LAAM**	0.128	**MINING**	0.124
GVR61	0.055	**SPAIN**	0.126	**SPAIN**	0.123
LAAM	0.045	**GVR61**	0.109	**YRSOPEN**	0.119
TROPPOP	0.045	**MINING**	0.097	**MUSLIM00**	0.114
AVELF	0.044	**MUSLIM00**	0.093	**BUDDHA**	0.108
MUSLIM00	0.043	**BUDDHA**	0.093	**AVELF**	0.105
BUDDHA	0.042	**AVELF**	0.089	**GVR61**	0.104
MINING	0.04	RERD	0.086	DENS60	0.086
OTHFRAC	0.034	TROPPOP	0.071	RERD	0.082
SPAIN	0.032	OPENDEC1	0.067	OTHFRAC	0.080
OPENDEC1	0.031	OTHFRAC	0.063	OPENDEC1	0.076
ABSLATIT	0.029	PRIEXP70	0.062	PRIGHTS	0.066
PRIEXP70	0.028	H60	0.061	GOVSH61	0.063
H60	0.027	GOVSH61	0.058	H60	0.061
GOVSH61	0.024	DENS60	0.055	TROPPOP	0.058
DENS60	0.022	PRIGHTS	0.053	PRIEXP70	0.053
FERTLDC1	0.022	ABSLATIT	0.052	GGCFD3	0.048
POP1560	0.018	PROT00	0.048	PROT00	0.046
PROT00	0.018	POP1560	0.046	HINDU00	0.045

Chapter 4

Conclusions and Discussion

In this thesis, we proposed two new noise models and approximation methods to the marginal likelihood. In particular, we concentrated on two common assumptions, and its relaxations allowing for noise:

- Sparsity assumption of the precision matrix in the Gaussian graphical model.
- Sparsity assumption of the coefficients in linear regression.

We relaxed these assumptions to allow for small negligible non-zero partial correlations and regression coefficients, respectively. Efficient estimation of the marginal likelihood and calibration to the actual degree of noise is key to our methods, and we discuss these general issues in the following.

4.1 Methods for marginal likelihood estimation

Not only for the noise models that we proposed in Chapter 2 and 3, but for most realistic Bayesian models, there is no closed-form analytic solution of the marginal likelihood. Therefore, efficient method for calculating the marginal likelihood are crucial. There are roughly five types of methods for the exact or approximate calculation of the marginal likelihood.

- Numerical integration, like Bayesian Quadrature (e.g. Osborne et al. (2012)).
- MCMC methods for explicit calculation, like Chib's method (Chib, 1995; Chib and Jeliazkov, 2001), CAME (Pajor et al., 2017), or SMC (Zhou et al., 2016).
- MCMC methods for implicit calculation, that means only estimation of marginal likelihood ratios (Bayes factors) through indicator variables (Green, 1995; Green and Hastie, 2009).

- Laplace approximation (e.g. Ando (2010)).
- Variational methods (Blei et al., 2017).

Numerical integration is only computationally feasible in low dimensions. On the other hand, MCMC methods' accuracy depend on the ability to guarantee convergence to the stationary distribution, and thus the ability to acquire samples from the true posterior distribution. However, in high dimensional parameter spaces, guaranteeing convergence to the stationarity distribution is a formidable task. Therefore, we think that in practice, MCMC methods can be considered more an approximation rather than being exact methods.

In Chapter 3, we side-stepped the explicit calculation of marginal likelihoods, and instead introduced a latent variable vector $\mathbf{z} \in \{0,1\}^p$ which indicates the selected model. By sampling from the posterior distribution of $\mathbf{z}$, we estimated the posterior model probabilities. In this case efficient sampling of the posterior distribution was possible, since we could integrate over the regression coefficients β_j, and therefore the reversible jump MCMC methodology Green (1995); Green and Hastie (2009) was not necessary. Since, in most cases only the ratios of marginal likelihoods (Bayes factors) are of interest, such implicit methods are often sufficient.

In certain situations, the Laplace approximation can be a viable alternative to exact methods, and enjoys asymptotic correctness, if certain regularity conditions are met. However, for the disjunct support priors introduced in Chapter 3, theses regularity conditions were not met. A Laplace approximation was also not appropriate in Chapter 2, since the parameters were in the space of positive definite matrices, whereas a Laplace approximation assumes the Euclidean space.

Variational methods are another alternative to MCMC methods. Variational methods minimize the distance between a family of approximate distributions and the target distribution (which is in most cases the posterior distribution). As a (pseudo) distance measure, commonly the KL-divergence is used due to computational convenience, leading to a lower bound (called evidence lower bound, ELBO) on the marginal likelihood. However, there are no theoretic guarantees on the tightness of this lower bound. Furthermore, the calculation of the ELBO involves the calculation of an expectation which is also not analytically tractable, in general, though Monte Carlo approximations can partly overcome this limitation (Kucukelbir et al., 2017). In Chapter 2, we circumvented some of the analytically intractable expectations that would have been needed for a full variational approximation, by using a type of mode matching between the approximate and target distribution.

