LUDWIG-MAXIMILIANS-UNIVERSITY MUNICH

DEPARTMENT OF STATISTICS

Variable selection using grouped horseshoe priors

Bachelor's Thesis

Department of Statistics

Author:

Tobias Pielok, B.Sc. (TUM)

Supervisor: Dr. Fabian Scheipl

Submission date: 27th February, 2019

Acknowledgement

To Fabian Scheipl, for giving me the opportunity to write about such an interesting topic and great support.

To my parents, for always believing in me.

Contents

Notation and Symbols

$\mathcal{B}_{\mathcal{M}} \bigotimes \mathcal{B}_{\mathcal{N}}$	product-σ-algebra generated by $\{M \times N : \, M \in \mathcal{B}_{\mathcal{M}}, N \in \mathcal{B}_{\mathcal{N}}\}$
$\boldsymbol{\theta} \in \mathbb{R}^d$	If $\boldsymbol{\theta}$ is a random variable: $\boldsymbol{\theta}$ maps to $\mathbb{R}^d$
$\mathrm{C}^{+}(0,a)$	standard half-Cauchy distribution on the positive reals with scale parameter a

1 Introduction

Often in data analysis the situation arises where not all parameters of a model can be well identified through the data. In these situation one possible way to estimate the model is via variable selection, which is a process where a subset of the most informative predictors is chosen, s.t. for this subset the model can be estimated for.

Variable selection can be carried out with a frequentist approach, e.g. lasso feature selection, or a Bayesian approach with so-called shrinkage priors. The most popular shrinkage priors are the spike-and-slab priors and the horseshoe priors. By using the horseshoe priors one applies global shrinkage to all variables, but allows for informative variables to locally escape the shrinkage. In this thesis the horseshoe priors are extended to the case of variables, which posses a grouped structure. Hence the priors are denoted as *grouped horseshoe priors*. In section (2) the Bayesian inference of the grouped horseshoe priors is rigorously introduced, and how this inference can it be executed computationally. In section (3) the grouped horseshoe priors are shown. With these priors it is possible to estimate additive models, for which in this thesis an suggestion for the optimal hyperparameter of the grouped horseshoe priors is derived in section (3.3). Also in this thesis the performance and shrinkage qualities of these priors are measured in two simulation studies (4) and benchmarked (5) in two real data sets against the other variable selection methods, which are mentioned above.

In this thesis it is shown, that the horseshoe prior in setting of grouped variables performs robustly in low and high sparsity data situations and can compete with other commonly used variable selection methods.

2 Basic Concepts

The foundation of Bayesian horseshoe priors is based on the inference of the posterior distribution of the parameters, which can be attained by combining the prior knowledge of the parameters and the likelihood of the data, and will be introduced in section 2.1. In order of carrying out this inference it is necessary to sample from the posterior distribution. In most cases this posterior distribution can be neither expressed in an analytical form nor are there standard samplers for it. Hence the *Hamiltonian Monte Carlo* (HMC) method and its derivative *No-U-Turn Sampler* (NUTS) will be shown in section 2.2, with which it is possible to sample from the posterior distribution even in a high dimensional case. To express in this Bayesian framework the structure of grouped variables, on which in section 3.2 the grouped horseshoe priors will be later applied, additive models will be introduced in section 2.3.

2.1 Bayesian inference

In Bayesian statistics the inference is carried out on the posterior distribution of a parameter θ, after taking into account the realization of data x. In order of defining this posterior distribution thoroughly some rigorous definitions are needed: Let $\mathcal{X}$ and $\mathcal{T}$ denote the *sample space* and *parameter space*, i.e. $\theta \in \mathcal{T}$ and $x \in \mathcal{X}$. Their σ-algebras will be called respectively $\mathcal{B}_{\mathcal{X}}$ and $\mathcal{B}_{\mathcal{T}}$. Also let the family of *sampling distribution* $\mathcal{P} := \{P_\theta : \ \theta \in \mathcal{T}\}$, where P_θ is a probability measure over $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$. With these definitions a general *statistical experiment* can be defined as $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$ and for any given probability measure μ on $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$ a *Bayesian experiment* can be identified with $(\mathcal{T} \times \mathcal{X}, \mathcal{B}_{\mathcal{T}} \bigotimes \mathcal{B}_{\mathcal{X}}, \Pi_{\mu,\mathcal{P}})$, where $\Pi_{\mu,\mathcal{P}}$ is the joint distribution of parameter-observation, s.t. $\forall T \in \mathcal{B}_{\mathcal{T}}, X \in \mathcal{B}_{\mathcal{X}}$:

$$\Pi_{\mu,\mathcal{P}}(T, X) = \int_T P_\theta(X) \mu(d\theta). \tag{2.1.1}$$

The measure μ will be called the *prior distribution* of the parameter. The *predictive distribution* of the observations $P_{\mu,\mathcal{P}}$ is the marginal distribution of $\Pi_{\mu,\mathcal{P}}$ on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$, s.t. $\forall X \in \mathcal{B}_{\mathcal{X}}$:

$$P_{\mu,\mathcal{P}}(X) = \Pi_{\mu,\mathcal{P}}(\mathcal{T}, X). \tag{2.1.2}$$

Eq. (2.1.1) can be rewritten with the predictive distribution (2.1.2) and the family of *posterior distributions* $\{\mu_{\mathcal{P}}(\cdot|\, x) :\ x \in \mathcal{X}\}$, s.t. $\forall T \in \mathcal{B}_{\mathcal{T}}, X \in \mathcal{B}_{\mathcal{X}}$

$$\Pi_{\mu,\mathcal{P}}(T,X) = \int_X \mu_{\mathcal{P}}(T|\, x) P_{\mu,\mathcal{P}}(dx). \tag{2.1.3}$$

Under the assumption that the Bayesian experiment $(\mathcal{T} \times \mathcal{X}, \mathcal{B}_{\mathcal{T}} \bigotimes \mathcal{B}_{\mathcal{X}}, \Pi_{\mu,\mathcal{P}})$ is dominated in the Bayesian setting (if not other stated, this will be assumed for the rest of the thesis) it can be shown (see [Polpo et al. (2015)]), that a posterior distribution for given x can also be expressed with the Radon-Nikodym derivative, s.t. $\forall T \in \mathcal{B}_{\mathcal{T}}$:

$$\mu_{\mathcal{P}}(T|\, x) = \frac{\int_T \frac{dP_\theta}{d\lambda}(x)\mu(d\theta)}{\int_{\mathcal{T}} \frac{dP_\theta}{d\lambda}(x)\mu(d\theta)}, \tag{2.1.4}$$

where $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, \mathcal{P})$ is dominated by a finite measure λ. The function $\frac{dP_\theta}{d\lambda}$ is called *likelihood function*. A class $\mathcal{P}_C$ of prior distributions is said to be a conjugate family for a class $\mathcal{F}$ of likelihood functions if $\mu_{\mathcal{P}}(T|\, x) \in \mathcal{P}_C\ \forall T \in \mathcal{B}_{\mathcal{T}}, X \in \mathcal{B}_{\mathcal{X}}$.
Building on these definitions inference concepts can be introduced: An adequate choice of a point estimate to determine the location of the parameter θ depends on the underlying loss function. If a quadratic loss function L is chosen, i.e. $L(\theta, \hat{\theta}) = ||\theta - \hat{\theta}||_2^2$, the expected loss $\mathbb{E}_{\theta\sim\mu}(L(\theta,\hat{\theta}))$ is minimized by the *posterior mean*

$$\hat{\theta} = E_{\theta|x\sim\mu_{\mathcal{P}}}(\theta|\, x), \tag{2.1.5}$$

if it exists (see [Berger (1985)]). A $100(1-\alpha)\%$ *credible set* is a set $C \subset \mathcal{T}$, s.t.

$$1-\alpha \leq P(C|\, x) = \mu_{\mathcal{P}}(C|\, x). \tag{2.1.6}$$

Which means that the parameter θ has the subjective probability of $(1-\alpha)$ that $\theta \in C$. Because C is not unique, additional constraints can be imposed on the solution. If the size of C, which can be defined as $S(C) := \mu(C)$, is minimized one gets under mild conditions the $100(1-\alpha)\%$ HDP credible set as described in [Berger (1985)], if there exists a posterior density $p_{\theta|\, x}$. The $100(1-\alpha)\%$ HDP credible set is defined as

$$C = \{\theta \in \mathcal{T} :\ p_{\theta|\, x}(\theta|\, x) \geq k(\alpha)\}, \tag{2.1.7}$$

where $k(\alpha)$ is the largest constant, s.t. C is a $100(1-\alpha)\%$ credible set.

