Stochastic nets
& Bayesian regularisation from constraints

Joshua J Bon

B.C.M. (Hons., Stat.), UWA, 2017
B.C.M. & B.Com., UWA, 2015

under the supervision of

Berwin A Turlach* Kevin Murray† Christopher Drovandi‡

* Department of Mathematics and Statistics, University of Western Australia
† School of Population and Global Health, University of Western Australia
‡ School of Mathematical Sciences, Queensland University of Technology

This thesis is presented for the degree of
Master of Philosophy
at the University of Western Australia.

February 8, 2019
Thesis declaration

I, Joshua J Bon, certify that:

This thesis has been substantially accomplished during enrolment in this degree.

This thesis does not contain material which has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution.

In the future, no part of this thesis will be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree.

This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text and, where relevant, in the Authorship Declaration that follows.

This thesis does not violate or infringe any copyright, trademark, patent, or other rights whatsoever of any person.

This thesis does not contain work that I have published, nor work under review for publication.

Signature:

Date: 08/02/2019
Abstract

Regularisation in Bayesian modelling uses classes of priors which encourage shrinkage on posterior quantities. This property can be beneficial for sparse or underdetermined problems, and reduce overfitting. In this thesis, we propose a new probabilistic interpretation for regularisation by way of constraints. We augment a given probability distribution with a stochastic constraint that restricts the support of the prior, emitting a “regularised” distribution as a result. This introduces the notion of probabilistic regularisation as an operator that acts on a probability distribution, rather than classes of priors which are considered to have desirable shrinkage properties. The operation is similar to exponential titling, but also has a strong connection to slice sampling.

Using stochastic constraints, we demonstrate that the classic duality between constrained and penalised optimisation has a Bayesian counterpart for regression problems. Consequently we are able to deepen the connection between the Bayesian and standard Lasso beyond maximum a posteriori equivalence. We explore several aspects of stochastic constraint priors, including full Bayesian computation and theoretical results which aid in understanding of how these distributions (or models with these distributions as priors) behave. Some new distributions are derived using stochastic constraint priors, and we argue that the framework naturally enables prior marginal and prior joint information to be joined into a singular coherent prior distribution.

We show how stochastic constraints underlie popular priors in the high-dimensional shrinkage literature, such as the (regularised) horseshoe, Dirichlet-Laplace, and R2-D2 priors. Considering these priors as stochastic constraints exposes an interesting aspect of their geometry which we can leverage to propose a zero-nonzero decision rule for sparse regression — closely related to the penalised credible regions method. We are also able to derive and implement an efficient sampler for high-dimensional regression models with stochastic constraint priors based on the illuminated geometry.
Acknowledgements

I have a great deal of gratitude to share with those who have supported me over the past two years. First of all, thank you to my wife, Mary, for the encouragement, tireless understanding, perspective, and care you give to me. I am a more well-rounded person thanks to you.

To my supervisors, Berwin, Kevin, and Chris, thanks for riding this rollercoaster with me. You managed to produce a smooth interpolation of the journey and provided me with invaluable knowledge, critique, and inspiration. I am incredibly grateful for the time you have put towards my development as a statistician.

To my “co-workers”, Andrew, Michael, and Lachlan, thanks for camaraderie, especially important to me over long distances. I look forward to working with all of you in the future. Thank you for the enthralling discussions and sound advice — not just in statistics, but in many aspects of life. Thank you to Andrew and Michael for looking over the thesis and helping to improve its exposition.

Thanks goes to my parents, Janene and Harry, and to my sister, Marney, for all your support over the years, particularly the ones spent writing this thesis.

I am a recipient of funding from several bodies and am very grateful for their financial support. This research was supported by an Australian Government Research Training Program (RTP) Scholarship, the Bruce and Betty Green Postgraduate Research Scholarship, and a scholarship from the Centre for Applied Statistics at University of Western Australia.

This work was also supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. The author is also grateful to the Centre for Applied Statistics for facilitating access to this service, with special thanks to Matthew for his help. I am indebted to Michael for his high quality instruction and assistance on working in a remote computing environment.

I am eternally grateful for the R language (R Core Team; 2018), and its wonderful community. I have relied on many amazing R packages in this thesis, and have tried to create a list of the main packages I used. The future package (Bengtsson; 2018) was instrumental in parallel evaluation of code, and the Matrix package (Bates and Maechler; 2018) handled matrix manipulations seamlessly, and with great efficiency, for me. I relied on several random number generators; extradistr (Wolodzko; 2018), GIGrvg (Leydold and Hormann; 2017), RZigZag (Bierkens; 2018), and mvtnorm (Genz et al.; 2018; Genz and Bretz; 2009).

For facilitating writing constrained Zig-Zag samplers, I would like to thank Rcpp (Eddelbuettel and François; 2011) and RcppEigen (Bates and Eddelbuettel; 2013) (I based this work on RZigZag). Many thanks to lasso2 (Osborne et al.; 2000;
Lokhorst et al.; 2018) for its constraint-based Lasso solver.

Thanks to knitr (Xie; 2015) and kableExtra (Zhu; 2018) for generating the tables in this thesis. On that note, thanks must also go to the L\LaTeX\ community. I am grateful for ROCR (Sing et al.; 2005) for generating prediction metrics, as well as tidybayes (Kay; 2018) and coda (Plummer et al.; 2006) for wrangling and summarising MCMC output.

Lastly, I would also like to acknowledge the tidyverse, specifically dplyr (Wickham et al.; 2018), purrr (Henry and Wickham; 2018), tidy (Wickham and Henry; 2018), ggplot2 (Wickham; 2016), scales (Wickham; 2018), tibble (Müller and Wickham; 2018), and magrittr (Bache and Wickham; 2014), for work flow, data wrapping, and graphics. I learned to love \textit{R} with base, but I am glad to be tidy.
Contents

Abstract iii

Acknowledgements v

List of Notation ix

1 Background and motivation 1
  1.1 Introduction .................................................. 1
  1.2 Reinterpreting the Lasso .................................. 3

2 Stochastic constraint random variables 7
  2.1 Univariate case ............................................. 7
  2.2 Multivariate case .......................................... 8
  2.3 Some classes of stochastic constraint distributions ........ 11
  2.4 Relation to other random variables and simulation methods . 12
    2.4.1 Scale mixture of uniform distributions ......... 13
    2.4.2 Slice sampling .................................. 14
    2.4.3 Skew random variables ............................ 15
    2.4.4 Exponentially tilted random variables .......... 15
  2.5 Example stochastic constraint distributions ............. 16
    2.5.1 Generalised Gaussian distribution ............. 16
    2.5.2 Horseshoe prior .................................. 17
    2.5.3 Regularised horseshoe and horseshoe+ prior .... 18
    2.5.4 Dirichlet-Laplace prior ........................... 18
    2.5.5 R2-D2 prior ...................................... 19

3 Some theoretical results 21
  3.1 Conditions for stochastic constraint distributions to be proper ... 21
  3.2 Results related to optimisation-based inference ............ 24
  3.3 Characterising shrinkage from stochastic constraints ........ 25

4 Simulating stochastic constraint random variables 29
  4.1 Rejection samplers ....................................... 29
  4.2 Gibbs samplers .......................................... 30
  4.3 Piecewise deterministic Markov processes ............... 32

5 On the geometry of global-local sparsity priors 39
  5.1 Exponential stochastic constraints ....................... 39
  5.2 Stochastic net priors ................................... 42
  5.3 Eliciting sparsity uncertainty from geometry of stochastic constraints 43
    5.3.1 Thresholding for the horseshoe prior ............ 43
    5.3.2 Penalised credible regions ....................... 44
    5.3.3 Contributions from stochastic constraints .......... 44
5.3.4 Pen-CR and SCVS variable selection comparison 47

6 Using stochastic constraints to construct priors 53
6.1 Horseshoe with informative prior means 54
6.2 Horseshoe with positive skew 55

7 Efficient sampling for Bayesian regression with sparse stochastic net priors 59
7.1 The general truncated Gibbs sampler 59
7.2 Partially truncated R2-D2 sampler 62

8 Discussion 71

Appendices 73

A Appendices 75
A.1 Proofs excluded from main text 75
A.1.1 Theorem 3.5 75
A.1.2 Theorem 3.8 75
A.1.3 Corollary 3.9 76
A.1.4 Lemma 3.10 77
A.1.5 Theorem 3.11 77
A.1.6 Corollary 3.12 78
A.1.7 Theorem 5.3 79
A.1.8 Proposition 6.1 80
A.1.9 Proposition 6.2 81
A.2 Further details on Monte Carlo samplers 82
A.2.1 Gibbs sampler for horseshoe prior from stochastic constraints 82
A.2.2 Power-adjusted inverse Gamma distribution 84
A.2.3 Poisson process switching times for Gaussian(-like) distributions 85
Notation

General nomenclature

For the benefit of the reader and to enable clearer understanding, we introduce some notation which is to be used in this thesis.

- Matrices and vectors are boldface, and matrices are in upper case, e.g. vector \( \mathbf{x} \) and matrix \( \mathbf{X} \).
- \( \mathbf{I}_n \) denotes the \( n \times n \) identity matrix. The dimensions of \( \mathbf{I} \) without subscript \( n \) can be determined by its context.
- A column vector of zeros is denoted by \( \mathbf{0} \), with size determined by its context.
- For a vector, \( \mathbf{x} \), \( x_i \) denotes the \( i \)th element of \( \mathbf{x} \), and \( \mathbf{x}(\cdot) \) denotes the vector of all but the \( i \)th element of \( \mathbf{x} \). For a matrix, \( \mathbf{X} \), \( X_{i,j} \) is the element of the \( i \)th row and \( j \)th column of \( \mathbf{X} \), and \( \mathbf{X}_i \) is the \( i \)th column of \( \mathbf{X} \), and \( \mathbf{X}(\cdot) \) is the matrix without the \( i \)th column.
- The transpose of a matrix, say \( \mathbf{A} \), is denoted by \( \mathbf{A}^\top \).
- The definite integral \( \int_D f(\mathbf{x}) \, d\mathbf{x} \) where \( \mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix}^\top \) and domain \( D \subseteq \mathbb{R}^n \) is the \( n \) dimensional integral such that

\[
\int_D f(\mathbf{x}) \, d\mathbf{x} = \int_D \cdots \int f(x_1, x_2, \ldots, x_n) \, dx_1 \, dx_2 \cdots \, dx_n.
\]

- The gradient vector of a multivariate function, say \( f(\mathbf{x}) \), is denoted by \( \text{grad} f(\mathbf{x}) \equiv \nabla f(\mathbf{x}) \), and the partial derivative of the \( i \)th variable of \( f(\mathbf{x}) \) is \( \nabla_i f(\mathbf{x}) \).
- The indicator function is denoted by \( 1(\cdot) \).
- Hadamard, or entry-wise, multiplication and division is denoted by \( \odot \) and \( \oslash \) respectively.
- If a probability density function, \( f_\theta(t) \), for random variable \( \theta \), is written without an argument, i.e. \( f_\theta \), it is taken to mean \( f_\theta = f_\theta(\theta) \).
- When discussing conditional distributions it is assumed the data, typically \( \mathbf{y} \) and \( \mathbf{X} \) are given, even if they do not appear explicitly in the conditioning.
Table of symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathbb{R} ), ( \mathbb{R}_+ )</td>
<td>The real line, and the positive real line</td>
</tr>
<tr>
<td>( \mathbb{Z}^* )</td>
<td>Non-negative integers</td>
</tr>
<tr>
<td>( S^n )</td>
<td>Simplex in ( n ) dimensions</td>
</tr>
<tr>
<td>( f_\theta(t) )</td>
<td>The density function of ( \theta ), a continuous random variable</td>
</tr>
<tr>
<td>( F_\theta(t) )</td>
<td>The cumulative distribution function of ( \theta )</td>
</tr>
<tr>
<td>( \overline{F}_\theta(t) )</td>
<td>The complementary cumulative distribution function of ( \theta )</td>
</tr>
<tr>
<td>( \equiv )</td>
<td>Equality in distribution</td>
</tr>
<tr>
<td>( x \preceq y )</td>
<td>Element-wise inequality for vectors ( x ) and ( y )</td>
</tr>
<tr>
<td>( \Pr(A) )</td>
<td>Probability of event ( A )</td>
</tr>
<tr>
<td>( \Gamma(\alpha) )</td>
<td>Gamma Function</td>
</tr>
<tr>
<td>( B(\alpha, \beta) )</td>
<td>Beta function, ( B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta) )</td>
</tr>
<tr>
<td>( B(\alpha) )</td>
<td>Multivariate Beta function, ( B(\alpha) = \prod_{i=1}^{n} \Gamma(\alpha_i)/\Gamma(\sum_{i=1}^{n} \alpha_i) )</td>
</tr>
<tr>
<td>( K_\alpha(x) )</td>
<td>Modified Bessel function of the second kind</td>
</tr>
<tr>
<td>(</td>
<td>x</td>
</tr>
<tr>
<td>(</td>
<td>x</td>
</tr>
<tr>
<td>(</td>
<td>x</td>
</tr>
<tr>
<td>(</td>
<td>x</td>
</tr>
<tr>
<td>(</td>
<td>X</td>
</tr>
<tr>
<td>(</td>
<td>S</td>
</tr>
<tr>
<td>( \delta(x) )</td>
<td>Dirac delta function</td>
</tr>
<tr>
<td>( x^n )</td>
<td>Vector ( x ) raised to the ( n )th power, element-wise</td>
</tr>
<tr>
<td>( \mathcal{O}(\cdot) )</td>
<td>Big-oh notation</td>
</tr>
<tr>
<td>( \phi(x) )</td>
<td>Probability density function of standard normal distribution</td>
</tr>
<tr>
<td>( \Phi(x) )</td>
<td>Cumulative distribution function of standard normal distribution</td>
</tr>
</tbody>
</table>
Table of probability distributions

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Name</th>
<th>Probability density</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N(\mu, \sigma^2)$</td>
<td>Gaussian, $\mathbb{R}$</td>
<td>$\mathcal{N}(x; \mu, \sigma^2) = (2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/(2\sigma^2)}$</td>
</tr>
<tr>
<td>$N(\mu, \Sigma)$</td>
<td>Gaussian, $\mathbb{R}^d$</td>
<td>$\mathcal{N}(x; \mu, \Sigma) = ((2\pi)^d</td>
</tr>
<tr>
<td>$U(a, b)$</td>
<td>Uniform, $[a, b]$</td>
<td>$1/(b - a)$</td>
</tr>
<tr>
<td>Beta($\alpha, \beta$)</td>
<td>Beta, $(0, 1)$</td>
<td>$\text{B}(\alpha, \beta)^{-1}x^{\alpha-1}(1-x)^{\beta-1}$</td>
</tr>
<tr>
<td>Exp($\lambda$)</td>
<td>Exponential, $\mathbb{R}_+$</td>
<td>$\lambda \exp{-\lambda x}$</td>
</tr>
<tr>
<td>Ca($\mu, \sigma$)</td>
<td>Cauchy, $\mathbb{R}$</td>
<td>$(\sigma/\pi)(\sigma^2 + (x - \mu)^2)^{-1}$</td>
</tr>
<tr>
<td>Ca$_+$($\sigma$)</td>
<td>Cauchy, $\mathbb{R}_+$</td>
<td>$(2\sigma/\pi)(\sigma^2 + x^2)^{-1}$</td>
</tr>
<tr>
<td>Dir($\alpha$)</td>
<td>Dirichlet, $\mathbb{S}^n$</td>
<td>$\prod_{i=1}^n x_i^{\alpha_i-1}/\text{B}(\alpha)$</td>
</tr>
<tr>
<td>Gam($\alpha, \beta$)</td>
<td>Gamma, $\mathbb{R}_+$</td>
<td>$\beta^\alpha\Gamma(\alpha)^{-1}x^{\alpha-1}\exp{-\beta x}$</td>
</tr>
<tr>
<td>GGam($\alpha, \beta, \kappa$)</td>
<td>Gen. gamma, $\mathbb{R}_+$</td>
<td>$\kappa\beta^\alpha\Gamma(\alpha/\kappa)^{-1}x_{\kappa-1}\exp{-\beta x^\kappa}$</td>
</tr>
<tr>
<td>Rayl($\sigma$)</td>
<td>Rayleigh, $\mathbb{R}_+$</td>
<td>$\sigma^2 x^{-1}\exp{-x^2/(2\sigma^2)}$</td>
</tr>
<tr>
<td>DE($\sigma$)</td>
<td>Laplace, $\mathbb{R}$</td>
<td>$(2\sigma)^{-1}\exp{-</td>
</tr>
<tr>
<td>SkN($\alpha$)</td>
<td>Skew normal, $\mathbb{R}$</td>
<td>$2\phi(x)\Phi(\alpha x)$</td>
</tr>
<tr>
<td>IGam($\alpha, \beta$)</td>
<td>Inv. gamma, $\mathbb{R}_+$</td>
<td>$\beta^\alpha\Gamma(\alpha)^{-1}x^{-\alpha-1}\exp{-\beta/x}$</td>
</tr>
<tr>
<td>Wald($\mu, \lambda$)</td>
<td>Inv. Gaussian, $\mathbb{R}_+$</td>
<td>$\lambda^{1/2}(2\pi x^3)^{-1/2}\exp{-\lambda(z - \mu)^2/(2\mu^2 x)}$</td>
</tr>
<tr>
<td>GIG($\alpha, \beta, \lambda$)</td>
<td>Gen. inv. Gaus., $\mathbb{R}_+$</td>
<td>$(\beta/\lambda)^{\alpha/2}(2K_0(\sqrt{\beta\lambda}))^{-1}x^{\alpha-1}\exp{-\beta x + \lambda/x}/2$</td>
</tr>
<tr>
<td>BP($\alpha, \beta$)</td>
<td>Beta prime, $\mathbb{R}_+$</td>
<td>$\text{B}(\alpha, \beta)^{-1}x^{\alpha-1}(1+x)^{-\alpha-\beta}$</td>
</tr>
</tbody>
</table>
CHAPTER 1

Background and motivation

1.1 Introduction

Priors in Bayesian models act as regularisers by concentrating probability to certain areas of the parameter space \textit{a priori}. Just as in optimisation-based regularisation, this acts to move or shrink parameter estimates towards chosen neighbourhoods of the support. Shrinkage or regularisation can incorporate contextual information, such as sparsity, into parameter estimates and often improves their inferential and predictive properties.

One of the first works on shrinkage was the seminal paper by Stein (1956). This paper established the inadmissibility of the maximum likelihood estimate (MLE) for the multivariate normal distribution with three or more parameters, and demonstrated the need for new estimators that encouraged shrinkage. Later, the James-Stein estimator (James and Stein; 1961) was proposed as a better alternative to the aforementioned MLE and is now the quintessential example of shrinkage estimation. An empirical Bayes interpretation of the James-Stein estimator was published shortly thereafter (Stein; 1962). The existence of natural Bayesian interpretations to many shrinkage estimators (or priors) has led to substantial interest from Bayesian methodologists.

Recently there has been renewed interest in Bayesian priors for high-dimensional problems. Several continuous priors have been proposed and justified by desirable properties with respect to the multivariate normal problem. Examples include the horseshoe prior (Carvalho et al.; 2010), Dirichlet-Laplace prior (Bhattacharya et al.; 2015), horseshoe+ prior (Bhadra et al.; 2017), regularised horseshoe (Piironen and Vehtari; 2017), and $R^2$-induced Dirichlet Decomposition (R2-D2) prior (Zhang et al.; 2017). Unlike the James-Stein estimator, which shrinks all parameters globally, these priors perform global-local shrinkage which shrinks the parameters by different amounts. The degree of global and local shrinkage is determined by combining the data and prior in a principled way, namely through Bayesian estimation. Polson and Scott (2010) catalogue an even larger number of priors for sparse shrinkage.

Geometrically, these priors succeed in obtaining desirable properties for sparsity by placing high prior probability on thin hyperellipsoids or cross-polytopes centred at zero. This induces approximate sparsity into the posterior distribution of the parameters subject to shrinkage. This geometry approximates some of the properties of discontinuous shrinkage priors such as the spike-and-slab type prior (Piironen and Vehtari; 2017), which place non-zero probability on a zero-subset of parameters (Mitchell and Beauchamp; 1988; George and McCulloch; 1993; Ishwaran and Rao; 2005; Castillo et al.; 2015). The result constitutes a mixture distribution between a discrete ‘spike’ at the origin and continuous ‘slab’ elsewhere. Whilst intuitively appealing for variable selection, these priors can be sensitive to prior choice of slab
width and inclusion probability, and result in a large model space which can be computationally difficult to explore.

In this thesis, a new probabilistic interpretation for regularisation is proposed. The parameters subject to shrinkage are augmented with a latent variable, and constrained to a subset of the new support. The corresponding joint distribution induces a regularised prior distribution as the marginal of the parameters, which we call the stochastic constraint (SC) prior (or distribution). A key concept developed herein is the notion of Bayesian regularisation as an operator that acts on a distribution, rather than as a class of priors with desirable shrinkage properties. Stochastic constraint distributions can be viewed as a latent-variable method, but their analogy to optimisation-based regularisation suggests that they can be used as a new family of priors that are intuitive and flexible. Due to their data-augmentation (or latent variable) interpretation, stochastic constraint distributions can be useful for generating Gibbs samplers. Moreover, they exemplify the interplay between Bayesian computational problem solving and discovering richer classes of probability distributions (Gelman; 2004).

Stochastic constraint priors form a flexible class of priors, encompassing the Bayesian Lasso (Park and Casella; 2008), Bridge priors (Polson et al.; 2014), normal-scale mixtures, such as the horseshoe prior (Carvalho et al.; 2010), and Laplace-scale mixtures, such as the Dirichlet-Laplace prior (Bhattacharya et al.; 2015) as special cases. They offer significant flexibility which can be exploited to propose new priors for a number of situations. They also show promise for better characterisation of existing shrinkage priors in the literature. For example, the SC representation contributes to understanding how closely the regularised and standard horseshoe are related.

The thesis is structured as follows. The remainder of Chapter 1 discusses stochastic constraint priors’ duality to standard Bayesian priors. We demonstrate how this duality can be an analog to the dual problem for regularisation in optimisation-based inference. This provides the underlying motivation for studying stochastic constraints. Chapter 2 introduces a framework for describing stochastic constraint random variables, relates them to other types of probability distributions, and considers some examples. Chapter 3 gives theoretical results on the properties of stochastic constraint distributions, and the characteristics of posterior distributions that use SC priors. An overview of several simulation methods for stochastic constraints is given in Chapter 4. In Chapter 5 the geometry of a class of sparsity priors is explored when viewed as a stochastic constraint distribution. A new method for determining a sparse representation of the posterior distribution is tested with the Dirichlet-Laplace prior, based on the geometric interpretation. Chapter 6 demonstrates analytically and by experimentation how stochastic constraints can be used to formulate priors that reflect both the marginal and joint prior information in Bayesian analysis. Two examples that extend the horseshoe prior are examined. Chapter 7 uses stochastic constraints in a Gibbs sampler to improve the accuracy and speed of high-dimensional linear regression. The R2-D2 prior is used to illus-
trate the new sampler which compares favourably to standard Gibbs sampling in a simulation study. Finally, Chapter 8 concludes with a discussion on possible areas of future research for stochastic constraint distributions.

1.2 Reinterpreting the Lasso

To begin, we illustrate the proposed framework with the Bayesian Lasso (Park and Casella; 2008). In its simplest form, this Lasso is a product of independent identically distributed (iid) Laplacian (or double-exponential) distributions with fixed, shared regularising parameter $\lambda$. The prior's density is

$$f_{\theta} = (\lambda/2)^p \exp\{-\lambda \|\theta\|_1\} \quad \text{for } \theta \in \mathbb{R}^p,$$  \hfill (1.1)

where $\|\cdot\|_1$ is the $L^1$-norm. An equivalent prior can be specified by augmenting a flat (improper) prior distribution over an unrestricted support, $\tilde{\theta} \in \mathbb{R}^p$, with an exponentially distributed constraint variable, $\tilde{\omega}$. Truncating the joint distribution $(\tilde{\theta}, \tilde{\omega})$ to the support where the $L^1$-norm of $\tilde{\theta}$ is less than $\tilde{\omega}$ recovers the Bayesian Lasso in the marginal distribution of $\tilde{\theta}$ after truncation. This can be represented with the following conditional-hierarchical structure

$$\begin{align*}
(\theta, \omega) & \overset{d}{=} (\tilde{\theta}, \tilde{\omega} \mid \|\tilde{\theta}\|_1 \leq \tilde{\omega}) \\
\tilde{\theta} & \sim f_{\tilde{\theta}} \propto 1, \quad \tilde{\theta} \in \mathbb{R}^p \\
\tilde{\omega} & \sim \text{Exp}(\lambda).
\end{align*}$$  \hfill (1.2)

The marginal distribution of $\theta$ from the hierarchy in Equations (1.2) to (1.4) is the iid product-Laplacian distribution with rate $\lambda$ as given in (1.1). It is the Bayesian Lasso prior for $\theta$ (Park and Casella; 2008). The conditional and hierarchical structure of the above prior can be described by the following process. Starting from the foundation of the hierarchy, two independent random variables, $\tilde{\theta}$ and $\tilde{\omega}$, are used to form a joint random variable $(\tilde{\theta}, \tilde{\omega})$. Following this, the support of the joint random variable is constrained by conditioning on the set $\|\tilde{\theta}\|_1 \leq \tilde{\omega}$. The conditioning induces dependence between the random variables and, in this case, shrinks the constrained $\tilde{\theta}$ variable. Finally, the conditional probability is used to define the distribution for the stochastic constraint random variable $(\theta, \omega)$. Although the density for $\tilde{\theta}$, $f_{\tilde{\theta}}$, is initially assigned an improper prior, the resulting density of $(\theta, \omega)$ is proper. We formally define flat priors in this framework as the limiting case of a normal distribution in Section 3.1.

It is important to note that a truncated random variable $(\theta' \mid \tilde{\omega})$ with density $f_{\theta' \mid \tilde{\omega}} \propto 1(\|\theta'\|_1 \leq \tilde{\omega})$, i.e. the truncation is conditional on $\tilde{\omega}$, will not result in the same distribution as the above stochastic constraint prior. This difference may seem trivial in this case, where $(\theta' \mid \tilde{\omega})$ is conditionally uniform, but it is a consequential distinction to emphasise when working with stochastic constraints in a more general setting. The critical difference pertains to where the constraint or truncation is made. When the truncation is made conditionally on $\tilde{\omega}$, rather than
Table 1.1: Optimisation-based penalty and constraint Lasso duality and the Bayes analog with the Bayesian Lasso and stochastic constraints.

<table>
<thead>
<tr>
<th>Regularisation type</th>
<th>Standard</th>
<th>Constraint</th>
</tr>
</thead>
</table>
| Optimisation-based  | \[
\begin{align*}
    \max_{\tilde{\theta} \in \mathbb{R}^p} \left\{ \log \pi(x|\tilde{\theta}) - \lambda \|\tilde{\theta}\|_1 \right\} \\
    \text{s.t. } \|\tilde{\theta}\|_1 \leq \tilde{\omega}
\end{align*}
\] | \[
\begin{align*}
    \max_{\tilde{\theta} \in \mathbb{R}^p} \log \pi(x|\tilde{\theta}) \\
    \text{s.t. } \|\tilde{\theta}\|_1 \leq \tilde{\omega}
\end{align*}
\] | Equivalent for some \( \lambda \) and \( \tilde{\omega} \) |
| Fully-Bayesian      | \[
\begin{align*}
    x|\hat{\theta} & \sim \pi(x|\hat{\theta}) \\
    \hat{\theta} & \overset{\text{iid}}{\sim} \text{DE}(1/\lambda)
\end{align*}
\] | \[
\begin{align*}
    x|\hat{\theta} & \sim \pi(x|\hat{\theta}) \\
    \hat{\theta} & \overset{\text{iid}}{\sim} f_{\theta} \propto 1 \\
    \tilde{\omega} & \sim \text{Exp}(\lambda) \\
    \text{s.t. } \|\hat{\theta}\|_1 \leq \tilde{\omega}
\end{align*}
\] | Equivalent marginal posterior for \( \theta \) |

Optimum equals MAP estimate

Jointly on \( \tilde{\theta} \) and \( \tilde{\omega} \), the result is not a stochastic constraint distribution, except in a few special circumstances. For example, the Bayesian Lasso can also be generated with \( (\theta'|\tilde{\omega}) \sim f_{\theta'|\tilde{\omega}} \) and \( \tilde{\omega} \sim \text{Gam}(2, \lambda) \), but by doing so we lose the interpretability and flexibility of the stochastic constraint representation.

A connection to the standard Lasso (Tibshirani; 1996) can also be made. For a general (centred) regression problem, the Lasso estimator is the solution to

\[
\begin{align*}
    \min_{\hat{\theta} \in \mathbb{R}^p} -\log \pi(x|\hat{\theta}) \quad \text{s.t. } \|\hat{\theta}\|_1 \leq \tilde{\omega}.
\end{align*}
\]

for some \( \tilde{\omega} > 0 \). Generally the constraint-formulated Lasso in (1.5) is presented in Lagrangian form for ease of computation. However comparing (1.5) to the right hand side of (1.2) emphasises their similarities. In fact, placing priors \( \hat{\theta} \sim f_{\hat{\theta}} \propto 1 \) and \( \tilde{\omega} \sim \text{Exp}(\lambda) \) on the standard constraint-formulated Lasso variables generates the prior structure in Equations (1.2) to (1.4). As such, there is a strong connection between the standard and Bayesian Lasso. Not only does the maximum \textit{a posteriori} estimate (MAP) correspond to the standard Lasso estimator (Tibshirani; 1996; Park and Casella; 2008), the Bayesian Lasso can be reproduced with specific priors on the constraint-form of the standard Lasso (1.5). This simple example demonstrates that stochastic constraint priors are the Bayesian analog of the duality between constrained optimisation and its equivalent Lagrangian formulation (under conditions discussed in Section 3.2). An overview of this connection is given in Table 1.1.