4.2 Robustness to small negligible noise

Our experiments in Chapter 2, as well as Chapter 3, show that often the performance differences between with and without noise models are most striking for large n, and less important for small sample sizes. This is similar to the findings in Miller and Dunson (2018), where they illustrate the phenomena with

a simple Bernoulli example: let us assume, we want to test for the hypothesis $H_0 : \theta = 0.5$, against the alternative $H_1 : \theta \neq 0.5$. Furthermore, assume that the true value of θ is not exactly 0.5, but due to noise actually 0.51. Using standard posteriors of the Bernoulli model[1] with a uniform prior, for samples sizes n roughly smaller or equal to 1000, the null hypothesis H_0 is favored correctly (in the sense that it is robust to noise). However, for large sample sizes $n > 10000$, the alternative hypothesis H_1 is going to be favored.

As a remedy to this problem, Miller and Dunson (2018) introduced an approach which they call c-posterior. Our approach, as well as the c-posterior approach, requires to specify a hyper-parameter controlling the robustness to such noise. We showed that standard values, as in Chapter 2, or estimates on performance (like mean squared error for regression tasks) as in Chapter 3, can lead to reasonable choices: when there is no noise they lead to similar performance as standard models, but in the case of noise, they can be considerably more accurate. As such, our proposed methods come with some guidance on the choice of hyper-parameters that proved to be useful. However, we admit that there are situations where default/automatic choices of the hyper-parameters are uncomfortable. But we note that this is similar to the specification of priors in general, where truly objective prior specification can be difficult (Berger et al., 2001). In such applications, it depends on the data analyst to strike a balance between ease of interpretability (model complexity) and fit to the data, as also suggested in (Miller and Dunson, 2018).

[1] That means not their proposed c-posteriors.

Appendices

Appendix A

Variable Clustering in the Gaussian Graphical Model

A.1 Convergence of 3-block ADMM

We can write the optimization problem in (2.4) as

$$\begin{aligned}
&\text{minimize } f_1(X_\epsilon) + f_2(X_1, \ldots, X_k) + f_3(Z) \\
&\text{subject to} \\
&\quad -X - \beta X_\epsilon + Z = 0\,, \\
&X_\epsilon, X_1, \ldots, X_k \succ 0\,,
\end{aligned}$$

with

$$\begin{aligned}
f_1(X_\epsilon) &:= \operatorname{trace}(A_\epsilon X_\epsilon) - a_\epsilon \cdot \log|X_\epsilon|\,, \\
f_2(X_1, \ldots, X_k) &:= \sum_{j=1}^{k} \Big(\operatorname{trace}(A_j X_j) - a_j \cdot \log|X_j|\Big)\,, \\
f_3(Z) &:= n \cdot \operatorname{trace}(SZ) - n \cdot \log|Z|\,.
\end{aligned}$$

First note that the functions f_1, f_2 and f_3 are convex proper closed functions. Since $X_\epsilon, X_1, \ldots, X_k \succ 0$, we have due to the equality constraint that $Z \succ 0$. Assuming that the global minima is attained, we can assume that $Z \preceq \sigma I$, for some large enough $\sigma > 0$. As a consequence, we have that $\nabla^2 f_3(Z) = Z^{-1} \otimes Z^{-1} \succeq \sigma^{-2} I$, and therefore f_3 is a strongly convex function. Analogously, we have that f_1 and f_2 are strongly convex functions, and therefore also coercive. This allows us to apply Theorem 3.2 in (Lin et al., 2018) which guarantees the convergence of the 3-block ADMM.

A.2 Derivation of variational approximation

Here, we give more details of the KL-divergence minimization from Section 2.4.2. Recall, that the remaining parameters $\nu_{g,\epsilon} \in \mathbb{R}$ and $\nu_{g,j} \in \mathbb{R}$ are optimized by minimizing the KL-divergence between the the factorized distribution g and the posterior distribution $p(\Sigma_\epsilon, \Sigma_1, \dots \Sigma_k | \mathbf{x}_1, ..., \mathbf{x}_n, \boldsymbol{\eta}, \mathcal{C})$. We have

