Bayesian Analysis (2025) **20**, Number 2, pp. 657–680

Posterior Shrinkage Towards Linear Subspaces

Daniel K. Sewell*

Abstract. It is common to hold prior beliefs that are not characterized by points in the parameter space but instead are relational in nature and can be described by a linear subspace. While some previous work has been done to account for such prior beliefs, the focus has primarily been on point estimators within a regression framework. We argue, however, that prior beliefs about parameters ought to be encoded into the prior distribution rather than in the formation of a point estimator. In this way, the prior beliefs help shape *all* inference. Through exponential tilting, we propose a fully generalizable method of taking existing prior information from, e.g., a pilot study, and combining it with additional prior beliefs represented by parameters lying on a linear subspace. We provide computationally efficient algorithms for posterior inference that, once inference is made using a non-tilted prior, does not depend on the sample size. We illustrate our proposed approach on an antihypertensive clinical trial dataset where we shrink towards a power law dose-response relationship, and on monthly influenza and pneumonia data where we shrink moving average lag parameters towards smoothness. Software to implement the proposed approach is provided in the R package `SUBSET` available on GitHub.

Keywords: exponential tilting; prior information; posterior inference.

1 Overview

Prior beliefs are often reflected in Bayesian analyses by shrinking estimates towards some point $\boldsymbol{\theta}_0$ of the parameter space Θ, such as those priors inducing sparsity (e.g., Park and Casella, 2008; Carvalho et al., 2010). However, at other times prior knowledge leads to beliefs that are not characterized by points in Θ but rather are relational in nature: In a regression framework we may believe a priori that the coefficients of a polychotomous ordinal factor covariate might reflect a linear shape; In a two sample binomial context, the response rates of the two corresponding populations may be believed to be near equal; When conducting an ANOVA we may believe a priori that the multiple populations have near homoscedastic responses. The prior beliefs in these examples and in other similar situations can be encoded by shrinking our estimates towards the intersection of the parameter space and a linear subspace.

Stein-type estimators have received much attention from statisticians since the landmark paper by James and Stein (James and Stein, 1961). Such estimators have been used repeatedly in regression settings to shrink the mean response curve towards a linear subspace contained within the span of the columns of the design matrix. An early application of this was the work by Blaker (1999), who developed Stein-type estimators that shrinks the mean response towards the space spanned by the principal components. More recently, Shin et al. (2020) and Wiemann and Kneib (2021) considered

*University of Iowa, 145 N. Riverside Dr., Iowa City, IA 52241, daniel-sewell@uiowa.edu

 https://doi.org/10.1214/24-BA1414

the context of spline regression. By applying what they termed a functional horseshoe prior, the authors were able to shrink the mean response curve towards a linear subspace (e.g., a simple polynomial relationship between y and X). While these results are exciting, they are limited to regression settings, only shrink the mean of the response vector, and require Gaussianity of the prior, which may be singular. Huber and Koop (2021) developed a somewhat similar approach for vector autoregressive models with a focus on shrinking towards factor models. An et al. (2009) considered Stein-type estimators in the generalized linear model setting, proving some theoretical results in the frequentist paradigm; see references therein for other landmark papers on Stein-type estimators.

Deviating away from shrinking the mean response in regression settings, Oman (1982) novelly focused on the regression coefficients directly. Using an empirical Bayesian approach, Oman developed a point estimator for the regression coefficients that provided shrinkage of the least squares estimate towards its projection onto the linear subspace of interest. Lee and Birkes (1994) proposed a subspace ridge estimator for the regression coefficients, a generalized ridge estimator shrinking the estimates towards a linear subspace; Lee and Birkes also showed how their point estimator can be derived through a three-stage hierarchical Bayesian approach with improper priors. Hansen (2016) expanded beyond the scope of regression settings (although their work still applies in that context), proposing a general purpose point estimator that is a weighted average of the unconstrained maximum likelihood estimator (MLE) and the MLE restricted to the subspace of interest. Finally, Floto et al. (2022) applied an exponentially tilted Gaussian prior for deep neural networks.

Prior work in this realm of shrinkage towards linear subspaces have focused on point estimation, and nearly all on the frequentist properties of the proposed estimators. However, if a researcher wants to shrink their estimates towards a linear subspace, it is because there is *prior information* regarding the plausible values of the parameters. Therefore, taking a Bayesian stance is the most sensible approach, where such prior information can naturally be incorporated into the prior belief distribution over the parameters. This is in contrast to (1) ad hoc- even if reasonable- adjustments of frequentist point estimates, and (2) loss functions that act to shrink posterior-based point estimates towards a subspace.