Viewing the Bayesian Lasso as a stochastic constraint reveals the components that construct such a prior. The pre-constrained distribution of each parameter is flat and independent, that is \( f_{\theta} \propto 1 \), to reflect an absence of marginal information.
and no \textit{a priori} dependence between the parameters. Following this specification of marginal information, we incorporate prior knowledge of sparsity with a probabilistic constraint that shrinks the parameters jointly. This is the constraint $\|\tilde{\theta}\|_1 \leq \tilde{\omega}$ with exponential prior on $\tilde{\omega}$. The marginal and joint information are then combined to form the stochastic constraint prior.

We argue that this specification is both intuitive and flexible for Bayesian analysis. Three components can be chosen; the base prior $f_{\tilde{\theta}}$, the constraint (or penalty) function, and the constraining random variable $\tilde{\omega}$. Each component has a unique role in specifying a multivariate distribution to represent the prior knowledge for a set of parameters. Interestingly, regularisation can be applied to a given distribution rather than being an inherent property of the chosen prior, mirroring the process by which regularisation is incorporated into optimisation-based inference. Moreover, the penalty function can be chosen to suit the type of shrinkage required, and to respect the support of the underlying random variables, $\tilde{\theta}$ and $\tilde{\omega}$. 
CHAPTER 2

Stochastic constraint random variables

Stochastic constraint random variables can be constructed via two main approaches. The first uses an augmented random variable to define the random variables in a hierarchical setup, as in Section 1.2. The second is an algebraic transformation, similar to that of exponential tilting (Esscher; 1932; Siegmund; 1976), which is a transformation of the probability density function, or more generally its probability measure. Section 2.4.4 discusses the connection between exponential tilting and stochastic constraint random variables.

We focus on the case of continuous variables, but similar stochastically constrained discrete variables can be defined. We use the language of standard probability theory for ease of access, but note that an account of stochastic constraints using measure theory may be useful future research. To begin, univariate stochastic constraint random variables are discussed, followed by multivariate stochastic constraint random variables.

2.1 Univariate case

The three key elements of a stochastic constraint random variable are the

- base random variable \( \tilde{\theta} \),
- constraining or constraint random variable \( \tilde{\omega} \), and
- penalty function \( r(\tilde{\theta}) \).

In the univariate case the stochastic constraint distribution is defined by the hierarchy

\[
(\theta, \omega \mid \lambda) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} \mid r(\tilde{\theta}) \leq \tilde{\omega}, \lambda)
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}}, \quad \tilde{\theta} \in \Theta \subseteq \mathbb{R}
\]

\[
(\tilde{\omega} \mid \lambda) \sim f_{\tilde{\omega} \mid \lambda}, \quad \tilde{\omega} \in \Omega \subseteq \mathbb{R},
\]

where \( f_x \) denotes the density function of a random variable \( x \). The penalty function \( r(\theta) \) maps \( \Theta \) to \( \Omega \) and controls the class of regularisation and, together with the constraining random variable \( \tilde{\omega} \), controls the shrinkage being applied to the base random variable. The hyperparameter \( \lambda \) is fixed for the time being.

The joint probability distribution corresponding to (2.1) is

\[
f_{\theta, \omega \mid \lambda} = \frac{f_{\tilde{\theta}}(\theta)f_{\tilde{\omega} \mid \lambda}(\omega)1(r(\theta) \leq \omega)}{Z(\lambda)},
\]

where \( Z(\lambda) \) is the normalising constant. Equivalently, we can define the stochastic
constraint distribution with by the density function
\[ f_{\theta|\lambda}(\theta) = \frac{f_{\tilde{\theta}}(\theta) \tilde{F}_{\tilde{\omega}|\lambda}(r(\theta))}{Z(\lambda)}, \quad (2.2) \]
where the constraining random variable \( \tilde{\omega} \) has complementary cumulative distribution function (CCDF) \( \tilde{F}_{\tilde{\omega}|\lambda}(\omega) \), support on \( \Omega \), and hyperparameter \( \lambda \). The normalising constant for the new random variable \( \theta | \lambda \) is
\[ Z(\lambda) = \int_{\Theta} f_{\tilde{\theta}}(\theta) \tilde{F}_{\tilde{\omega}|\lambda}(r(\theta)) \, d\theta, \]
and will be finite and greater than zero unless otherwise stated.

The marginal density of \( \theta | \lambda \) is plotted in Figure 2.1 for some simple univariate stochastic constraint variables. The figure illustrates the effect of applying two different stochastic constraints to a base unit normal distribution. The density of the base random variable is shown in panel (a) without constraint. In panels (b) and (c), a larger spike at zero corresponds to the increasing the probability mass near zero of the constraining random variable. The univariate Bayesian elastic net is provided as an explicit example in Example 2.1.

**Example 2.1** (Bayesian elastic net as a stochastic constraint distribution). Take \( \tilde{\theta} \sim N(\mu, \sigma^2) \) and \( \tilde{\omega} \sim \text{Exp}(\lambda_2) \) with parameters \( \mu, \sigma^2, \) and \( \lambda_2 \) fixed. The stochastic constraint transformation of \( \tilde{\theta} \) and \( \tilde{\omega} \), with penalty function \( r(\theta) = |\theta| \), results in the stochastic constraint distribution with density
\[ f_{\theta|\mu,\sigma^2,\lambda_2} \propto \exp \left( -\frac{(\theta - \mu)^2}{2\sigma^2} - \lambda_2|\theta| \right). \]
Setting \( \mu = 0 \) and \( \sigma^2 = (2\lambda_1)^{-1} \) generates the univariate Bayesian elastic net (Li and Lin; 2010), based on the original elastic net of Zou and Hastie (2005) with density
\[ f_{\theta|\lambda_1,\lambda_2} \propto \exp \left( -\lambda_1\theta^2 - \lambda_2|\theta| \right). \]
The density for \( \mu = 0, \sigma^2 = 1, \) and \( \lambda_2 = 1 \) corresponds to panel (b) in Figure 2.1.

### 2.2 Multivariate case

The general joint stochastic constraint prior can be formulated in a hierarchy as
\[ (\theta, \omega | \lambda) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} | r(\tilde{\theta}) \preceq \tilde{\omega}, \lambda) \quad (2.3) \]
\[ \tilde{\theta} \sim f_{\tilde{\theta}}, \quad \tilde{\theta} \in \Theta \subseteq \mathbb{R}^p \quad (2.4) \]
\[ (\tilde{\omega} | \lambda) \sim f_{\tilde{\omega}|\lambda}, \quad \tilde{\omega} \in \Omega \subseteq \mathbb{R}^q, \quad (2.5) \]
where \( r(\cdot) \) maps \( \Theta \) to \( \Omega \), \( f_{\tilde{\theta}} \) and \( f_{\tilde{\omega}} \) are probability density functions for \( \tilde{\theta} \) and \( \tilde{\omega} \) respectively, and \( \preceq \) denotes element-wise inequality. The hyperparameter \( \lambda \) can be
2.2. MULTIVARIATE CASE

Figure 2.1: Marginal densities of some example univariate stochastic constraint distributions. (a) Base random variable $\tilde{\theta} \sim N(0,1)$ only. (b) Base random variable with constraining random variable $\tilde{\omega} \sim \text{Exp}(1)$ and penalty function $|\theta|$. (c) Base random variable with constraining random variable $\tilde{\omega} \sim \text{Gam}(0.5, 1)$ and penalty function $|\theta|$. 
fixed or assigned a prior. The density of the general joint stochastic constraint prior is

\[ f_{\theta,\omega|\lambda} = \frac{f_{\tilde{\theta}}(\theta)f_{\tilde{\omega}|\lambda}(\omega)1(r(\tilde{\theta}) \preceq \tilde{\omega})}{Pr(r(\tilde{\theta}) \preceq \tilde{\omega}|\lambda)}. \]  

(2.6)

Assuming the elements of \( \tilde{\omega}|\lambda \) are independent, the marginal distribution of \( \theta \) derived from (2.3) to (2.5) is then

\[ f_{\theta|\lambda} = \frac{f_{\tilde{\theta}}(\theta)\prod_{j=1}^{q} F_{\tilde{\omega}_j|\lambda}(r_j(\theta))}{Pr(r(\tilde{\theta}) \preceq \tilde{\omega}|\lambda)}, \]  

(2.7)

where \( F_{\tilde{\omega}_j|\lambda} \) is the CCDF of \( \tilde{\omega}_j|\lambda \). We will assume independence for the conditional random variables \( \tilde{\omega}_j|\lambda \) unless otherwise stated.

It is possible to give an interpretation to a flat prior on \( \tilde{\theta} \) (e.g. \( f_{\tilde{\theta}} \propto 1 \) and \( \tilde{\theta} \in \mathbb{R}^p \)) by choosing \( \tilde{\theta} \sim N(0, \sigma^2 I) \) and taking the limit as \( \sigma^2 \to \infty \) on the joint density \( (\theta, \omega) \).

Section 3.1 discusses this further. In the limit, the joint stochastic constraint prior is

\[ f_{\theta|\lambda} = \frac{\prod_{j=1}^{q} F_{\tilde{\omega}_j|\lambda}(r_j(\theta))}{\int_{\Theta} \prod_{j=1}^{q} F_{\tilde{\omega}_j|\lambda}(r_j(\theta)) d\theta} \]  

(2.8)

if the normalising term is finite. This expression could have been obtained informally by setting \( f_{\tilde{\theta}} = 1 \) in (2.7) in both the numerator and denominator. In general we will call \( Z(\lambda) \) the normalising term whose value is

\[ Z(\lambda) = \begin{cases} 
\Pr(r(\tilde{\theta}) \preceq \tilde{\omega}|\lambda) & \text{if } f_{\tilde{\theta}} \text{ is proper}, \\
\int_{\mathbb{R}^p} \prod_{j=1}^{q} F_{\tilde{\omega}_j|\lambda}(r_j(\theta)) d\theta & \text{if } f_{\tilde{\theta}} \propto 1, \tilde{\theta} \in \mathbb{R}^p.
\end{cases} \]  

(2.9)

In the fully-independent case, \( \tilde{\theta} \) are also independent, the number of constraints and variables are equal so that \( p = q \), and \( r_i(\theta) = r_i(\theta_i) \) for all \( i = 1, 2, \ldots, p \). The marginal distribution for \( \theta_i|\lambda \) is given by

\[ f_{\theta|\lambda} = \frac{f_{\tilde{\theta}}(\theta)F_{\tilde{\omega}|\lambda}(r(\theta))}{Pr(r(\tilde{\theta}) \preceq \tilde{\omega}|\lambda)}. \]  

(2.10)

if \( f_{\tilde{\theta}}(\theta) \) is proper and after dropping the subscript for simplicity. If \( f_{\tilde{\theta}}(\theta) \) is flat and improper then we have

\[ f_{\theta|\lambda} = \frac{F_{\tilde{\omega}|\lambda}(r(\theta))}{\int_{\mathbb{R}} F_{\tilde{\omega}|\lambda}(r(\theta)) d\theta} \]  

(2.11)

if the normalising term is finite. Therefore, in the fully-independent case, the normalising term of the marginal distribution \( \theta_i|\lambda \) is

\[ Z(\lambda) = \begin{cases} 
\Pr(r(\tilde{\theta}) \preceq \tilde{\omega}|\lambda) & \text{if } f_{\tilde{\theta}} \text{ is proper}, \\
\int_{\mathbb{R}} F_{\tilde{\omega}|\lambda}(r(\theta)) d\theta & \text{if } f_{\tilde{\theta}} \propto 1, \tilde{\theta} \in \mathbb{R}.
\end{cases} \]  

(2.12)
The stochastic constraints discussed in this thesis may appear similar to random truncation (or soft-constraints). However, a random truncation on $\tilde{\theta}$, differs to the right hand side of Equation (2.3) as the truncation is imposed on the conditional distribution $(\tilde{\theta} | \tilde{\omega})$, rather than on the joint distribution of $(\tilde{\theta}, \tilde{\omega})$. For this reason, we use the terminology stochastic constraints to refer to the types of joint-support-restricted probability distributions we are interested in. We explore the link to scale mixtures of uniform distributions in Section 2.4.1.

2.3 Some classes of stochastic constraint distributions

Some useful classes of stochastic constraint distributions can be defined as follows.

Definition 2.2 (Global stochastic constraint distribution). A global stochastic constraint distribution is an SC distribution where shrinkage is applied to all the variables of the base distribution jointly. In other words, each component of the penalty function $r(\theta)$ is functionally dependent on all elements of $\theta$.

Example 2.3 (Global stochastic constraint distribution). Let $\tilde{\theta} \sim N(0, \sigma^2 I)$ and $\tilde{\omega} \sim \text{Gam}(a, b)$. An example of a global stochastic constraint distribution is the random variable generated by

$$(\theta, \omega | a, b) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} | \|\tilde{\theta}\|_1 \leq \tilde{\omega}, a, b),$$

whose marginal distribution is

$$f_{\theta|a,b} \propto \exp \left( -\frac{\theta^\top \theta}{2\sigma^2} \right) \Gamma \left( a, b\|\theta\|_1 \right),$$

where $\Gamma(a, t)$ is the regularised upper incomplete gamma function, i.e. the CCDF for $\tilde{\omega}$. The distribution is a global stochastic constraint distribution since the penalty function $r(\theta) = \|\theta\|_1$ is functionally dependent on all elements of $\theta$.

Definition 2.4 (Local stochastic constraint distribution). A local stochastic constraint variable is an SC distribution where shrinkage is applied to all the variables of the base distribution separately, and the constraining random variables are independent and identically distributed (perhaps conditional on hyperparameters). In other words, each component of the penalty function $r(\theta)$ is functionally dependent on only one the element of $\theta$, and $\tilde{\omega} | \lambda \overset{iid}{\sim} f_{\omega|\lambda}$.

Example 2.5 (Local stochastic constraint distribution). Let $\tilde{\theta} \sim N(0, \sigma^2 I)$ and $\tilde{\omega} \overset{iid}{\sim} \text{Gam}(a, b)$, both with number of elements $p$. An example of a local stochastic constraint distribution is the random variable generated by

$$(\theta, \omega | a, b) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} | \|\tilde{\theta}\| \leq \tilde{\omega}, a, b),$$
whose marginal distribution is

\[ f_{\theta|a,b} \propto \exp \left( -\frac{\theta^\top \theta}{2\sigma^2} \right) \prod_{i=1}^{p} \Gamma (a, b|\theta_i) . \]

This is a local stochastic constraint distribution since \( r(\theta) = [|\theta_1|, |\theta_2|, \ldots, |\theta_p|]^\top \), so each element of the penalty is functionally independent of all elements of \( \theta \), except for one (the \( i \)th element).

**Definition 2.6 (Separable penalty function).** A separable penalty function is one that is additively separable. That is, \( r(\theta) \) can be written as

\[ r(\theta) = q_1(\theta_1) + q_2(\theta_2) + \cdots + q_p(\theta_p), \]

where the \( q_i \) are some functions that satisfy the conditions for penalty functions. It may be that \( q_i(\cdot) \) functions depend on some parameters \( \lambda \) in which case we could write the separable function as

\[ r(\theta; \lambda) = q_1(\theta_1; \lambda) + q_2(\theta_2; \lambda) + \cdots + q_p(\theta_p; \lambda). \]

**Example 2.7 (Separable penalty function).** Let the support of the base distribution and constraining random variable be \( \Theta = \mathbb{R} \) and \( \Omega = \mathbb{R}_+ \) respectively. The penalty function \( r(\theta) = ||\theta||_2^2 \) is separable because it may be additively decomposed into a series of functions that only involve one of the parameters, namely

\[ r(\theta) = ||\theta||_2^2 = \theta_1^2 + \theta_2^2 + \cdots + \theta_p^2. \]

Importantly, each separated function is a penalty function for the base distribution and constraining random variable in its own right.

The relation between global and local stochastic constraints, as well as separable penalty function are discussed in Chapter 5.

2.4 **Relation to other random variables and simulation methods**

There are several special cases of stochastic constraint distributions that are equivalent to well-known random variables in the literature. Their origins and uses are varied, for example, exponentially tilted distributions are often used for rare-event simulation, whilst skew random variables have been proposed for modelling. Scale mixture of uniforms and slice samplers have been used to facilitate computation of complicated posterior distributions. These latent-variable methods are intricately related to stochastic constraint priors but are used to generate efficient samplers, rather than our primary aim of constructing new and useful distributions.
2.4. RELATION TO OTHER RANDOM VARIABLES AND SIMULATION METHODS

2.4.1 Scale mixture of uniform distributions

Scale mixture of uniform distributions have been used to construct tractable Gibbs samplers in Bayesian models (Walker et al.; 1997; Qin et al.; 1998b,a), and have been known since at least Feller (1971). In recent years, they have been found use again for shrinkage-type regressions (Mallick and Yi; 2014, 2018) and covariance matrix estimation (Wang and Pillai; 2013).

The existence of a scale mixture of uniform representation relies on the conditions in Theorem 2.8.

**Theorem 2.8** (Scale mixture of uniform representations. Feller (1971); Walker et al. (1997)). Suppose \( \theta \) is a real-valued random variable with continuous, unimodal, and symmetric distribution with mode zero, having density \( f_{\theta} \) for \( \theta \in \mathbb{R} \). Suppose \( f'_{\theta} \) exists for all \( \theta \), then \( f_{\theta} \) has the form

\[
 f_{\theta}(\theta) = \int_0^\infty \frac{1}{2\omega} 1(|\theta| < \omega) f_{\omega}(\omega) \, d\omega, \quad (2.13)
\]

where \( f_{\omega}(\omega) \propto -2\omega f'_{\theta}(\omega) \) is a density on \([0, \infty)\).

By representing the distribution \( f_{\theta} \) as an integral in Equation (2.13) we can view the random variable as having a distribution that is a scale mixture of uniforms. This can be presented in the hierarchy

\[
 (\theta | \omega) \sim U(-\omega, \omega) \\
 \omega \sim f_{\omega} \propto -2\omega f'_{\theta}(\omega),
\]

which has an equivalent stochastic constraint definition given by

\[
 (\theta, \omega) \overset{d}{=} \left( \tilde{\theta}, \tilde{\omega} \mid |\tilde{\theta}| \leq \tilde{\omega} \right) \\
 \tilde{\theta} \sim f_{\tilde{\theta}} \propto 1, \quad \tilde{\theta} \in \mathbb{R} \\
 \tilde{\omega} \sim f_{\tilde{\omega}} \propto -f'_{\tilde{\theta}}(\tilde{\omega}), \quad \tilde{\omega} \in \mathbb{R}^+.
\]

**Example 2.9** (Two representations of the Normal distribution). A scale mixture of uniforms representation of the unit Normal distribution for \( \theta \sim N(0, 1) \) can be written as

\[
 (\theta | \omega) \sim U(-\omega, \omega) \\
 \omega \sim GGam(3, 2^{-1/2}, 2),
\]

by Theorem 2.8, where \( GGam(\alpha, \beta, \kappa) \) is the generalised gamma distribution. Whilst an equivalent stochastic constraint formulation is

\[
 (\theta, \omega) \overset{d}{=} \left( \tilde{\theta}, \tilde{\omega} \mid \tilde{\theta}^2 \leq \tilde{\omega} \right) \\
 \tilde{\theta} \sim f_{\tilde{\theta}} \propto 1, \quad \tilde{\theta} \in \mathbb{R} \\
 \tilde{\omega} \sim Exp(1).
\]
Example 2.9 illustrates the relation between scale mixture of uniforms and an equivalent stochastic constraint distribution. It also reveals the added flexibility in the stochastic constraint formulation, namely, the ability to choose a base distribution and penalty function. Previously this attribute has not been important, since it does not contribute to constructing Gibbs samplers. But when constructing new distributions this flexibility is paramount.

2.4.2 Slice sampling

Slice sampling is an augmented-data or latent variable method which simulates uniform distributions over regions that have the correct target marginal probability law (Chen and Schmeiser; 1998; Damien et al.; 1999; Neal; 2003). It is a natural extension of the scale mixtures of uniform representation of random variables discussed in Section 2.4.1. In the one-dimensional we sample from the joint distribution defined by

$$f_{\theta, \omega} = \frac{1(g(\theta) \geq \omega)}{Z} \quad \text{for } (\theta, \omega) \in \Theta \times \Omega,$$

where \(\Theta \subseteq \mathbb{R}, \Omega = \{\omega : 0 \leq \omega \leq \max_{\theta \in \Theta} g(\theta)\}, g(\theta) \geq 0 \text{ for all } \theta \in \Theta, \text{ and } Z = \int_{\Theta} g(\theta) \, d\theta.$$
The marginal density for \(\theta\) is therefore

$$f_\theta = \int_0^{g(\theta)} Z^{-1} \, d\omega = \frac{g(\theta)}{Z},$$

which is the target density. A Gibbs sampler can be constructed from Equation (2.14) which can lead to a tractable simulation algorithm for an otherwise difficult density \(f_\theta\).

Replacing the identity condition in Equation (2.14) with \(1(-g(\theta) \leq -\omega)\) trivially generates a stochastic constraint distribution with hierarchy

\[
(\theta, \omega) \overset{d}{=} \left(\tilde{\theta}, \tilde{\omega} \mid -g(\tilde{\theta}) \leq \tilde{\omega}\right)
\]

\[
\tilde{\theta} \sim f_\tilde{\theta} \propto 1, \quad \tilde{\theta} \in \mathbb{R}
\]

\[
\tilde{\omega} \sim \text{Exp}(1),
\]

Elliptical slice sampling is a related simulation method introduced by Murray et al. (2010). It generates random samples from densities of the form \(f_\theta \propto N(\theta; 0, \Sigma) g(\theta)\) by slicing out the \(g(\theta)\) component of the target and sampling over slices on ellipses. The method is usually applied so that \(g(\theta)\) is the likelihood, while the remaining Gaussian term represents the prior distribution. However, the reverse has recently been used in Hahn et al. (2019, forthcoming). To see the connection with stochastic constraint random variables it is instructive to derive the sampler using the following formulation. The stochastic constraint hierarchy of

\[
(\theta, \omega) \overset{d}{=} \left(\tilde{\theta}, \tilde{\omega} \mid -\log g(\tilde{\theta}) \leq \tilde{\omega}\right)
\]

\[
\tilde{\theta} \sim N(0, \Sigma)
\]

\[
\tilde{\omega} \sim \text{Exp}(1),
\]
2.4. RELATION TO OTHER RANDOM VARIABLES AND SIMULATION METHODS

has marginal distribution equal to the target of \( f_\theta \propto N(\theta; 0, \Sigma)g(\theta) \). From the stochastic constraint decomposition, a Gibbs sampler derived with full conditionals

\[
\begin{align*}
\tau & \sim \text{Exp}(1) \\
(\omega | \theta) &= \tau - \log g(\theta) \\
(\theta | \omega) & \sim f_{\theta | \omega} \propto N(\theta; 0, \Sigma)1(\log g(\tilde{\theta}) \leq \tilde{\omega})?
\end{align*}
\]

A Metropolis-Hastings-within-Gibbs (MH-within-Gibbs) sampling strategy is then used by replacing the (possibly non-existent) exact conditional sampler for \( \theta | \omega \) with a Metropolis-Hastings (MH) step. The MH proposal is

\[
(\theta' | \theta) = \theta \cos u + x \sin u
\]

where \( x \sim N(0, \Sigma) \) and \( u \sim U(0, 2\pi) \), and is accepted if \( -\log g(\theta') < \omega \). The elliptical slice sampler arises by using an adaptive slice sampler for \( u \), conditional on \( \theta \) and \( \omega \). The slice sampler will eventually find a \( u \) satisfying the acceptance condition, so rejections need not occur.

2.4.3 Skew random variables

One definition of skew random variables was given by O’Hagan and Leonard (1976) (see Azzalini; 2005, for an overview) whereby skewness is induced on a base random variable by multiplying the density by the cumulative distribution function of a second random variable. In their setting both random variables must have densities that are symmetric about zero, leading to a density of the following form

\[
f_{\theta | \alpha}(\theta) = 2f_{\tilde{\theta}}(\theta)F_{\tilde{\omega} | \alpha}(\omega), \quad \theta \in \mathbb{R}, \tag{2.16}
\]

where \( \alpha \), generally a scale parameter for \( \tilde{\omega} \), controls the skewness. In particular, the scaling factor of 2 used in the above formula is due to the assumption that the underlying random variables are symmetric, and will be different otherwise.

Equation (2.16) is remarkably similar to the marginal distribution of stochastic constraint random variables. In fact, an equivalent distribution can be defined by

\[
(\theta, \omega | \alpha) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} | -\tilde{\theta} \leq \tilde{\omega}, \alpha) \quad \tilde{\theta} \sim f_{\tilde{\theta}}, \quad \tilde{\theta} \in \mathbb{R} \\
(\tilde{\omega} | \alpha) \sim f_{\tilde{\omega} | \alpha}, \quad \tilde{\omega} \in \mathbb{R},
\]

which belongs to the class of stochastic constraint distributions. The marginal distribution of \( \theta | \alpha \) will be proportional to the density in (2.16), and equal to it when \( f_{\tilde{\theta}} \) and \( f_{\tilde{\omega} | \alpha} \) are symmetric about zero.

2.4.4 Exponentially tilted random variables

Positive exponentially tilted random variables belong to the class of stochastic constraint variables. A positive
continuous random variable \( \tilde{\theta} \) with density \( f_{\tilde{\theta}} \) can be exponentially tilted by defining the new random variable \( \theta \) with density

\[
f_{\theta}(\theta) \propto f_{\tilde{\theta}}(\theta) \exp(-\lambda \theta),
\]

where \( \lambda > 0 \). In stochastic constraint terms the formulation is

\[
(\theta, \omega) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} \mid \tilde{\theta} \leq \tilde{\omega})
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}}, \quad \tilde{\theta} \in \mathbb{R}_+
\]

\[
\tilde{\omega} \sim \text{Exp}(\lambda).
\]

Exponentially tilted random variables on the negative axis can be similarly defined, but those on the entire real line do not have a simple stochastic constraint representation (perhaps not any). The issue is in introducing the CCDF \( \bar{F}_{\tilde{\omega}}(\theta) = \exp(-\lambda \theta) \) into Equation (2.17) via a stochastic constraint requires that \( \tilde{\omega} \) has the same support as \( \tilde{\theta} \); the real line. For the exponential distribution this is not the case.

### 2.5 Example stochastic constraint distributions

It is possible to write many priors for sparse regression and other related problems using the stochastic constraint framework. However, the SC reformulation does require one to carefully consider the new normalising constant of the stochastic constraint density. Particularly for Gibbs sampling, the hyperparameters of the stochastic constraint distribution will have different conditional distributions compared to their standard counterparts.

#### 2.5.1 Generalised Gaussian distribution

The generalised Gaussian, or exponential power distribution, can be defined with density function

\[
f_{\theta | \mu, \eta, \kappa} = \frac{\kappa}{2 \eta \Gamma(1/\kappa)} \exp \left\{ -\frac{|\theta - \mu|^\kappa}{\eta^\kappa} \right\}.
\]

(2.18)

It generalises the Normal \((\kappa = 2)\) and Laplacian \((\kappa = 1)\) distributions, providing a smooth interpolation between the two.

The zero-mean generalised Gaussian (2.18) has a stochastic constraint decomposition given by

\[
(\theta, \omega \mid \kappa, \eta) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} \mid |\tilde{\theta}|^\kappa \leq \tilde{\omega}, \kappa, \eta)
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}} \propto 1, \quad \tilde{\theta} \in \mathbb{R}
\]

\[
(\tilde{\omega} \mid \eta) \sim \text{Exp}(\eta^{-\kappa}).
\]

The associated probability density function is therefore

\[
f_{\theta, \omega \mid \eta, \kappa} = \frac{\eta^{-\kappa} \exp \left\{ -\eta^{-\kappa} \omega \right\} \mathcal{1}(|\theta|^\kappa \leq \omega)}{\mathcal{Z}(\eta, \kappa)},
\]

(2.19)
where the normalising constant is $Z(\eta, \kappa) = 2\kappa^{-1}\eta \Gamma(1/\kappa)$ which is clear by comparing the density in Equation (2.18) to the marginal $\int_{\mathbb{R}} f_\theta \omega_{\eta, \kappa} \, d\omega$.

The stochastic constraint form of the generalised Gaussian leads to a simple Gibbs sampler with full conditional distributions

$$
(\theta | \omega) \sim U(-\omega^{1/\kappa}, \omega^{1/\kappa})
$$

$$
\tau \sim \text{Exp}(\eta^{-\kappa})
$$

$$
(\omega | \theta) = |\theta|^\kappa + \tau
$$

in the univariate case. The elementary nature of the above conditional distributions is due to the memoryless property of exponential distribution and the flat prior used. We discuss the multivariate case with a Bayesian linear model in Example 4.1.

If $0 < \kappa < 1$ and the generalised Gaussian is being used as a prior for Bayesian linear regression then the model is known as the Bayesian Bridge. Polson et al. (2014) suggested several data-augmentation schemes in order to sample from this model. The Gibbs sampler for the Bayesian bridge model arising from a stochastic constraint decomposition is simpler than the methods found in Polson et al. (2014); the latent variable is exponential rather than a mixture of two gamma distributions, and the conditional $\theta$ is uniform rather than triangular. An equivalent Gibbs sampler was recently used in Mallick and Yi (2018) who noted both advantages and drawbacks compared to the recommended sampler in Polson et al. (2014).