$$
\begin{aligned}
KL(g||p) = &- \int g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j) \\
&\log \frac{p(\Sigma_\epsilon, \Sigma_1, \dots \Sigma_k, \mathbf{x}_1, ..., \mathbf{x}_n | \boldsymbol{\eta}, \mathcal{C})}{g_\epsilon(\Sigma_\epsilon) \cdot \prod_{j=1}^{k} g_j(\Sigma_j)} d\Sigma_\epsilon d\Sigma \\
&+ c \\
= &-\frac{1}{2} \mathbb{E}_{g_J, g_\epsilon}[n \cdot \log |(\Sigma^{-1} + \beta \Sigma_\epsilon^{-1})|] \\
&- \frac{1}{2} \mathbb{E}_{g_\epsilon}[(\nu_\epsilon + d + 1) \cdot \log |\Sigma_\epsilon^{-1}| \\
&- \operatorname{trace}((\Sigma_{\epsilon,0} + \beta n S)\Sigma_\epsilon^{-1})] - \operatorname{Entropy}[g_\epsilon] \\
&+ \sum_{j=1}^{k} \Big(- \frac{1}{2} \mathbb{E}_{g_j}[(\nu_j + d_j + 1) \cdot \log |\Sigma_j^{-1}| \\
&- \operatorname{trace}((\Sigma_{j,0} + n S_j)\Sigma_j^{-1})] - \operatorname{Entropy}[g_j] \Big) + c
\end{aligned}
$$

$$
\begin{aligned}
= &-\frac{1}{2} n \, \mathbb{E}_{g_J, g_\epsilon}[\log |\Sigma^{-1} + \beta \Sigma_\epsilon^{-1}|] \\
&+ \frac{1}{2} (\nu_\epsilon + d + 1) \, \mathbb{E}_{g_\epsilon}[\log |\Sigma_\epsilon|] \\
&+ \frac{1}{2} \operatorname{trace}((\Sigma_{\epsilon,0} + \beta n S) \, \mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) - \operatorname{Entropy}[g_\epsilon] \\
&+ \frac{1}{2} \sum_{j=1}^{k} (\nu_j + d_j + 1) \, \mathbb{E}_{g_j}[\log |\Sigma_j|] \\
&+ \frac{1}{2} \sum_{j=1}^{k} \operatorname{trace}((\Sigma_{j,0} + n S_j) \, \mathbb{E}_{g_j}[\Sigma_j^{-1}]) \\
&- \sum_{j=1}^{k} \operatorname{Entropy}[g_j] + c \,,
\end{aligned}
$$

where c is a constant with respect to g_ϵ and g_j. However, the term $E_{g_J, g_\epsilon}[\log |\Sigma^{-1} + \beta \Sigma_\epsilon^{-1}|]$ cannot be solved analytically, therefore we need to resort to some sort of approximation. Assuming that

$$
E_{g_J, g_\epsilon}[\log |\Sigma^{-1} + \beta \Sigma_\epsilon^{-1}|] \approx E_{g_J, g_\epsilon}[\log |\Sigma^{-1}|] \,,
$$

we get

$$
\begin{aligned}
KL(g||p) \approx & -\frac{1}{2} n \, \mathbb{E}_{g_J, g_\epsilon}[\log |\Sigma^{-1}|] \\
& + \frac{1}{2}(\nu_\epsilon + d + 1) \, \mathbb{E}_{g_\epsilon}[\log |\Sigma_\epsilon|] \\
& + \frac{1}{2} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S) \, \mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) - \mathrm{Entropy}[g_\epsilon] \\
& + \frac{1}{2} \sum_{j=1}^{k} (\nu_j + d_j + 1) \, \mathbb{E}_{g_j}[\log |\Sigma_j|] \\
& + \frac{1}{2} \sum_{j=1}^{k} \mathrm{trace}((\Sigma_{j,0} + n S_j) \, \mathbb{E}_{g_j}[\Sigma_j^{-1}]) \\
& - \sum_{j=1}^{k} \mathrm{Entropy}[g_j] + c \\
= & - \mathbb{E}_{g_\epsilon}[\log \Big(|\Sigma_\epsilon|^{-\frac{1}{2}(\nu_\epsilon + d + 1)} \\
& e^{-\frac{1}{2}\mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\Sigma_\epsilon^{-1})} \Big)] \\
& - \mathrm{Entropy}[g_\epsilon] - \sum_{j=1}^{k} \mathbb{E}_{g_j}[\log \Big(|\Sigma_j|^{-\frac{1}{2}(\nu_j + n + d_j + 1)} \\
& e^{-\frac{1}{2}\mathrm{trace}((\Sigma_{j,0} + n S_j)\Sigma_j^{-1})} \Big)] + \mathrm{Entropy}[g_j] + c \\
= & - \mathbb{E}_{g_\epsilon}[\log \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)] \\
& - \mathrm{Entropy}[g_\epsilon] \\
& - \sum_{j=1}^{k} \mathbb{E}_{g_j}[\log \mathrm{InvW}(\nu_j + n, \Sigma_{j,0} + n S_j)] \\
& + \mathrm{Entropy}[g_j] + c' \\
= & \, KL(g_\epsilon \,||\, \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)) \\
& + \sum_{j=1}^{k} KL(g_j \,||\, \mathrm{InvW}(\nu_j + n, \Sigma_{j,0} + n S_j)) \\
& + c' \, ,
\end{aligned}
$$

where we used that $\mathbb{E}_{g_J, g_\epsilon}[\log |\Sigma^{-1}|]$
$= - \sum_{j=1}^{k} \mathbb{E}_{g_j}[\log |\Sigma_j|]$, and c' is a constant with respect to g_ϵ and g_j.