2.2 HMC and NUTS

As already seen, for carrying out the Bayesian inference the posterior distribution $\mu_{\mathcal{P}}$ is needed. In general solving Eq. (2.1.3) or Eq. (2.1.4) for the posterior distribution is only in trivial cases analytically possible. Hence one tries to approximate the distribution by sampling from it. In most cases there is no standard sampler for $\mu_{\mathcal{P}}$. Assume there exists a corresponding probability density function $p(\theta)$ to $\mu_{\mathcal{P}}$. Then one way to overcome this problem is to use the HMC method, where a sequence of random samples are drawn from a known distribution using Hamiltonian dynamics. In HMC in order to sample a parameter θ, for which it holds that $\theta \sim \mu_{\mathcal{P}}$ and $\theta \in \mathbb{R}^d$, an auxiliary variable ρ is typically drawn from a multivariate normal distribution, s.t.

$$\rho \sim \mathcal{N}(0, \Sigma), \tag{2.2.1}$$

where $\rho \in \mathbb{R}^d$ and the covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. With the joint density $p(\rho, \theta)$ a so-called *Hamiltonian* can be defined with regards of the canonical distribution as

$$H(\rho, \theta) = -\log p(\rho, \theta) = \underbrace{-\log p(\rho \mid \theta)}_{=:T(\rho \mid \theta)} \underbrace{-\log p(\theta)}_{=:V(\theta)}, \tag{2.2.2}$$

where T is called kinetic energy and V the potential energy. In general a Hamiltonian can be defined as follows: Let $\rho(t), \theta(t)$ be functions, s.t. $\rho, \theta : \mathbb{R} \to \mathbb{R}^d$, and $H(\rho, \theta)$ be a scalar function sufficiently smooth. H is called a Hamiltonian if it holds that $\forall i \in \{1, \ldots, d\}$:

$$\begin{aligned} \frac{d\rho_i}{dt} &= \frac{\partial H}{\partial \theta_i}, \\ \frac{d\theta_i}{dt} &= -\frac{\partial H}{\partial \rho_i}. \end{aligned} \tag{2.2.3}$$

It can be shown that the Hamiltonian dynamics (2.2.3) is reversible and preserves volume, which can even hold for the discretized version of the differential equations (see Brooks et al. (2011)). With the *leapfrog integrator*, which uses a discrete time step ε, this can be achieved and Eq. (2.2.3) can be solved numerically stable (see Brooks et al. (2011)). Transitioning from a state (ρ, θ) to a new state (ρ^*, θ^*) in the leapfrog scheme, firstly a new ρ is drawn independently of θ and previous values of ρ and secondly L leapfrog steps are applied. One step of the leapfrog integrator can be summarized as follows:

1. A half-step update is made for the auxiliary variable ρ:
$$\rho \leftarrow \rho - \frac{\varepsilon}{2}\frac{\partial V}{\partial \theta}$$
2. With the new ρ the parameter θ is updated with a full-step:
$$\theta \leftarrow \theta + \varepsilon \Sigma \rho$$
3. Another half-step update of ρ is performed with the updated θ:
$$\rho \leftarrow \rho - \frac{\varepsilon}{2}\frac{\partial V}{\partial \theta}$$

After L leapfrog steps the proposal state (ρ^*, θ^*) is accepted as the new state with a probability of
$$\min(1, \exp(H(\rho,\theta) - H(\rho^*,\theta^*))).$$
Otherwise (ρ, θ) is returned as the new state and hence also serves as the new initialization for the next iteration. This is called the *Metropolis Acceptance Step*.

The series of θ returned by this scheme is also called a chain. As described in Durmus et al. (2017) under certain regularity conditions and control of the tail of the posterior distribution $\mu_{\mathcal{P}}$ this chain is irreducible and (Harris) recurrent and $\mu_{\mathcal{P}}$ is its so-called invariant distribution. For a chain with these properties and under some additional assumptions it can be shown that with the use of ergodic theorems, that the chain converges for almost every starting parameter θ_0 in distribution to $\mu_{\mathcal{P}}$ and the law of large numbers and the central limit theorem are still valid, s.t. inferences (2.1.5) and (2.1.7) can be carried out approximately on a converged chain (see Meyn and Tweedie (1993)).

The convergence speed of HMC is highly influenced by the hyper-parameters, i.e. the number of steps L, the step-size ε and the covariance matrix Σ. One scheme to tune the parameter L is NUTS. The advantage of using NUTS consists in its auto-tuning capability without the need to execute pre-runs. The general idea of NUTS can be described as follows: For the state (ρ, θ) the Hamiltonian dynamics are simulated randomly both forwards and backwards in time to guarantee time reversibility. In every iteration the steps taken in one direction are doubled. The algorithm is stopped for a current $(\tilde{\theta}, \tilde{\rho})$ when
$$\frac{d}{dt}\frac{||\tilde{\theta}-\theta||_2^2}{2} = (\tilde{\theta}-\theta)^T \frac{d}{dt}(\tilde{\theta}-\theta) = (\tilde{\theta}-\theta)^T \Sigma \tilde{\rho} < 0,$$

i.e. the expected squared jump distance would shrink and the dynamics could be described as an *U-turn*. From these simulated states new proposals are sampled (for further details refer to Hoffman and Gelman (2014)).

2.3 Bayesian approach to additive models

Here the basic terminology related with additive models from a Bayesian perspective is recalled and the notation that will be used in the following is fixed. As a reference it is relied on Fahrmeir and Kneib (2011).

To study additive models firstly an understanding of univariate polynomial smoothing is helpful. For n observation-pairs (x_i, y_i), where y is the output and x is its covariate, with univariate polynomial smoothing a *polynomial spline* $f(x)$ can be found, which fulfils

$$y_i = f(x_i) + \varepsilon_i, \tag{2.3.1}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ for an Gaussian observation model. This polynomial spline f of degree D over $M+1$ (not necessarily equidistant) knots can be mathematically equivalent expressed in different spline bases, s.t.

$$f(x) = \sum_{k=1}^{K} \gamma_k B_k(x), \tag{2.3.2}$$

where $K = D + M$. Commonly used spline bases are *Truncated power series* and *B-splines*. In this thesis the B-splines basis will be used, because of its superior numerical stability and its more adaptive Bayesian interpretation. Important to note is that B-splines form a local basis. With Eq. (2.3.2) Eq. (2.3.1) can be written as a linear model for all observations, s.t.

$$\mathbf{y} = \mathbf{X}\gamma + \varepsilon, \tag{2.3.3}$$

where $\mathbf{X}$ contains the basis function evaluated at the observed covariate $\mathbf{x}$.

With higher node (and polynomial) degree polynomial splines are more prone to overfitting. To counter this *P-splines* are used. Because for B-splines γ represents local regression coefficients, high differences of neighbouring γ_i should be penalized. Hence the so-called *penalized least-squares* (PLS) criterion can be formulated for B-splines with equidistant knots as

$$PLS(\lambda) = ||\mathbf{y}-\mathbf{X}\gamma||_2^2 + \lambda||\mathbf{D}_d\gamma||_2^2 = ||\mathbf{y}-\mathbf{X}\gamma||_2^2 + \lambda\gamma^T \underbrace{\mathbf{D}_d^T\mathbf{D}_d}_{=:\mathbf{K}}\gamma, \qquad (2.3.4)$$

where $\mathbf{D}_d \in \mathbb{R}^{(K-d)\times K}$ is the so-called *difference matrix* of order d and $\mathbf{K}$ the penalty matrix. The PLS criterion can be equivalently defined by using a d^{th} random walk prior distribution for γ, s.t.

$$\begin{aligned} p(\gamma,\sigma^2|\,\mathbf{y}) &\propto p(\mathbf{y}|\,\gamma,\sigma^2)p(\gamma|\,\tau^2)p(\sigma^2)p(\tau^2) \\ &\propto \exp(-\frac{1}{2\sigma^2}||\mathbf{y}-\mathbf{X}\gamma||_2^2)(\frac{1}{\tau^2})^{\text{rank}(\mathbf{K})/2}\exp(-\frac{1}{2\tau^2}\gamma^T\mathbf{K}\gamma)\cdot \\ &\quad \cdot p(\sigma^2)p(\tau^2), \end{aligned} \qquad (2.3.5)$$

where typically prior distributions of σ^2 and τ^2 are chosen, s.t.

$$\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma), \tau^2 \sim \text{IG}(a,b). \qquad (2.3.6)$$

Note because $\mathbf{K}$ has no full rank, which can be clearly seen from its construction, this means that for γ a degenerate Gaussian distribution is assumed. By introducing a parameter $\mathbf{u}$, for which it holds that

$$\mathbf{u} = \underbrace{\begin{bmatrix} I_d & 0 \\ 0 & \mathbf{D}_d \end{bmatrix}}_{=:\mathbf{S}}\gamma, \qquad (2.3.7)$$

i.e. the first d terms are the first d terms of γ and the last $K-d$ terms the d'th differences of γ, Eq. (2.3.5) can be reparametrized as

$$\begin{aligned} p(\mathbf{u},\sigma^2|\,\mathbf{y}) &\propto p(\mathbf{y}|\,\mathbf{u},\sigma^2)p(\mathbf{u}|\,\tau^2)p(\sigma^2)p(\tau^2) \\ &\propto \exp(-\frac{1}{2\sigma^2}||\mathbf{y}-\mathbf{X}\mathbf{S}^{-1}\mathbf{u}||_2^2)p(\mathbf{u}|\,\tau) \\ &\quad \cdot p(\sigma^2)p(\tau^2), \end{aligned} \qquad (2.3.8)$$

where $\mathbf{u}|\,\tau \sim N(0,\tau^2\mathbf{Z})$ with diagonal covariance matrix $\mathbf{Z}$.