### 2.5.2 Horseshoe prior

Conditional on the measurement error, $\sigma^2$, the horseshoe prior of Carvalho et al. (2010) takes the form

$$
\theta|\sigma^2, \tau, \lambda \sim N(0, \sigma^2 \tau^2 D\lambda^2)
$$

$$
\lambda \overset{\text{iid}}{\sim} \text{Ca}_+(1)
$$

$$
\tau \sim \text{Ca}_+(1),
$$

where $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_p]^\top$, and $D\lambda^2 = \text{diag}\{\lambda_1^2, \lambda_2^2, \ldots, \lambda_p^2\}$. This prior can be represented in the stochastic constraint framework as

$$
(\theta, \omega | \lambda, \tau, \sigma^2) \overset{d}{=} \left(\hat{\theta}, \hat{\omega} \mid \hat{\theta}^2 \preceq \hat{\omega}, \lambda, \tau, \sigma^2\right)
$$

$$
\hat{\theta} \sim f_{\hat{\theta}} \propto 1, \hat{\theta} \in \mathbb{R}^p
$$

$$
(\hat{\omega}_i | \lambda_i, \tau, \sigma^2) \sim \text{Exp}(2^{-1}[\lambda_i \tau \sigma]^{-2}) \quad (1 \leq i \leq p)
$$

$$
\lambda \overset{\text{iid}}{\sim} \text{Ca}_+(1)
$$

$$
\tau \sim \text{Ca}_+(1),
$$

where $\hat{\theta}^2$ is element-wise squaring, and $\preceq$ denotes an element-wise inequality as before. This representation is not unique, as replacing the penalty functions with $|\hat{\theta}_i| \leq \hat{\omega}_i$ and conditional distribution with $(\hat{\omega}_i | \lambda_i, \tau, \sigma^2) \sim \text{Rayl}(\lambda_i \tau \sigma)$ generates the same prior. However, the Gibbs sampler arising from the exponential representation is more convenient (see Appendix A.2.1).
2.5.3 Regularised horseshoe and horseshoe+ prior  The regularised horseshoe distribution (Piironen and Vehtari; 2017) can be generated by replacing the flat prior on $f_{\tilde{\theta}}$ in the stochastic constraint formulation of the horseshoe (2.21) with $f_{\tilde{\theta} \mid \alpha^2} \sim N(0, \alpha^2I)$. The hyperparameter $\alpha^2$ is generally given a prior also. The aim of replacing the flat prior is to apply a small amount of shrinkage to the largest coefficients of $\theta$, since the plain horseshoe was designed to leave these approximately equal to their ordinary least square estimates. One of the compelling use cases for such a prior is when encountering data separation in Bayesian logistic regression with the horseshoe as a prior.

One contribution of the stochastic constraints framework is to supply an intuitive explanation of the difference between the plain horseshoe and regularised horseshoe priors. The flat prior of the horseshoe indicates an absence of prior knowledge about each marginal coefficient in $\theta$. Whilst the stochastic constraint shrinks the coefficients jointly, in such a way that most are shrunk heavily towards zero and some are left unchanged. The regularised horseshoe maintains the joint shrinkage but adds independent shrinkage to each individual coefficient. This can be useful in some situations. This intuitive explanation is not a substitute for the mathematical exposition provided in Piironen and Vehtari (2017); still, having an interpretation like this may aid in proposing and developing new priors.

The horseshoe+ prior was proposed by Bhadra et al. (2017) and can be defined by replacing the unit half Cauchy prior on $\lambda_i$ in the standard (2.20) or stochastic constraint representation (2.21) with

\[(\lambda_i \mid \eta_i) \sim \text{Ca}(\eta_i)\]
\[\eta_i \sim \text{Ca}(1),\]

for $1 \leq i \leq p$. This prior increases the shrinkage applied to small regression coefficients compared to the horseshoe prior, and is intended for situations where the regression coefficients are “ultra-sparse”, meaning there are very few nonzero parameters compared to the parameter size $p$. The authors argue that the horseshoe+ prior is better able to separate signal from noise in the normal-means model.

2.5.4 Dirichlet-Laplace prior  The Dirichlet-Laplace (DL) prior (Bhattacharya et al.; 2015) is another joint shrinkage prior for regression models. It is based on a Laplacian-scale mixture kernel with hyperparameters chosen to encourage posteriors with most parameters shrunk to zero. The canonical DL prior, conditional on the regression variance $\sigma^2$, takes the form

\[(\theta_i \mid \phi_i, \tau, \sigma) \sim \text{DE}(\phi_i \tau \sigma), \quad (1 \leq i \leq p)\]
\[\phi_i \sim \text{Dir}(a, a, \ldots, a)\]
\[\tau \sim \text{Gam}(pa, 1/2),\]
whereas the stochastic constraint version is

\[
(\theta, \omega \mid \phi, \tau, \sigma^2) \overset{d}{=} \left( \tilde{\theta}, \tilde{\omega} \mid |\tilde{\theta}| \leq \tilde{\omega}, \phi, \tau, \sigma^2 \right)
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}} \propto 1, \tilde{\theta} \in \mathbb{R}^p
\]

\[
(\tilde{\omega} | \phi_i, \tau, \sigma) \sim \text{Exp}([\phi_i \tau \sigma]^{-1}) \quad (1 \leq i \leq p)
\]

\[
\phi_i \sim \text{Dir}(a, a, \ldots, a) \quad (1 \leq i \leq p)
\]

\[
\tau \sim \text{Gam}(pa, 1/2)
\]

where \(|\tilde{\theta}|\) denotes element-wise absolute value. Given the stochastic constraint representation, it is immediately clear how one might make an adjustment to generate a “regularised” version of this prior (in the sense of the regularised horseshoe), by replacing the flat base prior with a distribution that is weakly informative on the maximum scale of the estimators. Whilst we do not consider this here, a comparison between regularised versions of the horseshoe, horseshoe+, DL, and other priors for logistic regression with data separation would be worth future study.

\subsection*{2.5.5 R2-D2 prior}

The final example of a continuous shrinkage prior is the R2-D2 prior proposed by Zhang et al. (2017). The authors show that this prior has many desirable properties for linear regression, as well as theoretical and empirical evidence that it outperforms the horseshoe, horseshoe+, and DL priors in high-dimensional linear regression. The prior has a structure similar to the DL prior, and can be defined succinctly as

\[
(\beta_i | \lambda_i, \sigma^2) \sim \text{DE}((\lambda_i \sigma^2/2)^{1/2}) \quad (1 \leq i \leq p)
\]

\[
\lambda_i \sim \text{BP}(a, \pi, b) \quad (1 \leq i \leq p),
\]

with several equivalent representations given in Zhang et al. (2017). One stochastic constraint version of the prior is

\[
(\theta, \omega | \lambda, \sigma^2) \overset{d}{=} \left( \tilde{\theta}, \tilde{\omega} \mid |\tilde{\theta}| \leq \tilde{\omega}, \lambda, \sigma^2 \right)
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}} \propto 1, \tilde{\theta} \in \mathbb{R}^p
\]

\[
(\tilde{\omega} | \phi_i, \tau, \sigma) \sim \text{Exp}([\phi_i \tau \sigma]^{-1}) \quad (1 \leq i \leq p)
\]

\[
\lambda_i \sim \text{BP}(a, \pi, b) \quad (1 \leq i \leq p).
\]

Just as with the other shrinkage priors the key is to decompose the generalised Gaussian kernel into a stochastic representation. We explore the computational aspects of using the stochastic constraints representation for the continuous shrinkage priors in Chapter 7, and use the R2-D2 prior as a specific example.
Some theoretical results

In this section we note some general results about stochastic constraint random variables, to explore some of their properties and identify their uses.

3.1 Conditions for stochastic constraint distributions to be proper

Stochastic constraint random variables will always be proper and non-degenerate probability distributions if the base and constraining random variables are themselves proper random variables, and the constraint permits a sufficiently rich subset of joint parameter space. We describe the conditions in Theorem 3.1 below.

Theorem 3.1 (Stochastic constraints are proper distributions). Let \( f_{\tilde{\theta}} \) and \( f_{\tilde{\omega}} \) be densities for continuous random variables \( \tilde{\theta} \) and \( \tilde{\omega} \) with support \( \Theta \) and \( \Omega \) respectively. Define the joint support of the stochastic constraint random variable to be \( S = \{ (\theta, \omega) \in \Theta \times \Omega : r(\theta) \preceq \omega \} \) and let \( n = \dim(\Theta \times \Omega) \). Without loss of generality, let any hyperparameters be fixed. If \( f_{\tilde{\theta}} \) and \( f_{\tilde{\omega}} \) are proper, and there exists an \( n \)-ball, \( B(c, r) \), centred at \( c \in \mathbb{R}^n \) with radius \( r > 0 \) such that \( B(c, r) \subset S \) then the stochastic constraint distribution with base distribution \( f_{\tilde{\theta}} \), constraining distribution \( f_{\tilde{\omega}} \), and penalty \( r(\theta) \) is a proper, non-degenerate distribution.

Proof. The stochastic constraint distribution has density \( f_{\theta, \omega} = f_{\tilde{\theta}}(\theta)f_{\tilde{\omega}}(\omega)1(r(\theta) \preceq \omega)/Z \) where \( Z = \int_S f_{\tilde{\theta}}(\theta)f_{\tilde{\omega}}(\omega) \, d\theta \, d\omega \). In order to be proper and non-degenerate it suffices to show that \( 0 < Z < \infty \). For the upper limit, \( f_{\tilde{\theta}} \) and \( f_{\tilde{\omega}} \) are both proper and \( S \subset \Theta \times \Omega \) so \( Z \leq \int_\Theta \int_\Omega f_{\tilde{\theta}}(\theta)f_{\tilde{\omega}}(\omega) \, d\theta \, d\omega = 1 \). For the lower limit, \( B(c, r) \subset S \subset \Theta \times \Omega \) therefore \( Z > \int_{B(c, r)} f_{\tilde{\theta}}(\theta)f_{\tilde{\omega}}(\omega) \, d\theta \, d\omega > 0 \).

The existence of a ball \( B(c, r) \subset S \) guarantees that the penalty function does not restrict the joint support of the stochastic constraint to a measure zero set and hence does not produce a degenerate prior. It is a weak condition and is only violated in pathological examples, e.g. if the penalty function is equal to a constant.

We would also like to consider flat improper base priors or distributions. In Section 2.4.1 a stochastic constraint distribution with flat base prior was written as a scale mixture of uniform distributions. Here however, we wish to remain within the stochastic constraint framework and hence take a different route. To formalise our definition of a flat prior in the context of stochastic constraints, we define a flat prior to be the limit of a sequence of zero-mean normal distributions, as the variance approaches infinity.

Without loss of generality consider the stochastic constraint distribution, with (sup-
pressed) fixed hyperparameters, as

$$
(\theta, \omega) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} | r(\tilde{\theta}) \preceq \tilde{\omega})
$$

$$
\tilde{\theta} \sim f_{\tilde{\theta}} \propto \mathcal{N}(\tilde{\theta}; \mathbf{0}, \alpha \mathbf{I})g(\tilde{\theta})
$$

$$
\tilde{\omega} \sim f_{\tilde{\omega}}, \quad \tilde{\omega} \in \Omega,
$$

where \(\tilde{\theta}\) has length \(d\), \(f_{\tilde{\theta}}\) is a proper prior, and \(g(\tilde{\theta})\) determines the shape of the improper prior of interest. For example, a flat improper prior requires the shape defined by setting \(g(\tilde{\theta}) = 1\). The corresponding density function for (3.1) is

$$
f_{\theta, \omega} = (2\pi\alpha)^{-d/2} \exp\left\{-\frac{\theta^T \theta}{2\alpha}\right\} g(\theta) f_{\tilde{\omega}}(\omega) 1(r(\theta) \preceq \tilde{\omega}),
$$

(3.2)

where \(Z = \iint_{S} (2\pi\alpha)^{-d/2} \exp\left\{-\frac{\theta^T \theta}{2\alpha}\right\} g(\theta) f_{\tilde{\omega}}(\omega) d\theta d\omega\) and \(S = \{(\theta, \omega) \in \Theta \times \Omega : r(\theta) \preceq \tilde{\omega}\}\) as before.

**Definition 3.2** (Stochastic constraint distribution with flat base prior). Let \(g(\theta)\) be an improper prior, and let the structure

$$
(\theta, \omega) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} | r(\tilde{\theta}) \preceq \tilde{\omega})
$$

$$
\tilde{\theta} \sim f_{\tilde{\theta}} \propto g(\tilde{\theta}), \quad \theta \in \mathbb{R}^p
$$

$$
\tilde{\omega} \sim f_{\tilde{\omega}}, \quad \tilde{\omega} \in \Omega
$$

be formally defined as the limit as \(\alpha \to \infty\) of (3.1) or equivalently the same limit of the density in (3.2). In the limit the density is

$$
f_{\theta, \omega} = \frac{g(\theta) f_{\tilde{\omega}}(\omega) 1(r(\theta) \preceq \tilde{\omega})}{Z},
$$

(3.3)

as desired, with normalising constant \(Z = \iint_{S} g(\theta) f_{\tilde{\omega}}(\omega) d\omega d\theta\).

It is clear that the stochastic constraint distribution will be improper if the integral defining \(Z\) diverges. However, there are more instructive conditions that we can derive for particular cases.

**Proposition 3.3** (Proper stochastic constraints from flat base distributions). Let \(g(\theta) = 1\), for a flat improper base prior. Define \(V(\omega) = \int_{S_\omega} 1 d\theta\) where \(S_\omega = \{\theta \in \Theta : r(\theta) \preceq \omega\}\), i.e. the Euclidean volume of the region defined by \(S_\omega\) for a fixed \(\omega\). The normalising constant \(Z\) in (3.3) can be written as

$$
Z = \mathbb{E}(V(\omega)),
$$

if it exists.

**Proof.** The result is obtained by integrating \(\theta\) first in \(Z = \iint_{S} g(\theta) f_{\tilde{\omega}}(\omega) d\omega d\theta\) with \(g(\theta) = 1\). \(\square\)
Proposition 3.3 suggests a method to determine if a stochastic constraint variable is proper when it is defined with a flat improper base distribution and \( V(\omega) \) is difficult to integrate with respect to \( \tilde{\omega} \). For example, constructing a bounding region of \( S_\omega \), may be possible, with a simpler upper bound for \( V(\omega) \). If the upper bound of the volume can be integrated then the distribution will be proper.

Another route for a bounding function is possible if the \( \tilde{\omega} \) are independently distributed.

**Proposition 3.4** (Proper stochastic constraints from independent constraining variables). If \( \tilde{\omega} \) are independent then the normalising constant in (3.3) can be written as

\[
Z = \int_{\Theta} g(\theta) \prod_{j=1}^{q} \tilde{F}_{\tilde{\omega}}((r_j(\theta))) \, d\theta,
\]

if it exists.

**Proof.** The result is obtained by integrating \( \omega \) first in \( Z = \iiint S g(\theta) f_{\tilde{\omega}}(\omega) \, d\omega \, d\theta \).

If an upper bounding function can be found for the integrand then it may be possible to find a finite upper bound for the integral and hence show that the stochastic constraint distribution is proper. Furthermore, for simple stochastic constraint priors the following theorem is often useful.

**Theorem 3.5** (Flat, local and iid stochastic constraints). Assume the proposed local stochastic constraint distribution has flat improper base prior, i.e. \( g(\theta) = 1 \), and the \( q \) constraining random variables are independent and identically distributed by \( \tilde{\omega} \sim f_{\tilde{\omega}} \).

If each component of the penalty function is defined by \( r_i(\theta) = |\theta_i - \mu| \) for some fixed \( \mu \), where \( u : \mathbb{R}_+ \to \mathbb{R}_+ \), \( u(0) = 0 \), and \( u \) has inverse function \( v(t) \) then

\[
Z = 2^q \mathbb{E}(v(\tilde{\omega}))^q.
\]

**Proof.** A proof for Theorem 3.5 is given in Appendix A.1.1.

As an example, under the conditions of Theorem 3.5 we can say that if \( r_i(\theta) = |\theta_i - \mu|^k \) then \( Z = 2^q \mathbb{E}(\tilde{\omega}^{1/k})^q \), which is only a function of the fractional moments of \( \tilde{\omega} \). There is the special case for \( k = 1 \) where \( Z \) is also only a function of the first moment. This suggests the conjecture that stochastic constraint priors with (i) flat base distribution, (ii) penalties with similar behaviour to \( r_i(\theta) = |\theta_i - \mu|^k \) as \( \theta_i \to \infty \), and (iii) constraining variables with heavy tails (and hence the required fractional moment does not exist) will produce an improper prior for the marginal distribution of \( \theta_i \).
3.2 Results related to optimisation-based inference

The duality between constrained and penalised optimisation is well known in statistics, due to its role in defining regularised estimators and methods. For example, the Lasso estimator is proposed as a constrained regression problem in the seminal paper by Tibshirani (1996), but is generally estimated using an equivalent penalty-based optimisation, or in other words the Lagrangian form of the Lasso.

Even though the stochastic constraint characterisation of a prior closely resembles the duality between constraint and penalty versions of regularised regression, we would like to know under what conditions the MAP estimator is invariant between the stochastic constraint formulation and standard formulation.

Let \( \pi(y|\theta) \) be the model or likelihood of interest, and let \( r: \Theta \to \Omega \subseteq \mathbb{R}_+ \) be a penalty function. If the prior is a stochastic constraint distribution with \( f_{\theta,\omega|\lambda} \propto f_{\theta}(\theta)f_{\omega|\lambda}(\omega)1(r(\theta) \leq \omega) \) then the MAP estimate is defined by

\[
(\theta^*, \omega^*) = \arg \max_{(\theta, \omega) \in S} \{ \pi(y|\theta)f_{\theta}(\theta)f_{\omega|\lambda}(\omega) \}
\]

where \( S = \{(\theta, \omega) \in \Theta \times \Omega : r(\theta) \leq \omega \} \). \hfill (3.4)

Proposition 3.6 (Maximum a posteriori estimate with stochastic constraint prior). If the constraint random variable \( \tilde{\omega}|\lambda \) has density which is decreasing on \( \Omega \), then the optimum for \( \theta \) in (3.4) is equivalent to

\[
\theta^* = \arg \max_{\theta \in \Theta} \{ \pi(y|\theta)f_{\theta}(\theta)f_{\tilde{\omega}|\lambda}(r(\theta)) \}
\]

whilst the optimum for \( \omega^* \) is

\[
\omega^* = r(\theta^*).
\]

A proof can be constructed by noting that a decreasing \( f_{\tilde{\omega}|\lambda}(\omega) \) is minimised for the smallest \( \omega \) belonging to \( S \), which is therefore \( \omega = r(\theta) \). In comparison, when a stochastic constraint prior has had the constraining random variable integrated out, the MAP is

\[
\theta^* = \arg \max_{\theta \in \Theta} \{ \pi(y|\theta)f_{\theta}(\theta)\bar{F}_{\tilde{\omega}|\lambda}(r(\theta)) \}.
\]

In general, the marginalised MAP in (3.6) and full MAP in (3.5) will only be equivalent if \( f_{\tilde{\omega}|\lambda} = c\bar{F}_{\tilde{\omega}|\lambda} \) for some constant \( c \).

Corollary 3.7 (Maximum a posteriori estimate equivalence for exponential distribution). If \( \tilde{\omega} | \lambda \sim \text{Exp}(\lambda) \) then optimisations (3.4), (3.5), and (3.6) will have equivalent estimates for \( \theta \).

Corollary 3.7 is true since the density of an exponential is decreasing and satisfies \( f_{\tilde{\omega}|\lambda} = c\bar{F}_{\tilde{\omega}|\lambda} \) for constant \( c = 1/t \) where \( t \) is the rate of the exponential distribution.
3.3. CHARACTERISING SHRINKAGE FROM STOCHASTIC CONSTRAINTS

It is well known that MAP estimates are not invariant under transformations. As demonstrated in Corollary 3.7 however, the MAP estimate for stochastic constraints with an exponential constraining variable are invariant to marginalisation. Stochastic constraints that can be reformulated to use an exponential constraining variable also share this property.

3.3 Characterising shrinkage from stochastic constraints

**Theorem 3.8** (Posterior normalising constant I). Let \((\theta, \omega | \lambda)\) have prior \(f_{\theta, \omega | \lambda}\) as is in (2.3) to (2.5), and let \(\pi(y | \theta)\) be the model of interest. Consider the conditional regularised posterior \(p(\theta | y, \lambda) = \int_{\Omega} \pi(y | \theta) f_{\theta, \omega | \lambda}(\theta) d\omega / Z(\lambda)\) and unregularised posterior \(\tilde{p}(\theta | y) = \pi(y | \theta) f_{\tilde{\theta}}(\theta) / \tilde{Z}\). If \(\tilde{Z} \in (0, \infty)\), i.e. the unregularised posterior is proper, then the (conditional) ratio of normalising constants \(Z(\lambda) / \tilde{Z}\) can be written as

\[
\frac{Z(\lambda)}{\tilde{Z}} = \Pr(r(\tilde{\theta}) \preceq \tilde{\omega} | y, \lambda).
\]

Additionally, if \(\lambda \in \Lambda\) is given a prior \(f_{\lambda}\), then

\[
\frac{Z}{\tilde{Z}} = \Pr(r(\tilde{\theta}) \preceq \tilde{\omega} | y),
\]

where \(\tilde{\omega}\) has marginal distribution \(f_{\tilde{\omega}} = \int_{\Lambda} f_{\omega | \lambda}(\omega) f_{\lambda}(\lambda) d\lambda\) and

\[
Z = \int_{\Lambda} \int_{\Theta} \int_{\Omega} \pi(y | \theta) f_{\tilde{\theta}}(\theta) f_{\omega | \lambda}(\omega) 1(r(\theta) \preceq \omega) d\omega d\theta d\lambda.
\]

**Proof.** A proof is contained in Appendix A.1.2.

Generally, normalising constant ratios for models with different priors do not have such neat formulae or interpretable meanings – aside from an importance sampling interpretation. However, the stochastic constraint distribution structure results in the ratio having the following meaning: It is the probability that a random draw from the unregularised posterior is within the constraint region, \(r(\tilde{\theta}) \preceq \tilde{\omega}\), where \(\tilde{\omega}\) is drawn independent of draw from its’ (possibly hierarchal) prior. Theorem 3.8 is useful to characterise how the shrinkage from stochastic constraints operates in terms of posterior expectations. We explore this in Corollary 3.9 below.

**Corollary 3.9** (Forward transform). Following from Theorem 3.8 the regularised posterior expectation of \(g(\theta)\) (i.e. the expectation with respect to \(p(\theta | y)\)) can be written as

\[
E(g(\theta) | y) = \frac{E \left[ g(\tilde{\theta}) \Pr(r(\tilde{\theta}) \preceq \tilde{\omega} | \theta, \lambda | y) \right]}{\Pr(r(\tilde{\theta}) \preceq \tilde{\omega} | y)}
\]

(3.7)

for any function \(g(\theta)\) where \(E(g(\theta) | y)\) exists.

**Proof.** The proof appears in Appendix A.1.3.
Corollary 3.9 demonstrates the relation between regularised posterior expectation and unregularised posterior expectation. Two forces are acting on the unregularised posterior expectation $E(g(\tilde{\theta})|y)$. Firstly, a continuous weighting, $\Pr(r(\tilde{\theta}) \preceq \tilde{\omega}|\tilde{\theta}, \lambda)$, shrinks the parameter $\tilde{\theta}$ towards the null space of $r(\theta)$, and is dependent on the distribution choice of the constraint random variable $\tilde{\omega}$. However, a second factor resists the shrinkage in the expectation. The probability $\Pr(r(\tilde{\theta}) \preceq \tilde{\omega}|y)$ offsets this shrinkage, by taking into account the posterior probability that the unconstrained random variable $r(\tilde{\theta})$ is less than $\tilde{\omega}$ (element-wise) prior to the constraint being imposed.

**Lemma 3.10** (Posterior normalising constant II). Under the same conditions as Theorem 3.8, the ratio of normalising constants $\tilde{Z}/Z(\lambda)$ can be written as

$$\frac{\tilde{Z}}{Z(\lambda)} = E\left[\Pr(r(\theta) \preceq \tilde{\omega}|\theta, \lambda)^{-1} | y, \lambda\right].$$

**Proof.** A proof appears in Appendix A.1.4. \qed

Once again this result unexpected and due to the structure of stochastic constraint priors.

We can use Lemma 3.10 to prove the following corollary.

**Theorem 3.11** (Backward transform). Assume that $E(g(\tilde{\theta}) | y)$ exists, then under the conditions of Theorem 3.8,

$$E(g(\tilde{\theta}) | y) = \frac{E\left[\frac{g(\theta)}{\Pr(r(\theta) \preceq \omega|\theta, \lambda)} | y, \lambda\right]}{E\left[\Pr(r(\theta) \preceq \tilde{\omega}|\theta, \lambda)^{-1} | y, \lambda\right]} \quad \text{for some } \lambda \in \Lambda.$$  

Moreover if $\lambda$ has an assigned prior we can write

$$E(g(\tilde{\theta}) | y) = \frac{E\left[\frac{g(\theta)}{\Pr(r(\theta) \preceq \omega|\theta, \lambda)} | y\right]}{E\left[\Pr(r(\theta) \preceq \tilde{\omega}|\theta, \lambda)^{-1} | y\right]}.$$  

**Proof.** A proof appears in Appendix A.1.5. \qed

This corollary is useful as it demonstrates a back-transformation from our regularised posterior to the unregularised posterior. In particular, we can calculate the shrinkage factor of the model parameters without directly calculating $E(\tilde{\theta}|y)$ in the unregularised model. The shrinkage factor, for a function $h$ defined as

$$\kappa_h = E(h(\tilde{\theta})|y) \otimes E(h(\tilde{\theta})|y), \quad (3.8)$$
can be expressed as using only expectations of the regularised posterior
\[
\kappa = \mathbb{E}\left[ \Pr(r(\theta) \leq \hat{\omega} | \theta, \lambda)^{-1} | y \right] \left( \mathbb{E}(h(\theta) | y) \odot \mathbb{E} \left[ \frac{h(\theta)}{\Pr(r(\theta) \leq \hat{\omega} | \theta, \lambda)} | y \right] \right)
\]  
(3.9)
using the results in Theorem 3.11, where \( \odot \) denotes element-wise division. The shrinkage factor can be calculated using the posterior draws from \( \pi(\theta, \lambda | y) \) without needing to fit and calculate the expectation from the unregularised model.

**Corollary 3.12** (Shrinkage factor for local stochastic constraints). Assume that \( \mathbb{E}(\hat{\theta} | y) \) exists, then under the conditions of Theorem 3.8, if the model has a local stochastic constraint prior, and the posterior of each \( \hat{\theta}_j \) are independent (the posterior under the base distribution), the shrinkage coefficient of some measurable function \( h(\theta_j) \) can be expressed as
\[
\kappa_{h,i} = \frac{\mathbb{E}(h(\theta_i) | y)}{\mathbb{E}(h(\theta_i))} = \frac{\mathbb{E}(h(\theta_i)) \mathbb{E} \left[ \hat{F}_{\hat{\omega}_i | \lambda} (r_i(\theta_i))^{-1} | y \right]}{\mathbb{E} \left[ h(\theta_i) \hat{F}_{\hat{\omega}_i | \lambda} (r_i(\theta_i))^{-1} | y \right]}
\]
\[
= \left( \frac{\text{Cov} \left[ h(\theta_i), \hat{F}_{\hat{\omega}_i | \lambda} (r_i(\theta_i))^{-1} | y \right]}{\mathbb{E}(h(\theta_i)) \mathbb{E} \left[ \hat{F}_{\hat{\omega}_i | \lambda} (r_i(\theta_i))^{-1} | y \right] + 1} \right)^{-1}.
\]  
(3.10)

A proof is given in Appendix A.1.6. The result also holds conditional on \( \lambda \).

Corollary 3.12 describes the shrinkage for models with independent posterior distributions under the base prior \( f_{\theta} \). For example, the normal-means model, with flat or independent normal prior satisfies this condition, but a linear regression with correlated predictors will not. Due to its limited scope, Corollary 3.12 is mainly useful as an illustration of the shrinkage occurring from stochastic constraints in stylised examples.

Equation (3.10) can also be rewritten using the (generalised) coefficient of variation (ratio of standard deviation to mean) as
\[
\kappa_{h,i} = \left( \hat{\text{cv}}[h(\theta_i)] \Hat{\text{cv}}[\hat{F}_{\hat{\omega}_i | \lambda} (r_i(\theta_i))^{-1}] \hat{\text{cor}}[h(\theta_i), \hat{F}_{\hat{\omega}_i | \lambda} (r_i(\theta_i))^{-1}] + 1 \right)^{-1},
\]
where \( \hat{\text{cv}}[g(\theta_i)] = \sqrt{\mathbb{E}(g(\theta_i)^2 | y) / \mathbb{E}(g(\theta_i) | y)} \) and \( \hat{\text{cor}}[g(\theta_i), m(\theta_i)] = \text{cor}[g(\theta_i), m(\theta_i) | y]. \)

This aids further interpretability. For example, if the noise-to-signal ratio (the CV) is close to zero, i.e. low noise and high signal, then \( \kappa_{h,i} \approx 1 \) whereas if the noise-to-signal ratio is sufficiently high, then the shrinkage depends on the other two terms.

The corresponding result for the horseshoe prior (Carvalho et al.; 2010) in the normal-means models with \( h(\theta_i) = \theta_i, \sigma^2 = 1, \) and \( \tau = 1 \) is
\[
\kappa_i = 1 - k_i \text{ where } k_i \sim \text{Beta}(1/2, 1/2),
\]  
(3.11)
which is relatively simple to derive because \( \mathbb{E}(h(\hat{\theta}_i) | y) = y_i \) in this model. The results of this section extend the applicability of calculating shrinkage coefficients in several
directions. The nature of priors that have a stochastic constraint interpretation is that the base prior generally dictates the marginal information available for that prior. In many cases the base prior is taken to be flat, but when it has a proper density, we argue that the shrinkage coefficients described here are more appropriate than replacing $E(\tilde{\theta}_i \mid y)$ in the denominator with a maximum likelihood estimate. This is because the shrinkage that takes place due to the stochastic constraint is generally the one we would like to know the effect of – whereas the base prior is the starting place.

We also note that the shrinkage factors inspired by the horseshoe result in (3.11) are not applicable to models where the coefficients are correlated. They may be good approximations if there is low correlation but this is not usually the case for high-dimensional analysis. The shrinkage factors discussed also apply to functions of the parameters, not just their individual means. Hence the shrinkage from stochastic constraint priors can be calculated for a wide variety of situations. We explore alternatives in Section 5.3.
Simulating stochastic constraint random variables

Stochastic constraint distributions can be simulated in a number of ways. If the marginal distribution, \( f_\theta \), is a known distribution or is relatively uncomplicated, then it may be efficient to use general samplers such as Hamiltonian Monte Carlo (see Betancourt; 2017, and references therein). In this section we describe simulation strategies for the stochastic constraints that utilise their specific structure. These methods can be more efficient than general samplers for some stochastic constraint distributions, or more appropriate for some situations such as Gibbs samplers. Investigating the potential samplers also illuminate the characteristics of stochastic constraint random variables. For comparison, the algorithms are described in such a way that they produce only a single (approximate) random draw from the target distribution, but can be trivially adapted for multiple draws.