From the above expression, we see that we can optimize the parameters of g_ϵ

and g_j independently from each other. The optimal parameter $\hat{\nu}_{g,\epsilon}$ for g_ϵ is

$$
\begin{aligned}
\hat{\nu}_{g,\epsilon} &= \arg\min_{\nu_{g,\epsilon}} KL(g_\epsilon \,||\, \mathrm{InvW}(\nu_\epsilon, \Sigma_{\epsilon,0} + \beta n S)) \\
&= \arg\min_{\nu_{g,\epsilon}} (\nu_\epsilon + d + 1)\, \mathbb{E}_{g_\epsilon}[\log|\Sigma_\epsilon|] \\
&\quad + \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\, \mathbb{E}_{g_\epsilon}[\Sigma_\epsilon^{-1}]) - 2 \cdot \mathrm{Entropy}[g_\epsilon] \\
&= \arg\min_{\nu_{g,\epsilon}} (\nu_\epsilon + d + 1)\Big(- d\log 2 + d\log(\nu_{g,\epsilon} + d + 1) \\
&\quad + \log|\hat{\Sigma}_\epsilon| - \sum_{i=1}^{d} \psi\Big(\frac{\nu_{g,\epsilon} - d + i}{2}\Big)\Big) \\
&\quad + \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon} + d + 1} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\hat{\Sigma}_\epsilon^{-1}) \\
&\quad - 2\log\Gamma_d(\frac{\nu_{g,\epsilon}}{2}) - \nu_{g,\epsilon} d - d(d+1)\log(\nu_{g,\epsilon} + d + 1) \\
&\quad + (\nu_{g,\epsilon} + d + 1) \sum_{i=1}^{d} \psi\Big(\frac{\nu_{g,\epsilon} - d + i}{2}\Big) \\
&= \arg\min_{\nu_{g,\epsilon}} p(\nu_\epsilon + d + 1)\log(\nu_{g,\epsilon} + d + 1) \\
&\quad - (\nu_\epsilon + d + 1) \sum_{i=1}^{d} \psi\Big(\frac{\nu_{g,\epsilon} - d + i}{2}\Big) \\
&\quad + \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon} + d + 1} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\hat{\Sigma}_\epsilon^{-1}) \\
&\quad - 2\log\Gamma_d(\frac{\nu_{g,\epsilon}}{2}) - \nu_{g,\epsilon} d - d(d+1)\log(\nu_{g,\epsilon} + d + 1) \\
&\quad + (\nu_{g,\epsilon} + d + 1) \sum_{i=1}^{d} \psi\Big(\frac{\nu_{g,\epsilon} - d + i}{2}\Big) \\
&= \arg\min_{\nu_{g,\epsilon}} \frac{\nu_{g,\epsilon}}{\nu_{g,\epsilon} + d + 1} \mathrm{trace}((\Sigma_{\epsilon,0} + \beta n S)\hat{\Sigma}_\epsilon^{-1}) \\
&\quad - 2\log\Gamma_d(\frac{\nu_{g,\epsilon}}{2}) - \nu_{g,\epsilon} d + d\nu_\epsilon \log(\nu_{g,\epsilon} + d + 1) \\
&\quad + (\nu_{g,\epsilon} - \nu_\epsilon) \sum_{i=1}^{d} \psi\Big(\frac{\nu_{g,\epsilon} - d + i}{2}\Big)\,.
\end{aligned}
$$

And analogously, we have

$$\begin{aligned}\hat{\nu}_{g,j} = \arg\min_{\nu_{g,j}} \; & \frac{\nu_{g,j}}{\nu_{g,j}+d_j+1}\mathrm{trace}((\Sigma_{j,0}+nS_j)\hat{\Sigma}_j^{-1}) \\ & - 2\log\Gamma_{d_j}(\frac{\nu_{g,j}}{2}) - \nu_{g,j}d_j \\ & + d_j(\nu_j+n)\log(\nu_{g,j}+d_j+1) \\ & + (\nu_{g,j}-\nu_j-n)\sum_{i=1}^{d_j}\psi\Big(\frac{\nu_{g,j}-d_j+i}{2}\Big)\,.\end{aligned}$$