The purpose of this paper is to provide the Bayesian practitioner highly generalizable, easily implemented, and computationally efficient methods to incorporate prior information which can be encoded by shrinking towards a linear subspace. Our approach applies shrinkage on any parameters of interest, unlike some prior work that focuses solely on the mean structure, and while certainly applicable to regression settings, it can be applied to any parametric setting. We emphasize that our approach *incorporates* additional information into the prior, rather than *replaces* other prior information with a specific prior distribution that conveniently induces shrinkage. Further, our approach can be applied to any proper prior, not just Gaussian priors. By adjusting the prior- and hence the posterior- rather than the point estimator, all posterior inferential statements account for prior information regarding the linear subspace.

The remainder of the paper is as follows. In Section 2, we discuss our proposed method to exponentially tilt an existing prior to append a priori knowledge or beliefs

about unknown parameters lying in or near linear subspaces. In Section 3 we provide computationally efficient methods of obtaining posterior inference under the tilted prior, including methods for when the linear subspace itself is not fully known. In Section 4 we evaluate our method in a simulation study. In Section 5, we illustrate our methods on two real datasets: a clinical trial evaluating antihypertensive drug treatment where we shrink towards a power law dose-response relationship, and on monthly influenza and pneumonia data where we shrink moving average lag parameters towards smoothness. Section 6 provides a brief summary and discussion.

2 Exponentially tilted priors

2.1 Introduction

Consider the typical Bayesian data analysis set up: Let y denote the observed data with corresponding likelihood $\pi(y|\theta)$ parameterized by some p-dimensional unknown parameter vector $\theta \in \Theta \subseteq \Re^p$. Let $\pi_0(\theta)$ denote the *base prior density*, which reflects our prior beliefs about plausible values of θ.

While the practitioner is usually a practiced hand at setting up a prior distribution to reflect information on location or degree of uncertainty about the true value of θ, often the information we feel more assured about is much more difficult to encode in π_0. Consider the simplest of examples: if we believe a priori that a placebo mean will be 1 and a treatment mean will be 5, then we can center our prior π_0 on $\theta = (\theta_{placebo}, \theta_{trmnt})$ at $(1, 5)$. But what ought one to do if, on the other hand, we have little or no prior information on the exact values of the treatment effects but believe that the placebo will likely have some non-zero effect and that the treatment arm will have 5 times the effect? In such a case we believe that there ought to be some x such that the true θ isn't too far from $(x, 5x)$; in other words, we believe the true θ should be somewhere around the linear subspace defined by the span of $(1, 5)'$.

In the simplest cases such as the example given above, it may be tempting to consider a reparameterization (e.g., $\theta_{trmnt} = 5\theta_{placebo}$ or $\theta_{trmnt} = 5\theta_{placebo} + noise$). However, there are distinct disadvantages to using reparameterization as a general approach to encoding relational information on the parameters. First, because of the lack of generalizability of reparameterization, determining how to perform estimation must occur on a case-by-case basis, and in some instances estimation may prove challenging or computationally onerous. Second, it is not obvious how to integrate relational information through a reparameterization with other non-relational prior information, such as that obtained from a pilot study or literature review. Third, it may not be possible to perform such a reparameterization when the parameter space is bounded. Fourth, reparameterization is often a hard constraint that may not be correctly specified; this is equivalent to a degenerate prior which violates Cromwell's Rule which, when misspecified, cannot retrieve the true parameter values regardless of sample size. In what follows, we propose a fully generalizable approach that overcomes these issues.

2.2 SUBSET priors

Consider the general case where there is some $L \in \Re^{p\times q}$, such as a design matrix (e.g., $L = (1, 1, -1)'$ with $q = 1$ in Example 1 below), that gives rise to a linear subspace $\widetilde{\mathcal{L}} := \text{span}(L)$, and we believe a priori that θ lies on or near $\mathcal{L} := \widetilde{\mathcal{L}} \cap \Theta$. We aim, then, to add this knowledge to our other prior information on θ, i.e., take our base prior π_0 and adjust it in such a way so as to put smaller prior probability over regions away from $\mathcal{L}$. Towards that, we propose using the following exponentially tilted prior π_ν:

$$\pi_\nu(\theta) := \frac{1}{Z_{\nu,\phi}}\pi_0(\theta)e^{-\frac{\nu}{2}\theta'(I_p-P(\phi))\theta}, \tag{1}$$
$$\text{where } Z_{\nu,\phi} := \mathbb{E}_{\pi_0}\left(e^{-\frac{\nu}{2}\theta'(I_p-P(\phi))\theta}\right),$$