4.1 Rejection samplers

Rejections samplers for stochastic constraints can be an appropriate method for simulating stochastic constraint random variables. The only change to the underlying base and constraining random variable is a truncation of their joint support, hence rejecting samples that fall outside of this new support will result in samples from the target distribution.

The general algorithm for simulating a singe realisation of a stochastic constraint random variable is Algorithm 1.

**Algorithm 1:** Rejection sampler for stochastic constraint variable

**Input:** \( N \), maximum iterations.

\[ t \leftarrow 1 \]
\[ r \leftarrow 1 \]

**while** \( t \leq N \) and \( r = 1 \) **do**

- Sample \( \tilde{\theta} \sim f_\theta \)
- Sample \( \tilde{\omega} \sim f_\omega \)
- **if** all \( r(\tilde{\theta}) \preceq \tilde{\omega} \) **then**
  - \((\theta, \omega) \leftarrow (\tilde{\theta}, \tilde{\omega})\)
  - \( r \leftarrow 0 \)
  - \( t \leftarrow t + 1 \)

**return** \((\theta, \omega)\) if a sample has been accepted \((r = 0)\), otherwise a warning message.

The rejection sampler in Algorithm 1 is not appropriate for improper densities for \( f_\theta \). In this case a Gibbs sampler or latent-variable approach can be used.

The rejection sampler for stochastic constraint variables is likely to be inefficient except in simple cases. The acceptance probability is simply the integral of the
probability distributions, \( f_\theta \) and \( f_\omega \), over the constrained region \( S = \{ (\theta, \omega) \in \Theta \times \Omega : r(\theta) \preceq \omega \} \) given by

\[
\int \int_S f_\theta(\theta)f_\omega(\omega) \, d\theta \, d\omega.
\]

Particularly in the multivariate case, the proportion of samples that fall into \( S \) and are accepted will be low due to the curse of dimensionality. There are also many occasions where a rejection sampler for a univariate SC will be too inefficient for practical use. Hence, alternative simulation strategies are also needed.

### 4.2 Gibbs samplers

Gibbs sampling is a simulation technique popular in Bayesian statistics (Geman and Geman; 1984). It is a Markov chain Monte Carlo (MCMC) technique that constructs a series of samples whose stationary distribution is equivalent to the target distribution under mild conditions. A Gibbs sampler can be generated by converting a multivariate target distribution into several conditional distributions, or by exploiting a latent structure in the target distribution.

Stochastic constraint variables are latent-variable structures by design, and fit neatly into the Gibbs sampling framework. A general algorithm is given in Algorithm 2, which serves as an example as many refinements can be made.

**Algorithm 2:** Gibbs sampler for stochastic constraint variable

**Input:** \( N \), maximum iterations, \( \omega[0] \), initial value.

\[
t \leftarrow 1
\]

\[
\text{while } t \leq N \text{ do}
\]

\[
\text{Sample } \theta[t] \mid \omega[t-1] \sim f_{\theta|\omega[t-1]}(\theta) \propto f_\theta(\theta)1(r(\theta) \preceq \omega[t-1])
\]

\[
\text{Sample } \omega[t] \mid \theta[t] \sim f_{\omega|\theta[t]}(\omega) \propto f_\omega(\omega)1(r(\theta[t]) \preceq \omega)
\]

\[
t \leftarrow t + 1
\]

return \((\theta[N], \omega[N])\)

Algorithm 2 is presented as a general example of Gibbs sampler for stochastic constraint distributions. The conditional distributions in Algorithm 2 may not have a known perfect sampler, in which case additional sampling strategies will be needed.

Several refinements to Algorithm 2 can be made to develop a more efficient sampler. For example, with a multivariate stochastic constraint variable it may be beneficial to extend the Gibbs sampler along the margins, i.e. condition on all of the components of \( \theta \) and \( \omega \). It is also easy to extend the Gibbs sampler to include hyperparameters of the stochastic constraint random variable. In any case, care must be taken with the output as the Markov chain defined by \( \{(\theta[i], \omega[i])\}_{i=1}^N \) may not have reached its
stationary distribution in \( N \) iterations. Unlike the rejection sampler in Algorithm 1, the Gibbs sampler is approximate not exact.

A special case of the Gibbs sampler arises when the constraining random variables are independent exponential distributions. The memoryless property of the exponential distribution results in a conditional distribution that is easy to sample from. The implementation is described in Algorithm 3 where \( \tilde{\omega}_i \sim \text{Exp}(\lambda_i) \).

**Algorithm 3:** Gibbs sampler with exponential constraining variable

**Input:** \( N \), maximum iterations. \( \omega^{[0]} \), initial value.

\[
t \leftarrow 1 \\
\text{while } t \leq N \text{ do} \\
\quad \text{Sample } \theta^{[t]} | \omega^{[t-1]} \sim f_{\theta | \omega^{[t-1]}}(r(\theta) \leq \omega^{[t-1]}) \\
\quad \text{for } i \in \{1, 2, \ldots, q\} \text{ do} \\
\quad \quad \text{Sample } \tau_i \sim \text{Exp}(\lambda_i) \\
\quad \quad \omega^{[t]}_i \leftarrow \tau_i + r_i(\theta^{[t]}) \\
\quad t \leftarrow t + 1 \\
\text{return } (\theta^{[N]}, \omega^{[N]})
\]

Algorithm 3 arises because the conditional density for \( \omega_i | \theta \) is

\[
f_{\omega_i | \theta} \propto \lambda_i \exp(-\lambda_i \omega_i)1(r_i(\theta) \leq \omega_i)
\]

with independent exponential distributions. Moreover, this density is also a shifted exponential distribution due to the left truncation. The truncation resulting in a shifted distribution is unique to the exponential distribution in the class of continuous distributions and due to its memoryless property.

Finding the conditional distributions \( \omega | \theta \) for use in a Gibbs sampler is also instructive for how to generate the constraining random variables if the \( \theta \) have been generated without relying on the latent-structure from the stochastic constraints. If a marginal sample of \( \theta \) has been generated we can simulate the \( \omega \) using the same conditional distribution for \( \omega | \theta \) that we used in the Gibbs sampler. This is useful for some scenarios that will be discussed in Chapter 5.

**Example 4.1** (Gibbs sampler for linear model with generalised Gaussian stochastic constraints prior). Consider a multivariate linear regression with likelihood given as \( \pi(y | X, \theta, \sigma^2) = \mathcal{N}(y; X\theta, \sigma^2I) \). If the prior for \( \theta \) are given by iid generalised Gaussian stochastic constraint priors on \( (\theta, \omega | k, \sigma^2) \) with \( \eta_i = (\sigma^2k_i)^{\psi/\kappa} \) (see Section 2.5.1) then the prior density is

\[
f_{\theta, \omega | k, \sigma^2} = \prod_{i=1}^p \exp\left\{ -\frac{\omega_i}{(\sigma^2k_i)^{\psi/\kappa}} \right\} 1(\{\theta_i|\leq \omega_i \)
\]

where \( \kappa \) and \( \psi \) are fixed and positive. Let \( \sigma^2 \) have prior with density \( f_{\sigma^2} \). When the number of observations \( n \) is greater than the number of coefficients \( p \), and \( X^\top X \) is
invertible, an appropriate Gibbs sampler has full conditional posterior distributions

\[
\begin{align*}
(\theta | \omega, \sigma^2, k) &\sim N(\hat{\theta}, \sigma^2 (X^\top X)^{-1}) \text{ s.t. } |\theta_i|^\kappa \leq \omega_i, \quad 1 \leq i \leq p \\
(\omega_i | \theta, \sigma^2, k) &= \varphi_i + |\theta_i|^\kappa \\
\varphi_i &\sim \text{Exp}(\sigma^2 k_i^{-\psi}) , \quad 1 \leq i \leq p \\
(\sigma^2 | \theta, \omega, k) &\sim f_{\sigma^2|\theta,\omega,k}
\end{align*}
\]

where \( \hat{\theta} = (X^\top X)^{-1} X^\top y \). The hyperparameter \( k \) may be assigned a prior, but for now consider it fixed.

Adjusting Algorithm 3 to include a sampling step for the full conditional distribution of \( \sigma^2 \) will result in a sampler for the generalised Gaussian model.

In Example 4.1, if \( \sigma^2 \) has inverse Gamma prior \( \text{IGam}(a,b) \) then the density \( f_{\sigma^2|\theta,\omega,k} \) is proportional to

\[
f_{\sigma^2|\theta,\omega,k} \propto (\sigma^2)^{-n/2-\psi(1/\kappa+1)-a-1} \exp\left\{ -\frac{b}{2} \left( y - X\theta \right)^\top \left( y - X\theta \right) - \sum_{i=1}^p \omega_i/k_i^\psi \right\}.
\]

When \( \psi = 1 \), the density corresponds to an inverse Gamma distribution

\[
\sigma^2|\theta \sim \text{IGam} \left( n/2 + p\psi(1/\kappa+1) + a, b + \frac{1}{2} \left( y - X\theta \right)^\top \left( y - X\theta \right) + \sum_{i=1}^p \omega_i/k_i^\psi \right).
\]

Another special case occurs when the improper prior \( f_{\sigma^2} \propto \sigma^{-2} \) is used for \( \sigma^2 \). This prior is also covered by setting \( a = b = 0 \) in the density for \( \sigma^2 \).

There are two main difficulties with this Gibbs sampler in Example 4.1. Firstly, simulating from a multivariate truncated normal distribution can be challenging. However, there have been some significant advances in this area, such as Botev (2017) who describes exact simulation for the truncated normal under linear restrictions using minimax tilting. This method is sufficient in many but not all scenarios, so we describe an alternative in Section 4.3. Further discussion is contained in Chapter 7 where we consider scalable Bayesian regression with shrinkage priors.

The second difficulty is simulating from \( f_{\sigma^2|\theta,\omega,k} \) when \( \psi \neq 1 \). In this case the density is non-standard. Rejection sampling using the inverse gamma in (4.1) as an envelope distribution is a possible approach, but the acceptance probability can be too low, which we show in Appendix A.2.2. The conditional density for \( \sigma^2 \) is a natural candidate for slice sampling, which is also discussed in Appendix A.2.2.

### 4.3 Piecewise deterministic Markov processes

Piecewise deterministic Markov processes (PDMP) are a class of continuous time Markov processes that can be used to simulate from complicated target distributions (see Fearnhead et al.; 2018, for an overview). They use state and velocity
components to generate a non-reversible Markov process, where the state coincides with the target distribution. The marginal state Markov process has the correct stationary distribution for the target. Notable examples within this class are the Zig-Zag process (Bierkens et al.; 2017) and the Bouncy Particle Sampler (Bouchard-Côté et al.; 2018). Whilst the realisations of the stochastic process they simulate are continuous, taking a set of points along the continuous path generates samples that can be used like any other MCMC samples. It is also possible to use the entire path to calculate expectations.

PDMPs create a random trajectory over the support of the target distribution by moving in a deterministic (usually linear) manner but changing direction at random times. The Markov process generated is a set of positions and velocities \( \{ \mathbf{x}(t), \mathbf{v}(t) \}_{t \geq 0} \in S \times V \). The stochasticity of the change points are such that the limiting distribution is the target. Whilst several Markov chain Monte Carlo methods are able to simulate stochastic constraints, PDMP are appealing because they easily extend to distributions with restricted support. Bierkens et al. (2018a) describes how to construct PDMPs on restricted domains and produce the correct limiting distributions.

The change points, or events times, in PDMPs are simulated with the first arrival times of inhomogeneous Poisson processes whose rates are derived from the derivative of the (unnormalised) log-density of the target distribution. When the support of the target distribution is restricted, a reflection occurs if the trajectory hits the boundary before an event. The reflection is determined in such a way to ensure that the stationary distribution is not disturbed. A generic description of a PDMP for stochastic constraint variables is given in Algorithm 4.

The number \((k)\) and type of Poisson processes, as well as velocity type and kernel \(Q_i(\mathbf{x}, \mathbf{v})\) varies with each PDMP. We refer to the function updating the velocity as a kernel because it may be a function (point mass distribution), or a non-degenerate probability distribution, either discrete or continuous. The boundary kernel \(Q_B(\mathbf{x}, \mathbf{v})\) may represent a point mass or non-degenerate probability distribution.

The Zig-Zag process is a PDMP that operates by moving around the support of the target distribution with velocity \( \mathbf{v} \in V = \{-1, 1\}^{p+q} \). At the event times of the process, a single element of the velocity is switched by negation. The number of Poisson process to control these switches is \( k = p + q \), the total number of elements in the stochastic constraint variable. In comparison, the Bouncy Particle Sampler has \( \mathbf{v} \in \mathbb{R}^{p+q} \) and uses a single Poisson process where updates in velocity are based on Newtonian elastic collision.

To expand upon Algorithm 4, we start by considering the role and structure of the Poisson process. The arrival times of the Poisson process dictate the changes to velocity which ultimately control the trajectory and samples of the process. The intensity functions are linked to the target distribution by the gradient (partial
Algorithm 4: General PDMP sampler for stochastic constraint variables

Input: $T$, maximum time. $x^{[0]} = (\theta^{[0]}, \omega^{[0]}) \in S$, initial value. $v^{[0]} \in V$, initial velocity. $S = \{(\theta, \omega) \in \Theta \times \Omega : r(\theta) \preceq \omega\}$, support (or boundary $\partial S$).

$t \leftarrow 0$
while $t \leq T$
do
  for $i \in \{1, 2, \ldots, k\}$ do
    Sample $\tau_i \sim$ first arrival Poisson process with intensity $\lambda_i(s; x^{[t]}, v^{[t]})$
    Find next boundary $\tau_B \leftarrow \inf \{\tau > 0 : x^{[t]} + \tau v^{[t]} \notin S\}$
    $i' \leftarrow \arg \min \{\tau_1, \tau_2, \ldots, \tau_k\}$
    $\tau' \leftarrow \min \{\tau_1, \tau_2, \ldots, \tau_k\}$
    if $\tau_B \leq \tau'$ then
      $x^{[t+\tau_B]} \leftarrow x^{[t]} + \tau_B v^{[t]}$
      $v^{[t+\tau_B]} \leftarrow Q_B(x^{[t+\tau_B]}, v^{[t]})$ (sample if kernel non-degenerate)
      $t \leftarrow t + \tau_B$
    else
      $x^{[t+\tau']} \leftarrow x^{[t]} + \tau' v^{[t]}$
      $v^{[t+\tau']} \leftarrow Q_{i'}(x^{[t+\tau']}, v^{[t]})$ (sample if kernel non-degenerate)
      $t \leftarrow t + \tau'$
  return $x^{[T]} = (\theta^{[T]}, \omega^{[T]})$

derivatives) of the negative log-density. For the conical Zig-Zag sampler (Bierkens et al.; 2019) the intensity function of the inhomogeneous Poisson process governing the switching times is given by

$$\lambda_i(s; x, v) = \lambda'_i(x + sv, v), s > 0$$
$$\lambda_i(x, v) = \max \left\{0, v_i \nabla_i u(x)\right\}, 1 \leq i \leq p + q$$

where $u(x) = u(\theta, \omega) = -\log f_{\theta, \omega}$.

whereas the Bouncy Particle sampler (Bouchard-Côté et al.; 2018) has only a single inhomogeneous Poisson process ($k = 1$) with intensity

$$\lambda_1(s; x, v) = \lambda'_1(x + sv, v), s > 0$$
$$\lambda'_1(x, v) = \max \left\{0, v^T \nabla u(x)\right\}$$

but also uses an homogenous Poisson process to refresh the velocity and ensure that the invariant distribution is reducible (Bouchard-Côté et al.; 2018).

The first arrival time of the Poisson processes with intensity $\lambda_i(s; x, v)$ can be simulated by solving

$$z > 0 \text{ s.t. } \int_0^z \lambda_i(s; x^{[t]}, v^{[t]}) \, ds = y \quad (4.2)$$

where $y$ is simulated from an exponential distribution with unit rate, and $(x^{[t]}, v^{[t]})$ are determined in Algorithm 4. It may be possible to solve (4.2) analytically or
numerically, however this may be inefficient. In this case, thinning (Lewis and Shedler; 1979) can be used as an alternative simulation method to the first arrival.

The velocity kernel for the Zig-Zag process changes the sign of a single element of \( v \in \{-1, 1\}^{p+q} \) by

\[
Q_i(x, v) = b_i(v)
\]

where \( b_i(v) \) applies the following rule to the \( j \)th component of \( v \)

\[
[b_i(v)]_j = \begin{cases} 
v_j & \text{if } i \neq j \\
-v_j & \text{if } i = j.
\end{cases}
\]

In comparison the Bouncy Particle sampler uses velocity kernel

\[
Q_1(x, v) = v - 2v^T \nabla u(x) / \| \nabla u(x) \|
\]

with no dependence on the \( i \)th component since all elements of the velocity are updated simultaneously.

The last component of the PDMP to consider for stochastic constraint variables is the boundary kernel, which dictates the change in velocity when the trajectory of the process reaches the boundary. The boundary kernel is chosen so that the invariant distribution is unchanged. Two main conditions governing the boundary kernel are given in Bierkens et al. (2018a) for support restricted to an open, pathwise-connected subset of \( \mathbb{R}^{p+q} \) with Lipschitz boundary. The first condition dictates that the boundary kernel satisfies a detailed balance equation on the boundary. We defer discussing these conditions until they are needed to choose the kernel in Example 4.2.

**Example 4.2** (Zig-Zag sampler for linear model with Lasso stochastic constraints prior). Consider the model described in Example 4.1 with \( \kappa = 1 \) and \( \sigma^2 \) fixed for simplicity. Then we have a double-exponential prior for a multivariate linear model with known variance. The prior has density

\[
f_{\theta, \omega | k, \sigma^2} = \prod_{i=1}^{p} \frac{\exp \left\{ -\frac{\omega_i}{\sigma^2 k_i} \right\} 1(|\theta_i| \leq \omega_i)}{2(\sigma^2 k_i)^{2\psi}}
\]

The gradient vector of the negative log posterior density is

\[
-\nabla \log \pi(\theta, \omega | y, k, \sigma^2) = -\left[ \frac{\partial \log \pi}{\partial \theta} \right] \left[ \frac{\partial \log \pi}{\partial \omega} \right] = \left[ \sigma^{-2} \left( X^T X \theta - X^T y \right) \right]
\]

where the \(-\psi \) exponent is applied element wise. We specifically avoid refactoring the Gaussian likelihood of the posterior into \( \mathcal{N}(\theta; \hat{\theta}, \sigma^2(X^T X)^{-1}) \) as \( X^TX \) may not be invertible – truncating the support by a random amount ensures the overall density is proper.
The switching times for the $\theta$ variable, $\tau_i$ for $1 \leq i \leq p$, can be simulated analytically as they have gradient equivalent to that of a (possibly degenerate) Gaussian log-density. The calculation amounts to solving a quadratic equation and checking the solutions are admissible. We provide the details in Appendix A.2.3.

The elements of the gradient related to $\omega$ are constant, which results in a constant intensity function, and so the resulting Poisson process is homogenous. Therefore the first arrival for the process governing switching times for $\omega_i$ can be simulated by an exponential distribution with rate

$$\eta_i = \begin{cases} v_i (\sigma^2 k_i)^{-\psi}, & \text{if } v_i = 1 \\ 0, & \text{if } v_i = -1. \end{cases}$$

Therefore if $v_i = -1$ no switch will occur in the velocity of the $i$th component of $\omega$ until a boundary is hit.

The last switching events to consider are when the trajectory reaches the boundary of the stochastic constraint variables’ support. The support $S$ is the intersection of $2p$ half-spaces, $S = \bigcap_{1 \leq i \leq p} H_i^{(1)} \cap H_i^{(-1)}$ where

$$H_i^{(1)} = \{ x = (\theta, \omega) \in \mathbb{R}^p \times \mathbb{R}^p : \theta_i - \omega_i \leq 0 \}$$

and

$$H_i^{(-1)} = \{ x = (\theta, \omega) \in \mathbb{R}^p \times \mathbb{R}^p : -\theta_i - \omega_i \leq 0 \}.$$

The outwards unit normal $n(x)$, can be written as

$$n(x) = \begin{cases} n_i^{(1)}, & \text{if } \theta_i - \omega_i = 0 \\ n_i^{(-1)}, & \text{if } \theta_i + \omega_i = 0 \end{cases}$$

for $x = (\theta, \omega) \in \partial S$ (assuming hitting corners has zero probability) where $n_i^{(1)}$ and $n_i^{(-1)}$, the normals for the boundary of each half-space, has elements

$$[n_i^{(h)}]_{j} = 2^{-1/2} \times \begin{cases} h, & \text{if } i = j \\ -1, & \text{if } i + p = j \\ 0, & \text{otherwise} \end{cases}$$

for $1 \leq i, j \leq p$. Let $Q_B(x, v, ds)$ be the boundary kernel to be used for this Zigzag process. Note that previously we suppressed the argument $ds$ which maps the support to probability mass (if discrete) or density (if continuous). The detailed balance equation that needs to be satisfied is

$$Q_B(x, s, dt) \rho(ds) = Q_B(x, t, ds) \rho(dt) \text{ for } x \in \partial S$$

where $\rho(ds)$ is the invariant distribution. Using the boundary kernel

$$Q_B(x, v, ds) = \delta(b_{i+p}(v) - ds) \tag{4.4}$$
where \( \delta(x) \) is the Dirac delta function and \( b_{i+p}(v) \) is the bit negation defined in (4.3) which negates the velocity related to the \( i \)th component of \( \omega \) (i.e. the \( i + p \)th component of \( x = (\theta, \omega) \)). A small amount of algebra shows that kernel (4.4) satisfies the detailed balance equation – since it always hits the boundary at a perpendicular angle (with this constraint) reflecting back in this direction is the simplest way to proceed.

The second condition that must be satisfied can be written as

\[
\sum_{u \in V} (n(x) \cdot u) Q_B(x, v, \, \text{d}u) = -v \cdot n(x) \quad \text{for } (x, v) \in \partial S \times V
\]

which simplifies to

\[
n(x) \cdot b_{i+p}(v) = -v \cdot n(x)
\]

for this particular boundary kernel. Which can be shown to be true on the boundary of each half space defining \( S \) and for every velocity in \( V = \{-1, 1\}^{p+q} \).
Sparsity priors with a global-local structure such as the horseshoe (Carvalho et al.; 2010) or Dirichlet-Laplace (Bhattacharya et al.; 2015) priors, among others, have a particular geometry that has not been explored in the existing literature. This geometry is exposed when the priors are expressed as stochastic constraints, analogously to studying the Lasso as a constrained optimisation rather than a penalised one.

So far we have discussed stochastic constraints as local constraints on parameters which may be related by a shared global shrinkage parameter. However, through further manipulation, some local sparsity priors can be expressed with a single global stochastic constraint and a penalty function on all parameters. In particular, a stochastic constraint variable with an exponential constraint distribution will satisfy this condition.

\section{Exponential stochastic constraints}

Consider the class of local stochastic constraint priors with an exponential constraint variable, conditional on hyperparameters $k$. The hierarchal representation is

\begin{align}
(\theta, \omega \mid k) & \overset{d}{=} \left( \hat{\theta}, \hat{\omega} \mid r_i(\hat{\theta}_i) \leq \hat{\omega}_i, 1 \leq i \leq p, k \right) \\
\hat{\theta} & \sim f_{\hat{\theta}}, \quad \hat{\theta} \in \Theta \\
\hat{\omega}_i \mid k_i & \overset{iid}{\sim} \text{Exp}(k_i^{-1}), \quad 1 \leq i \leq p.
\end{align}

The hyperparameters $k$ are generally also given a prior or may be defined as a function of several random variables, but these are suppressed because we work with the conditional distribution in this chapter. Conditional on $k$, an equivalent global stochastic constraint prior for (5.1) is given in Proposition 5.1.

\textbf{Proposition 5.1.} Let $(\theta, \omega \mid k)$ have a local stochastic constraint prior. If the prior has an exponential constraint variable for $\hat{\omega}$ as in (5.1), then the global stochastic constraint prior defined by

\begin{align}
(\theta, \omega \mid k) & \overset{d}{=} \left( \hat{\theta}, \hat{\omega} \left| \sum_{i=1}^{p} \frac{r_i(\hat{\theta}_i)}{k_i} \leq \hat{\omega}, k \right. \right) \\
\hat{\theta} & \sim f_{\hat{\theta}}, \quad \hat{\theta} \in \Theta \\
\hat{\omega} & \sim \text{Exp}(1)
\end{align}

will have an identical marginal distribution $(\theta \mid k)$ to the local stochastic constraint prior.
Proof. The density function of the marginal, $\theta | k$, can be written as

$$f_{\theta | k} \propto f_{\theta}(\theta) \exp \left( - \sum_{i} r_i(\theta_i) k_i \right)$$

for both (5.1) and (5.2). Hence the hierarchical distributions in (5.1) and (5.2) describe equivalent marginal distributions.

Proposition 5.1 describes the equivalence between local and global representations of stochastic constraint distributions with an exponential constraining variable. For the popular continuous sparsity priors, (5.2) can be expressed as

$$(\theta, \omega | k) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} | \| \tilde{\theta} \otimes k \|_\nu \leq \tilde{\omega}, k)$$

$${\tilde{\theta}} \sim f_{\tilde{\theta}} \propto 1, \quad \tilde{\theta} \in \mathbb{R}^p$$

$$\tilde{\omega} \sim \text{Exp}(1)$$

where $\nu = 2$ (horseshoe prior) or $\nu = 1$ (Dirichlet-Laplace and R2-D2 priors), and the prior on $k$ is chosen appropriately depending on the required distribution. The global representation shows that these priors restrict the parameters to an anisotropic $L^{-\nu}$ ball with asymmetry controlled by the length-scale $k$. For example, in the Dirichlet-Laplace prior $k_i = \phi_i \tau \sigma$ and $\nu = 1$ which implies the $L$-1 ball has radius $\omega \tau \sigma$, conditional on these parameters and $k$. This interpretation is possible since the $\phi$ have a Dirichlet prior and hence its components sum to one. Overall, this shows that the parameters are jointly constrained to the $L$-1 ball, and the conditional radius of this ball is the product of two (not one) global shrinkage parameters and the standard deviation.

We formalise the global form of stochastic constraints for the generalised Gaussian distribution in Example 5.2.

Example 5.2 (Global stochastic constraint representation of generalised Gaussian). Let each parameter $\theta_i$ for $1 \leq i \leq p$ be independent and identically distributed (conditionally on hyperparameters) by the generalised Gaussian distribution in (2.18). Assume $k_i = \eta^c$ for the scale hyperparameter in the same equation. The density for $\theta_i$ with a stochastic constraint decomposition (see Section 2.5.1) can then be written as

$$f_{\theta_i, k_i} = \frac{1}{2 \Gamma(1 + \nu)} k_i^{1+\nu} \left( \frac{|\theta_i|^\nu \leq \omega_i}{\omega_i} \right) \exp \left( -\frac{\omega_i}{k_i} \right)$$

(5.3)

jointly with $\omega_i$ and conditional on $k_i$ with fixed $\nu$ for $1 \leq i \leq p$. Let $k_i$ be a function of some hyperparameters say $k_i = k(\lambda_i, \tau, \sigma^2)$ where we generally give $\lambda_i, \tau, \sigma$ priors.

The prior structure in (5.3) includes popular sparsity priors such as the horseshoe, Dirichlet-Laplace and R2-D2 priors. From Proposition 5.1, after normalising, the
5.1. EXPONENTIAL STOCHASTIC CONSTRAINTS

Local stochastic constraint prior in (5.3) can be expressed as

\[
f_{\theta, \omega|k} = \frac{1}{2^p \Gamma(1 + \nu)^p} \prod_{i=1}^{p} k_i^{1/\nu} \exp\left(-\omega \sum_{i=1}^{p} k_i^{-1} |\theta_i|^\nu \leq \omega \right) \exp(-\omega).
\]  

(5.4)

A global stochastic constraint prior can only be represented by a local equivalent if the penalty function is separable. The result appears trivial when considering the refactoring of the marginal distribution, with \( f_{\tilde{\theta}}(\theta) \propto 1 \), as

\[
f_{\theta|k} \propto \exp\left( -\sum_i r_i(\theta) \right) = \prod_i \exp\left( -\frac{r_i(\theta)}{k_i} \right),
\]

but it is interesting that it holds for any permissible base prior \( f_{\tilde{\theta}} \), whether the elements of \( \tilde{\theta} \) are correlated or independent.

The exponential distribution as the constraint random variable is unique in this sense, as it is the only random variable that exhibits this local to global equivalence. This is a direct result of the memoryless property of the exponential distribution, resulting in Theorem 5.3.

**Theorem 5.3** (Equivalence between global and local stochastic constraints). For any global stochastic constraint prior with separable penalty function \( r(\theta; k) = \sum_{i=1}^{p} q_i(\theta_i; k) \geq 0 \) where all \( q_i(\theta_i) \geq 0 \), there exists an equivalent local stochastic constraint prior representation (equivalent marginal distributions) with identical constraint distribution if and only if the constraint variables can be expressed as exponential random variables (conditional on hyperparameters).

**Proof.** A proof is given in Appendix A.1.7.

The dual representation of global and local stochastic constraints with exponential constraining distributions allows us to consider popular sparsity priors in the literature as global stochastic constraints. This interpretation is useful for sparsity priors because it provides an intuitive understanding of the geometry behind these objects. This is very similar to the role the constraint formulation of regularisation in optimisation-based inference plays in understanding the geometry.