A.3 Spectral clustering for variable clustering

Let $S \in \mathbb{R}^{d\times d}$ denote the sample covariance matrix of the observed variables. Under the assumption that the observations are drawn i.i.d. from a multivariate normal distribution, with mean $\mathbf{0}$ and precision matrix $X + \beta X_\epsilon$, the log-likelihood[1] of the data is given by

$$\frac{n}{2}(\log|X+\beta X_\epsilon| - \mathrm{trace}((X+\beta X_\epsilon)S))\,,$$

where n is the number of observations. We assume that X is block sparse, i.e. a permutation matrix P exists such that P^TXP is block diagonal. If we knew the number of blocks k, then we could estimate the block matrix X (and thus the variable clustering) by the following optimization problem.

Optimization Problem 1:

$$\begin{aligned}&\underset{X\succ 0}{\text{minimize}} - \log|X+\beta X_\epsilon| + \mathrm{trace}((X+\beta X_\epsilon)S) \\ &\text{subject to} \\ &\text{X is block sparse with exactly } k \text{ blocks}\,,\end{aligned}$$

where βX_ϵ is assumed to be a constant matrix with small entries. We claim that this can be reformulated, for any $q > 0$, as following.

Optimization Problem 2:

$$\begin{aligned}&\underset{X\succ 0}{\text{minimize}} - \log|X+\beta X_\epsilon| + \mathrm{trace}((X+\beta X_\epsilon)S) \\ &\text{subject to} \\ &L_{ii} = \sum_{k\neq i}|X_{ik}|^q\,, \\ &L_{ij} = -|X_{ij}|^q \;\text{ for } i\neq j, \\ &rank(L) = p-k\,.\end{aligned}$$

[1] Up to a constant that does not depend on X.

Proposition 1. *Optimization problem 1 and 2 have the same solution. Moreover, the k dimensional null space of L can be chosen such that each basis vector is the indicator vector for one variable block of X.*

Proof. First let us define the matrix $\tilde{X}$, by $\tilde{X}_{ij} := |X_{ij}|^q$. Then clearly, iff X is block sparse with k blocks, so is $\tilde{X}$. Furthermore, $\tilde{X}_{ij} \geq 0$, and L is the unnormalized Laplacian as defined in (Von Luxburg, 2007). We can therefore apply Proposition (2) of (Von Luxburg, 2007), to find that the dimension of the eigenspace of L corresponding to eigenvalue 0, is exactly the number of blocks in $\tilde{X}$. Also from Proposition (2) of (Von Luxburg, 2007) it follows that each such eigenvector $\mathbf{e}_k \in \mathbb{R}^d$ can be chosen such that it indicates the variables belonging to the same block, i.e. $\mathbf{e}_k(i) \neq 0$, iff variable i belongs to block k. □ □

Using the nuclear norm as a convex relaxation for the rank constraint, we have

$$\begin{aligned}
&\underset{X \succeq 0}{\text{minimize}} - \log|X + \beta X_\epsilon| + \text{trace}((X + \beta X_\epsilon)S) + \lambda_k ||L||_* \\
&\text{subject to} \\
&L_{ii} = \sum_{k \neq i} |X_{ik}|^q \,, \\
&L_{ij} = -|X_{ij}|^q \;\; \text{for } i \neq j \,.
\end{aligned}$$

with an appropriately chosen λ_k. By the definition of L, we have that L is positive semi-definite, and therefore $||L||_* = \text{trace}(L)$. As a consequence, we can rewrite the above problem as

$$\begin{aligned}
X^* := \underset{X \succeq 0}{\arg\min} &- \log|X + \beta X_\epsilon| + \text{trace}((X + \beta X_\epsilon)S) \\
&+ \lambda_k \sum_{i \neq j} |X_{ij}|^q \,.
\end{aligned}$$

Finally, for the purpose of learning the Laplacian L, we ignore the term βX_ϵ and set it to zero. This will necessarily lead to an estimate of X^* that is not a clean block matrix, but has small non-zero entries between blocks. Nevertheless, spectral clustering is known to be robust to such violations (Ng et al., 2002). This leads to Algorithm 2 in Section 2.4.3.