and $P(\phi)$ is the projection matrix associated with $\widetilde{\mathcal{L}}$, which may depend on a user-specified hyperparameter ϕ. (In 3.2 we will discuss estimating ϕ, but until that time ϕ will be assumed to be a known constant and will not play a role. Thus, we will drop ϕ from the notation until that time.) Eq. (1) shows concretely how we achieve our aim: π_ν takes the overall shape determined by the base prior π_0 and through the exponential tilting term penalizes π_0 in areas which lie on or near the orthogonal subspace of $\mathcal{L}$ (see Figure 1b), implicitly then upweighting areas near $\mathcal{L}$. We will henceforth refer to π_ν as the SUBSET (**SUB**space **S**hrinkage via **E**xponential **T**ilting) prior.

Proposition 1. *If π_0 is a valid probability density function (pdf), the SUBSET prior π_ν is also a valid pdf.*

The above proposition immediately follows from the fact that $0 < e^{-\frac{\nu}{2}\theta'(I_p-P)\theta} \leq 1$ $\forall\theta$. To illustrate the exponentially tilted prior, we provide three examples below.

Example 1. Suppose our base prior over $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ is a multivariate normal distribution centered at zero with spherical covariance matrix, i.e., $\boldsymbol{\theta} \sim N(\mathbf{0}, \tau I_3)$, where τI_3 is the precision matrix.[1] This base prior reflects where we believe $\boldsymbol{\theta}$ is centered at $(\mathbf{0})$, and the degree of uncertainty we have about $\boldsymbol{\theta}$ (τ). Further, suppose that we have reason to believe that θ_1, θ_2, and θ_3 are all roughly equal to each other in magnitude, but that θ_3 is of the opposite sign as θ_1 and θ_2. Then we have

$$\mathcal{L} = \text{span}\begin{pmatrix}1\\1\\-1\end{pmatrix}, \qquad P = \frac{1}{3}\begin{pmatrix}1 & 1 & -1\\1 & 1 & -1\\-1 & -1 & 1\end{pmatrix},$$

and the SUBSET prior is $\boldsymbol{\theta} \sim N\big(\mathbf{0}, \tau I_3+\nu(I_3-P)\big)$, so that $\text{Corr}(\theta_1, \theta_2) = \nu/(3\tau+\nu)$ and similarly $\text{Corr}(\theta_1, \theta_3) = \text{Corr}(\theta_2, \theta_3) = -\nu/(3\tau + \nu)$. We can control the magnitude of the correlation between the variables, then, by increasing or decreasing the exponential tilting parameter ν.

[1] In this paper, we will always denote a normal distribution in terms of its precision, rather than variance, so that $N(a, b)$ represents a normal distribution with mean a and variance b^{-1}.

Example 2. Suppose we are in a two-sample normal context, i.e., $y_{ij} \overset{ind}{\sim} N(\mu_i, 1/\sigma_i^2)$, $i = 1, 2$, $j = 1, \ldots, n_i$. In this setting it is common to assume homoscedasticity. This assumption, equivalent to setting $\Pr(\sigma_1^2 \neq \sigma_2^2) = 0$, is highly unlikely to be true in any real context, although it is most often reasonable to assume that there will be near homogeneity, that is, the two variances will be roughly, though not exactly, equal. Instead of making the homogeneity assumption, we can set a base prior on (σ_1^2, σ_2^2) and shrink this prior towards the linear subspace spanned by $(1, 1)'$. Figure 1 shows the bivariate base prior set as the product of two independent half-t distributions with 2 degrees of freedom (1a), how that base prior is rescaled due to the exponential tilting term in (1) with $\nu = 2$ (1b), and the corresponding SUBSET prior that shrinks towards homoscedasticity (1c). The hyperparameter ν controls the degree of shrinkage, which can be seen to dictate the a priori correlation between the two variances in (1d).

Example 3. Suppose we are in the two-sample binomial context, i.e., $y_i \sim Bin(n_i, p_i)$, $i = 1, 2$. We use as our base prior a product of independent Jeffreys reference priors, i.e., $Beta(1/2, 1/2)$. However, we believe a priori that it is likely that the response rate p_1 is twice that of p_2, and hence we wish to push our prior density mass away from regions in $(0, 1)^2$ that do not reflect this, i.e., we wish to shrink our prior probability towards $\text{span}((2, 1)') \cap (0, 1)^2$. Figure 2 shows the Jeffreys priors (a) and the SUBSET prior (b).