The special role played by the exponential distribution is important to consider. When a stochastic constraint has a global interpretation we can infer a tradeoff between magnitudes of parameter values. It guarantees that some parameters will need to be shrunk, and if some values are large, it necessitates that others are small. Local stochastic constraint priors with exponential constraining variables can also be expressed as global stochastic constraints and therefore possess the global property. But, because they also have a local representation, posterior sampling is more likely to be manageable. Overall, using exponential constraining variables benefits shrinkage and provide computational tractability.
5.2 Stochastic net priors

In light of the geometrical aspects revealed for stochastic constraint variables with exponential constraining distributions in Section 5.1, we now define the class of stochastic net priors, which encompass a large number of continuous sparsity priors in the literature. They have been implicitly discussed throughout this thesis; here we give them a formal definition.

**Definition 5.4** (Stochastic net priors). For \( \nu > 0 \), let the stochastic net prior be defined by the following structure

\[
(\theta, \omega \mid \tau, \lambda) \overset{d}{=} \left( \tilde{\theta}, \tilde{\omega} \mid \|\tilde{\theta} \otimes \lambda\|_{\nu} \leq \tau \tilde{\omega}, \tau, \lambda \right)
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}}, \quad \tilde{\theta} \in \mathbb{R}^{p} \\
\lambda \overset{\text{iid}}{\sim} f_{\lambda}, \quad \lambda \in \mathbb{R}^{p}_{+} \\
\tau \sim f_{\tau}, \quad \tau \in \mathbb{R}_{+} \\
\tilde{\omega} \sim \text{Exp}(1).
\]

The name is inspired by elastic net regularisation (Zou and Hastie; 2005), and describes how the shrinkage is enacted through the constraint. Specifically, the constraint

\[\|\tilde{\theta} \otimes \lambda\|_{\nu} \leq \tau \tilde{\omega},\]

describes a region for \( \tilde{\theta} \) which is a flexible and adaptive net in the sense described by Zou and Hastie (2005). However, in this case the parameters that control the net have prior distributions, and are hence stochastic.

Each component of the prior (5.5) serves a particular role:

- The base prior \( f_{\tilde{\theta}} \) incorporates prior information about the coefficients before joint shrinkage is induced. This may be marginal prior information, but a jointly defined prior is also possible.

- The \( \lambda_i \) define the relative length-scale of each \( \theta_i \) (i.e. \( |\theta_i/\lambda_i|^{\nu} \leq \tau \tilde{\omega} \) if \( \theta_j = 0 \) for \( i \neq j \)) and hence determines the amount of differential shrinkage on \( \theta_i \).

- The global shrinkage parameter, \( \tau \), controls how large the scaled norm of \( \theta \) can grow. It is often dependent on the model error \( \sigma^2 \).

- The stochastic constraint variable, \( \tilde{\omega} \), is used to define the underlying mechanism of the random truncation. Choosing an exponential distribution allows relatively easy sampling from the local stochastic constraint representation described in Section 5.1.

Stochastic net priors were implicitly discussed in Section 5.1 without parametrisation of the hyperparameters \( \lambda_i \) and \( \tau \), and they encompass many popular continuous
shrinkage priors. For example, the horseshoe and regularised horseshoe prior can be viewed as stochastic constraint priors (see Sections 2.5.2 and 2.5.3), and stochastic nets. In particular, the regularised horseshoe prior is found by replacing the flat base distribution for $\tilde{\theta}$ in the SC horseshoe with iid normal distributions. The Bayesian Lasso, Bayesian elastic net, horseshoe+, Dirichlet-Laplace, and R2-D2 priors, among others, all belong to this class.

5.3 Eliciting sparsity uncertainty from geometry of stochastic constraints

Whilst continuous shrinkage priors have many desirable properties, they do not emit a formal decision rule or probability that can inform model selection. That is, unlike their discrete counterparts (for example spike and slab priors, see Ishwaran and Rao; 2005, for overview), no direct posterior inclusion probability is possible (the posterior probability the parameter is nonzero). Instead, heuristic strategies have been suggested to address this, which we outline below before exploring the additional insights afforded by the stochastic constraint representation of continuous sparsity priors.

5.3.1 Thresholding for the horseshoe prior

Thresholding is a generic term which is used to describe a variety of decision rules that use a cut-off point (hard-thresholding) or shrinkage combined with cut-off point (soft-thresholding) to obtain exact zeros from estimators. For posterior distributions with continuous shrinkage priors, a popular thresholding rule was suggested by Carvalho et al. (2010) along with the horseshoe prior.

The thresholding rule in Carvalho et al. (2010) (we will refer to it simply as thresholding for the remainder of this section) is derived under the normal-means model $y_i \sim N(\theta_i, \sigma^2)$ for $1 \leq i \leq n$. Using the horseshoe as prior distribution, the posterior expectation of $\theta_i$ is

$$E(\theta_i \mid y) = (1 - \hat{\kappa}_i)y_i$$

(5.6)

where $\hat{\kappa}_i$ is the posterior expectation of a function of the hyperparameter(s). The thresholding rule is motivated by the following comparison. If the normal-means model were fit with a spike and slab prior having a sufficiently heavy-tailed slab distribution, then the resulting posterior expectation is approximately

$$E(\theta_i \mid y) \approx \hat{w}_iy_i$$

(5.7)

where $\hat{w}_i$ is the posterior inclusion probability, i.e. $Pr(\theta_i \neq 0 \mid y)$. Comparing (5.6) to (5.7) leads to the decision rule naming $\theta_i$ a signal when $(1 - \hat{\kappa}_i) \geq 0.5$ and noise otherwise. Or in other words the thresholding function for the expected value of the coefficient $\hat{\theta}_i$ is

$$\varsigma(\hat{\theta}_i, \hat{\kappa}_i) = \begin{cases} \hat{\theta}_i, & \text{if } 1 - \hat{\kappa}_i \geq 0.5 \\ 0, & \text{if } 1 - \hat{\kappa}_i < 0.5. \end{cases}$$
Datta et al. (2013) show that thresholding with the horseshoe prior has desirable asymptotic optimality properties, whilst Bhadra et al. (2017) extended the rule for the horseshoe+ prior. Although popular and useful in multiple testing scenarios, thresholding has a number of shortcomings. Firstly, it does not extend to linear models since the correlation from the $X^\top X$ matrix interferes with the simplicity of $\mathbb{E}(\theta_i \mid y)$ in (5.6). Secondly, the comparison used to motivate the result holds for the horseshoe prior, not necessarily other continuous shrinkage priors. Thirdly, their applicability is limited to Gaussian likelihoods and it is difficult to see if they could be extended usefully beyond this.

5.3.2 Penalised credible regions Bondell and Reich (2012) propose penalised credible regions for variable selection in high dimensions. Their method attempts to coerce the posterior distribution for a set of coefficients from a continuous shrinkage prior into a sparse summary. They begin by considering the following optimisation problem

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^p} \|\beta\|_0 \text{ s.t. } \beta \in C_\alpha,$$  

(5.8)

where $C_\alpha$ is the $(1 - \alpha) \times 100\%$ posterior credible region based on the chosen prior distribution. When applied to global-local shrinkage priors, their method uses the estimated posterior mean $\hat{\beta}$ and covariance matrix $\hat{\Sigma}$ of the regression coefficients (Zhang and Bondell; 2018) with a relaxed (and approximated) version of the optimisation in (5.8), given as

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^p} (\beta - \hat{\beta})^\top \hat{\Sigma}^{-1} (\beta - \hat{\beta}) + \lambda_\alpha \|\beta \odot \hat{\beta}\|_1$$  

(5.9)

which still results in sparse solutions due to the $L_1$ penalty. By varying the level of $\alpha$, or equivalently $\lambda_\alpha$, a sequence of sparse solutions can be generated, just like the Lasso. The penalty parameter $\lambda_\alpha$ can then be chosen by cross-validation.

Unlike thresholding, the penalised credible regions method is designed for linear models, not just the normal-means model. The recent combination of this method using the Dirichlet-Laplace prior was shown to have good selection consistency properties in Zhang and Bondell (2018). In particular, this combination has selection consistency for $p$ increasing in dimension as $p = o(n)$, whereas Bondell and Reich (2012) showed that this was not true for joint selection under the normal prior they used.

5.3.3 Contributions from stochastic constraints Inspired by the previous work of Bondell and Reich (2012) and Zhang and Bondell (2018), we use the geometric properties of continuous shrinkage priors to explore a new method of variable selection from posterior samples of continuous shrinkage priors. We propose a new post-estimation procedure for extracting this sparsity information from priors of the form in (5.4) when viewed as global stochastic constraints (Section 5.1). We investigate our method using high-dimensional multivariate Gaussian model models.
5.3. ELICITING SPARSITY UNCERTAINTY FROM GEOMETRY OF STOCHASTIC CONSTRAINTS

From one perspective, sparsity is already encoded in global stochastic constraints priors. To see this, consider the model with stochastic constraints as

\[ (y \mid \beta) \sim N(\alpha 1 + X\beta, \sigma^2 I) \]

\[ (\beta, \omega \mid k) \sim f_{\beta, \omega \mid k} \propto 1 \left( \sum_{i=1}^{p} k_{i}^{-1} |\beta_i|^\nu \leq \omega \right) \exp(-\omega) \]

(5.10)

where \( k_i = k(\lambda_i, \tau, \sigma^2) \), and \( \lambda_i, \tau \) and \( \sigma^2 \) are given priors to make the model equivalent to the chosen shrinkage prior, e.g. the Dirichlet-Laplace prior. We separate the intercept coefficient \( \alpha \) to distinguish that it should not be included in multivariate shrinkage. As such the intercept, \( \alpha \), is suppressed in the remainder of this section. Conditional on hyperparameters contained in \( k \), and stochastic constraint variable \( \omega \), the posterior distribution of (5.10) is proportional to

\[ \pi(\beta \mid y, \omega, k) \propto \exp\left\{ -\frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \right\} \text{ s.t. } \sum_{i=1}^{p} k_{i}^{-1} |\beta_i|^\nu \leq \omega, \]

(5.11)

a truncated multivariate normal distribution. The matrix \( X^\top X \) will be singular when \( p > n \), but the resulting distribution is proper on the restricted support. We can leverage this representation for sparsity by drawing our attention to the mode of this conditional distribution. The conditional posterior mode, as opposed to the (conditional) posterior mean, of (5.11) may have sparse solutions if \( \nu \leq 1 \). When \( \nu = 1 \), the optimisation in (5.11) is equivalent to Lasso with differential shrinkage on the coefficients in \( \beta \). If the stochastic constraint representation of the prior uses \( \nu \leq 1 \) then the mode of (5.11) will be sparse so long as their is enough shrinkage induced by the prior (in the same way that the Lasso will not shrink if the penalty parameter is insufficient to induce zero coefficients). The shrinkage induced by global-local stochastic constraints, such as the Dirichlet-Laplace prior, are generally enough to induce this behaviour, as we show in Section 5.3.4.

We propose the following procedure, dubbed stochastic constraint variable selection (SCVS), for performing variable selection using posterior samples from (5.10).

**Procedure 5.5** (Stochastic constraint variable selection).

1. Run a sampler for chosen hierarchal model with global stochastic constraint prior to generate posterior samples \( \{\beta^{[t]}, \omega^{[t]}, \lambda^{[t]}, \tau^{[t]}, (\sigma^2)^{[t]}\}_{t=1}^{N} \). If the constraint variable \( (\omega) \) is not generated by the sampler use its full conditional distribution to sample (see Section 4.2 for example).

2. For each sample \( (\beta^{[t]}, \omega^{[t]}, (\sigma^2)^{[t]}) \) find mode of (5.11) and record which \( \beta \) are non-zero in the solution.

3. Calculate the proportion of times a given parameter is nonzero in Step 2. This is the estimated posterior pseudo-inclusion probability (PIP). Correlations between inclusions are also possible to calculate.
4. If a binary decision is required, select those parameters who have PIP greater than some value $z_p \in (0, 1)$, typically $z_p \geq 0.5$ but can be chosen by cross-validation.

To solve (5.11) for step 2 we consider the case where $\nu = 1$. As it is in many cases, if $\nu = 1$ then the mode of (5.11) is the Lasso optimisation problem in constraint formulation with differential weights $k_i$ in the constraint. This optimisation can be written as

$$
\arg\min_{\beta \in \mathbb{R}^p} (y - X\beta)^\top (y - X\beta) \text{ s.t. } \sum_{i=1}^{p} k_i^{-1} |\beta_i| \leq \omega,
$$

but is also equivalent to

$$
\arg\min_{\beta' \in \mathbb{R}^p} (y - X'\beta')^\top (y - X'\beta') \text{ s.t. } \|\beta'\|_1 \leq \omega
$$

where $\beta' = K^{-1}\beta$, $X' = XK$.

Most algorithms for the Lasso use implementations in terms of the penalty function. Whilst there is a direct connection between the penalty tuning parameter, say $\lambda$, and $\omega$; it is data dependent. It requires computation of a solution-path to determine the equivalent $\lambda$ from $\omega$. This is unnecessary as we can use the algorithm detailed in Osborne et al. (2000), who exploit the constrained formulation to find solutions to (5.13). The algorithm is provided in the R package lassoR (Lokhorst et al.; 2018) with C code underlying the implementation.

Procedure 5.5 can be viewed as an amalgamation of a differential Lasso and a full Bayesian model. From the Lasso perspective, the individual penalty parameters have a random distribution, which we utilise to calculate a distribution for which variables are non-zero (in some sense at least, they are not true posterior inclusion probabilities). Whereas other model selection techniques generally use cross-validation for example, here the posterior distribution of $\omega$ serves this purpose. From a Bayesian perspective, the procedure estimates the posterior distribution of the optima of (5.12) which we use in our decision rule.

At least superficially, it appears that Procedure 5.5 is closer in terms of approximation to the true model than the posterior credible region method. We present some simulations to compare the two algorithms in Section 5.3.4. However, proving selection consistency for the SCVS method is still under investigation. Some benefits of SCVS over the aforementioned competitors are:

- It can handle both normal-means and linear regression (unlike thresholding),
- There is no need for post-hoc tuning or cross-validation, as this is controlled by the prior within the Bayesian inference itself (unlike PCR),
- It can easily adapt when the base distribution in the stochastic constraint is not flat, and
• It may be possible to extend to non-Gaussian models where the Lasso penalty is able to select sparse solutions.

Whilst the pseudo-inclusion probabilities are not true posterior inclusion probabilities, they are true probabilities in the sense that they estimate the probability of a binary event indicator.

5.3.4 Pen-CR and SCVS variable selection comparison To compare the penalised credible region (Pen-CR) and stochastic constraint variable selection (SCVS) variable selection methods we fit linear regression models with the Dirichlet-Laplace prior. The tuning hyperparameter for the Dirichlet-Laplace prior is chosen by the $R^2$ method described in Zhang and Bondell (2018), and the prior on $\sigma^2$ is improper with density $f_{\sigma^2} \propto \sigma^{-2}$. The two variable selection methods are then applied to the posterior draws from the model. The posterior is approximated with 5000 draws sampled from a Gibbs sampler (Bhattacharya et al.; 2015; Zhang and Bondell; 2018) with an additional 5000 burn-in samples.

We perform two simulation studies across a range of settings. In the first simulation study (Simulation I) we use similar settings to Zhang and Bondell (2018). The observations, $y$, are simulated with an $n$-dimensional multivariate normal distribution given by

$$y \sim N(X\beta, I_n)$$

where $X = [x_1, x_2 \cdots x_p]^\top$, $n = 60$, and $p$ is chosen from $p \in \{50, 500, 1000\}$. The rows, $x_i$, of the design matrix, $X$, are simulated by drawing from a zero-mean multivariate normal distribution with autocorrelation $\rho_1$ chosen from $\rho_1 \in \{0.5, 0.9\}$ such that any row $i$ has elements with correlation $\text{cor}(x_{ij}, x_{ik}) \in \rho_1^{j-k}$. The sparsity of the true $\beta$ is controlled by $\beta^* = [0_{10}^\top, b_1^\top, 0_{30}^\top, b_2^\top, 0_{p-50}^\top]^\top$ where the nonzero elements $b_1$ and $b_2$, each have five elements that are simulated from independent uniform distributions on $(0, 1)$.

The second simulation study (Simulation II) is based on Zhang et al. (2017) and has the following differences to Simulation I. The number of parameters is chosen from $p \in \{50, 100, 250\}$, and the each row of the design matrix, $x_i$, is simulated from a zero-mean multivariate normal distribution with variance-covariance matrix $\Sigma$ have diagonal elements $\Sigma_{ii} = 1$ for $1 \leq i \leq p$ and off-diagonal elements equal to $\Sigma_{ij} = \rho_2$ for $1 \leq i, j \leq p$ and $i \neq j$ where $\rho_2 \in \{0.25, 0.5, 0.75, 0.9\}$. The nonzero elements of $\beta^*$ are simulated independently by zero-mean student-t distributions with degree of freedom three. Each setting for simulations I and II are replicated 250 times.

The performance of the SCVS procedure is assessed using the sum squared error of the posterior mean estimate of $\beta$, computed by $\text{SSE} = (\hat{\beta} - \beta^*)^\top (\hat{\beta} - \beta^*)$, where $\hat{\beta}$ is the posterior mean estimate of $\beta$ from the Gibbs sampler, and $\beta^*$ is the true value of $\beta$ used to simulate the data. The SSE for Simulation I and II are given in Table 5.1 and Table 5.3, respectively. These summaries are calculated using a rule to choose a singular estimate from the penalised credible regions (Pen-CR) method.
Table 5.1: Simulation I. Sum of squared errors (SSE) (mean with standard deviation in parenthesis) for penalised credible regions (Pen-CR) and stochastic constraints variable selection (SCVS) methods of variable selection. Simulation replications are for sample size $n$, parameter size $p$, design matrix correlation $\rho_1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>$\rho_1$</th>
<th>Pen-CR</th>
<th>SCVS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>CV-$\lambda_\alpha$ prop $\geq 0.5$</td>
<td>prop $\geq 0.5$</td>
</tr>
<tr>
<td>50</td>
<td>0.25</td>
<td>0.74 (0.46)</td>
<td>0.61 (0.37)</td>
<td>0.54 (0.27)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.97 (0.59)</td>
<td>0.78 (0.46)</td>
<td>0.63 (0.34)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>2.00 (1.25)</td>
<td>1.38 (0.84)</td>
<td>1.07 (0.62)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>3.52 (1.90)</td>
<td>2.14 (1.09)</td>
<td>1.69 (0.89)</td>
</tr>
<tr>
<td>60</td>
<td>0.25</td>
<td>1.13 (0.68)</td>
<td>1.12 (0.59)</td>
<td>1.21 (0.65)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.52 (0.93)</td>
<td>1.30 (0.74)</td>
<td>1.36 (0.75)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>2.72 (1.70)</td>
<td>1.96 (1.16)</td>
<td>2.01 (1.10)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>5.07 (2.62)</td>
<td>3.04 (1.56)</td>
<td>3.14 (1.53)</td>
</tr>
<tr>
<td>1000</td>
<td>0.25</td>
<td>1.61 (0.92)</td>
<td>1.71 (0.85)</td>
<td>2.07 (1.06)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>1.69 (1.01)</td>
<td>1.48 (0.81)</td>
<td>1.77 (0.89)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>2.83 (1.62)</td>
<td>1.97 (1.11)</td>
<td>2.34 (1.08)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>5.10 (2.80)</td>
<td>2.83 (1.57)</td>
<td>3.17 (1.52)</td>
</tr>
</tbody>
</table>

and stochastic constraints variable selection method respectively. For Pen-CR we show the results of two rules. The first chooses the parameter $\lambda_\alpha$ on the solution path of the Lasso in (5.9) by cross-validation. The sparse estimate returned is the corresponding Lasso solution, under column “CV-$\lambda_\alpha$”. The second rule finds the proportion of times a coefficient is nonzero across the solution path. The estimate returned is the posterior mean where coefficients with nonzero proportion less than half set to zero (“prop $\geq 0.5$”). The rule for the SCVS method calculates the proportion of times an element is nonzero from (5.12) over the posterior distribution. The estimate returned is the posterior mean where coefficients with nonzero proportion less than half are set to zero (“prop $\geq 0.5$”).

The receiver operating characteristic (ROC) and precision-recall (PRC) area under the curve (AUC) summaries are presented for simulation I and II in Table 5.2 and Table 5.4 respectively. The ROC curve is the curve defined by plotting the true positive rate (y-axis) against the false positive rate (x-axis) along the tuning parameter of the method. The tuning parameter for the Pen-CR method is the Lasso penalty parameter $\lambda_\alpha$, whilst the parameter for the SCVS method is the $z_p$ given in Procedure 5.5 as the cutoff for the proportion of times each coefficient is nonzero from the posterior samples. The PRC curve is found by plotting the true positive rate (or recall, x-axis) against the precision (y-axis). The AUC for this plot gives a measure of the trade-off between precision and sensitivity.

The SSE summaries for Simulation I in Table 5.1 show a mixed comparison between stochastic constraint variable selection and penalised credible regions. In the $p = 50$ setting SCVS outperforms both rules for Pen-CR and the relative performance
5.3. ELICITING SPARSITY UNCERTAINTY FROM GEOMETRY OF STOCHASTIC CONSTRAINTS

Table 5.2: Simulation I. Receiver operating characteristic (ROC) and precision-recall (PRC) area under the curve (AUC) values (mean with standard deviation in parenthesis) for penalised credible regions (Pen-CR) and stochastic constraints variable selection (SCVS) methods of variable selection. Simulation replications are for sample size $n$, parameter size $p$, design matrix correlation $\rho_1$.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>ROC-AUC</th>
<th>PRC-AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Pen-CR</td>
<td>SCVS</td>
</tr>
<tr>
<td>50</td>
<td>0.25</td>
<td>0.87 (0.07)</td>
<td>0.88 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.87 (0.08)</td>
<td>0.88 (0.08)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.86 (0.08)</td>
<td>0.87 (0.08)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.85 (0.09)</td>
<td>0.85 (0.08)</td>
</tr>
<tr>
<td>60</td>
<td>500</td>
<td>0.84 (0.07)</td>
<td>0.87 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.85 (0.08)</td>
<td>0.90 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.84 (0.08)</td>
<td>0.90 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.78 (0.10)</td>
<td>0.89 (0.08)</td>
</tr>
<tr>
<td>1000</td>
<td>0.25</td>
<td>0.81 (0.08)</td>
<td>0.85 (0.08)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.84 (0.09)</td>
<td>0.91 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.84 (0.10)</td>
<td>0.94 (0.06)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.80 (0.10)</td>
<td>0.92 (0.06)</td>
</tr>
</tbody>
</table>

 increases marginally as the autocorrelation $\rho_1$ increases. When $p = 500$ and $p = 1000$ “prop $\geq 0.5$” rule for the Pen-CR method outperforms the other two rules. And the SCVS rule only performs the same or better against the cross-validation Pen-CR rule for $\rho_1 \geq 0.5$. Whilst this suggests the Pen-CR “prop $\geq 0.5$” rule should be favoured based on the SSE, it is difficult to theoretically justify its usage as it simply chooses parameters that were nonzero on 50% or more occasions along the Lasso solution path.

The AUC results for Simulation I reveal a relatively even comparison between the Pen-CR and SCVS ROC metric for $p = 50$, and a modest decline in the mean precision-recall metric for stochastic constraints method compared to Pen-CR. For $p = 500$ and $p = 1000$, the SCVS method begins to outperform the Pen-CR in terms of the ROC metric, especially for $\rho_1 \geq 0.75$, but we note the high standard errors which suggests variability in this assessment. The PRC-AUC value remains lower for the SCVS method, but the two values are noticeably closer for $p = 500$ and $p = 1000$, as well as when the correlation is higher. In general, the SCVS performs better over the ROC, whilst the Pen-CR performs better on PRC. The tradeoff between ROC and PRC indicates that the SCVS method is better able to identify small nonzero values at the expense of occasionally misidentifying zero as nonzero. Whilst the Pen-CR method too often sets some nonzero parameters to zero.

The second simulation has higher overall correlation within the design matrix and nonzero parameters drawn from a more variable distribution. The SCVS method generally performs better than the Pen-CR in terms of mean SSE for all dimensions
Table 5.3: Simulation II. Sum of squared errors (SSE) (mean with standard deviation in parenthesis) for penalised credible regions (Pen-CR) and stochastic constraints variable selection (SCVS) methods of variable selection. Simulation replications are for sample size \( n \), parameter size \( p \), design matrix correlation \( \rho^2 \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p )</th>
<th>( \rho^2 )</th>
<th>Pen-CR</th>
<th>SCVS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>( CV-\lambda_\alpha )</td>
<td>prop ( \geq 0.5 )</td>
</tr>
<tr>
<td>50</td>
<td>0.25</td>
<td>3.82 (4.68)</td>
<td>3.30 (5.75)</td>
<td>1.26 (1.25)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>4.01 (3.78)</td>
<td>3.29 (3.31)</td>
<td>1.67 (1.19)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>5.28 (7.08)</td>
<td>4.41 (6.25)</td>
<td>3.18 (5.83)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>6.96 (5.63)</td>
<td>6.48 (5.23)</td>
<td>5.61 (4.43)</td>
</tr>
<tr>
<td>60</td>
<td>0.25</td>
<td>3.45 (3.29)</td>
<td>2.94 (3.00)</td>
<td>1.35 (1.05)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>3.73 (3.47)</td>
<td>3.15 (3.04)</td>
<td>1.98 (1.64)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>5.02 (3.95)</td>
<td>4.72 (3.97)</td>
<td>3.69 (4.14)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>7.57 (5.11)</td>
<td>7.54 (4.86)</td>
<td>6.87 (3.92)</td>
</tr>
<tr>
<td>250</td>
<td>0.25</td>
<td>3.33 (3.03)</td>
<td>3.01 (2.91)</td>
<td>2.08 (2.10)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>4.37 (4.30)</td>
<td>4.06 (3.92)</td>
<td>3.20 (3.05)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>5.41 (4.10)</td>
<td>5.32 (3.98)</td>
<td>4.88 (3.37)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>9.50 (7.91)</td>
<td>9.78 (7.72)</td>
<td>9.65 (7.13)</td>
</tr>
</tbody>
</table>

Table 5.4: Simulation II. Receiver operating characteristic (ROC) and precision-recall (PRC) area under the curve (AUC) values (mean with standard deviation in parenthesis) for penalised credible regions (Pen-CR) and stochastic constraints variable selection (SCVS) methods of variable selection. Simulation replications are for sample size \( n \), parameter size \( p \), design matrix correlation \( \rho^2 \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( p )</th>
<th>( \rho^2 )</th>
<th>ROC-AUC</th>
<th>PRC-AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Pen-CR</td>
<td>SCVS</td>
</tr>
<tr>
<td>50</td>
<td>0.25</td>
<td>0.81 (0.11)</td>
<td>0.89 (0.08)</td>
<td>0.71 (0.10)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.82 (0.10)</td>
<td>0.87 (0.08)</td>
<td>0.69 (0.10)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.79 (0.10)</td>
<td>0.82 (0.08)</td>
<td>0.62 (0.10)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.74 (0.10)</td>
<td>0.76 (0.09)</td>
<td>0.51 (0.12)</td>
</tr>
<tr>
<td>60</td>
<td>0.25</td>
<td>0.81 (0.10)</td>
<td>0.88 (0.08)</td>
<td>0.67 (0.11)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.80 (0.10)</td>
<td>0.85 (0.08)</td>
<td>0.62 (0.11)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.76 (0.10)</td>
<td>0.80 (0.08)</td>
<td>0.52 (0.14)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.71 (0.10)</td>
<td>0.74 (0.10)</td>
<td>0.39 (0.14)</td>
</tr>
<tr>
<td>250</td>
<td>0.25</td>
<td>0.79 (0.10)</td>
<td>0.85 (0.08)</td>
<td>0.60 (0.12)</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.78 (0.10)</td>
<td>0.83 (0.08)</td>
<td>0.55 (0.13)</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.75 (0.10)</td>
<td>0.78 (0.09)</td>
<td>0.43 (0.14)</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0.70 (0.09)</td>
<td>0.72 (0.10)</td>
<td>0.30 (0.14)</td>
</tr>
</tbody>
</table>
simulated, $p = \{50, 100, 250\}$. However, the high standard deviation for both methods makes the comparison unreliable for simulations where $p = 250$ or $\rho_2 \geq 0.75$.

The AUC metrics in Simulation II demonstrate the tradeoff between ROC and PRC when changing from the Pen-CR to the SCVS method of variable selection, much like the first simulation. Whilst the losses in PRC AUC are similar to Simulation I, the gains to the ROC are tempered by the higher overall design matrix correlation.

Overall, the simulated comparison between stochastic constraints variable selection and penalised credible regions is mixed. One conclusion is apparent; SCVS includes more nonzero coefficients, whilst Pen-CR chooses less. The result being that SCVS under-shrinks too often and Pen-CR over-shrinks too often. It is also important to note that the use of the “prop $\geq 0.5$” cutoff in the SSE evaluations is an arbitrary one as neither the Pen-CR or SCVS versions are inclusion probabilities. This means that cross-validation could be done to select the cutoff for the SCVS method which may improve the mean SSE values. However, we argue that the benefit of the SCVS method is its descriptiveness. For example a comparison between two outputs of the SCVS and Pen-CR method form Simulation I.
CHAPTER 6

Using stochastic constraints to construct priors

Stochastic constraints provide a flexible framework for constructing priors for Bayesian analysis. The separation of marginal and joint parameter prior information is a natural setup, and constructing such priors with a latent variable constraint gives an intuitive description of the prior. This separation principle is similar in spirit to that of Kessler et al. (2015). We first discuss the elements of stochastic constraint distributions that are useful for prior construction, then provide two illustrative examples to highlight some of these key features.

The structure underlying stochastic constraint variables is intuitive and aids in flexible prior construction. Starting from a base prior $f_{\tilde{\theta}}$, to incorporate (further) shrinkage into the prior, constrain the support of the parameter — and to include uncertainty about the constraint, use a random variable. This can be particularly useful for constructing multivariate priors. For example, the base prior can be independent for each $\tilde{\theta}_i$, but we can apply shrinkage to a function of all the parameters, using $r(\theta)$, to ensure sparsity or some other regularisation goal.