Now considering the case of multiple covariates $\mathbf{x}_1,\ldots,\mathbf{x}_p$ for an output $\mathbf{y}\in\mathbb{R}^n$ it is possible to extend the smoothing splines analogously to this multivariate setting, but by doing this the number of parameter rises quickly and so does the amount of data, which would be needed to identify these. Hence one commonly used technique is to restrict the model to an additive structure, s.t. the so-called additive model[1] for $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{U}\beta + \sum_{i=1}^{m_s} \mathbf{f}_i(\mathbf{x}_i) + \varepsilon, \tag{2.3.9}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ and $\mathbf{U} \in \mathbb{R}^{n \times m_p}$ denotes the model matrix of predictors $\mathbf{u}_1, \ldots \mathbf{u}_{m_l}$, which are modelled linearly, with associated parameter vector $\beta \in \mathbb{R}^{m_p}$ and $\mathbf{f}_i$ the vector of one dimensional smoothing function to the corresponding covariate $\mathbf{x}_i$, which is defined $\forall i \in \{1, \ldots, m_s\}$. Every $\mathbf{f}_i$ can expressed as

$$\mathbf{f}_i(\mathbf{x}_i) = (f_i(x_{i1}), \ldots, f_i(x_{in}))^T = (\sum_{k=1}^{K_i} \gamma_{ik} B_k^i(x_{i1}), \ldots, \sum_{k=1}^{K_i} \gamma_{ik} B_k^i(x_{in}))^T. \tag{2.3.10}$$

To ensure the identifiability of the single $\mathbf{f}_i$ an artificial centring constraint is introduced, s.t. $\forall i \in \{1, \ldots, m_s\}$ it holds that

$$\sum_{j=1}^{n} f_i(x_{ij}) = 0. \tag{2.3.11}$$

The linearity of Eq. (2.3.10) can be expressed, s.t.

$$\mathbf{f}_i(\mathbf{x}_i) = \mathbf{X}_i \gamma_i. \tag{2.3.12}$$

With Eq. (2.3.12) the additive model (2.3.9) can be written as

$$\mathbf{y} = \mathbf{U}\beta + \sum_{i=1}^{m_s} \mathbf{X}_i \gamma_i + \varepsilon. \tag{2.3.13}$$

To estimate the regression coefficients of the additive model (2.3.13) without overfitting the extended PLS criterion, which is to be minimized, based on the estimated output $\eta = \mathbf{U}\beta + \sum_{i=1}^{m_s} \mathbf{X}_i \gamma_i$ can be used, which can be formulated as

$$\mathrm{PLS}(\gamma_1, \ldots, \gamma_{m_s}, \beta) = (\mathbf{y} - \eta)^T (\mathbf{y} - \eta) + \sum_{i=1}^{m_s} \lambda_i \gamma_i^T \mathbf{K}_i \gamma_i, \tag{2.3.14}$$

[1] In this thesis only additive model containing one dimensional splines will be used, but in general additive model can be also composed from higher dimensional smoothing functions.

where $\mathbf{K}_i$ is the penalty matrix of $\mathbf{f}_i$. In the Bayesian framework the extended PLS criterion can be modelled, s.t. it holds for the posterior distribution of the regression coefficients that

$$\begin{aligned}p(\gamma_1,\ldots,\gamma_{m_s},\beta,\tau_1^2,\ldots,\tau_{m_s}^2,\sigma^2|\,\mathbf{y}) \propto (\frac{1}{\sigma^2})^{n/2}\exp(-\frac{1}{2\sigma^2}(\mathbf{y}-\eta)^T(\mathbf{y}-\eta))\cdot\\ \cdot p(\beta)\cdot\prod_{i=1}^{m_s}(\frac{1}{\tau_i^2})^{\mathrm{rank}(K)/2}\exp(-\frac{1}{2\tau_i^2}\gamma_i^T\mathbf{K}_i\gamma_i)\cdot\\ \prod_{i=1}^{m_s}p(\tau_i^2)\cdot p(\sigma^2),\end{aligned} \tag{2.3.15}$$

i.e. the posterior mode of $\gamma_i,\ldots,\gamma_{m_s}$ fulfils the PLS criterion. Typically a weakly informative prior distribution is used for β, s.t. $\beta \sim N(\mathbf{c},\mathbf{C})$, where $\mathbf{C}$ is a 'large' covariance matrix. Because in this setting the sample density of $\mathbf{y}$ is normal and by using the fact that the normal distribution is a conjugate family for these, it can be easily shown that the full conditional $\beta|\cdot \sim N(\mu_\beta,\Sigma_\beta)$ with

$$\Sigma_\beta = \sigma^2(\mathbf{U}^T\mathbf{U}+\sigma^2\mathbf{C}^{-1})^{-1}, \tag{2.3.16}$$

$$\mu_\beta = \frac{1}{\sigma^2}\Sigma_\beta(\mathbf{U}^T(\mathbf{y}-\eta+\mathbf{U}\beta)+\sigma^2\mathbf{C}^{-1}\mathbf{c}). \tag{2.3.17}$$

In a similar fashion for the reparametrization (2.3.7) it can be shown that for $\mathbf{u}_l = \mathbf{S}_l\gamma_l$ it holds that $\mathbf{u}_l|\cdot \sim N(\mu_{u_l},\Sigma_{u_l})$ with

$$\Sigma_{u_l} = \sigma^2(\tilde{\mathbf{X}}_l^T\tilde{\mathbf{X}}_l+\frac{\sigma^2}{\tau_l}\mathbf{Z}_l^{-1})^{-1}, \tag{2.3.18}$$

$$\mu_{u_l} = \frac{1}{\sigma^2}\Sigma_{u_l}\tilde{\mathbf{X}}_l^T(\mathbf{y}-\sum_{i=1,i\neq l}^{m_s}\tilde{\mathbf{X}}_i\mathbf{u}_i-\mathbf{U}\beta), \tag{2.3.19}$$

where $\tilde{\mathbf{X}}_l = \mathbf{X}_l\mathbf{S}_l^{-1}$.

3 Bayesian horseshoe

When applying the Bayesian horseshoe technique for variable selection so-called horseshoe prior distribution is assumed for the parameters of interest, which main idea consist in using a global shrinkage parameter while allowing for an *informative* parameter to escape the shrinkage through a local shrinkage parameter. The mathematical foundations of this type of prior distribution is described in section 3.1. In section 3.2 these ideas will be extended to the case of grouped variables. In situations where the parameters can not be identified well through the data, the hyperprior choice of the global shrinkage parameter can influence the results strongly. Because of this a possible hyper prior choice is introduced in section 3.3. The implementation of the horseshoe priors, which will be applied on simulated data in section 4, uses ideas from section 3.4.

3.1 Mathematical foundations

In Piironen and Vehtari (2017) the horseshoe prior is introduced as an estimation model for an observed p-dimensional vector $\mathbf{b}|\,\beta \sim N(\beta, \Sigma)$, where β is assumed to be sparse. In this model it is assumed $\forall i \in \{1, \ldots, p\}$ that

$$\beta_i|\,\lambda_i, \tau \sim N(0, \lambda_i^2\tau^2), \quad \lambda_i \sim \mathrm{C}^+(0,1), \quad \tau \sim \mathrm{C}^+(0,1). \tag{3.1.1}$$

To understand the properties of this model better a linear Gaussian regression model is assumed for section 3.1, s.t. for an observed output $\mathbf{y}$ it holds that

$$\mathbf{y} = U\beta + \varepsilon, \tag{3.1.2}$$

with $\varepsilon \sim N(0, \sigma^2)$. In the case of the linear Gaussian model (3.1.2) if follows from Eq. (3.1.1) that the prior distribution of β is s.t. $\beta \sim N(0, \tau^2\Delta)$, where Δ is a diagonal matrix with λ_i^2 as its diagonal entries. By plugging these values into Eq. (2.3.17) one gets that the posterior expected mean of β

$$\mu_\beta = (\mathbf{U}^T\mathbf{U} + \sigma^2\frac{1}{\tau^2}\Delta^{-1})^{-1}\mathbf{U}^T\mathbf{y}. \tag{3.1.3}$$

Since Δ is a diagonal matrix Eq. (3.1.3) can be rewritten as

$$\mu_\beta = \tau^2\Delta(\sigma^2(\mathbf{U}^T\mathbf{U})^{-1} + \tau^2\Delta)^{-1}\underbrace{(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T\mathbf{y}}_{=:\hat{\beta}}, \tag{3.1.4}$$

where $\hat{\beta}$ denotes the maximum likelihood solution. In the case of zero mean and uncorrelated predictors with unit variance it holds that $\mathbf{U}^T\mathbf{U} \approx nI$, s.t. Eq. (3.1.4) can be component-wisely approximated as

$$\mu_{\beta_j} = (1-\kappa_j)\hat{\beta}_j, \tag{3.1.5}$$

where

$$\kappa_j = \frac{1}{1+n\sigma^{-2}\tau^2\lambda_j^2} \tag{3.1.6}$$

is the so-called shrinkage factor for the j'th component of β. This name makes sense, since $n\sigma^{-2}\tau^2\lambda_j^2 > 0$ it follows that $\kappa_j \in (0,1)$ and hence the limiting case of $\kappa = 1$ represents full shrinkage, where $\mu_{\beta_j} = 0$, and the other limiting case of $\kappa = 0$ no shrinkage, where $\mu_{\beta_j} = \hat{\beta}_j$. Consequently, one sees from Eq. (3.1.6) that with smaller n/σ^2, i.e. the ratio of the size of the data to the variance of the errors[2], shrinkage is more likely. This relation also holds for τ and λ, s.t. an informative variable β_k can escape the global shrinkage with a '*large*' associated λ_k. By using Eq. (3.1.1) and applying the transformation theorem it can be shown, that the conditional density of κ_j

$$p(\kappa_j|\,\tau,\sigma) = \frac{1}{\pi}\frac{\sigma^{-1}\tau\sqrt{n}}{(n\sigma^{-2}\tau^2-1)\kappa_j+1}\frac{1}{\sqrt{\kappa_j}\sqrt{1-\kappa_j}}. \tag{3.1.7}$$

For $n\sigma^{-2}\tau^2 = 1$ it holds that $\kappa_j \sim \text{Beta}(\frac{1}{2},\frac{1}{2})$ and its a horseshoe resembling density can be seen in Fig. 1, from which the name of this method was derived. While under $n\sigma^{-2}\tau^2 = 1$ equal probability mass is distributed on shrinkage and no shrinkage, under $n\sigma^{-2}\tau^2 = 0.1$ more mass is distributed towards 1, which means that shrinkage is favoured.