Example 4. Consider the heteroscedastic weighted regression model where for $i = 1, 2, \ldots, N$

$$y_i \overset{ind}{\sim} N(X_{1,i}\beta, 1/\sigma_i^2),$$
$$\sigma_i^2 = X_{2,i}\gamma,$$

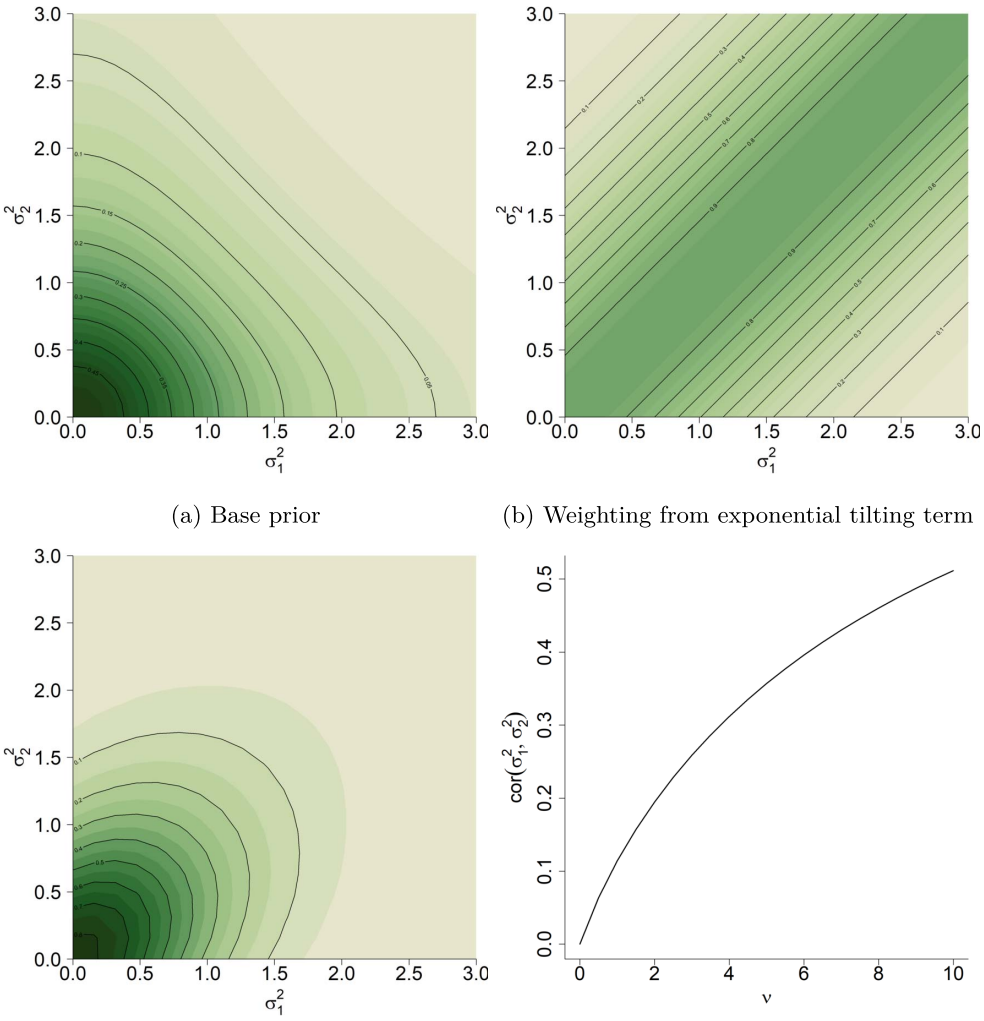
where $X_{1,i}$ and $X_{2,i}$ are covariates used to model the mean and variance structures respectively ($X_{1,i}$ may equal $X_{2,i}$). If we do not wish to make the linear relationship between the regression weights and the covariates $X_{2,i}$ a hard constraint, we may instead consider letting the σ_i^2 follow, e.g., an inverse gamma distribution, and use a SUBSET prior that shrinks $(\sigma_1^2, \ldots, \sigma_N^2)$ to the column space of $(X_{2,1}', \ldots, X_{2,N}')'$.

The next theorem demonstrates that the SUBSET prior leads to a higher concentration of posterior probability mass near the linear subspace everywhere along that subspace.