We can also view stochastic constraints as an operator which acts on a base distribution. For example, perhaps a suitable base prior is known for each parameter in a model and we would like to formulate a joint prior which uses these. If we wish to add information regarding the sparsity of the parameter vector, then adding a stochastic constraint to the joint prior distribution can incorporate this information, whilst respecting the original base prior.

The base distribution can be chosen to control the characteristics of the prior before shrinkage. This may include the shape of the distribution, the variable’s expected magnitude, or perhaps correlation between the variables. The constraining variable is also free to be chosen although, as mentioned in Chapter 5, using an exponential distribution does have desirable properties. We emphasise that the ability to adapt SC priors does not reduce the interpretability of their structure. This is an appealing quality for a framework used to construct priors for Bayesian inference.

Many scenarios call for shrinkage of parameter estimates towards zero, incorporating assumptions about sparsity or regularising ill-posed problems. However, regularisation and shrinkage can take more complex forms. For example, shrinkage towards subspaces and bounded subsets has been explored previously (Oman; 1982; George; 1986b,a) and there has also been recent work on Bayesian shrinkage towards subspaces using a horseshoe-inspired prior (Shin et al.; 2016). Stochastic constraints are also flexible in this respect. For an SC with constraining variable and penalty function on the real line, shrinkage is induced towards the nullspace of $r(\theta)$, which is a very flexible definition. Relatedly, the penalty function can (and should) be chosen to respect the support of the base random variable.

As previously mentioned, the constraint is applied to the joint distribution, not
the distribution of the base variable conditional on the constraint variable. Consequently, stochastic constraint variables can always be written, up to a proportional constant, as the product of the base density and the constraining variable’s CCDF composed with the penalty function. This is not the case for the aforementioned alternative, otherwise known as soft truncation, which can easily generate an intractable prior.

As discussed in Section 2.5, sparsity-inducing priors have been extensively studied for the standard normal-means model and linear regression. Besides sparsity, it is possible that other prior knowledge regarding the parameters is available to practitioners. However, the forms taken by the aforementioned sparsity priors are not flexible enough to accommodate many types of additional information. The stochastic constraint formulation of continuous sparsity priors exposes the base distribution, which can be changed and exploited to construct a prior that better reflects previous knowledge of the variable. Additional constraints can also be imposed to incorporate prior information.

We explore the flexibility of stochastic constraints in two simple examples. In the first we show analytically that the stochastic constraints framework can define shrinkage priors with marginal information that shrinks the parameters correctly, contrasting a naive approach, in which the shrinkage does not act towards zero. In the second example we specify a base prior distribution that has a positive skew to reflect prior information that the parameters are positive. This latter scenario is investigated using a simulation study.

### 6.1 Horseshoe with informative prior means

Stochastic constraint priors can enact shrinkage onto marginally informative distributions. We use the normal-means model and simple horseshoe prior (Carvalho et al.; 2010) to illustrate our findings analytically, but the idea is applicable to all stochastic constraint priors, including stochastic nets (5.5).

To begin, consider a naive attempt to concentrate probability mass about a set of known means, $\mu_i$, when using the horseshoe prior.

**Proposition 6.1.** Let $(\theta, \omega)$ have standard horseshoe prior as in (2.20) with simplified hierarchical structure $\tau = \sigma = 1$. If we use the normal-means model for the data, $y_i \sim N(\theta_i, 1)$ for $1 \leq i \leq n$, then the conditional posterior expectation of $\theta_i$ can be stated as

$$E(\theta_i, k_i | y) = (1 - k_i) y_i$$

where $k_i \sim \text{Beta}(1/2, 1/2)$ is the prior distribution. Changing the conditional prior on $\theta_i$ to $\theta_i | \lambda_i \sim N(\mu_i, \lambda_i^2)$ for fixed $\mu_i$ leads to the new conditional posterior expectation on $\theta_i$

$$E(\theta_i, k_i | y) = (1 - k_i) y_i + k_i \mu_i.$$

6.2. HORSESHOE WITH POSITIVE SKEW

Proof. The proof appears in Appendix A.1.8.

The first expectation in Proposition 6.1 is well known in the literature and motivates the name of the horseshoe prior (Carvalho et al.; 2010), whilst the second expectation is a simple extension of the first.

Proposition 6.1 demonstrates that incorporating prior information, by way of known mean \( \mu_i \) changes the horseshoe prior so that shrinkage occurs towards the conditional mean \( \mu_i \), rather than towards zero. Specifically, if \( k_i \approx 1 \) then this naive approach leads to \( \mathbb{E}(\theta_i|y, k_i) \approx \mu_i \) instead of the desired behaviour of shrinkage towards zero.

To incorporate both marginal information and high-dimensional shrinkage into a prior, a better approach is to update the base prior in the SC horseshoe.

**Proposition 6.2.** Let \( (\theta, \omega) \) have a stochastic constraint horseshoe prior as in (2.21), but with base prior \( f_{\hat{\theta}i} = \mathcal{N}(\hat{\theta}_i; \mu_i, \phi_i) \) rather than flat, where \( \mu_i \) and \( \phi_i \) are known from prior knowledge. Let \( \tau = \sigma = 1 \) for a simplified hierarchal structure. For the normal-means model, \( y_i \overset{iid}{\sim} \mathcal{N}(\theta_i, 1) \) for \( 1 \leq i \leq n \), the posterior expectation can be stated as

\[
\mathbb{E}(\theta_i|y, k_i) = \frac{(1 - k_i)(y_i + \mu_i\phi_i^{-1})}{1 + (1 - k_i)\phi_i^{-1}}.
\]

Proof. A proof is presented in Appendix A.1.9.

The SC horseshoe prior enables marginal prior information, \( f_{\hat{\theta}_i} \), to be incorporated into a joint sparsity structure. Note that if \( k_i \approx 1 \) then \( \mathbb{E}(\theta_i|y, k_i) \approx 0 \) and if \( k_i \approx 0 \) then we recover the posterior mean of the normal-means model with prior \( f_{\hat{\theta}} \) and no joint sparsity regularisation. The standard horseshoe is recovered in the limit of \( \sigma^2 \rightarrow \infty \).

This prior is able to both shrink each individual \( \theta_i \) towards \( \mu_i \), whilst shrinking the norm of the parameters \( \theta \) towards zero. Thereby including both information about the individual means and information about the sparsity of \( \theta \).

[Talk more about motivation: weak information about mean etc, high-dimensional setting]

### 6.2 Horseshoe with positive skew

Another example of prior information that may need to be incorporated into a sparse linear regression problem is non-negativity in the parameters. To illustrate, we take the horseshoe prior, in stochastic constraint form, and replace the marginal (flat) base distribution with a skewed improper density. The horseshoe prior is symmetric by design, but this modification helps to reflect our belief that the parameters of the model are positive with high probability.
We could choose a base distribution from the class of skew symmetric random variables described in Section 2.4.3, namely \( f_{\tilde{\beta}_i} = 2N(\tilde{\beta}_i; 0, \phi) \bar{F}_{\tilde{\eta}}|_{\alpha}(\tilde{\beta}_i - \tilde{\beta}_i) \) where \( \tilde{\eta}_i \sim N(0, \alpha) \). This would produce a positive-skewed version of the regularised horseshoe prior. However, we let \( \phi \to \infty \), which is equivalent to using improper base prior \( f_{\tilde{\beta}_i} = \bar{F}_{\tilde{\eta}}|_{\alpha}(-\tilde{\beta}_i) \) for \( \tilde{\beta}_i \in \mathbb{R} \) and \( 1 \leq i \leq p \). The resulting prior has hierarchical structure

\[
(\beta, \omega \mid \lambda, \tau, \sigma^2) \overset{d}{=} (\tilde{\beta}, \tilde{\omega} \mid \lambda, \tau, \sigma^2, \tilde{\beta}^2 \leq \tilde{\omega})
\]

\[
(\tilde{\beta}_i \mid \alpha) \sim f_{\tilde{\beta}_i \mid \alpha} \propto \bar{F}_{\tilde{\eta}}|_{\alpha}(-\tilde{\beta}_i) \quad (1 \leq i \leq p)
\]

\[
(\tilde{\omega}_i \mid \lambda_i, \tau, \sigma^2) \sim \text{Exp}(2^{-1}[\lambda_i \tau \sigma]^2)
\]

\[
\lambda_i \sim \text{Ca}_+(1)
\]

\[
\tau \sim \text{Ca}_+(1)
\]

\[
\sigma^2 \sim f_{\sigma^2} \propto \sigma^{-2},
\]

where the (improper) density for each \( \tilde{\beta}_i \) can be simplified to \( f_{\tilde{\beta}_i \mid \alpha} \propto \Phi(\tilde{\beta}_i / \alpha) \), and \( \Phi(x) \) is the cumulative distribution function for the standard normal distribution. To encourage prior mass in the positive space we can fix \( \alpha > 0 \), or incorporate a prior on \( \alpha \) with support on the positive axis.

Interestingly, because the improper prior for \( f_{\tilde{\beta}_i \mid \alpha} \) is derived from the limit of a skew-symmetric distribution we can rewrite the first two lines of (6.1) with additional latent variables \( \tilde{\eta}_i \) as

\[
(\beta, \omega \mid \lambda, \tau, \sigma^2) \overset{d}{=} (\tilde{\beta}, \tilde{\omega} \mid \lambda, \tau, \sigma^2, \max\{\tilde{\eta}_i, -\tilde{\omega}_i^{1/2}\} \leq \tilde{\beta}_i \leq \tilde{\omega}_i^{1/2}, 1 \leq i \leq p)
\]

\[
\tilde{\beta} \sim f_{\tilde{\beta}} \propto 1
\]

\[
\tilde{\eta}_i \overset{iid}{\sim} N(0, \alpha).
\]

The representation in (6.2) leads to a Gibbs sampler equivalent to a standard SC horseshoe Gibbs sampler (see Appendix A.2.1) but with the additional constraints \( \tilde{\eta}_i \leq \tilde{\beta}_i \), and full conditional distribution for \( \eta_i \)

\[
(\eta_i \mid \cdot) \sim N(0, \alpha) \text{ s.t. } \eta_i \leq \beta_i,
\]

i.e. a truncated normal distribution for all \( \eta_i \) where \( 1 \leq i \leq p \).

To test the effectiveness of the prior compared to a standard horseshoe prior, we ran simulations in settings based on Zhang et al. (2017), using linear models where all nonzero parameters were positive. The observations, \( y \), are simulated with an \( n \)-dimensional multivariate normal distribution given by

\[
y \sim N(X\beta, I_n)
\]

where \( X = [x_1 \ x_2 \ \cdots \ x_p]^{\top} \), \( n = 60 \), and \( p \) is chosen from \( p \in \{50, 100, 250\} \). The rows, \( x_i \), of the design matrix, \( X \), are simulated by drawing from a zero-mean multivariate normal distribution with covariance matrix \( \Sigma \). The matrix \( \Sigma \) has diagonal
Table 6.1: Posterior mean SSE and AUC values (mean across simulations with standard deviation in parenthesis) comparing positive-skew horseshoe (Skew) to regular horseshoe (Reg.). The SSE values are grouped by coefficient sizes. The sparsity of the true \( \beta \) is controlled by \( \beta^* = [0_{10}^\top, b_1^\top, 0_{30}^\top, b_2^\top, 0_{p-50}^\top]^\top \) where the nonzero components, \( b_1 \) and \( b_2 \), each have five elements that are simulated by taking the absolute value of a randomly generated zero-mean student-t distributions with three degrees of freedom. Each setting is replicated 250 times and \( \alpha = 0.25 \).

The performance of the models are assessed using the sum squared error of the posterior mean estimate of \( \beta \), computed by \( \text{SSE} = (\hat{\beta} - \beta^*)^\top (\hat{\beta} - \beta^*) \), where \( \hat{\beta} \) is the posterior mean estimate of \( \beta \) from the Gibbs sampler, and \( \beta^* \) is the true value of \( \beta \) used to simulate the data. Table 6.1 contains the SSE and AUC results of the simulation study. The SSE values are summarised by coefficient sizes. The groupings are \( \{0\} \) for true zeros, \( (0, 0.5] \) for small nonzero coefficients, \( (0.5, \infty) \) for medium to large coefficients, and \( [0, \infty) \) to summarise all coefficients. The AUC

| n  | p  | \( \rho \) | HS prior | SSE by groups. True \( |\beta| \) in | AUC |
|----|----|--------|----------|---------------------------------|-----|
|    |    |        |          | \( \{0\} \) | \( (0, 0.5] \) | \( (0.5, \infty) \) | \( [0, \infty) \) |
| 50 | 0.25 | Skew | 0.16 (0.09) | 0.09 (0.08) | 0.30 (0.21) | 0.55 (0.25) | 0.92 (0.07) |
|    |     | Reg. | 0.25 (0.19) | 0.10 (0.09) | 0.28 (0.19) | 0.63 (0.31) | 0.91 (0.07) |
|    | 0.50 | Skew | 0.21 (0.14) | 0.12 (0.09) | 0.48 (0.28) | 0.81 (0.33) | 0.91 (0.06) |
|    |     | Reg. | 0.35 (0.25) | 0.13 (0.10) | 0.44 (0.26) | 0.92 (0.37) | 0.90 (0.07) |
|    | 0.75 | Skew | 0.31 (0.18) | 0.15 (0.11) | 1.00 (0.58) | 1.45 (0.65) | 0.87 (0.07) |
|    |     | Reg. | 0.61 (0.39) | 0.18 (0.14) | 0.84 (0.53) | 1.63 (0.72) | 0.85 (0.08) |
|    | 0.90 | Skew | 0.58 (0.43) | 0.14 (0.14) | 2.52 (1.28) | 3.25 (1.54) | 0.83 (0.08) |
|    |     | Reg. | 1.25 (0.84) | 0.21 (0.24) | 2.18 (1.23) | 3.64 (1.71) | 0.80 (0.09) |
| 100| 0.25 | Skew | 0.18 (0.11) | 0.13 (0.09) | 0.39 (0.28) | 0.69 (0.34) | 0.91 (0.07) |
|    |     | Reg. | 0.34 (0.23) | 0.13 (0.09) | 0.34 (0.26) | 0.82 (0.39) | 0.90 (0.07) |
|    | 0.50 | Skew | 0.21 (0.14) | 0.16 (0.13) | 0.65 (0.41) | 1.01 (0.48) | 0.90 (0.06) |
|    |     | Reg. | 0.48 (0.33) | 0.16 (0.14) | 0.55 (0.36) | 1.19 (0.57) | 0.88 (0.07) |
|    | 0.75 | Skew | 0.33 (0.23) | 0.16 (0.12) | 1.53 (0.90) | 2.02 (1.03) | 0.87 (0.07) |
|    |     | Reg. | 0.89 (0.61) | 0.16 (0.13) | 1.23 (0.73) | 2.28 (1.12) | 0.85 (0.07) |
|    | 0.90 | Skew | 0.52 (0.37) | 0.18 (0.14) | 3.18 (1.64) | 3.88 (1.85) | 0.83 (0.08) |
|    |     | Reg. | 1.51 (1.00) | 0.21 (0.24) | 2.49 (1.40) | 4.20 (1.93) | 0.79 (0.09) |
| 250| 0.25 | Skew | 0.15 (0.10) | 0.16 (0.12) | 0.63 (0.44) | 0.94 (0.49) | 0.91 (0.06) |
|    |     | Reg. | 0.41 (0.20) | 0.15 (0.11) | 0.53 (0.39) | 1.09 (0.50) | 0.89 (0.07) |
|    | 0.50 | Skew | 0.18 (0.16) | 0.20 (0.14) | 0.93 (0.64) | 1.32 (0.70) | 0.89 (0.07) |
|    |     | Reg. | 0.55 (0.36) | 0.19 (0.13) | 0.76 (0.53) | 1.51 (0.73) | 0.87 (0.07) |
|    | 0.75 | Skew | 0.27 (0.24) | 0.22 (0.15) | 1.99 (1.06) | 2.48 (1.12) | 0.87 (0.08) |
|    |     | Reg. | 1.09 (0.78) | 0.21 (0.16) | 1.53 (0.86) | 2.83 (1.24) | 0.84 (0.08) |
|    | 0.90 | Skew | 0.48 (0.38) | 0.23 (0.20) | 4.73 (2.46) | 5.44 (2.65) | 0.82 (0.08) |
|    |     | Reg. | 2.12 (1.66) | 0.23 (0.20) | 3.82 (2.18) | 6.17 (3.24) | 0.78 (0.09) |
is calculated using the t-score metric \(|\hat{\mu}_i|/\hat{\sigma}_i\) where \(\hat{\mu}_i\) is the posterior mean of the regression coefficients and \(\hat{\sigma}_i\) is the posterior standard deviation.

Using the skew horseshoe prior there is a noticeable reduction in the mean SSE for true zero coefficients, \(\beta \in \{0\}\), compared to the standard horseshoe prior. The discrepancy is larger as \(\rho\) increases. The AUC value is better for the positive-skew horseshoe across the simulations, but only marginally so. This indicates that the incorporation of the skew towards positive values helps better estimate the value of zero coefficients but, does not make much of a difference in identifying them (at least relative to the false positive rate).

The SSE for the medium to large coefficients, \(\beta \in (0, 0.5]\), was indistinguishable between the two models, but performance better for the large coefficients, \(\beta \in (0.5, \infty)\), for the standard horseshoe model. This suggests that the large coefficients are biased by placing high probability mass on positive parameters, an unfortunate effect from the skew-positive horseshoe prior.

Overall, the skew-positive horseshoe prior performs better on average than the standard horseshoe prior in terms of SSE, but the standard deviation of these estimates are relatively high. The potential gain over the true zero coefficients is mostly offset by the loss in the medium to high coefficients. To address the loss in the medium to large coefficients, it may be beneficial to place a prior on the \(\alpha\) term rather than set it equal to a fixed constant. Moreover, each \(\beta_i\) could be given an \(\alpha_i\) positive-skewness parameter for further flexibility. This is left for future research.
CHAPTER 7

Efficient sampling for Bayesian regression with sparse stochastic net priors

Regression with stochastic net priors can generate a conditional distribution structure for better performing Gibbs sampling. The efficiency arises from targeting a truncated density rather than a density defined by a traditional hierarchy for some of the parameters. This is especially effective for continuous sparsity priors, since a large proportion of the parameters are truncated very close to zero in each step of the algorithm. Exploiting this property can reduce expensive matrix operations, and increase the speed of the Gibbs sampler.

7.1 The general truncated Gibbs sampler

Consider a Gaussian regression model with sparsity prior, where the prior has a scale mixture of normals representation. The Gibbs samplers derived from these priors generally rely on the scale mixture of normals representation for sampling. As such, the sampling step for \( \beta \) conditioned on all other parameter is

\[
(\beta | \cdot) \sim N(A X^\top y, \sigma^2 A),
\]

\[
A = (X^\top X + D)^{-1},
\]

where \( D \) is updated in each iteration. For general matrices \( X \) and \( D \), the inversion to determine \( A \) is a costly procedure in high dimensions, \( O(p^3) \) where \( p \) is the number of regression coefficients. Both Johndrow et al. (2018), who study MCMC sampling for the horseshoe prior, and Hahn et al. (2019, forthcoming), who study MCMC for arbitrary priors, identify the inversion of \( A \) as the main bottleneck for Gibbs samplers in medium and high dimensions with a linear Gaussian model. Whilst they provide their own strategies for reducing computations for this step, we explore the complementary role stochastic constraints can play to further reduce this burden. Moreover, the method we outline has wider applicability than the priors and models considered by Johndrow et al. (2018) and Hahn et al. (2019, forthcoming).

To illustrate, consider the equivalent stochastic net prior, in local constraint form

\[
(\beta, \omega | k) \overset{d}{=} \left( \begin{array}{c} \tilde{\beta}, \tilde{\omega} \mid k, |\tilde{\beta}|^\nu \leq \tilde{\omega} \\ \tilde{\beta} \sim f_{\tilde{\beta}}, \quad \tilde{\beta} \in \mathbb{R}^p \\ \tilde{\omega}_i \sim \text{Exp}(k_i), \quad (1 \leq i \leq p) \end{array} \right)
\]

where \( k_i \) is a function of hyperparameters \( k_i = k(\lambda_i, \tau, \sigma^2) \), and \( \lambda_i, \tau \) and \( \sigma^2 \) are assigned some prior distributions for \( 1 \leq i \leq p \). We assume that \( f_{\tilde{\beta}} \) is either flat or iid normal (conditional on hyperparameters, if any), but proceed with \( f_{\tilde{\beta}} \propto 1 \) as most steps can be updated for a normal density trivially. The sampling step of the
CHAPTER 7. EFFICIENT SAMPLING FOR BAYESIAN REGRESSION WITH SPARSE STOCHASTIC NET PRIORS

The derived Gibbs algorithm for $\beta$ is

$$
(\beta | \omega, \sigma^2) \sim f_{\beta | \omega, \sigma^2} \propto \exp \left\{ -\frac{1}{2\sigma^2} (y - X\beta)^\top (y - X\beta) \right\} \quad \text{s.t. } |\beta_i|^{\nu} \leq \omega_i,
$$

(7.2)
a truncated normal distribution. Note that there is no dependence on $k$, other than $\sigma^2$, for this conditional distribution. Generally, truncated distributions increase the sampling difficulty, but here we use the constraints together with the sparsity dictated by the prior to our advantage.

If we can sample from the conditional distribution (7.2) and avoid inverting matrices the size of $A$ we may be able to implement a more efficient Gibbs sampler. Below we outline an exact Gibbs sampler (in the sense that no approximations are made to the conditional sampling steps), under the assumption that $X^\top X$ can be computed and stored.

The general idea to sample from (7.2) is to exploit the magnitude information contained in the constraints $\omega$. In each iteration of the Gibbs sampler we form a partition, $P$, from $\{1, 2, \ldots, p\}$, where $P = \{I, J\}$ such that

$$
\omega_s \leq \epsilon < \omega_t \quad \text{(for all } s \in I \text{ and } t \in J),
$$
a subset of $\omega$ less than or equal to some cutoff $\epsilon$, and a subset with elements greater than $\epsilon$. More generally a larger partition could be formed, but we focus on this simple case. Let $\beta_I$ and $\beta_J$ be vectors with elements corresponding to the indices contained in $I$ and $J$ respectively, and let $\beta_{(j)}$ be the vector $\beta$ without the $j$th element. Instead of sampling from (7.2) directly we use conditional samplers based on the partition $P$,

$$
(\beta_I | \beta_J, \omega, \sigma^2) \sim f_{\beta_I | \beta_J, \omega, \sigma^2} \propto f_{\beta_I} \quad \text{s.t. } |\beta_i|^{\nu} \leq \omega_i \text{ for } i \in I
$$

and

$$
(\beta_J | \beta_I, \omega, k) \sim f_{\beta_J | \beta_I, \omega, k}.
$$

The conditional distribution of the second group $\beta_J$ may also use a truncated sampler, as the first does, or we can collapse over the constraints $\omega$ leading to a sampling step similar to (7.1), but with computational complexity $O(|J|^3)$ rather than $O(p^3)$.

For sparsity priors that encourage a large number of coefficients to be close to zero in each iteration, the size of group $J$, can be a small fraction of $p$.

The coefficients in $\beta_I$ will have magnitude $|\beta_i|^{\nu} \leq \epsilon$ so the joint support of $\beta_I | \beta_J$ will be a very small hyperrectangle (an $|I|$-orthotope). For a sufficiently small $\epsilon$, the density over this support will be near uniform and there will be little dependence between the parameters within the group. Hence, we can use a full conditional one-dimensional sampler without adding too much correlation between parameters. Moreover, $\beta_I$ will have little dependence on $\beta_J$ due to this tight constraint. These two properties lead to the observation that, for sufficiently small $\epsilon$, 

$$
(\beta_i | \beta_{(i)}, \omega, \sigma^2) \overset{d}{=} (\beta_i | \omega, \sigma^2) \text{ for } i \in I.
$$

(7.3)
Therefore sampling each $\beta_i$ for $i \in I$ with full conditional distribution will not substantially increase the Markov chain's autocorrelation relative to a block sampling step for $\beta \mid \omega, k$. In light of this, the strategy to reduce computations is to choose the sampler for $(\beta_i \mid \beta_{(i)}, \omega, \sigma^2)$ such that the computational complexity is at most linear in $p$, and therefore iterating through each $i \in I$ costs $O(p|I|)$. For example, in Section 7.2 we use a conditional Metropolis-Hastings (MH) sampler with uniform proposal for each parameter $\beta_i$ with $i \in I$.

If a uniform MH sampler is used for each $\beta_i$ for $i \in I$, the acceptance ratio will be very close to one when a sufficiently small $\epsilon$ is chosen. To see this consider the MH log-ratio for $\beta_i$ with uniform sampler, $\log \alpha$, with value

$$\log \alpha (\beta'_i; \beta, \sigma^2) = -\frac{1}{2\sigma^2} \left( X_i^\top X_i (\beta'_i - \beta_i)^2 + 2 \left( \beta_{(i)}^\top X_{(i)} X_i - X_i^\top y \right) (\beta'_i - \beta_i) \right),$$

(7.4)

where $\beta'_i$ is the uniform proposal from $\text{U}(-\omega^{1/\nu}_i, \omega^{1/\nu}_i)$, $X_i$ is the $i$th column of $X$, and $X_{(i)}$ is $X$ without the $i$th column. The parameter constraints restrict the absolute change in value of $\beta_i$ to $|\beta'_i - \beta_i| \leq 2\epsilon^{1/\nu} \leq 2\epsilon^{1/\nu}$ which causes the log-ratio to be close to zero for small $\epsilon$.

If $X^\top X$ and $X^\top y$ can be pre-computed and stored then the matrices in (7.4) are obtained by subsetting the aforementioned matrices. The only matrix operation required is $\beta_{(i)}^\top X_{(i)} X_i$, which is $O(p)$ as required.

If the approximation (7.3) is good enough then sampling $(\beta_i \mid \beta_{(i)}, \omega, k)$ for $i \in I$ followed by the sampling from $(\beta_J \mid \beta_I, \omega, k)$ should result in a joint distribution that is close to the blocked distribution $(\beta \mid \omega, k)$. In other words, we can sample the $\beta_i \mid \beta_{(i)}$ conditionally (in a very specific way) without gaining much autocorrelation. Whilst the method is an approximation to the blocked sampling step $(\beta \mid \omega, k)$, it is not an approximation in the sense that the implied MH-within-Gibbs sampler is incorrect. Sampling conditionally from the $\beta_{(i)}$ and $\beta_j$ as specified still results in a correct MH-within-Gibbs sampler.

The remaining coefficients, $\beta_J$, are sampled from $(\beta_J \mid \beta_I, \omega, k)$ for which a number of methods are available. Some of the possibilities are:

1. We can collapse over the $\omega$ (marginalise the stochastic constraint) and sample from $(\beta_J \mid k, \beta_I)$. In cases where $\nu = 2$, this distribution will be a conditional multivariate Gaussian, which will be relatively efficient only if $|J| \ll p$ due to the required matrix inversions.

2. If $|J| < \min(n, 100)$ then it is possible to use the exact truncated Gaussian sampler proposed by Botev (2017) to sample from $(\beta_J \mid \beta_I, \omega, k)$ directly. Again, the matrix inversion required this samplers will only be efficient if $|J| \ll p$.

3. For situations that do not fit the above, an efficient MCMC sampler for
Using option 1 with a one-dimensional uniform MH sampler for the $\beta_I$ leads to a reduction in complexity from $O(p^3)$ to $O(|J|^3 + p|I|)$, so the benefit of the decomposition depends on the relative size of $|J|$ to $p$. This sampler is valid for any cutoff choice $\epsilon$. As such, we may choose it to fix the size of $|J|$ in which case the algorithm will be $O(p^2)$ in each iteration.

The proposed MH-within-Gibbs sampler exploits the sparsity induced by the shrinkage priors when most components are tightly constrained and the remaining are not constrained at all. This is typical of the sparsity priors in the literature, especially when $p \gg n$. Below we outline such a sampler for the R2-D2 prior and describe some promising results from a simulation study.

### 7.2 Partially truncated R2-D2 sampler

To illustrate the usefulness of the partially truncated Gaussian sampler described in the previous section, we derive a new Gibbs sampler for the R2-D2 prior. Whilst there are several equivalent ways to incorporate the constraints into the R2-D2 prior, we choose the closest to the original structure for posterior sampling in Zhang et al. (2017) for a fair comparison. The posterior sampling hierarchy for a linear Gaussian regression model with R2-D2 prior is

\[
\begin{align*}
(y \mid \beta, \sigma^2) &\sim N(X\beta, \sigma^2 I) \\
(\beta_i \mid \psi, \phi, \gamma, \sigma^2) &\sim N(0, \psi_i \phi_i \gamma \sigma^2 / 2), \quad (1 \leq i \leq p) \\
\psi &\sim \text{Exp}(1/2) \\
\phi &\sim \text{Dir}(a_\pi, a_\pi, \ldots, a_\pi) \text{ s.t. } a_\pi = a / p \\
(\gamma \mid \eta) &\sim \text{Gam}(a, \eta) \\
\eta &\sim \text{Gam}(b, 1) \\
\sigma^2 &\sim \text{IGam}(c, d),
\end{align*}
\]

for given parameters $a$, $b$, $c$, and $d$, while the stochastic constraints version will be the same but with $\beta_i$ replaced by

\[
(\beta_i, \omega_i \mid \cdot) \sim f_{\beta_i, \omega_i} \propto \exp \left\{ -\frac{\omega_i}{\psi_i \phi_i \gamma \sigma^2} \right\} 1(\beta_i^2 \leq \omega_i)
\]
7.2. PARTIALLY TRUNCATED R2-D2 SAMPLER

for $1 \leq i \leq p$. The full conditional distributions for the standard Gibbs sampler are

- $(\beta | \cdot) \sim N(\mu, \sigma^2 V)$
- $(\sigma^2 | \cdot) \sim IGam(c + (n + p)/2, d + \chi)$
- $(\psi_i^{-1} | \cdot) \sim \text{Wald} \left( |\beta_i|^{-1}(\phi_i \gamma \sigma^2/2)^{1/2}, 1 \right)$ (1 \leq i \leq p)
- $(\gamma | \cdot) \sim \text{GIG} \left( \frac{2}{\sigma^2} \sum_{i=1}^{p} \frac{\beta_i^2}{\psi_i \phi_i}, 2\eta, a - p/2 \right)$ (7.5)
- $(\eta | \cdot) \sim \text{Gam}(a + b, 1 + \gamma)$
- $(\phi_i | \cdot) = \frac{\varphi_i}{\|\varphi\|_1}$, where $\varphi_i \sim \text{GIG} \left( \frac{2\beta_i^2}{\psi_i \sigma^2}, 2\eta, a_\pi - 1/2 \right)$.

as given in Zhang et al. (2017), where

- $\mu = VX^\top y,$
- $V = (X^\top X + S^{-1})^{-1},$
- $S = (\gamma/2) \text{diag}\{\psi_1 \phi_1, \psi_2 \phi_2, \cdots, \psi_p \phi_p\},$ and
- $\chi = \beta^\top S^{-1} \beta + (y - X\beta)^\top (y - X\beta)/2.$

Our proposed MH-within-Gibbs sampler is given in Algorithm 5 and uses these conditional distributions, but changes the sampling step for $\beta$ as described in Section 7.1. Based on this sampling scheme, the conditional distribution for each constraint parameter $\omega_i$ is

- $(\omega_i | \cdot) = \beta_i^2 + \tau_i$, where $\tau_i \sim \text{Exp}(1/(\psi_i \phi_i \gamma \sigma^2)).$ (7.6)

The algorithm uses the mean and variance of the conditional distribution of $\beta_J$ defined by

- $\mu_{J|I} = V_J (X_J^\top y - P_{J,I} \beta_I)$
- $V_J = (P_{J,J})^{-1},$

where $P = X^\top X + S^{-1}$ is the precision matrix and $P_{K,L}$ denotes the matrix created by taking only the $k \in K$ rows and $l \in L$ columns. Note also that $X_J$ is the matrix $X$ with only $j \in J$ columns. We now outline the simulation study used to assess the computational gains from the proposed truncated Gibbs sampler. Following this we discuss the results.