[2] For a Gaussian observation model the variance of the errors σ^2 quantifies how informative the data is.

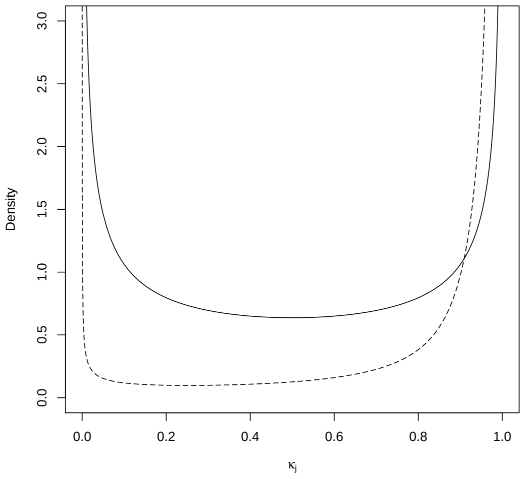


Figure 1: Conditional density of the shrinkage factor κ_j for $n\sigma^{-2}\tau^2 = 1$ (solid) and for $n\sigma^{-2}\tau^2 = 0.1$ (dashed).

3.2 Extending horseshoe prior to the case of grouped variables

In this chapter the general idea of Xu et al. (2016) is used and applied on additive models with penalized splines. In these additive models two natural group structures arise:

1. A categorical predictor with $m_{G_1}+1$ levels can be modelled with m_{G_1} dummy coded variables, which represent a group of m_{G_1} linear predictors.

2. A smoothing function with parameter $\mathbf{u} \in \mathbb{R}^{K_1}$ also can be seen as a group of K_1 parameters.

Because one usually wants a smooth function defined on the whole domain and does not want to select categorical predictors level-wise, the horseshoe prior can straightforwardly be extended for this grouped structures, s.t. for an additive model as defined in Eq. (2.3.9), where the g linear predictors are categorical with $m_{G_i}+1$ levels and where the parameter vector $\mathbf{u}_j \in \mathbb{R}^{K_j}$ of the m_s smoothing functions,

$$\beta|\,\tau^2,\lambda_{11}^2,\dots,\lambda_{1g}^2 \sim N(0,\tau^2\mathbf{D}_{\lambda_1}), \tag{3.2.1}$$

$$\forall j \in \{1,\dots,m_s\} \quad \mathbf{u}_j|\,\tau,\lambda_{2j} \sim N(0,\tau^2\lambda_{2j}^2\mathbf{Z}_j), \tag{3.2.2}$$

where $\mathbf{D}_{\lambda_1} = \mathrm{diag}(\lambda_{11}^2 I_{m_{G_1}}, \dots, \lambda_{1g}^2 I_{m_{G_g}})$ and

$$\lambda_{1i} \sim \mathrm{C}^+(0,1), \quad \lambda_{2i} \sim \mathrm{C}^+(0,1), \quad \tau \sim \mathrm{C}^+(0,1).$$

This means every group structure has one local shrinkage parameter. Hence the whole group is either shrunken as a whole or stays unshrunken.

3.3 Hyperprior choice for global shrinkage parameter

In Piironen and Vehtari (2017) it is shown that for the global shrinkage parameter the choice of scale 1 can produce suboptimal results in settings where τ can not well be identified by the data, e.g high ratio of number of parameter to number of data or very noisy data. Hence the authors propose to introduce τ_0 as an hyper-parameter, s.t. $\tau \sim \mathrm{C}^+(0,\tau_0)$ and show that an good way to choose this parameter for a Gaussian observation model is by using an guess of the number of relevant variables p_0 of of all the D to be shrunken variables, s.t.

$$\tau_0 = \frac{p_0}{D - p_0} \frac{\sigma}{\sqrt{n}}, \tag{3.3.1}$$

where σ^2 is the variance of errors and n the number of observations.
In this thesis it will be shown, that this choice of τ_0 is still reasonable in the setting of grouped variables as discussed in section 3.2:
Now consider the grouped variables setting of section 3.2. Every group has its own local shrinkage coefficient κ_{ij}. From the prior distribution of κ_{ij} (3.1.7)[3] it follows that

$$\mathbb{E}(\kappa_{ij} | \, \tau, \sigma) = \underbrace{\frac{1}{1 + \sigma^{-1}\tau\sqrt{n}}}_{=:\mu_\kappa}. \tag{3.3.2}$$

[3] Since the structures of Eq. (2.3.17) and Eq. (2.3.19) are similar and $\mathbf{Z}$ is a diagonal covariance matrix the derivation of κ in the setting of a smooth function can be carried out in the same manner as in section (3.1).

Since κ_{ij} is typically either near zero or near one it motivates the definition of effective number of parameter m_{eff} as

$$m_{\text{eff}} = \sum_{i=1}^{g}(1-\kappa_{1i})m_{G_i} + \sum_{j=1}^{m_s}(1-\kappa_{2j})K_j. \tag{3.3.3}$$

By taking the expectation of Eq. (3.3.3) it follows that

$$\mathbb{E}(m_{\text{eff}}|\ \tau,\sigma) = (1-\mu_\kappa)\underbrace{(\sum_{i=1}^{g} m_{G_i} + \sum_{j=1}^{m_s} K_j)}_{=:D} = \frac{\sigma^{-1}\tau\sqrt{n}}{1+\sigma^{-1}\tau\sqrt{n}}D. \tag{3.3.4}$$

By solving Eq. (3.3.4) for τ for an guess of the effective number of parameter p_0 the same proposal as presented in Eq. (3.3.1) is derived. One can see from Eq. (3.3.4) that m_{eff} and τ are clearly linked and if one wants to keep the prior on m_{eff} consistent for different data situations, τ should scale with $\sqrt{n}/\sigma$. The suggested integration of τ_0 in the model is done by using $\tau \sim \mathrm{C}^+(0,\tau_0)$, because then the median of $\tau = \tau_0$ and the heavy tails of the half-cauchy distribution allows for a higher adaptivity to the data. The integration can be done in other ways than using $\tau \sim \mathrm{C}^+(0,\tau_0)$, e.g. fixing τ to τ_0 or using $\tau \sim N^+(0,\tau_0)$. The effects of these other approaches are discussed in Piironen and Vehtari (2017).

3.4 Practical aspects

When using NUTS (2.2) to sample from the posterior distribution of a horseshoe prior model, so-called *divergent transition* can occur (as described in Piironen and Vehtari (2015)), where the step size of the leapfrog integrator is so big that some features of the target distribution can not be depicted, i.e. the estimation becomes biased. To encounter this the general step size can be made smaller or a reparametrization, s.t. the geometry of the posterior distribution is simplified, can be used or both. For this thesis the parametrization as suggested in Peltola et al. (2014)(codes at `https://github.com/to-mi/stan-survival-shrinkage`) is used, where a parameter ν with $\nu \sim \mathrm{C}^+(0,\nu_0)$ is not directly sampled from the half-cauchy distribution, but instead auxiliary parameters r_1, r_2 are introduced, s.t.

$$\nu = r_1\sqrt{r_2}. \tag{3.4.1}$$

When $r_1 \sim N(0, v_0)$ and $r_1 \sim \mathrm{InvG}(0.5, 0.5)$ it follows that $v \sim \mathrm{C}^+(0, v_0)$. By sampling v indirectly with r_1 and r_2 the number of divergent transition is lowered, but still a relatively small step size is needed.

4 Simulations

To test the capabilities of the grouped horseshoe priors separately, two simulation studies are conducted. On the one hand a model only consisting of factor variables is estimated and different choices of the hyperparameter τ_0 are compared in section 4.1, while on the other hand a pure smooth function model is estimated in section 4.2 and compared to estimations done by other variable selection methods such as spikeSlabGAM (see Scheipl (2010)) and GAMSEL (see Chouldechova and Hastie (2015)), where an overlap group-lasso penalty is used.