Appendix B

Disjunct Support Prior for Variable Selection in Regression

B.1 Slice sampler

First, let us introduce the auxiliary random variable U, and the following joint distribution:

$$p(U, \sigma_1^2) = \begin{cases} \frac{1}{L} \cdot \text{Inv-}\chi^2(\sigma_1^2|\tilde{\nu}, \tilde{\eta}^2) & \text{if } 0 < U < h(\sigma_1^2), \\ 0 & \text{else.} \end{cases},$$

where L is an appropriate normalization constant. We then have that

$$\begin{aligned} p(\sigma_1^2) &= \int_0^{h(\sigma_1^2)} p(u, \sigma_1^2) du \\ &= \frac{1}{L} \cdot \text{Inv-}\chi^2(\sigma_1^2|\tilde{\nu}, \tilde{\eta}^2) \int_0^{h(\sigma_1^2)} 1 du \\ &= \frac{1}{L} \cdot \text{Inv-}\chi^2(\sigma_1^2|\tilde{\nu}, \tilde{\eta}^2) \big[u\big]_0^{h(\sigma_1^2)} \\ &= \frac{1}{L} \cdot h(\sigma_1^2) \cdot \text{Inv-}\chi^2(\sigma_1^2|\tilde{\nu}, \tilde{\eta}^2) \\ &\propto h(\sigma_1^2) \cdot \text{Inv-}\chi^2(\sigma_1^2|\tilde{\nu}, \tilde{\eta}^2) \,. \end{aligned}$$

In order to sample from the joint distribution $p(U, \sigma_1^2)$, we employ a Gibbs sampler, where

$$p(U|\sigma_1^2) = \text{Uniform}([0, h(\sigma_1^2)]) \,,$$

and

$$p(\sigma_1^2|u) = \begin{cases} \frac{1}{\tilde{L}} \cdot \text{Inv-}\chi^2(\sigma_1^2|\tilde{\nu},\tilde{\eta}^2) & \text{if } h(\sigma_1^2) > u, \\ 0 & \text{else.} \end{cases},$$

for an appropriate normalization constant $\tilde{L}$.

B.2 Asymptotic approximation of $p(\mathbf{y}_n|X_n, S)$

In order to approximate the marginal likelihood $p(\mathbf{y}_n|X_n, S)$, we use the Laplace approximation from Theorem 1 in (Kass et al., 1990). The likelihood function of the normal linear model is Laplace regular (see proof in Kass et al. (1990)), which means that the conditions on the likelihood function in Theorem 1 (Kass et al., 1990) hold. Let us denote by Θ_S the Cartesian product of the support of the priors $p(\boldsymbol{\beta}|S)$ and $p(\sigma_r^2)$ (for technical reasons we may exclude the points at δ and $-\delta$ to make Θ_S an open subset of $\mathbb{R}^{d+1}$). Since the densities of the Cauchy distribution, the normal distribution, and the scaled inverse chi-square distribution, are four times continuously differentiable, we have that the priors $p(\boldsymbol{\beta}|S)$ and $p(\sigma_r^2)$ are four times continuously differentiable on its support.

Let $\hat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimate (MLE) for $\log p(\mathbf{y}_n|X_n, \boldsymbol{\theta})$. Note that by the consistency of the MLE, we have that $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_t$ (see for example Theorem 4.17. in Shao (2003)), therefore for any open ball around $\boldsymbol{\theta}_t$, denoted by $\mathcal{B}(\boldsymbol{\theta}_t)$, we have $P(\hat{\boldsymbol{\theta}}_n \in \mathcal{B}(\boldsymbol{\theta}_t)) \to 1$, and therefore $P(\hat{\boldsymbol{\theta}}_n \in \Theta_S) \to 1$.

Therefore, all conditions of Theorem 1 (Kass et al., 1990) are met. Let us define $p(\boldsymbol{\theta}|S) := p(\boldsymbol{\beta}|S) \cdot p(\sigma_r^2)$, and $h(\boldsymbol{\theta}) := -\frac{1}{n} \log p(\mathbf{y}_n|X_n, \boldsymbol{\theta})$. Next, applying Theorem 7 and 1 from (Kass et al., 1990), we have almost surely that [1]

$$\begin{aligned} p(\mathbf{y}_n|X_n, S) &= \int_{\Theta_S} p(\mathbf{y}_n|X_n, \boldsymbol{\theta}) p(\boldsymbol{\theta}|S) d\boldsymbol{\theta} \\ &= (2\pi)^{d+1} \cdot \left[\det(n \cdot \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n))\right]^{-\frac{1}{2}} \cdot p(\mathbf{y}_n|X_n, \hat{\boldsymbol{\theta}}_n) \cdot (p(\hat{\boldsymbol{\theta}}_n|S) + O(n^{-1})) \,. \end{aligned}$$

Furthermore, we have that

$$\begin{aligned} \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} h(\hat{\boldsymbol{\theta}}_n) &= -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}}_n) \\ &\xrightarrow{a.s.} -\mathbb{E}_{\mathbf{x},y} \left[\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(y|\mathbf{x}, \hat{\boldsymbol{\theta}}_n) \right]. \end{aligned}$$