The simulation study is similar to those in previous sections. The observations, $y$, are simulated with an $n$-dimensional multivariate normal distribution given by

- $y \sim N(X\beta, I_n).$
CHAPTER 7. EFFICIENT SAMPLING FOR BAYESIAN REGRESSION WITH SPARSE POSTERIOR MEAN ESTIMATE OF $\beta$

The performance of the samplers is assessed using the sum squared error of the results are extra repetitions.

We compare two versions of the proposed Gibbs sampler to the standard sampler given in Zhang et al. (2017). The versions of the new sampler vary by changing the cutoff, $\varepsilon \in \{0.1, 0.2\}$, which controls whether a parameter is sampled by uniform or multivariate normal distribution. Note that changing $\varepsilon$ does not affect the standard sampler, so the repeated rows for this Gibbs sampler in the tables summarising the results are extra repetitions.

The performance of the samplers is assessed using the sum squared error of the posterior mean estimate of $\beta$, computed by $SSE = (\hat{\beta} - \beta^*)^T (\hat{\beta} - \beta^*)$, where $\hat{\beta}$ is

Algorithm 5: Truncated R2-D2 sampler

**Input:** $N$, maximum iterations. Matrices $X^T X$ and $X^T y$. R2-D2 prior constants $\{a, b, c, d\}$. Cutoff $\varepsilon$.

$t \leftarrow 1$
Initiate $\{\beta, \sigma^2, \psi, \gamma, \eta, \phi\}^{[1]}$

while $t \leq N$ do

$t \leftarrow t + 1$
Sample $(\omega_i^{[l]} | \{\beta_i, \sigma^2, \psi_i, \gamma_i, \phi_i\}^{[l-1]})$ with (7.6), for $1 \leq i \leq p$

$I \leftarrow \{1 \leq i \leq p, \text{s.t. } \omega_i^{1/2} \leq \varepsilon\}$
$J \leftarrow \{1 \leq i \leq p, \text{s.t. } \omega_i^{1/2} > \varepsilon\}$

$\beta^* \leftarrow \beta^{[l-1]}$

for $i \in I$ do

Sample $(\beta_i^{[l]} | \omega_i^{[l]}) \sim U(-\omega_i^{1/2}, \omega_i^{1/2})$ (proposal)
Sample $u \sim U(0, 1)$
$\alpha_i \leftarrow \alpha(\beta_i^{[l]}; \beta^*, (\sigma^2)^{[l-1]})$, defined in (7.4)
if $\alpha_i > u$ then

$\beta_i^{[l]} \leftarrow \beta_i^{[l]}$ (accept with prob. $\min\{1, \alpha_i\}$)

end if

end for

Sample $(\beta_j^{[l]} | \beta_j^{[l]}, \sigma^2, \psi, \gamma, \phi)^{[l-1]} \sim N(\mu_{j[l]}, \sigma^2 V_j)$
Sample $\{\sigma^2, \psi, \gamma, \eta, \phi\}^{[l]}$ with conditional distributions in (7.5)

return $\{\beta, \sigma^2, \psi, \gamma, \eta, \phi\}_{i=1}^N$

where $X = [x_1, x_2, \ldots, x_p]^T$, $n = 100$, and $p$ is one of $p \in \{500, 1000, 2500, 5000\}$. The rows, $x_i$, of the design matrix, $X$, are simulated by drawing from a zero-mean multivariate normal distribution with autocorrelation $\rho$ chosen from $\rho \in \{0.1, 0.5\}$ such that any row $i$ has elements with correlation $\text{cor}(x_{ij}, x_{ik}) \in \rho|j-k|$. The error term, $\sigma^2$, is fixed at one. The sparsity of $\beta$ is controlled by setting the true value to $\beta^* = [0_{10}^T, b_1^T, 0_{30}^T, b_2^T, 0_{p-50}^T]^T$ where the nonzero components, $b_1$ and $b_2$ are

$$b_1 = [2, 2, -5, -5, -5]^T,$$

$$b_2 = [-2, -2, -2, 5, 5]^T.$$

The fixed hyperparameters for the R2-D2 prior are set to $a = p/n$ and $b = 1/2$, respectively. Each simulation setting is replicated 100 times for 2000 samples and the first 1000 samples are discarded as burnin.
7.2. PARTIALLY TRUNCATED R2-D2 SAMPLER

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>ρ</th>
<th>ε</th>
<th>SSE by groups. True</th>
<th>β</th>
<th>in</th>
<th>{0}</th>
<th>(0, \infty)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Standard</td>
<td></td>
<td></td>
<td>Standard</td>
<td>Truncated</td>
</tr>
<tr>
<td>500</td>
<td>0.1</td>
<td>1</td>
<td>0.1</td>
<td>4.39 (0.61)</td>
<td>3.76 (0.57)</td>
<td>25.11 (6.10)</td>
<td>22.41 (5.75)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2</td>
<td>4.29 (0.58)</td>
<td>4.20 (0.55)</td>
<td>24.59 (6.19)</td>
<td>23.45 (5.98)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
<td>3.35 (0.48)</td>
<td>2.88 (0.42)</td>
<td>21.27 (5.41)</td>
<td>19.46 (5.00)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2</td>
<td>3.44 (0.44)</td>
<td>3.43 (0.45)</td>
<td>21.89 (4.80)</td>
<td>21.25 (4.84)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.1</td>
<td>1</td>
<td>0.1</td>
<td>3.71 (0.84)</td>
<td>4.11 (0.56)</td>
<td>50.75 (19.40)</td>
<td>48.67 (10.67)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2</td>
<td>3.80 (0.60)</td>
<td>5.65 (0.60)</td>
<td>49.98 (13.18)</td>
<td>60.51 (10.68)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
<td>3.02 (0.35)</td>
<td>3.25 (0.39)</td>
<td>41.01 (7.89)</td>
<td>41.35 (8.10)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.2</td>
<td>3.12 (0.31)</td>
<td>4.73 (0.45)</td>
<td>41.04 (6.83)</td>
<td>49.74 (8.10)</td>
<td></td>
</tr>
<tr>
<td>2500</td>
<td>0.1</td>
<td>1</td>
<td>*0.14 (0.46)</td>
<td>3.62 (0.31)</td>
<td>*138.03 (22.59)</td>
<td>100.27 (12.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*0.12 (0.42)</td>
<td>5.19 (0.49)</td>
<td>*138.88 (21.42)</td>
<td>119.57 (6.81)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*0.53 (0.67)</td>
<td>3.18 (0.31)</td>
<td>*111.86 (40.94)</td>
<td>83.29 (10.92)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*0.48 (0.66)</td>
<td>5.15 (0.40)</td>
<td>*115.50 (40.37)</td>
<td>104.72 (9.49)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>0.1</td>
<td>1</td>
<td>**0.0</td>
<td>2.58 (0.28)</td>
<td>**145.0</td>
<td>130.38 (5.17)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>**0.0</td>
<td>3.19 (0.38)</td>
<td>**145.0</td>
<td>137.12 (1.58)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>**0.0</td>
<td>2.70 (0.30)</td>
<td>**145.0</td>
<td>119.26 (7.89)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>**0.0</td>
<td>3.82 (0.47)</td>
<td>**145.0</td>
<td>132.53 (3.06)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.1: Posterior mean SSE values (mean across simulations with standard deviation in parenthesis) comparing the standard Gibbs sampler to the truncated Gibbs sampler. The SSE values are grouped by coefficient sizes. * indicates the sampler had numerical issues in some chains. ** indicates that the sampler had numerical failure in all chains.

the posterior mean estimate of β given by the the MCMC scheme, and β* is the true value of β used to simulate the data. We compute the average and standard deviation of the SSE across the replications for each combination of p, ρ and ϵ. These results are reported in Table 7.1. One difference is that the truncated sampler has smaller SSE for zero-parameters for p = 500 and ϵ = 0.1, but this could be an artefact of how the Markov chains evolves in early samples. The biggest difference between the samplers is seen when p = 2500 and p = 5000. For p = 2500, the standard sampler begins to experience some numerical error. Investigating the individual chains showed that some (but not all) chains had all elements of β very close to zero. For p = 5000, the numerical issue was exacerbated and none of the chains were unable to move away from zero in the standard sampler. Comparatively, the truncated Gibbs sampler showed no signs of numerical issues for p = 2500 or p = 5000. These numerical issues were not reported in (Zhang et al.; 2017), but the models they consider are at most p = 1000, for which we also find no issues. We suspect the complication arises from ill conditioning of the precision matrix in the distribution for β. We experimented with several versions of the multivariate normal sampler, but were not able to resolve this issue.

Table 7.2 contains the results showing average time taken to run the Gibbs samplers,
Table 7.2: Time taken to run 2000 samples in minutes (mean across simulations with standard deviation in parenthesis) comparing the standard Gibbs sampler to the truncated Gibbs sampler. * indicates the sampler had numerical issues in some chains. ** indicates that the sampler had numerical failure in all chains.

<table>
<thead>
<tr>
<th>n</th>
<th>p</th>
<th>ρ</th>
<th>ε</th>
<th>Computer time (minutes)</th>
<th>Standard</th>
<th>Truncated</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>4.22 (1.02)</td>
<td>5.84 (0.82)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>3.18 (1.04)</td>
<td>4.84 (0.78)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>3.79 (0.70)</td>
<td>5.56 (0.28)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>3.65 (0.88)</td>
<td>5.78 (0.36)</td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>6.01 (0.82)</td>
<td>8.30 (2.04)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>7.10 (1.43)</td>
<td>10.02 (3.36)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>5.81 (1.18)</td>
<td>8.08 (2.04)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>5.80 (1.09)</td>
<td>7.91 (1.67)</td>
<td></td>
</tr>
<tr>
<td>2500</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>*14.20 (0.99)</td>
<td>11.93 (0.93)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>*14.02 (1.72)</td>
<td>12.20 (1.14)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>*14.12 (0.91)</td>
<td>12.01 (0.79)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>*14.30 (1.62)</td>
<td>12.22 (1.02)</td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>**64.71 (3.48)</td>
<td>29.95 (1.00)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>**64.53 (3.74)</td>
<td>30.90 (1.36)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.1</td>
<td>0.1</td>
<td>**64.48 (3.74)</td>
<td>29.82 (1.15)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.1</td>
<td>**64.57 (3.74)</td>
<td>30.56 (1.40)</td>
<td></td>
</tr>
</tbody>
</table>

whilst Figure 7.1 shows a graphical representation. Each chain was run in parallel on a single core of an Intel Xeon E5-2690V3 “Haswell” processor (12-core, 2.6 GHz), contained within the Pawsey Centre’s XC40 Series Supercomputer (Magnus). The standard Gibbs sampler was marginally faster for $p = 500$ and $p = 1000$, but its relative performance drops significantly for larger $p$. For $p = 5000$, the truncated Gibbs sampler takes less than half the time of standard Gibbs sampler (approximately 30 minutes compared to 65 for 2000 samples) and, as mentioned previously, did not have numerical issues. We note that the implementation of the truncated sampler in R is particularly naive, and it ought to be possible to make it faster than the standard sampler for all values of $p$.

The diagnostics reported in Table 7.3 for the partially truncated Gibbs algorithm reveal some interesting properties of the proposed sampler. Firstly, the uniform sampling component had high acceptance rates in the MH step, ranging between 0.70 and 0.88 on average. When the cutoff for the coefficient groupings, $\epsilon$, was 0.1, the sampler had a better acceptance rate on average. Secondly, the proportion of coefficients in group $I$ increased as the number of parameters, $p$, increased. When $p = 500$, only 42-56% of parameters are allocated to the uniform sampler per iteration, but when $p = 5000$ this increases to 82-86%. If the proportion were not increasing as $p$ did, then by $p = 5000$ the truncated sampler would be facing the same numerical issues as the standard Gibbs sampler. The fixed threshold for $\epsilon$
Figure 7.1: Average computation time of 2000 samples from the R2-D2 prior across 100 repetitions. The standard sampler had numerical instability for \( p \in \{2500, 5000\} \).

compared to increasing \( p \) is the likely cause of this effect, as increasing \( p \) for priors that induce this type of sparsity enforces additional shrinkage on each parameter.

To demonstrate that the number of iterations chosen in the study, 2000, is sufficient, we repeated experiment for \( p = 5000, n = 100, \varepsilon = 0.1 \) case with 12,000 iterations, discarding the first 9,000 as burn-in. This was replicated 100 times for each \( \rho \in \{0.1, 0.5\} \). The result for the mean SSE was close to the SSE reported in Table 7.1 (\( p = 5000 \)), while the time taken per iteration was almost exactly the same — approximately half an hour per 2000 iterations. The average MH acceptance rate increased to 0.88 in the case of \( \rho = 0.2 \) and was unchanged for \( \rho = 0.1 \). The proportion of \( \beta_i \) in group \( I \) was close to 0.86 for both settings. The distribution of effective sample sizes for nonzero \( \beta \) across all repetitions had 5%, 25%, 75%, and 95% quantiles at 300, 591, 2100, and 3000 respectively.

While we consider a fixed cutoff, \( \varepsilon \), a dynamic cutoff would also result in a valid Gibbs sampler. A potential rule could be to choose \( \varepsilon \) so that a fixed number of pa-
parameters are chosen in each iteration. Or we could draw the size from a distribution based on particular properties of the problem size and posterior concentration. This way we can guarantee the size of groups $I$ and $J$ and better control the computational complexity of the sampler which is $\mathcal{O}(|J|^3 + p|I|)$ in each iteration. Other improvements on the algorithm are also possible. Instead of a uniform sampler for the coefficients in group $I$, a triangular or quadratic-type distribution could be used, so long as the computational complexity does not increase. A univariate truncated normal might also be an option. Lastly, other models could be fit with a partially truncated Gibbs sampler. In principle, the method of separating coefficients based on their predicted bound ($\omega_i$) to sample small and large coefficients separately can be applied to models with non-Gaussian errors but this possibility has not yet been explored. Stochastic constraints appear to be very flexible in this regard.

Overall, the truncated Gibbs sampler succeeds by focussing computational efforts on parameters which are likely to be relatively large in any given iteration. The R2-D2 prior was only one example of its use, as it can be used for all of the continuous sparsity priors described in this thesis. It is able to sample when $p$ is large even when the standard Gibbs sampler has numerical issues and fails. Moreover, it does not approximate the Markov chain in any way, that is, it is an exact Gibbs sampler. Whilst we have not demonstrated that the partially truncated method is able to scale

<table>
<thead>
<tr>
<th>$n$</th>
<th>$p$</th>
<th>$\rho$</th>
<th>$\epsilon$</th>
<th>MH accept rate</th>
<th>Proportion in group $I$</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.1</td>
<td>0.1</td>
<td>0.82 (0.001)</td>
<td>0.42 (0.01)</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.70 (0.001)</td>
<td>0.55 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.82 (0.001)</td>
<td>0.43 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.70 (0.002)</td>
<td>0.56 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1000</td>
<td>0.1</td>
<td>0.1</td>
<td>0.83 (0.001)</td>
<td>0.62 (0.01)</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.73 (0.001)</td>
<td>0.68 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.83 (0.001)</td>
<td>0.62 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.73 (0.002)</td>
<td>0.68 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2500</td>
<td>0.1</td>
<td>0.1</td>
<td>0.86 (0.001)</td>
<td>0.79 (0.01)</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.77 (0.002)</td>
<td>0.79 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.86 (0.001)</td>
<td>0.79 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.76 (0.002)</td>
<td>0.79 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5000</td>
<td>0.1</td>
<td>0.1</td>
<td>0.88 (0.002)</td>
<td>0.86 (0.01)</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.78 (0.002)</td>
<td>0.82 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.88 (0.002)</td>
<td>0.85 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>0.78 (0.001)</td>
<td>0.82 (0.01)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7.3: Average acceptance rate of Metropolis-Hastings step in Algorithm 5, and average proportion of coefficients in group $I$ (mean across simulations with standard deviation in parenthesis). Proportion in group $I$ calculated by $|I|/p$. Elements of this group are sampled by a uniform sampler rather than a multivariate normal distribution.
the R2-D2 prior to the dimension of genome-wide association study \((p \geq 50000)\), as Johndrow et al. (2018) do for the horseshoe sampler, this technique might be useful in combination with other methods to do so, in particular, if an approximation can be used when sampling the large coefficients. As such, further testing is required for the partially truncated Gibbs sampler on dimensions higher than \(p = 5000\), and future research on approximation methods, akin to Johndrow et al. (2018), is needed.
CHAPTER 8

Discussion

In this thesis, stochastic constraint random variables are shown to be a flexible and intuitive basis for prior construction in Bayesian modelling. Furthermore, outside of constructing priors, two novel uses were discussed to motivate their study in a wider context. The first was sparsity quantification using the geometrical properties of priors illuminated by stochastic constraints. Whilst the second used the parameter magnitude information stochastic constraints reveal to facilitate designing a more efficient Gibbs sampler for sparsity priors.

The exponential distribution played a special role in this work; facilitating easier Gibbs sampling, and enabling dual representation of priors as local and global stochastic constraints. These benefits are a direct result of the exponential distribution’s memoryless property. We have undertaken some research into stochastic constraint distributions with discrete constraining variables, particularly the geometric distribution which shares the memoryless property with the exponential. This class of distributions may have applications in spike-and-slab priors, but is left for future research.

There are other connections to existing research that have not been examined in this thesis. For example, the recent work on a relaxation method for constraints in Bayesian inference developed by Duan et al. (2018). The priors considered in this work can be formulated as a stochastic constraint using an exponential constraining variable. Stochastic constraints provide a secondary interpretation for these relaxed priors, that is, shrinkage (in this case a relaxed constraint) is applied to a base distribution by probabilistically restricting the joint support of the augmented prior. Further investigation of the connection is needed, but may lead to a deeper understanding of stochastic constraints and the relaxed priors of Duan et al. (2018). The soft multivariate truncated normal distribution recently proposed by Souris et al. (2018) also falls into this category.

There is an important aspect of sampling stochastic constraint variables that has not been considered in this thesis. If a hyperparameter of the SC, say $\lambda$, is assigned a prior distribution and the normalising constant, $Z(\lambda)$, does not have a simple form, then a Gibbs sampler, or in fact any sampler, may have difficulty sampling $\lambda$. This is referred to as a doubly-intractable problem in Bayesian analysis. If $Z(\lambda)$ is a simple integral then it may be permissible to use numerical methods, but this is not feasible in high dimensions (i.e. large $p$). We mention some possible remedies for this situation but the details are left for future research. If exact sampling of the stochastic constraint conditional on $\lambda$ is possible, then the exchange algorithm in Murray et al. (2006) is a viable option. More advanced techniques are also available if exact sampling is not possible, for example Russian roulette sampling (see Lyne et al.; 2015, for an overview). Sampling methods for hyperparameters of stochastic constraint distributions with intractable normalising constants are worthy of further investigation as it will increase their scope, whilst maintaining practicality.
A general framework for stochastic constraints was developed in Chapter 2 along with a clear language to describe them. Specifically, the work in this thesis relies on the conditional-hierarchical structure to probabilistically describe how the constraints are applied to the joint distribution. Chapter 3 contains results on when stochastic constraints induce proper distributions, characteristics of these SC distributions, as well as properties of Bayesian models that use the SC prior for inference. We also considered how far the analog of duality between constraint and penalty regularised inference extends to the Bayesian setting.

The simulation of stochastic constraint distributions was considered in Chapter 4, and a number of strategies were proposed. The most effective algorithm will vary by situation, but PDMPs are the most promising for more general SC distributions in high dimensions. There is scope for improvement of the rejection sampler described in Section 4.1, in particular for a univariate stochastic constraint, taking the inverse CDF transform of both variables leads to a uniform distribution on a subset of $[0, 1]^2$. If this area can be approximated with bounding functions then better samplers can be developed.

There is further research to be done on the geometrical properties of sparsity priors with a stochastic constraint representation. For example, investigation of the surface area and volume that these priors have. This may help to characterise the differences between the popular sparsity priors in the literature. The fundamental contribution of Chapter 5 was to draw attention to the geometric properties of sparsity priors, which has not been considered in this way before. Hence, there are opportunities to utilise this sparsity information for other decision rule methods in future research.

We have only considered two new priors in Chapter 6. The scope for creating tailored priors with stochastic constraints, eloquently combining marginal and joint prior knowledge, is wide. However, under several circumstances using stochastic constraint priors does come at a computational cost. Therefore, a good motivation to improve or replace existing priors is needed before embarking on creating these new distributions.

Finally, there are many avenues to expand upon Chapter 7 to create more efficient samplers for priors that have a stochastic constraint formulation. Several direct changes are discussed at the end of the chapter, but the general technique for separating coefficients by magnitude is widely applicable. If such a separation reduces the computation time for a given model then this is a very useful technique. It can also be combined with other methods to generate scalable samplers for Bayesian inference.
Appendices
A.1 Proofs excluded from main text

A.1.1 Theorem 3.5

**Proof.** Since the stochastic constraint is local and the constraining variables are iid, the normalising constant decomposes into identical components

\[ Z = z^q \] where \[ z = \int_{\mathbb{R}} \int_{\mathbb{R}^+} f_{\omega}(\omega) 1(u(|\theta - \mu|) \leq \omega) \, d\omega \, d\theta \]

The integral \( z \) can be rewritten as

\[ z = \int_{S} f_{\omega}(\omega) \, d\omega \, d\theta \]

where the region to integrate over is \( S = \{(\theta, \omega) \in \mathbb{R} \times \mathbb{R}^+ : u(|\theta - \mu|) \leq \omega\} \) which can be equivalently stated as

\[ S = \{(\theta, \omega) \in \mathbb{R} \times \mathbb{R}^+ : |\theta - \mu| \leq v(\omega)\} = \{(\theta, \omega) \in \mathbb{R} \times \mathbb{R}^+ : \mu - v(\omega) \leq \theta \leq \mu + v(\omega)\} \]

since \( v \), the inverse of \( u \), is assumed to exist. Therefore \( z \) can be written as

\[ z = \int_{0}^{\infty} f_{\omega}(\omega) \int_{\mu - v(\omega)}^{\mu + v(\omega)} 1 \, d\theta \, d\omega \]

\[ = 2 \int_{0}^{\infty} f_{\omega}(\omega) v(\omega) \, d\omega \]

\[ = 2 \mathbb{E}(v(\tilde{\omega})) \] and hence \( Z = 2^q \mathbb{E}(v(\tilde{\omega}))^q \). \( \square \)

A.1.2 Theorem 3.8

**Proof.** Consider the posterior distributions for a model with likelihood \( \pi(y|\theta) \) using two different priors given by:

(i) The base prior \( f_{\theta}(\theta) \), and

(ii) The stochastic constraint prior \( f_{\theta,\omega|\lambda}(\theta, \omega) \propto f_{\theta}(\theta) f_{\omega|\lambda}(\omega) 1(r(\theta) \preceq \omega) \).

Let the posterior distribution related to the base prior (i) be

\[ \tilde{p}(\theta|y) = \tilde{Z}^{-1} \pi(y|\theta) f_{\theta}(\theta) \] where \( \tilde{Z} = \int_{\Theta} \pi(y|\theta) f_{\theta}(\theta) \, d\theta \)
and assume $\tilde{p}(\theta|y)$ is a proper density. On the other hand, let the posterior for the stochastic constraint prior (ii) be
\[
p(\theta, \omega|y, \lambda) = Z(\lambda)^{-1} \pi(y|\theta)f_\theta(\theta)f_\omega|\lambda(\omega)1(r(\theta) \leq \omega)
\]
where $Z(\lambda) = \int_\Theta \int_\Omega \pi(y|\theta)f_\theta(\theta)f_\omega|\lambda(\omega)1(r(\theta) \leq \omega) \, d\omega \, d\theta$

The product of the first two densities in $Z(\lambda)$ equate to the unnormalised posterior with respect to the base prior. Therefore
\[
Z(\lambda) = \int_\Theta \int_\Omega \tilde{p}(\theta|y)f_\omega|\lambda(\omega)1(r(\theta) \leq \omega) \, d\omega \, d\theta
\]
or more simply
\[
Z(\lambda) = \tilde{Z} \Pr(r(\tilde{\theta}) \leq \tilde{\omega}|y, \lambda).
\]
If $\lambda \in \Lambda$ has a prior $\lambda \sim f_\lambda$ then we can remove the conditioning on $\lambda$ to obtain
\[
Z = \tilde{Z} \Pr(r(\tilde{\theta}) \leq \tilde{\omega}|y)
\]
by integrating both sides of (A.1) with respect to $f_\lambda$, where $\tilde{\omega}$ has marginal distribution $f_\tilde{\omega} = \int_\Lambda f_\omega|\lambda(\omega)f_\lambda(\lambda) \, d\lambda$
\[
Z = \int_\Lambda \int_\Theta \int_\Omega \pi(y|\theta)f_\theta(\theta)f_\omega|\lambda(\omega)1(r(\theta) \leq \omega) \, d\omega \, d\theta \, d\lambda
\]
and by noting that $\tilde{\theta}$ has no dependence on $\lambda$. [double check this assertion] \hfill \square

A.1.3 Corollary 3.9

Proof. Following notation and result in Appendix A.1.2 the marginal posterior expectation of $g(\theta)$ with respect to the stochastic constraint prior can be written as
\[
E(\theta|y) = Z^{-1} \int_\Theta \int_\Lambda \int_\Omega g(\theta) \pi(y|\theta)f_\theta(\theta)f_\omega|\lambda(\omega)1(r(\theta) \leq \omega) f_\lambda(\lambda) \, d\omega \, d\lambda \, d\theta
\]
\[
= Z^{-1} \int_\Theta \int_\Lambda g(\theta) \pi(y|\theta)f_\theta(\theta) \int_\Omega f_\omega|\lambda(\omega)1(r(\theta) \leq \omega) \, d\omega \, f_\lambda(\lambda) \, d\lambda \, d\theta
\]
\[
= \frac{\tilde{Z}}{Z} \int_\Theta g(\theta) \frac{\pi(y|\theta)f_\theta(\theta)}{Z} \int_\Lambda \Pr(r(\theta) \leq \tilde{\omega}|\lambda) \, d\lambda \, d\theta
\]
\[
= \frac{\int_\Theta g(\theta) \Pr(r(\theta) \leq \tilde{\omega}|\lambda) \tilde{p}(\theta|y) \, d\theta}{\Pr(r(\theta) \leq \tilde{\omega}|y)}
\]
\[
= \frac{E\left[g(\theta) \Pr(r(\theta) \leq \tilde{\omega}|\lambda)\right]|y}{\Pr(r(\theta) \leq \tilde{\omega}|y)}
\]
as required. The same proof can be constructed for $E(\theta|y, \lambda)$ by removing the integral related to $\lambda$. \hfill \square
A.1. PROOFS EXCLUDED FROM MAIN TEXT

A.1.4 Lemma 3.10

**Proof.** Following from definition and notation in Theorem 3.8, first note that the conditional unregularised posterior \( p(\theta \mid y, \lambda) \) can be written as is equivalent to

\[
p(\theta \mid y, \lambda) = Z(\lambda)^{-1} \int_{\Theta} \pi(y \mid \theta) f_\theta(\theta) f_{\lambda}(\omega) 1(r(\theta) \leq \omega) \, d\omega
\]

\[
= Z(\lambda)^{-1} \pi(y \mid \theta) f_\theta(\theta) \Pr(r(\theta) \leq \omega \mid \theta, \lambda).
\]

The normalising constant of the unregularised posterior, \( \tilde{Z} \), can be written as

\[
\tilde{Z} = \int_{\Theta} \pi(y \mid \theta) f_\theta(\theta) \, d\theta
\]

\[
= Z(\lambda) \int_{\Theta} \frac{\pi(y \mid \theta) f_\theta(\theta) \Pr(r(\theta) \leq \omega \mid \theta, \lambda)}{Z(\lambda)} \Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \, d\theta
\]

\[
= Z(\lambda) \int_{\Theta} p(\theta \mid y, \lambda) \Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \, d\theta
\]

\[
= Z(\lambda) \mathbb{E}(\Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \mid y, \lambda)
\]

Therefore

\[
\frac{\tilde{Z}}{Z(\lambda)} = \mathbb{E}(\Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \mid y, \lambda)
\]  

(A.2)

as required. \( \square \)