Both simulation studies are special cases of the additive model (2.3). In general when $\mathbf{U}, \mathbf{X}_1, \ldots, \mathbf{X}_{m_s}$ are created by a data generating process (DGP) the "true" predictor η can be evaluated for given coefficient vectors $\beta, \gamma_1, \ldots, \gamma_{m_s}$ as $\eta = \mathbf{U}\beta + \sum_{i=1}^{m_s} \mathbf{X}_i \gamma_i$ and the response $\mathbf{y}$ as $\mathbf{y} = \eta + \varepsilon$. The number of all variables with non zero influence will be denoted as D. To control the difficulty of estimating the coefficient vectors and η the so-called signal-to-noise ratio SNR is used. The SNR is defined as $\mathrm{SNR} = n\, \mathrm{sd}_\eta^2 / \sum_{i=1}^n \varepsilon_i^2$, where $\mathrm{sd}_\eta = \sqrt{\sum_{i=1}^n (\bar{\eta} - \eta_i)^2 / n}$, i.e. SNR is the ratio of the systematic variability (*signal*) over the unsystematic one caused by the Gaussian error terms ε. To simulate a certain SNR level the response $\mathbf{y}$ can be sampled as $y_i \sim N(\eta_i, \mathrm{sd}_\eta^2 / \mathrm{SNR})$.

For every model 3 different prior specifications are compared:

- '*strict*' prior with $p_0 = 1$,
- '*optimal* prior with $p_0 =$ true number of non zero influence coefficients,
- '*loose*' prior with $p_0 = D - 1$.

In this chapter for every prior 4 chains of a size of 500 elements are simulated. For both simulations an adequate model based on the "true" predictors is estimated and its MSE, which will be called OracleMSE, is compared to the MSEs produced by the other methods.

4.1 Linear predictor scenario

In this simple setting the shrinkage property of the grouped horseshoe prior is investigated in the case of categorical predictors solely. As a reference a linear model only based on the non-zero influence variables is estimated, which will be referred as the ”oracle”-model for chapter (4.1). The DGP is described in section (4.1.1) and the results are discussed in (4.1.2).

4.1.1 Data generation

The DGP of this setting can be described as follows:

- $n = 100$ observations,
- $\text{SNR} = 0.1, 1, 5$,
- 9 categorical variables are defined as $\forall i \in \{1,2,3\}$
  - the ’*small-sized*’ groups G_{i1} with levels g_{11}, g_{12}, g_{13},
  - the ’*medium-sized*’ groups G_{i2} with levels $g_{21}, \dots, g_{25}$,
  - the ’*large-sized*’ groups G_{i3} with levels $g_{31}, \dots, g_{39}$,
- the 9 to the G_{ij} associated coefficient vectors are defined as
  - $\beta_{11} = (0, 1, 2)$,
  - $\beta_{12} = (0, 1, 4/3, 5/3, 2)$,
  - $\beta_{13} = (0, 1, 8/7, 9/7, \dots 2)$,
  - $\forall i \in \{2,3\}, j \in \{1,2,3\}\ \beta_{ij} = 0$,
- two subscenarios are defined as
  - a ’*low sparsity*’ scenario: Generate 6 covariates from G_{i1}, G_{i2}, G_{i3} with $i \in \{1,2\}$, i.e. 3 of which have zero influence, s.t. the true linear predictor is

$$\eta = \sum_{j=1}^{3} \mathbf{1}_{\{g_{1j}\}}(x_1) \cdot \beta_{11,j} + \sum_{j=1}^{5} \mathbf{1}_{\{g_{2j}\}}(x_2) \cdot \beta_{12,j} + \sum_{j=1}^{9} \mathbf{1}_{\{g_{3j}\}}(x_3) \cdot \beta_{13,j},$$

  - a '*high sparsity*' scenario: Generate 9 covariates from G_{i1}, G_{i2}, G_{i3} with $i \in \{1,2,3\}$, i.e. 6 of which have zero influence, s.t. the true linear predictor remains as

$$\eta = \sum_{j=1}^{3} \mathbf{1}_{\{g_{1j}\}}(x_1) \cdot \beta_{11,j} + \sum_{j=1}^{5} \mathbf{1}_{\{g_{2j}\}}(x_2) \cdot \beta_{12,j} + \sum_{j=1}^{9} \mathbf{1}_{\{g_{3j}\}}(x_3) \cdot \beta_{13,j},$$

- m = 100 replications per setting.

The predictive MSE is evaluated on test data sets with 5000 observations. The quality of shrinkage is compared with the MSE of the parameter estimates $\hat{\beta}_i$, which is defined as

$$\frac{1}{m} \sum_{i=1}^{m} |\beta - \hat{\beta}_i|_2^2.$$

4.1.2 Results

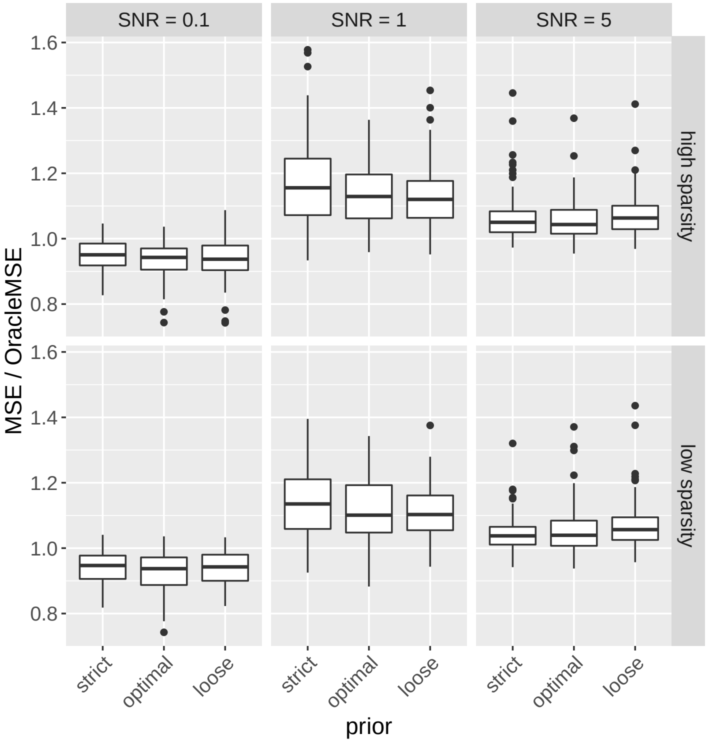
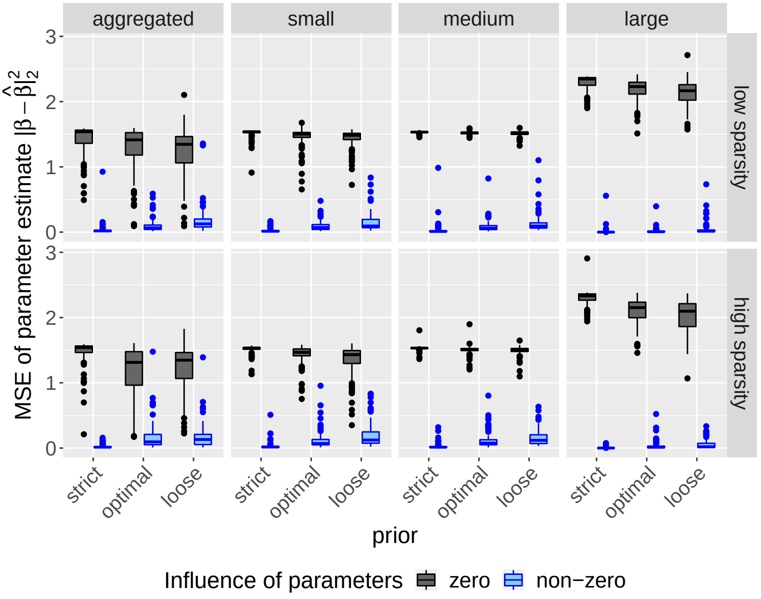


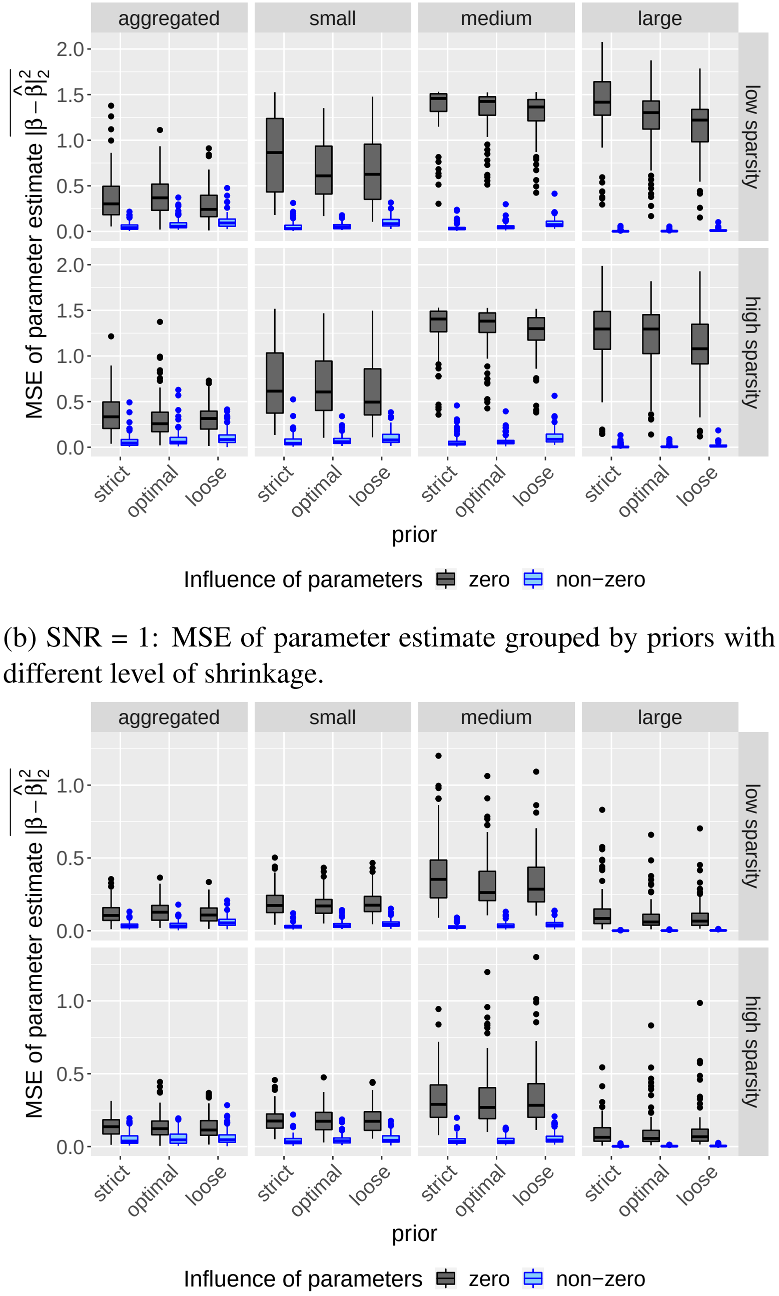
Figure 2: MSE / OracleMSE grouped by priors with different level of shrinkage.