Since $\boldsymbol{\theta} \mapsto \mathbb{E}_{\mathbf{x},y} \left[\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(y|\mathbf{x}, \boldsymbol{\theta}) \right]$ is a continuous function, and $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_t$, we have by the continuous mapping theorem that

$$\mathbb{E}_{\mathbf{x},y} \left[\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(y|\mathbf{x}, \hat{\boldsymbol{\theta}}_n) \right] \xrightarrow{p} \mathbb{E}_{\mathbf{x},y} \left[\frac{\partial^2}{\partial^2 \boldsymbol{\theta}} \log p(y|\mathbf{x}, \boldsymbol{\theta}_t) \right],$$

[1] We use here the notation $\det(X)$ for the determinant of a matrix X in order to avoid confusion with the absolute value function.

and since the matrix $-\mathbb{E}_{\mathbf{x},y}\left[\frac{\partial^2}{\partial^2\boldsymbol{\theta}}\log p(y|\mathbf{x},\boldsymbol{\theta}_t)\right]$ is positive definite with every entry in $O(1)$, we have that $\log\det(\frac{\partial^2}{\partial^2\boldsymbol{\theta}}h(\hat{\boldsymbol{\theta}}_n)) \in O_p(1)$. In summary, we have

$$\begin{aligned}\log p(\mathbf{y}_n|X_n, S) &= (d+1)\log(2\pi) - \frac{d+1}{2}\log n \\ &\quad - \frac{1}{2}\log\det(\frac{\partial^2}{\partial^2\boldsymbol{\theta}}h(\hat{\boldsymbol{\theta}}_n)) + \log p(\mathbf{y}_n|X_n, \hat{\boldsymbol{\theta}}_n) + \log(p(\hat{\boldsymbol{\theta}}_n|S) + O(n^{-1})) \\ &= -\frac{d+1}{2}\log n + \log p(\mathbf{y}_n|X_n, \hat{\boldsymbol{\theta}}_n) + O_p(1)\,.\end{aligned}$$

Bibliography

Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In Kitagawa G. Parzen E., Tanabe K., editor, *Reprint in Breakthroughs in statistics, 1992*, pages 610–624. Springer, 1973.

Bruce Alberts, Alexander Johnson, Julian Lewis, David Morgan, and Martin Raff. *Molecular biology of the cell*. Garland Science, 2014.

Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 3. Wiley New York, 2004.

Tomohiro Ando. *Bayesian model selection and statistical modeling*. Chapman and Hall/CRC, 2010.

David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

Maria J Bayarri, James O Berger, Anabel Forte, G García-Donato, et al. Criteria for Bayesian model choice with application to variable selection. *Annals of Statistics*, 40(3):1550–1577, 2012.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

James O Berger and Mohan Delampady. Testing precise hypotheses. *Statistical Science*, pages 317–335, 1987.

James O Berger, Luis R Pericchi, JK Ghosh, Tapas Samanta, Fulvio De Santis, JO Berger, and LR Pericchi. Objective Bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series*, pages 135–207, 2001.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112 (518):859–877, 2017.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction

method of multipliers. *Foundations and Trends in Machine Learning*, 3(1): 1–122, 2011.

Richard P Brent. Algorithms for finding zeros and extrema of functions without calculating derivatives. Technical report, Stanford University, Department of Computer Science, 1971.

Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455, 1998.

Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.

Bradley P Carlin and Thomas A Louis. *Bayesian methods for data analysis*. CRC Press, 2008.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1), 2017.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Siddhartha Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96(453): 270–281, 2001.

Hugh Chipman, Edward I George, Robert E McCulloch, Merlise Clyde, Dean P Foster, and Robert A Stine. The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, pages 65–134, 2001.

Jyotishka Datta and David B Dunson. Bayesian inference on quasi-sparse count data. *Biometrika*, 103(4):971–983, 2016.

Emilie Devijver and Mélina Gallopin. Block-diagonal covariance selection for high-dimensional Gaussian graphical models. *Journal of the American Statistical Association*, 113(521):306–314, 2018.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, et al. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

Carmen Fernandez, Eduardo Ley, and Mark FJ Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427, 2001.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Gonzalo Garcia-Donato and Miguel A Martinez-Beneito. On sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013.

Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

Peter J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

Peter J Green and David I Hastie. Reversible jump MCMC. *Genetics*, 155(3): 1391–1403, 2009.

P Richard Hahn and Carlos M Carvalho. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448, 2015.

Chris Hans, Adrian Dobra, and Mike West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516, 2007.

Kei Hirose, Hironori Fujisawa, and Jun Sese. Robust sparse Gaussian graphical modeling. *Journal of Multivariate Analysis*, 161:172–190, 2017.

Seyed Mohammad Javad Hosseini and Su-In Lee. Learning sparse Gaussian graphical models with overlapping blocks. In *Advances in Neural Information Processing Systems*, pages 3801–3809, 2016.