A.1.5 Theorem 3.11

**Proof.** Proceed as in previous proofs under conditions of Theorem 3.8 and by using \( p(\theta \mid y, \lambda) \) in Appendix A.1.4. The expectation of \( g(\theta) \) with respect to the unregularised posterior is

\[
\mathbb{E}(g(\theta) \mid y) = \tilde{Z}^{-1} \int_{\Theta} g(\theta) \pi(y \mid \theta) f_\theta(\theta) \, d\theta
\]

\[
= \frac{Z(\lambda)}{\tilde{Z}} \int_{\Theta} g(\theta) \frac{\pi(y \mid \theta) f_\theta(\theta) \Pr(r(\theta) \leq \omega \mid \theta, \lambda)}{Z(\lambda)} \Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \, d\theta
\]

\[
= \frac{Z(\lambda)}{\tilde{Z}} \int_{\Theta} g(\theta) p(\theta \mid y, \lambda) \Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \, d\theta
\]

By using (A.2) we can therefore write

\[
\mathbb{E}(g(\theta) \mid y) \mathbb{E}(\Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \mid y, \lambda) = \mathbb{E}\left[g(\theta) \Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \mid y, \lambda\right]
\]

(A.3)

which results in the first part of Theorem 3.11 after rearrangement. The second part follows by using the law of iterated expectation on (A.3) with \( \lambda \sim f_\lambda \) with result

\[
\mathbb{E}(g(\theta) \mid y) = \frac{\mathbb{E}\left[g(\theta) \Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \mid y\right]}{\mathbb{E}(\Pr(r(\theta) \leq \omega \mid \theta, \lambda)^{-1} \mid y)}.
\]

\( \square \)
A.1.6 Corollary 3.12

Proof. For a local stochastic constraint prior, with base distribution resulting in a posterior where each \((\tilde{\theta}_i \mid y, \lambda)\) is independent, the normalising constant can be expressed as

\[
\Pr(r(\theta) \leq \tilde{\omega} \mid y, \lambda) = \prod_{i=1}^{p} \Pr(r_i(\tilde{\theta}_i) \leq \tilde{\omega}_i \mid y, \lambda) = \prod_{i=1}^{p} \mathbb{E}(F_{\tilde{\omega}_i \mid \lambda}(r_i(\tilde{\theta}_i)) \mid y, \lambda). \tag{A.4}
\]

If \(g(\theta) = g(\theta_j)\) then from Corollary 3.9 and (A.4) we have

\[
\mathbb{E}(g(\theta_j) \mid y, \lambda) = \frac{\mathbb{E} \left[ g(\tilde{\theta}_j) \prod_{i \neq j}^{p} F_{\tilde{\omega}_i \mid \lambda}(r_i(\tilde{\theta}_i)) \mid y, \lambda \right]}{\prod_{i=1}^{p} \mathbb{E}(F_{\tilde{\omega}_i \mid \lambda}(r_i(\tilde{\theta}_i)) \mid y, \lambda)} = \frac{\mathbb{E} \left[ g(\tilde{\theta}_j) F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j)) \mid y, \lambda \right]}{\prod_{i=1}^{p} \mathbb{E}(F_{\tilde{\omega}_i \mid \lambda}(r_i(\tilde{\theta}_i)) \mid y, \lambda)} = \frac{\mathbb{E} \left[ g(\tilde{\theta}_j) F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j)) \mid y, \lambda \right]}{\mathbb{E}(F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j)) \mid y, \lambda)}.
\]

From which we can derive

\[
\mathbb{E} \left[ h(\tilde{\theta}_j) \mid y, \lambda \right] = \mathbb{E}(F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j)) \mid y, \lambda) \mathbb{E}(h(\theta_j) F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j))^{-1} \mid y, \lambda) \tag{A.5}
\]

by letting \(g(\theta_j) = h(\theta_j) F_{\tilde{\omega}_j \mid \lambda}(r_j(\theta_j))^{-1}\), and

\[
\mathbb{E}(F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j)) \mid y, \lambda) = \mathbb{E}(F_{\tilde{\omega}_j \mid \lambda}(r_j(\theta_j))^{-1} \mid y, \lambda)^{-1} \tag{A.6}
\]

by letting \(h(\theta_j) = 1\), and finally

\[
\mathbb{E} \left[ h(\tilde{\theta}_j) \mid y, \lambda \right] \mathbb{E}(F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j))^{-1} \mid y, \lambda) = \mathbb{E}(h(\theta_j) F_{\tilde{\omega}_j \mid \lambda}(r_j(\theta_j))^{-1} \mid y, \lambda) \tag{A.7}
\]

by combining (A.5) and (A.6). We can also write the equality without conditioning on \(\lambda\) by noting that \(\mathbb{E} \left[ h(\tilde{\theta}_j) \mid y \right] = \mathbb{E} \left[ h(\tilde{\theta}_j) \mid y, \lambda \right]\) and using the law of iterated expectations with (A.7) to get

\[
\mathbb{E} \left[ h(\tilde{\theta}_j) \mid y \right] \mathbb{E}(F_{\tilde{\omega}_j \mid \lambda}(r_j(\tilde{\theta}_j))^{-1} \mid y) = \mathbb{E}(h(\theta_j) F_{\tilde{\omega}_j \mid \lambda}(r_j(\theta_j))^{-1} \mid y) \tag{A.8}
\]

Defining \(\kappa_{h,i} = \mathbb{E}(h(\theta_i) \mid y) / \mathbb{E}(h(\tilde{\theta}_i) \mid y)\), if the relevant expectations exist, it can be written as

\[
\kappa_{h,i} = \frac{\mathbb{E}(h(\theta_i) \mid y) \mathbb{E} \left[ F_{\tilde{\omega}_i \mid \lambda}(r_i(\theta_i))^{-1} \mid y \right]}{\mathbb{E} \left[ h(\theta_i) F_{\tilde{\omega}_i \mid \lambda}(r_i(\theta_i))^{-1} \mid y \right]} = \frac{\mathbb{E}(h(\theta_i) \mid y) \mathbb{E} \left[ F_{\tilde{\omega}_i \mid \lambda}(r_i(\theta_i))^{-1} \mid y \right]}{\mathbb{E} \left[ h(\theta_i) \mid y \right] \mathbb{E} \left[ F_{\tilde{\omega}_i \mid \lambda}(r_i(\theta_i))^{-1} \mid y \right]} = \frac{\mathbb{E} \left[ h(\theta_i) \mid y \right] \mathbb{E} \left[ F_{\tilde{\omega}_i \mid \lambda}(r_i(\theta_i))^{-1} \mid y \right]}{\mathbb{E}(h(\theta_i) \mid y) \mathbb{E} \left[ F_{\tilde{\omega}_i \mid \lambda}(r_i(\theta_i))^{-1} \mid y \right] + 1}.
\]
using (A.8) and the identity \( \text{Cov}(x,y) = \mathbb{E}(xy) - \mathbb{E}(x)\mathbb{E}(y) \) where \( x \) and \( y \) are two random variables. We can also write \( \kappa_{h,i} \) in terms of the coefficient of variation

\[
\kappa_{h,i} = (\hat{\text{cv}}[\hat{h}(\theta_i)] \hat{\text{cv}}[\hat{w}_i|\lambda(r_i(\theta_i))]^{-1} \hat{\text{cor}}[\hat{h}(\theta_i), \hat{F}_i|\lambda(r_i(\theta_i))]^{-1} + 1)^{-1}
\]

where \( \hat{\text{cv}}[g(\theta_i)] = \sqrt{\mathbb{E}(g(\theta_i)^2|y)/\mathbb{E}(g(\theta_i)|y)} \) and \( \hat{\text{cor}}[g(\theta_i), m(\theta_i)] = \text{cor}[g(\theta_i), m(\theta_i)|y] \).

### A.1.7 Theorem 5.3

**Proof.** Consider the global stochastic constraint distribution

\[
(\theta, \psi \mid k) \overset{d}{=} (\tilde{\theta}, \tilde{\psi} \mid r(\theta; k) \leq \tilde{\psi}, k)
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}}, \quad \tilde{\theta} \in \Theta \subseteq \mathbb{R}^p
\]

\[
\tilde{\psi} \sim f_w, \quad \tilde{\psi} \in W \subseteq \mathbb{R}_+
\]

with separable penalty function \( r(\theta; k) \geq 0 \) defined as

\[
r(\theta; k) = q_1(\theta_1; k) + q_2(\theta_2; k) + \cdots + q_p(\theta_p; k)
\]

where \( q_i(\theta_i; k) \geq 0 \) also. The corresponding marginal density is

\[
f_{\theta | k} \propto f_{\tilde{\theta}}(\theta)\tilde{F}_w[r(\theta; k)].
\]

(A.10)

Now consider a local stochastic constraints prior which uses each separable element of \( r(\theta; k) \) as penalty functions

\[
(\theta, \omega \mid k) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} \mid q_i(\tilde{\theta}_i; k) \leq \tilde{\omega}_i \text{ for } 1 \leq i \leq p, k)
\]

\[
\tilde{\theta} \sim f_{\tilde{\theta}}, \quad \tilde{\theta} \in \Theta \subseteq \mathbb{R}^p
\]

\[
\tilde{\omega} \overset{\text{iid}}{\sim} f_w, \quad \tilde{\omega}_i \in W \subseteq \mathbb{R}_+, 1 \leq i \leq p
\]

(A.11)

which has marginal density

\[
f_{\theta | k} \propto f_{\tilde{\theta}}(\theta) \prod_{i=1}^p \tilde{F}_w[q_i(\tilde{\theta}_i; k)].
\]

(A.12)

In order for the global and local marginal distributions to be equivalent, i.e. (A.10) and (A.12) respectively, then

\[
\tilde{F}_w[r(\theta; k)] = C \prod_{i=1}^p \tilde{F}_w(q_i(\tilde{\theta}_i; k))
\]

(A.13)

must be true for some positive constant \( C \). However, given the equality must hold for \( r(\theta; k) = 0 \) and hence \( q_i(\tilde{\theta}_i; k) = 0 \), for \( 1 \leq i \leq p \), we can deduce \( C = 1 \).
To prove \( \tilde{\psi}, \tilde{\omega}_1, \tilde{\omega}_2, \ldots, \tilde{\omega}_p \overset{\text{iid}}{\sim} \text{Exp}(k) \) if (A.11) exists such that (A.13) is true for \( C = 1 \), the following identity is useful.

\[
\tilde{F}_w[r(\theta; k)] = \Pr[w \geq r(\theta; k)] = \Pr[w \geq q_1(\tilde{\theta}_1; k)] \Pr[w \geq q_1(\tilde{\theta}_1; k)]
\]

\[
= \Pr\left[w \geq \prod_{i=2}^{p} q_i(\tilde{\theta}_i; k) \right] \Pr[w \geq q_1(\tilde{\theta}_1; k)]
\]

\[
= \Pr\left[w \geq \prod_{i=2}^{p} q_i(\tilde{\theta}_i; k) \mid w \geq q_2(\tilde{\theta}_2; k) \right] \Pr[w \geq q_2(\tilde{\theta}_2; k)] \Pr[w \geq q_1(\tilde{\theta}_1; k)]
\]

\[
= \Pr\left[w \geq \prod_{i=3}^{p} q_i(\tilde{\theta}_i; k) \right] \Pr[w \geq q_2(\tilde{\theta}_2; k)] \Pr[w \geq q_1(\tilde{\theta}_1; k)]
\]

\[
: \quad = \prod_{i=1}^{p} \Pr[w \geq q_i(\tilde{\theta}_i; k)] = \prod_{i=1}^{p} F_w[q_i(\tilde{\theta}_i; k)]
\]

if and only if \( w \) has exponential distribution (only considering continuous distributions). This identity is due to repeated application of the memoryless property, which is unique to the exponential in the class of continuous distributions. Therefore, for (A.13) to be true \( w \) must be exponentially distributed.

The converse follows by noting that if \( w \sim \text{Exp}(t) \) then (A.13) holds so the marginal distributions in (A.10) and (A.12) are equivalent. \( \square \)

A.1.8 Proposition 6.1

**Proof.** Consider the normal-means model with likelihood

\[
\pi(y|\theta) \propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2}(y_i - \theta_i)^2\right\}
\]

and priors given by

\[
f_{\theta_i|\lambda_i} \propto \exp\left\{-\frac{1}{2} \frac{(\theta_i - \mu_i)^2}{\lambda_i^2}\right\}, \text{ and}
\]

\[
f_{\lambda_i} \propto (1 + \lambda_i^2)^{-1}
\]

where \( y \in \mathbb{R}^p \), \( \theta_i \in \mathbb{R}^p \), and \( \lambda \in \mathbb{R}^+_p \). The conditional posterior distribution
for \((\theta_i \mid \lambda_i)\) is therefore

\[
\pi(\theta_i \mid y, \lambda_i) = \mathcal{N}(\theta_i; \bar{\mu}_i, 1 - k_i)
\]

with \(\bar{\mu}_i = (1 - k_i)y_i + k_i \mu_i\)
and \(k_i = (\lambda_i^2 + 1)^{-1}\)

for \(1 \leq i \leq p\).

Under the same transformation, namely \(k_i = (\lambda_i^2 + 1)^{-1}\) with normalising term \(|\frac{\partial \lambda_i}{\partial k_i}| = \frac{1}{2}(1 - k_i)^{-0.5}k_i^{-1.5}\), the implied prior distribution for \(k_i\) has density proportional to

\[
f_{k_i} \propto k_i^{-1/2}(1 - k_i)^{-1/2},
\]

or in other words \(k_i \sim \text{Beta}(1/2, 1/2)\).

The conditional posterior expectation of \(\theta_i\) is then

\[
E(\theta_i \mid y, k_i) = \bar{\mu}_i = (1 - k_i)y_i + k_i \mu_i.
\]

We can recover the result for the original horseshoe by taking \(\mu_i = 0\). \qed

### A.1.9 Proposition 6.2

**Proof.** Consider the normal-means model with likelihood

\[
\pi(y \mid \theta) \propto \prod_{i=1}^{n} \exp \left\{ -\frac{1}{2} (y_i - \theta_i)^2 \right\}
\]

and priors given by the stochastic constraint hierarchy

\[
(\theta, \omega \mid \lambda) \overset{d}{=} (\tilde{\theta}, \tilde{\omega} \mid \lambda, \tilde{\theta}_i^2 \leq \tilde{\omega}_i, i = 1, 2, \ldots, p)
\]

\[
\tilde{\theta}_i \sim \mathcal{N}(\mu_i, \phi_i), i = 1, 2, \ldots, p
\]

\[
(\tilde{\omega}_i \mid \lambda_i) \sim \text{Exp}(2^{-1}[\lambda_i]^{-2}), i = 1, 2, \ldots, p
\]

\[
\lambda \overset{\text{iid}}{\sim} \text{Ca}_+(1)
\]

i.e. the stochastic constraint horseshoe prior as in (2.21) with flat base prior replaced with iid Gaussian base prior and \(\tau = \sigma = 1\). The marginal prior for each \(\theta_i \mid \lambda_i\) can be written as

\[
f_{\theta_i \mid \lambda_i} \propto \exp \left\{ -\frac{1}{2} \frac{(\theta_i - \mu_i)^2}{\sigma_i^2} \right\} \exp \left\{ -\frac{\theta_i^2}{2\lambda_i^2} \right\}
\]

and the prior for \(\lambda_i\) has density \(f_{\lambda_i} \propto (1 + \lambda_i^2)^{-1}\). The conditional posterior distribution has density proportional to

\[
\pi(\theta_i \mid y, \lambda_i) \propto \exp \left\{ -\frac{1}{2} \left[ (y_i - \theta_i)^2 + \frac{(\theta_i - \mu_i)^2}{\phi_i} + \frac{\theta_i^2}{\lambda_i^2} \right] \right\}
\]
which implies a normal distribution
\[
\pi(\theta_i|y, \lambda_i) = N(\theta_i; \mu_i, \sigma_i^2)
\]
with \(\mu_i = \frac{(1 - k_i)(y_i + \mu_i\phi^{-1})}{1 + (1 - k_i)\phi^{-1}},\)
\(\sigma_i^2 = \frac{1 - k_i}{1 + (1 - k_i)\phi^{-1}},\)
and \(k_i = (\lambda_i^2 + 1)^{-1}\)
for \(1 \leq i \leq p\). The implied prior for \(k_i\) under this transformation is \(k_i \sim \text{Beta}(1/2, 1/2)\) as seen in Appendix A.1.8. The conditional posterior expectation is therefore
\[
E(\theta_i | y, k_i) = \frac{(1 - k_i)(y_i + \mu_i\phi^{-1})}{1 + (1 - k_i)\phi^{-1}}.
\]

\[\]

A.2 Further details on Monte Carlo samplers

A.2.1 Gibbs sampler for horseshoe prior from stochastic constraints

An example of a Gibbs sampler for the horseshoe prior formulated as a stochastic constraint prior is given below. The horseshoe prior with linear model can be written as
\[
(y|X, \beta, \sigma^2) \sim N(X\beta, \sigma^2 I_n)
\]
\[
(\beta_i, \omega_i | \lambda_i, \tau, \sigma^2) \sim \text{Exp}(\frac{2\omega_i}{\lambda_i^2\tau^2\sigma^2})\text{ I}(\beta_i^2 \leq \omega_i)
\]
\[
\sigma^2 \sim \sigma^{-2} d\sigma^2
\]
\[
\lambda_i \sim \text{Ca}(1)
\]
\[
\tau \sim \text{Ca}(1)
\]
for \(1 \leq i \leq p\), using a stochastic constraints representation of the horseshoe. If \(X^\top X\) has an inverse then the following full conditionals define a Gibbs sampler for the horseshoe:
\[
(\beta | \cdot) \sim N(\hat{\beta}, \sigma^2(X^\top X)^{-1}) \text{ s.t. } \beta^2 \leq \omega \text{ where } \hat{\beta} = (X^\top X)^{-1}X^\top y
\]
\[
(\omega_i | \cdot) \sim \text{Exp}(2^{-1}(\lambda_i\tau\sigma)^{-2}) \text{ s.t. } \omega_i \geq \beta_i^2
\]
\[
(\sigma^2 | \cdot) \sim \text{IGam} \left( \frac{n + 3p}{2} ; \frac{1}{2}(y - X\beta)^\top(y - X\beta) + \frac{1}{2\tau^2} \sum_{i=1}^{p} \frac{\omega_i}{\lambda_i^2} \right)
\]
\[
(\lambda_i | \cdot) \sim f_{\lambda_i} \propto \frac{\exp\{-k_i/\lambda_i^2\}}{\lambda_i^3(1 + \lambda_i^2)}
\]
\[
(\tau | \cdot) \sim f_{\tau} \propto \frac{\exp\{-q/\tau^2\}}{\tau^{3p}(1 + \tau^2)}
\]
where \( k_i = \omega_i / (2 \tau^2 \sigma^2) \) and \( q = \sum_{i=1}^{\infty} \frac{\omega_i}{\lambda^2} \).

Sampling the conditional distribution for \( \lambda_i \) is relatively straightforward. We require a sample, say \( \lambda \) from distribution 
\[
 f_\lambda \propto \exp\left\{ -k / \lambda^2 \right\} \lambda^3 (1 + \lambda^2).
\]

Consider the distribution with density and cumulative distribution functions
\[
 f_z(z) = \frac{z(z+1)^{-1} \exp\{-kz\}}{k^{-1} - \exp\{k\} E_1(k)},
\]

\[
 F_z(z) = 1 - \frac{\exp\{-kz\} - k \exp\{k\} E_1(k + kz)}{1 - k \exp\{k\} E_1(k)}
\]

for \( z \geq 0 \) and \( k > 0 \), where \( E_1(x) = \int_x^\infty \exp\{-t\}/t \, dt \) for \( x > 0 \), (a representation of the exponential integral on the positive real line). A random variable with distribution \( z \) is equivalent to that of \( \lambda \) by the transformation
\[
 \lambda = z^{-2}.
\]

So to draw from \( \lambda \) we may draw from \( z \) and then use the transformation. To draw a random sample from \( z \), we recommend the inverse CDF method. Although a numerical solver is required in this case, we can increase the speed by deriving bounds for the CDF for a given \( k \). By noting that \( E_1(k + kz) \leq E_1(k) \) as \( z \geq 0, k \geq 0 \) and the exponential integral is a decreasing function, we can show that the CDF has upper bound
\[
 F_z(z) \leq 1 - \frac{1 - \exp\{-kz\}}{1 - k \exp\{k\} E_1(k)}. \tag{A.14}
\]

On the other hand, \( E_1(x) > 0 \) for \( x \geq 0 \) so a lower bound for the CDF is
\[
 F_z(z) \geq 1 - \frac{\exp\{-kz\}}{1 - k \exp\{k\} E_1(k)}. \tag{A.15}
\]

Now, when solving \( z = \{x \geq 0 : y = F_z(x)\} \) where \( y \sim U(0,1) \), the solution \( z \) has bounds
\[
 -\frac{\log(1 - Ky)}{k} \leq z < -\frac{\log K + \log(1 - y)}{k}
\]

where \( K = 1 - k \exp\{k\} E_1(k) \)

by using the inequalities in (A.14) and (A.15).

Using the same transformation for the conditional density of \( \tau \) is not as fruitful as the case was for the \( \lambda_i \). The density function is equal to
\[
 f_{|\tau|} = \frac{2 \exp\{-q/\tau^2 - q\} \tau^{-3p/2} (1 + \tau^2)^{-1}}{\Gamma(3p/2 + 1/2) \int_q^\infty t^{(3p-1)/2-1} \exp(-t) \, dt}
\]
Therefore, we use the decomposition given in Makalic and Schmidt (2016) for $\tau$, where the positive Cauchy prior on $\tau$ is decomposed into the hierarchy

\[
(\tau^2 \mid \eta) \sim \text{IGam}(1/2, 1/\eta) \\
\eta \sim \text{IGam}(1/2, 1)
\]

which gives rise to the following additional full conditional distributions for the Gibbs sampler

\[
(\tau^2 \mid \cdot) \sim \text{IGam}(3(p - 1)/2, 1/\eta + q/2) \\
(\eta \mid \cdot) \sim \text{IGam}(1/2, 1 + 1/\tau^2).
\]

When $p \geq n$ we can use a draw from the Zig-zag sampler with switching times described in Appendix A.2.3, to replace full conditional for $\beta$.

### A.2.2 Power-adjusted inverse Gamma distribution

To sample from a univariate random variable, say $x$, with density equal to

\[
f_x(x) = Z^{-1} x^{-a-1} \exp \left\{ -\frac{b}{x} - \frac{c}{x^\alpha} \right\}
\]

where $a, b, c, \alpha > 0$ and $Z$ is the normalising constant, a rejection sampler may be appropriate. Let $y \sim \text{IGam}(a, b)$ with density

\[
f_y = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp \left\{ -\frac{b}{x} \right\}
\]

be the enveloping distribution. The bounding constant $M$ is

\[
M = \max_{x \in \mathbb{R}_+} \frac{f_x(x)}{f_y(x)} = \max_{x \in \mathbb{R}_+} \frac{\Gamma(a) \exp\left\{ -\frac{c}{x^\alpha} \right\}}{b^a Z} = \frac{\Gamma(a)}{b^a Z}.
\]

When $\alpha = 1/2$, the normalising constant, $Z$, for the target distribution is

\[
Z = \frac{2^{1-2a} \Gamma(2a)U\left(a, \frac{1}{2}, \frac{c^2}{4b}\right)}{b^a}
\]

where $U(a, b, z)$ is the confluent hypergeometric function (of the second kind). In this case the bounding constant $M$ simplifies to

\[
M = \frac{\sqrt{\pi}}{\Gamma\left(\frac{1}{2} + a\right)U\left(a, \frac{1}{2}, \frac{c^2}{4b}\right)}.
\]

In most applications of interest, this constant, which implies an acceptance rate of $1/M$ rejects too frequently. For example, as $a \to \infty$ the confluent hypergeometric function $U(a, b, z) \to 0$ for fixed $b$ and $z$, so the acceptance rate goes to zero.
Given we are often interested in high-dimensional problems, where \( a = \mathcal{O}(p) \) (see Example 4.1), this acceptance rate is not appropriate. Instead of a rejection sampler we can use a further stochastic constraint decomposition (an latent-variable approach) for \( f_x \). The joint distribution \((x, w)\) with density

\[
f_{x,w} = \frac{1}{c^2} x^{-a-1} \exp \left\{ -\frac{b}{x} \right\} \exp \left\{ -\frac{w}{c} \right\} 1(x^{-\alpha} \leq w)
\]

has the same marginal distribution as \( x \). It is a stochastic constraint random variable with inverse Gamma base, exponential constraint, and penalty function \( x^{-\alpha} \). The Gibbs sampler is simply

\[
(x | w) \sim \text{IGam}(a, b), \quad \text{s.t. } x^{-\alpha} \leq w
\]

\[
(w | x) = t + x^{-\alpha} \quad \text{where}
\]

\[
t \sim \text{Exp}(c^{-1})
\]

which can be transformed to use a right truncated Gamma distribution (rather than inverse) to get the Gibbs sampler

\[
(x | w) = y^{-1} \quad \text{where}
\]

\[
y \sim \text{Gam}(a, b), \quad \text{s.t. } y \leq w^{1/\alpha}
\]

\[
(w | x) = t + y^{\alpha} \quad \text{where}
\]

\[
t \sim \text{Exp}(c^{-1}).
\]

This is simply an application of slice sampling equivalent to using a stochastic constraint decomposition.

A.2.3 Poisson process switching times for Gaussian(-like) distributions Let the density we wish to simulate be \( f_\theta \) be

\[
f_\theta \propto \exp\{- (\theta^T A \theta - 2b^T \theta + c)/2\}, \quad \theta \in S \subseteq \mathbb{R}^p
\]  

(A.16)

and assume the support \( S \) is such that the density is proper. If \( S = \mathbb{R}^p \) and the exponent of \( f_\theta \) can be factored into \( (\theta - \mu)^T \Sigma (\theta - \mu) \) where \( \Sigma \) is non-degenerate, then we have a multivariate Gaussian distribution, but we work with (A.16) to be more general.

We wish to construct a PDMP for (A.16). If \( S \subseteq \mathbb{R}^p \), assume the boundary kernel has been chosen correctly. The gradient of the negative log-density is

\[
\nabla u(\theta) = -\nabla \log f_\theta = A \theta - b
\]

which, for the Zig-Zag sampler, results in an intensity function for the \( i \)th Poisson process

\[
\lambda_i(s; \theta, v) = \max\{0, v_i \nabla_i u(\theta + sv)\}
\]

\[
= \max\{0, v_i ([A(\theta + sv)]_i - b_i)\}
\]

\[
= \max\{0, v_i (s[A v]_i - b_i + [A \theta]_i)\}
\]

\[
= \max\{0, sm_i + n_i\}
\]
where \( m_i = v_i[A\mathbf{v}]_i, \ n_i = v_i([A\theta]_i - b_i) \), and \( \mathbf{v} \) is the current velocity of the PDMP and \([\ ]_i\) extracts the \( i \)th element of the vector. To simulate the first arrival time we need to solve

\[
 z > 0 \text{ s.t. } \int_0^z \lambda_i(s; \theta, \mathbf{v}) \, ds = y \tag{A.17}
\]

for a given \( y \), which is simulated with an exponential with unit rate. Let \( r_{i*} = \max\{0, -n_i/m_i\} \), the solution can be worked out analytically by simplifying the integral

\[
\int_0^z \lambda_i(s; \theta, \mathbf{v}) \, ds = \int_0^z \max\{0, sm_i + n_i\} \, ds
\]

\[
= \begin{cases} 
\int_0^{r_{i*}} sm_i + n_i \, ds, & \text{if } m_i < 0 \\
\int_{r_{i*}}^z sm_i + n_i \, ds, & \text{if } m_i \geq 0
\end{cases}
\]

\[
= \begin{cases} 
\frac{m_i}{2} \min\{r_{i*}, z\}^2 + n_i \min\{r_{i*}, z\}, & \text{if } m_i < 0 \\
\frac{m_i}{2} (z^2 - r_{i*}^2) + n_i (z - r_{i*}), & \text{if } m_i \geq 0
\end{cases}
\]

\[
= \begin{cases} 
\frac{m_i}{2} z^2 + n_i z, & \text{if } m_i < 0 \text{ and } z < r_{i*} \\
\frac{m_i}{2} r_{i*}^2 + n_i r_{i*}, & \text{if } m_i < 0 \text{ and } z \geq r_{i*} \\
\frac{m_i}{2} (z^2 - r_{i*}^2) + n_i (z - r_{i*}), & \text{if } m_i \geq 0
\end{cases}
\]

for \( z > 0 \). Using this simplification to solve (A.17) we find that the first arrival time is

\[
\tau_i = \begin{cases} 
\min\{z \in (0, r_{i*}) : \frac{m_i}{2} z^2 + n_i z - y = 0\}, & \text{if } m_i < 0 \\
\min\{z \in (0, \infty) : \frac{m_i}{2} (z^2 - r_{i*}^2) + n_i (z - r_{i*}) - y = 0\}, & \text{if } m_i \geq 0
\end{cases}
\]

which can be solved using the quadratic formula and checking which solutions are smallest and satisfy the bounds on \( z \). If no solution exists, the first arrival time (or switch time) for the \( i \)th variable is infinite, and the \( i \)th velocity will not switch on the current trajectory.


**URL**: https://CRAN.R-project.org/package=mvtnorm


**URL**: https://doi.org/10.1080/10618600.2018.1482762

**URL**: https://CRAN.R-project.org/package=purrr


**URL**: http://mjskay.github.io/tidybayes/


**URL**: https://CRAN.R-project.org/package=GIGrvg


URL: https://CRAN.R-project.org/package=lasso2


URL: https://CRAN.R-project.org/package=tibble


URL: https://journal.r-project.org/archive/


URL: https://CRAN.R-project.org/package=scales

URL: https://CRAN.R-project.org/package=dplyr

URL: https://CRAN.R-project.org/package=tidyr

URL: https://CRAN.R-project.org/package=extraDistr

URL: https://yihui.name/knitr/


URL: https://CRAN.R-project.org/package=kableExtra