Theorem 2. *Let $\mathcal{L}$, P, π_0, and π_ν be as defined above. Define $\mathcal{B}_\epsilon(\mathcal{L}) := \{\theta : \theta'(I-P)\theta < \epsilon\}$. Let $\pi_0(\theta|y)$ and $\pi_\nu(\theta|y)$ denote the posterior density function of θ under the base prior and the SUBSET prior respectively, and let $\Pr_0(\theta \in \mathcal{A}|y)$ and $\Pr_\nu(\theta \in \mathcal{A}|y)$ denote the posterior probability that $\theta \in \mathcal{A}$ for some region $\mathcal{A} \subset \Theta$ under the base and SUBSET priors respectively.*

For any $\mathcal{L} \neq \emptyset$ and $\epsilon > 0$ such that

$$0 < \int_{\mathcal{B}_\epsilon} \pi_j(\theta|y)d\theta < 1, \qquad j \in \{0, \nu\},$$



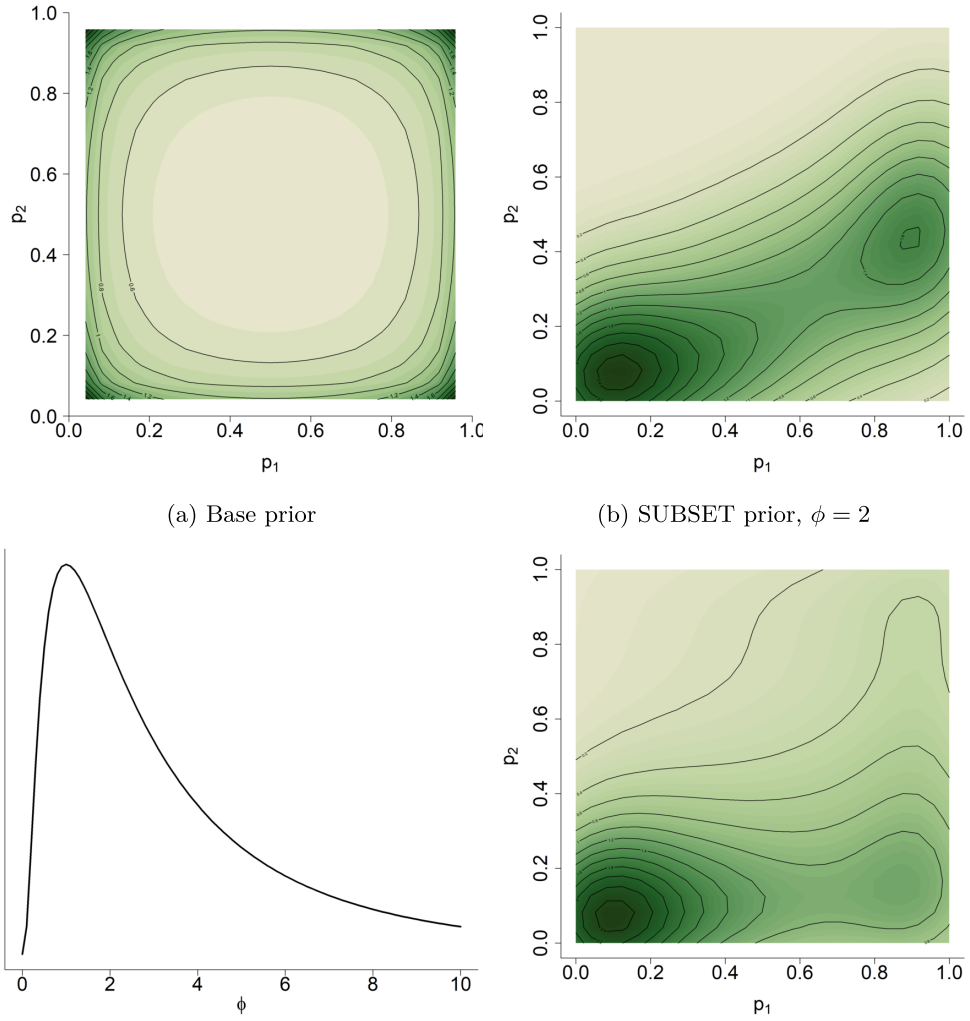
(a) Base prior

(b) Weighting from exponential tilting term

(c) SUBSET prior

(d) Correlation in π_ν as a function of ν

Figure 1: Illustrative Example 2 of applying the SUBSET prior to shrink towards homoscedasticity. (a) Product of independent half-t distributions with 2 degrees of freedom. (b) Weighting of the parameter space introduced by the exponential tilting. (c) SUBSET prior using the independent half-t's as the base prior and setting $\nu = 2$. (d) Prior correlation between σ_1^2 and σ_2^2 as a function of ν.



(a) Base prior

(b) SUBSET prior, $\phi = 2$

(c) $\phi \sim \ell N(1, 1)$

(d) SUBSET prior, unknown ϕ

Figure 2: Illustrative Example 3 of applying the SUBSET prior to shrink towards the space where the population 1 response rate is ϕ times that of population 2. The base prior is the product of independent Jeffrey's priors. In (b) and (d), the tilting parameter ν was fixed at 50.

we have

$$\Pr_\nu(\theta \in \mathcal{B}_\epsilon|y) > \Pr_0(\theta \in \mathcal{B}_\epsilon|y).$$

The proof can be found in the Supplementary Material (Sewell, 2023).