In this simulation study there are two main objectives. Firstly to measure how the strictness of the shrinkage affects the overall quality and secondly if the shrinkage is influenced by the size of the groups. The coefficient vectors have been scaled, s.t. they span the same interval, in order of making the groups comparable.

In Fig. (2) it can be seen that the distributions of the MSEs for SNR = 0.1, 5 differ not greatly. Also the different level of sparsity does not seem to have an structural effect on these distributions. Solely for SNR = 1 it can be observed, that with looser shrinkage the MSEs are drawn to slightly smaller values. For SNR = 0.1 it can be observed in Fig. (3a) that the large-sized group performs worse in terms of MSE of the parameter estimates. For SNR = 1 the strictness of the prior influences the parameter MSE of the large-sized groups the most, s.t. for a loose prior the large-sized groups perform better then the medium-sized groups. For the small-sized groups the lowest parameter MSEs can be observed in this setting, but also a comparably large range of parameter MSEs can be seen for these groups. By comparing all figures of Fig. (3) it can be concluded that for smaller SNR values with stricter prior the parameter MSE of non-zero influence variable descends and the parameter MSE of zero influence variable ascends slightly. For SNR $\geq$ 5 the strictness of the prior impacts lesser the MSEs. Additionally it can be observed that with a higher SNR the MSEs of the parameter estimates get smaller. Also it can be stated that while the size of the group has an effect on the performance in nearly all settings the level of sparsity has only an minor effect in some cases, i.e. the grouped horseshoe priors appears to be quite robust against different levels of sparsity.



(a) SNR = 0.1: MSE of all groups of parameter estimate grouped by priors with different level of shrinkage.



(b) SNR = 1: MSE of parameter estimate grouped by priors with different level of shrinkage.

(c) SNR = 5: MSE of parameter estimate of the medium groups grouped by priors with different level of shrinkage.

Figure 3: MSE of parameter estimate grouped by priors with different level of shrinkage for SNR = 0.1, 1, 5.

4.2 Smooth functions scenario

In order of testing the performance of the grouped horseshoe priors in the setting of smooth functions the simulation study in 4.1.5 of Scheipl (2010) is replicated. As a reference a conventional GAM (as implemented in Wood (2008)) only based on the non-zero influence variables is estimated, which will be referred as the ”oracle”-model for chapter (4.2). The performance is compared with a Bayesian alternative model approach spikeSlabGAM (see Scheipl (2010)) and the non Bayesian inference based method GAMSEL (see Chouldechova and Hastie (2015)). The DGP is described in section (4.2.1) and the results are discussed in (4.2.2).

4.2.1 Data generation

The DGP of this setting can be described as follows:

- $n = 200$ observations,
- $\text{SNR} = 5, 20$,
- 4 functions, which enter the linear predictor, are defined as
  - $f_1(x) = x$,
  - $f_2(x) = x + \frac{(2x-2)^2}{5.5}$,
  - $f_3(x) = -x + \pi \sin(\pi x)$,
  - $f_4(x) = 0.5x + 1.5\phi(2(x-0.2)) - \phi(x+0.4)$, where ϕ is the standard normal density function.
- Two subscenarios are defined as
  - a '*low sparsity*' scenario: Generate 16 covariates, 12 of which have non-zero influence, s.t. the true linear predictor is

$$\begin{aligned}\eta = {} & f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4) + \\ & + 1.5(f_1(x_5) + f_2(x_6) + f_3(x_7) + f_4(x_8)) + \\ & + 2(f_1(x_9) + f_2(x_{10}) + f_3(x_{11}) + f_4(x_{12}))).\end{aligned}$$

  - a '*high sparsity*' scenario: Generate 20 covariates, 4 of which have non-zero influence, s.t. the true linear predictor is

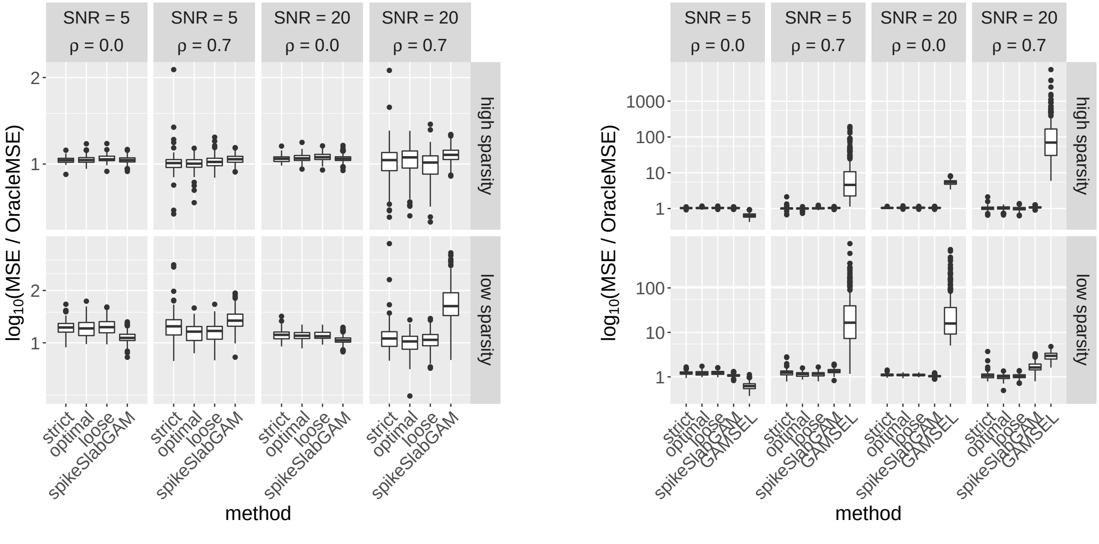
$$\eta = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_4(x_4).$$

- The covariates are either
  - $\overset{\text{i.i.d.}}{\sim} \mathcal{U}[-2,2]$ or
  - from an AR(1) process with correlation $\rho = 0.7$.
- m = 100 replications per setting.

The predictive MSE is evaluated on test data sets with 5000 observations. The quality of shrinkage is compared with the MSE of the estimated function outputs$\hat{f}_i$, which is defined as

$$\frac{1}{m}\sum_{i=1}^{m} |f - \hat{f}_i|_2^2.$$

4.2.2 Results



(a) MSE / OracleMSE grouped by priors with different level of shrinkage with GAMSEL excluded

(b) MSE / OracleMSE grouped by priors with different level of shrinkage with GAMSEL included

Figure 4: MSE / OracleMSE grouped by priors.

In this section grouped horseshoe priors with different levels of shrinkage are compared to each other, spikeSlabGAM, and GAMSEL in the smooth functions setting. For the horseshoe priors and spikeSlabGAM every covariate is fitted with a p-spline of order 2 and degree 3 with 20 knots. Default setting consisting of 3 chains of length 500 are used for spikeSlabGAM. For GAMSEL splines of degree 6 with nominal number of basis elements of 10 are fitted. The regularization parameter λ of GAMSEL is chosen from 50 10-fold cross validations. It is also evaluated how well the function approximation is carried out by the different priors. In Fig. (4a) it can be seen that the level of shrinkage does not seem to affect the performance in terms of the MSE of the grouped horseshoe priors. Also it can be observed that the grouped horseshoe priors perform quite robust compared to the other methods. Especially in the case of correlated covariates this can be noticed, where in the low sparsity scenario the MSEs of the horseshoe priors are lower distributed than the MSEs of spikeSlabGAM. In the other cases spikeSlabGAM performs slightly better than the horseshoe priors. For this simulation GAMSEL performs in terms of MSE far worse as it can be seen in Fig. (4b), e.g many extreme high values, except for uncorrelated covariates, where for SNR = 5 a clearly better performance in both sparsity scenarios can be observed. The MSEs of the centred function output estimates seem to have no noteworthy differences for different levels of shrinkage as it can be verified in Fig. (5), except for a slightly better performance of the more loose shrinkage for SNR = 5 and $\rho = 0.7$. Which coincides with the MSE distribution of the different shrinkage levels observed in Fig. (4).