Hemant Ishwaran, J Sunil Rao, and Others. Spike and slab variable selection: frequentist and Bayesian strategies. *Annals of Statistics*, 33(2):730–773, 2005.

Valen E Johnson and David Rossell. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170, 2010.

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

Robert E Kass, Luke Tierney, and Joseph B Kadane. The validity of posterior expansions based on Laplace's method. *Bayesian and likelihood methods in statistics and econometrics*, 7:473, 1990.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sadanori Konishi, Tomohiro Ando, and Seiya Imoto. Bayesian information criteria and smoothing parameter selection in radial basis function networks. *Biometrika*, 91(1):27–43, 2004.

Alp Kucukelbir, Dustin Tran, Rajesh Ranganath, Andrew Gelman, and David M Blei. Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474, 2017.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2): 365–411, 2004.

Alex Lenkoski and Adrian Dobra. Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *Journal of Computational and Graphical Statistics*, 20(1):140–157, 2011.

Feng Liang, Rui Paulo, German Molina, Merlise A Clyde, and Jim O Berger. Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

Tianyi Lin, Shiqian Ma, and Shuzhong Zhang. Global convergence of unmodified 3-block ADMM for a class of convex minimization problems. *Journal of Scientific Computing*, 76(1):69–88, 2018.

Hedibert F Lopes and Nicholas G Polson. Bayesian hypothesis testing: Redux. *arXiv preprint arXiv:1808.08491*, 2018.

Benjamin M Marlin and Kevin P Murphy. Sparse Gaussian graphical models with unknown block structure. In *International Conference on Machine Learning*, pages 705–712. ACM, 2009.

Benjamin M Marlin, Mark Schmidt, and Kevin P Murphy. Group sparse priors for covariance estimation. In *Conference on Uncertainty in Artificial Intelligence*, pages 383–392. AUAI Press, 2009.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Jeffrey W Miller and David B Dunson. Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, pages 1–13, 2018.

Andrew Y Ng, Michael I Jordan, Yair Weiss, and Others. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856, 2002.

Anthony O'Hagan. Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–118, 1995.

Michael Osborne, Roman Garnett, Zoubin Ghahramani, David K Duvenaud, Stephen J Roberts, and Carl E Rasmussen. Active learning of model evidence using Bayesian quadrature. In *Advances in neural information processing systems*, pages 46–54, 2012.

Anna Pajor et al. Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis*, 12(1):261–287, 2017.

Konstantina Palla, Zoubin Ghahramani, and David A Knowles. A nonparametric variable clustering model. In *Advances in Neural Information Processing Systems*, pages 2987–2995, 2012.

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

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

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, pages 814–822, 2014.

Christian Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science & Business Media, 2007.

Veronika Ročková and Edward I George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506): 828–846, 2014.

Xavier Sala-i Martin, Gernot Doppelhofer, and Ronald I Miller. Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American Economic Review*, pages 813–835, 2004.

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

James G Scott and James O Berger. An exploration of aspects of Bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162, 2006.

James G Scott and Carlos M Carvalho. Feature-inclusion stochastic search for Gaussian graphical models. *Journal of Computational and Graphical Statistics*, 17(4):790–808, 2008.

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

Jun Shao. *Mathematical Statistics*. Springer, 2003.

Siqi Sun, Yuancheng Zhu, and Jinbo Xu. Adaptive Variable Clustering in Gaussian Graphical Models. In *International Conference on Artificial Intelligence and Statistics*, pages 931–939, 2014.

Siqi Sun, Hai Wang, and Jinbo Xu. Inferring block structure of Graphical models in exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 939–947, 2015.

Kean Ming Tan, Daniela Witten, and Ali Shojaie. The cluster graphical lasso for improved estimation of Gaussian graphical models. *Computational Statistics & Data Analysis*, 85:23–36, 2015.

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

C Andy Tsao. A note on Lindley's paradox. *Test*, 15(1):125–139, 2006.

Mark van der Wilk, Matthias Bauer, ST John, and James Hensman. Learning invariances using the marginal likelihood. In *Advances in Neural Information Processing Systems*, pages 9938–9948, 2018.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Andrew M Walker. On the asymptotic behaviour of posterior distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(1): 80–88, 1969.

Sumio Watanabe. A widely applicable Bayesian information criterion. *Journal of Machine Learning Research*, 14(Mar):867–897, 2013.

David L Weakliem. *Hypothesis testing and model selection in the social sciences*. Guilford Publications, 2016.

Yuhong Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

Arnold Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.

Yan Zhou, Adam M Johansen, and John AD Aston. Toward automatic model comparison: an adaptive sequential Monte Carlo approach. *Journal of Computational and Graphical Statistics*, 25(3):701–726, 2016.