By influencing the posterior through the prior, all posterior inference, not just point estimation, is affected by the prior belief that θ lives on or near the linear subspace. Point estimation, too, is affected by using a SUBSET prior, but through the appropriate mechanism of affecting the Bayes risk through the posterior.

Remark 1. Lee and Birkes (1994) proposed a "subspace ridge" estimator, $\hat{\beta}_{SRDG}$, in a linear regression setting ($y|X\beta, \sigma^2 \sim N(X\beta, I_n/\sigma^2)$) which shrinks the estimates of β towards a linear subspace with projection matrix P. If the commonly implemented improper prior $\pi(\beta, \sigma^2) \propto 1/\sigma^2$ is used as a base prior and ν is set to k/σ^2 for some $k > 0$, then the posterior mean obtained from the SUBSET prior would be equivalent to the subspace ridge estimator, namely

$$\mathbb{E}_{\pi_\nu}(\beta|y) = \big(X'X + k(I - P)\big)^{-1} X'y = \hat{\beta}_{SRDG}.$$

Note that the focus of Lee and Birkes (1994) was on point estimation, rather than obtaining a posterior distribution which incorporated the prior knowledge that the regression coefficients ought to lie on or near the subspace.

Remark 2. It is easy to accommodate shrinkage of subsets of parameters towards different subspaces, as well as have some parameters not shrunk at all. That is, suppose we have a partition of size $(R + 1)$ of a p-dimensional set of parameters $\theta = (\theta_0, \theta_1, \theta_2, \ldots, \theta_R)$, where θ_0 of dimension p_0 is not being shrunk towards a linear subspace, and for $r = 1, \ldots, R$, θ_r of dimension p_r is being shrunk towards the linear subspace with $p_r \times p_r$ projection matrix P_r. Then we can set P in (1) to be the block diagonal matrix with diagonal blocks equal to $I_{p_0}, P_1, \ldots, P_R$.

3 Estimation

3.1 Fixed $\mathcal{L}$

Posterior inference can, of course, be accomplished in all of the usual ways. However, as we will describe shortly, when shrinking the posterior mass towards a linear subspace it is beneficial to evaluate the effect of various levels of shrinkage, i.e., values of ν, as well as to compare the results to those obtained from no shrinkage. With this in mind, it becomes clear that an efficient estimation method is needed. That is, in many- if not most- real data analyses, posterior sampling of $\theta|y$ is computationally expensive, and if MCMC is used convergence diagnostics must be obtained; performing this repeatedly for many values of ν quickly becomes untenable in most situations. The following importance sampler overcomes this issue by only requiring posterior samples of $\theta|y$ under the base prior to be obtained once, followed by very fast importance sampling for each value of ν whose computational cost does not, for example, scale with the sample size nor with the length of a MCMC burn-in period.

Importance sampler

Suppose we have obtained posterior samples $\{\theta_{y,k}\}_{k=1}^{K_y}$ under the base prior, i.e., from

$$\pi_0(\theta|y) \propto \pi(y|\theta)\pi_0(\theta). \tag{2}$$

as well as samples $\{\theta_{0,k}\}_{k=1}^{K_0}$ taken directly from the base prior π_0. We propose using $\pi_0(\cdot|y)$ as the importance distribution to obtain samples from the posterior under the SUBSET prior

$$\pi_\nu(\theta|y) \propto \pi(y|\theta)\pi_0(\theta)e^{-\frac{\nu}{2}\theta'(I-P)\theta}. \tag{3}$$

Under this importance distribution, the unnormalized importance weights $w_{y,k}(\nu)$ corresponding to $\theta_{y,k}$, are simply

$$w_{y,k}(\nu) \propto \exp\left\{-\frac{\nu}{2}\theta'_{y,k}(I-P)\theta_{y,k}\right\}. \tag{4}$$

In choosing ν, one has (at least) three choices. First, ν can be selected manually a priori based on obtaining a prior distribution over θ matching one's prior beliefs.

Second, one can do a sensitivity analysis a posteriori, evaluating the effects of various values of ν. This is very fast to compute since (1) $w_{y,k}(\nu)$ does not depend on sample size, and (2) $w_{y,k}(\nu) = [w_{y,k}(1)]^\nu$, and hence (4) need only be computed once for $\nu = 1$, and then these unnormalized weights can be exponentiated to obtain the weights for other values of ν. If one were to take this approach, it is important to ensure that the effective sample size does not decrease below an acceptable threshold, for as ν increases, the posterior under the SUBSET prior will shift farther away from $\pi_0(\cdot|y)$ leading to weight degeneracy.