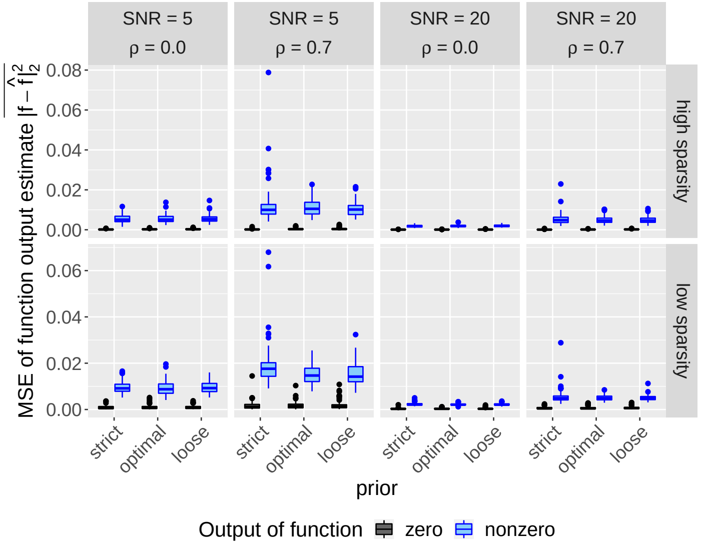


Figure 5: MSE of function output estimate.

5 Benchmarking

In this chapter the methods of section (4) are applied on real data sets and their performance is compared. Both data sets are modelled as Gaussian, i.e. the output is numerical and the error terms are assumed to be normal distributed. In section (5.1) the output is estimated based only on categorical predictors and in section (5.2) smooth functions are estimated for all numerical input variables. In this chapter the number of all coefficients will be denoted as D.

5.1 Automobile

This data set is from the Univerisity of California Irvine Machine Learning Repository [Dua and Karra Taniskidou (2017)]. The data was compiled by Jeffrey C. Schlimmer from the following sources:

- 1985 Model Import Car and Truck Specifications, 1985 Ward's Automotive Yearbook.
- Personal Auto Manuals, Insurance Services Office, 160 Water Street, New York, NY 10038
- Insurance Collision Report, Insurance Institute for Highway Safety, Watergate 600, Washington, DC 20037

In this data set information of different cars and their price, as described in section (5.1.1), is contained. The performance of estimating the price based on only the categorical predictors is evaluated in section (5.1.2).

5.1.1 Data description and modifications

The automobile data set consists of 205 observations with 26 attributes. Of these 26 attributes only the 10 nominal predictors are selected, s.t. the performance can be compared with the gglasso method (see Yang and Zou (2014)). The numerical outcome variable of interest is the price of the car, which ranges from 5118 to 45400$.

The nominal predictors for the price of one car are the make with 22 different kinds, whether it runs on diesel or gas, whether its engine is naturally aspirated or turbocharged, if it has two or four doors, the car body design with 5 different kinds, whether is has front-wheel drive, back-wheel drive or four-wheel drive, whether the engine is located in the front or the rear, different engine-types with 7 different kinds, the number of cylinder modelled as categorical variable with 7 types, the fuel system with 8 different kinds. Only complete cases are used, s.t. only 199 observation are used for the estimation of the models.

To create two levels of artificial sparsity in the data the 10 predictors are 1-times/2-times duplicated, permuted and added to the data set, s.t. the newly added data is distributed like the original data, but without relation to the outcome.

5.1.2 Comparisons and Results

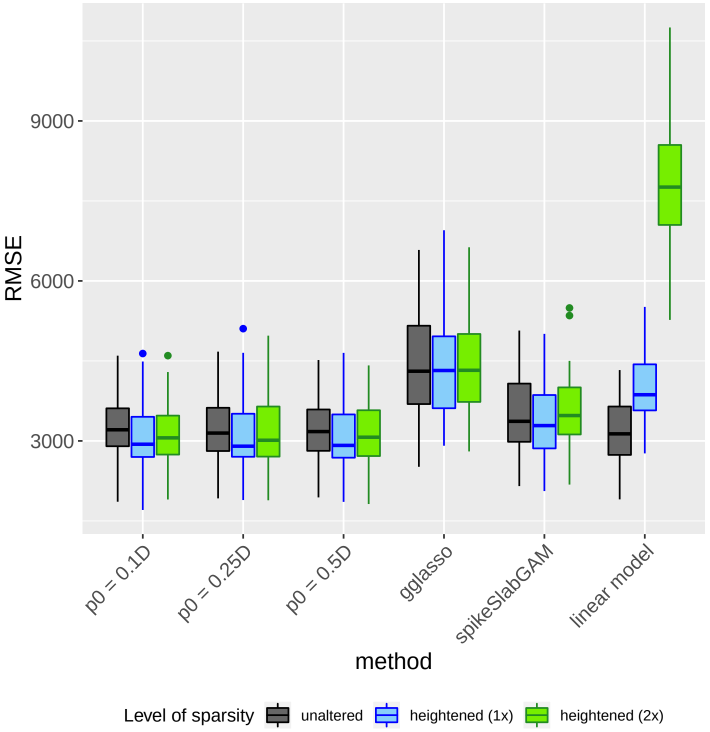


Figure 6: RMSE of prediction of 50 shuffle splits.

For this simulation 3 different horseshoe priors are evaluated with $p_0 = 0.1D, 0.25D, 0.5D$ (note, that D changes with added sparsity). For every prior 4 chains with a size of 200 elements are simulated. Default setting consisting of 3 chains of length 500 are used for spikeSlabGAM. The regularization parameter λ of gglasso is chosen from 100 5-fold cross validations. The performance of these methods is compared with their RMSEs, which are drawn from 50 valid[4] shuffle splits, i.e. 50 times the data set is randomly divided into training and test data with ratio 4:1.

In Fig. (6) it can be observed that the horseshoe priors differ only slightly for all levels of shrinkage. Since for the unaltered data the RMSEs of the grouped horseshoe priors and the linear model follow a similar distribution, it is assumed that the data is not sparse. All methods seem to be robust against higher levels of sparsity, except for the linear model as one would expect. For this data set the groups horseshoe priors perform best in terms of the RMSE, followed by spikeSlabGAM. Only for sparsity two-times heightened the performance of gglasso is better than the linear model, but still worse than the Bayesian approaches.

5.2 Boston housing

The data have been taken from the UCI Repository Of Machine Learning Databases [Dua and Karra Taniskidou (2017)] and are based on Harrison and Rubinfeld (1978). In this data set information of owner-occupied homes and their neighbourhood in suburbs of Boston and their median value, as described in section (5.2.1), is contained. The performance of estimating the median value based on only the numerical predictors is evaluated in section (5.2.2).

5.2.1 Data description and modifications

The Boston housing data set consists of 506 observations with 14 attributes. Of these 14 attributes only the 12 continuous predictors are selected, s.t. the performance can be compared with the GAMSEL method (see Chouldechova and Hastie (2015)). The numerical outcome variable of interest is the median value of owner-occupied homes in $1000, which ranges from 5 to 50.

The continuous predictors for the median value of owner-occupied homes are the per capita crime rate by town, the proportion of residential land zoned for lots over 25,000 sq.ft., the proportion of non-retail business acres per town, the nitrogen oxides concentration, the average number of rooms per dwelling, proportion of owner-occupied units built prior to 1940, weighted mean of distances to five Boston employment centres, full-value property-tax rate, pupil-teacher ratio by town, the proportion of African-

[4] A split is accepted, if the test data contains only levels, which are also present in the training data.

American by town, the lower status of the population.
To create an artificial sparsity in the data the 12 predictors are duplicated, permuted and added to the data set, s.t. the newly added data is distributed like the original data, but without relation to the outcome.

5.2.2 Comparisons and Results

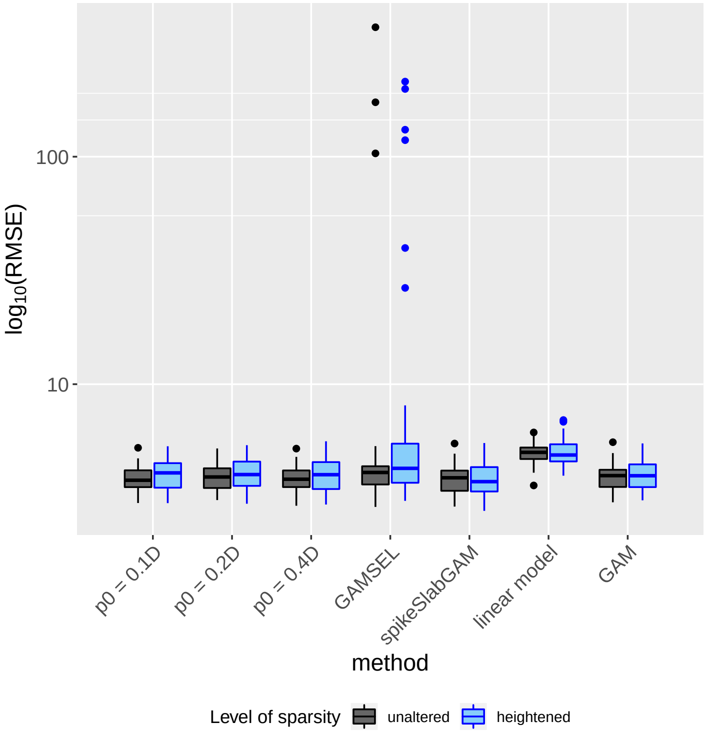


Figure 7: RMSE of prediction of 10 times repeated 5-fold cross validation.