Third, one can use Bayes factors to determine the value of ν. The Bayes factor for the SUBSET prior vs. base prior can be written as

$$\frac{\pi_\nu(y)}{\pi_0(y)} = \frac{\mathbb{E}_{\pi_0}\left(e^{-\frac{\nu}{2}\theta'(I-P)\theta}\middle|y\right)}{\mathbb{E}_{\pi_0}\left(e^{-\frac{\nu}{2}\theta'(I-P)\theta}\right)} \tag{5}$$

(see Supplementary Material for derivation; Sewell, 2023). This can then be approximated via

$$\frac{\pi_\nu(y)}{\pi_0(y)} \approx \frac{\frac{1}{K_y}\sum_{k=1}^{K_y} w_{y,k}(\nu)}{\frac{1}{K_0}\sum_{k=1}^{K_0} w_{0,k}(\nu)}, \tag{6}$$

where $w_{0,k}(\nu) := \exp\{-\frac{\nu}{2}\theta'_{0,k}(I-P)\theta_{0,k}\}$. Again, since $w_{0,k}(\nu) = [w_{0,k}(1)]^\nu$ and $w_{y,k}(\nu) = [w_{y,k}(1)]^\nu$, it is relatively fast to numerically maximize (6) as a function of ν. This importance sampling algorithm using Bayes factor to select ν is provided in Algorithm 1.

Algorithm 1: Importance sampler for posterior sampling under the SUBSET prior, returning the draws from the importance distribution and their importance weights, where the shrinkage parameter ν is selected by largest Bayes factor.

Input: Posterior samples $\{\theta_{y,k}\}_{k=1}^{K_y}$ under the base prior; Prior samples $\{\theta_{0,k}\}_{k=1}^{K_0}$; Projection matrix P.
for $k = 1, \ldots, K_0$ **do**
  Compute $w_{0,k}(1) = \exp\{-\frac{1}{2}\theta_{0,k}'(I-P)\theta_{0,k}\}$
end
for $k = 1, \ldots, K_y$ **do**
  Compute $w_{y,k}(1) = \exp\{-\frac{1}{2}\theta_{y,k}'(I-P)\theta_{y,k}\}$
end
Function `helper(`ν`)`**:**
  return $\frac{\frac{1}{K_y}\sum_{k=1}^{K_y}[w_{y,k}(1)]^{\nu}}{\frac{1}{K_0}\sum_{k=1}^{K_0}[w_{0,k}(1)]^{\nu}}$
Use univariate optimizer to find $\nu^* := \underset{\nu}{\text{argmax}}$ `helper` (ν)
Compute $w_{y,k}(\nu^*) := [w_{y,k}(1)]^{\nu^*}$
return $\{\theta_{y,k}, w_{y,k}(\nu^*)\}_{k=1}^{K_y}$

Large sample approximation

For large sample sizes (where posterior sampling can become tedious), many important and historical works have proved the asymptotic normality of the posterior distribution (Bernstein, 1917). While certain settings have required further development of these theorems (e.g., Shen, 2002, develops such results for semi- and non-parametric posteriors), fairly general conditions for where this holds are given in Chen (1985) and will be our focus moving forward.

If the conditions outlined by Chen hold, then under π_0

$$\theta|y, \pi_0 \dot{\sim} N(m_n, \Omega_n),$$

where m_n is the posterior mode and Ω_n is the Hessian of the negative log posterior at m_n. In this setting, we can approximate the exponentially tilted posterior in the following way:

$$\begin{aligned}
\pi_\nu(\theta|y) &\propto \pi(y|\theta)\pi_0(\theta)e^{-\frac{\nu}{2}\theta'(I_p-P)\theta} \\
&\dot{\propto} \exp\left\{-\frac{1}{2}\left[\theta'\Big(\Omega_n + \nu(I_p-P)\Big)\theta - 2\theta'\Omega_n m_n\right]\right\}, \\
\Rightarrow \theta|y, \pi_\nu &\dot{\sim} N(\tilde{m}_n, \widetilde{\Omega}_n), \qquad (7) \\
\widetilde{\Omega}_n &:= \Omega_n + \nu(I_p - P), \\
\tilde{m}_n &:= \widetilde{\Omega}_n^{-1}\Omega_n m_n.
\end{aligned}$$

This implies that once the posterior mode and Hessian under the base prior are computed (either analytically, numerically, or via posterior sampling), there is negligible computational cost required to obtain an approximate posterior under exponential tilting with tilting parameter ν.