For this simulation 3 different horseshoe priors are evaluated with $p_0 = 0.1D, 0.2D, 0.4D$ (note, that D changes with added sparsity). For every prior 4 chains with a size of 100 elements are simulated. Default setting consisting of 3 chains of length 500 are used for spikeSlabGAM. For GAMSEL splines of maximum number of spline basis function of 10 with a degree of freedom of 5 are fitted. The regularization parameter λ of GAMSEL is chosen from 50 10-fold cross validations. Also for reference a linear model and GAM with splines of degree 6 are fitted.
The performance of these methods is compared with their RMSEs, which are drawn from 10 5-fold cross validations.
Since the GAM performs better in terms of RMSE than the linear model as it can be

seen in Fig. (7) it can be concluded that there are non-linear effects in the data. In this setting spikeSlabGAM has the lowest RMSE values, which are nearly followed by the RMSE values of the GAM, for both levels of sparsity. The horseshoe priors perform slightly worse than spikeSlabGAM, but still robust. No difference can be noted for the different levels of shrinkage. GAMSEL has a similar performance than the grouped horseshoe priors in this data set, but with some far higher extreme values as it can be seen in Fig. (7).

6 Conclusions

As this thesis shows the grouped horseshoe priors are able to estimate robustly point estimates. It is especially robust comparably to the other presented methods against different levels of sparsity as it can be seen in the simulation study (4.2) and the benchmarks (5). The linear predictor simulation (4.1) suggests that there is an influence of the size of the groups, which should be investigated further. For this thesis the parametrization u of γ as shown in section (2.3) was used. The influence of choosing an other parametrization could be investigated. Also more studies could be done of when a chain is sufficiently long for a good point estimate, since it could be shown in (5.1) that chains of only length 200 produced competitive results. In most cases the hyperparameter choice did not affect the performance in terms of predictive MSE, but in some data situations the *optimal* prior offered a good comprise between estimation error of the zero and non-zero influence variables, as it could be seen in section (4.1).
Also it must be noted that a drawback of using the full Bayesian Inference is that it is quite time-consuming as compared to frequentist methods like gglasso or GAMSEL.
Overall the results of the grouped horseshoe priors look promising and more work should be invested in extending it to the general additive model case and to a broader family of smooth functions.

7 Appendix - Stan code

7.1 Linear predictor

```
data {
  int num_data;                // number of observations
  vector[num_data] Y;          // outcome variable
  real<lower=0> scale_global;  // scale for tau
  int<lower=0> num_linparam;   // overall number of predictor coefficients
  int<lower=0> num_groups;     // number of groups
  int<lower=0> group_ids[num_linparam,1]; // assignments of levels
  matrix[num_data, num_linparam] X; // model matrix of the inputs
  real<lower=1> nu_global;     // degree of freedom for the half-t prior for tau
  real<lower=1> nu_local;      // degree of freedom for the half-t prior for lambdas
}

parameters {
  real beta0;                           // intercept
  real logsigma;                        // log of noise std
  // auxiliary variables
  vector[num_linparam] z;
  real<lower=0> r1_global;
  real<lower=0> r2_global;
  vector<lower=0>[num_groups] r1_localB;
  vector<lower=0>[num_groups] r2_localB;
}

transformed parameters {
  real sigma;                                        // noise std
  real<lower=0> tau;                                 // global shrinkage parameter
  vector<lower=0>[num_groups] lambdaB;               // local shrinkage parameter
  vector[num_linparam] beta;                         // regression coefficients
  vector[num_data] Y_hat = rep_vector(0, num_data);  // eta
  sigma = exp(logsigma);
  tau = r1_global * sqrt(r2_global);
  lambdaB = r1_localB .* sqrt(r2_localB);
  for (m in 1:num_linparam) {
      beta[m] = z[m]* lambdaB[group_ids[m,1]]*tau ;
  }
  Y_hat = Y_hat + X*beta + beta0;
}

model {
  r1_global ~ normal(0.0, scale_global*sigma);
  r2_global ~ inv_gamma(0.5*nu_global, 0.5*nu_global);
  r1_localB ~ normal(0.0, 1);
  r2_localB ~ inv_gamma(0.5*nu_local, 0.5*nu_local);

  z ~ normal(0, 1);
  Y ~ normal(Y_hat, sigma);
}
```

7.2 Smooth functions

```
data {
  int<lower=0> num_data;              // number of observations
  vector[num_data] Y;                 // outcome
  int<lower=0> num_splines;           // number of splines
  int<lower=0> num_params;            // number of predictor coefficients
  int<lower=0> len_Bs;                // length of flattened transformed model model matrices Bs
  int<lower=0> bStart[1, num_splines];// starting indices of singular B
  int<lower=0> bEnd[1, num_splines];  // ending indices of singular B
  vector[len_Bs] Bs;                  // flattened transformed model model matrices Bs
  int<lower=0> gStart[1, num_splines];// starting indices of coefficients of a singluar predictor
  int<lower=0> gLen[1, num_splines];  // number of coefficients of a singluar predictor
  int<lower=0> gEnd[1, num_splines];  // ending indices of coefficients of a singluar predictor

  real<lower=1> nu_global;            // degree of freedom for the half-t prior for tau
  real<lower=1> nu_local;             // degree of freedom for the half-t prior for u
  real<lower=0> scale_global;         // scale for tau
}

parameters {
  real beta0;                         // intercept
  real logsigma;                      // log noise std
  // auxiliary variables
  vector[num_params] z;
  real<lower=0> r1_global;
  real<lower=0> r2_global;
  vector<lower=0>[num_splines] r1_localG;
  vector<lower=0>[num_splines] r2_localG;
}

transformed parameters {
  real sigma;                        // noise std
  real tau;                          // global shrinkage parameter
  vector[num_splines] lambda;        // local shrinkage parameter

  vector[num_params] u;              // coefficients of all predictors
  vector[num_data] Y_hat = rep_vector(0, num_data); // eta
  tau = r1_global * sqrt(r2_global);
  lambda = r1_localG .* sqrt(r2_localG);

  sigma = exp(logsigma);

  for (i in 1:num_splines){
      u[gStart[1,i]:gEnd[1,i]] = z[gStart[1,i]:gEnd[1,i]] * lambda[i]* tau;
      Y_hat = Y_hat + to_matrix(Bs[bStart[1,i]:bEnd[1,i]],
          num_data, gLen[1,i]) * u[gStart[1,i]:gEnd[1,i]];

  }
  Y_hat = Y_hat + beta0;
}
```

```
model {
  r1_global ~ normal(0.0, scale_global*sigma);
  r2_global ~ inv_gamma(0.5*nu_global , 0.5*nu_global);

  r1_localG ~ normal(0.0, 1);
  r2_localG ~ inv_gamma(0.5*nu_local , 0.5*nu_local);

  z ~ normal(0,1);
  Y ~ normal(Y_hat , sigma);
}
```

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer, 2 edition.

Brooks, S., Gelman, A., Jones, G. L., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC.

Chouldechova, A. and Hastie, T. (2015). Generalized additive model selection.

Dua, D. and Karra Taniskidou, E. (2017). UCI machine learning repository.

Durmus, A., Moulines, E., and Saksman, E. (2017). On the convergence of Hamiltonian Monte Carlo. *arXiv e-prints*.

Fahrmeir, L. and Kneib, T. (2011). *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Oxford University Press.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102.

Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1593–1623.

Meyn, S. and Tweedie, R. (1993). *Markov chains and stochastic stability*. Springer.

Peltola, T., Havulinna, A. S., Salomaa, V., and Vehtari, A. (2014). Hierarchical bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the Eleventh UAI Conference on Bayesian Modeling Applications Workshop - Volume 1218*, BMAW'14, pages 79–88, Aachen, Germany, Germany. CEUR-WS.org.

Piironen, J. and Vehtari, A. (2015). Projection predictive variable selection using stan+r.

Piironen, J. and Vehtari, A. (2017). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior.

Polpo, A., Louzada, F., Rifo, L. L. R., Stern, J. M., and Lauretto, M. (2015). *Interdisciplinary Bayesian Statistics*. Springer.

Scheipl, F. (2010). Bayesian regularization and model selection in structured additive regression models. Dr. Hut Verlag.

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):495–518.

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). Bayesian grouped horseshoe regression with application to additive models. In *AI 2016: Advances in Artificial Intelligence*, pages 229–240. Springer International Publishing.

Yang, Y. and Zou, H. (2014). A fast unified algorithm for solving group-lasso penalize learning problems. *Statistics and Computing*, 25(6):1129–1141.

List of Figures

List of Abbreviations

HMC Hamiltonian Monte Carlo

NUTS No-U-Turn Sampler

HDP highest posterior density

PLS penalized least-squares

DGP data generating process

SNR signal-to-noise ratio

MSE mean squared error

RMSE root-mean-square error

Statement

I hereby declare that this thesis is my own original work and that all sources have been acknowledged.

Munich, Tuesday 26th February, 2019

____

Tobias Pielok