The selection of ν using Bayes factors can be done in a similar fashion to Algorithm 1. The only change necessary is to replace the numerator in the `helper` function to be the expectation evaluated analytically:

$$\mathbb{E}_{\pi_0}\left(e^{-\frac{\nu}{2}\theta'(I-P)\theta}\middle|y\right) = |\Omega_n|^{\frac{1}{2}}|\widetilde{\Omega}_n|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(m_n'\Omega_n m_n - \tilde{m}_n'\widetilde{\Omega}_n\tilde{m}_n\right)\right\}.$$

3.2 Estimating $\mathcal{L}$

The subspace towards which we should shrink our posterior is not always precisely defined. Continuing Example 3, we may be confident that p_1 is larger than p_2, and while we may be confident that p_1 is somewhere in the neighborhood of 2 times larger than p_2, we might be hard pressed to specify that factor exactly.

Consider a more general setting, where the subspace is the span of some column vectors which in turn are a function of unknown parameters ϕ taking values in Φ, and let $P(\phi)$ denote the corresponding projection matrix. In the preceding example we were considering $\mathcal{L} = \text{span}((\phi, 1)') \cap (0, 1)^2$ for $\phi = 2$, but we may relax this so that $\phi \in (0, \infty)$. Figure 2 shows a log-normal prior over ϕ (c) and the resulting marginal prior over the two population response rates (d).

In the case of unknown ϕ, a 2-block Metropolis-Hastings-within-Gibbs sampler can be effectively employed to obtain posterior samples in a computationally efficient manner, where we alternate between sampling θ and sampling ϕ. This is due to the fact that for fixed θ, if we draw ϕ^* from its prior independently from our Markov chain's current value ϕ^{curr}, it can be shown that a Metropolis-Hastings sampling step accepts ϕ^* with probability

$$\min\left\{1, \exp\left\{-\frac{\nu}{2}\theta^{(k)\prime}(P(\phi^{curr}) - P(\phi^*))\theta^{(k)}\right\} \cdot \frac{Z_{\nu,\phi^{curr}}}{Z_{\nu,\phi^*}}\right\}. \tag{8}$$

Importance sampler (unknown ϕ)

Let ϕ take a finite number of values with corresponding prior probability mass function π_ϕ. In the case where ϕ is better considered continuous with probability density function $\tilde{\pi}_\phi$, given that ϕ will not be of primary importance, we assume that we may sufficiently approximate this by taking a sequence of quantiles $\{\phi_q\}_{q=1}^Q$ and taking $\pi_\phi(\phi_q) \propto \tilde{\pi}_\phi(\phi_q)$.

Unlike the case with fixed ϕ, now we must worry about the SUBSET normalizing constant $Z_{\nu,\phi}$, which typically will not have a closed form solution. However, since $Z_{\nu,\phi}$ is

an expectation with respect to the base prior on θ, we may obtain an arbitrarily accurate Monte Carlo estimate $\widehat{Z}_{\nu,\phi}$ by taking samples from $\pi_0(\theta)$, which will typically be easy to accomplish (as was done in Algorithm 1). Hence in our 2-block Gibbs sampler, for a given θ we can draw ϕ^* from π_ϕ and accept it with probability given in (8), substituting $Z_{\nu,\phi}$ with its Monte Carlo estimate $\widehat{Z}_{\nu,\phi}$.

For a fixed $\phi \in \{\phi_q\}_{q=1}^{Q}$, we can implement the importance sampler described in Section 3.1 to approximate the full conditional of $\theta|y,\phi$ by the atomized approximation

$$\hat{\pi}(\theta|y,\phi=\phi_q) = \sum_{k=1}^{K_y} \tilde{w}_{(\phi_q)y,k}\delta_{\theta_{y,k}}(\theta), \tag{9}$$

where $\{\theta_{y,k}, w_{(\phi_q)y,k}\}_{k=1}^{K}$ are the samples and importance weights obtained through fixing $\phi=\phi_q$ and computing the weights via (4), and $\tilde{w}_{(\phi_q)y,k}$ is the normalized importance weight for the k^{th} sample. The importance sampler is extremely fast, and draws from the importance distribution along with their weights can be obtained for each of the finite values of ϕ before running the Gibbs sampler. The resulting 2-block Gibbs sampler is described in Algorithm 2.

Large sample approximation (unknown ϕ)

Suppose that the posterior arising from the base prior can again be well approximated with a normal distribution. Given a fixed value of ϕ, the full conditional distribution of θ is the normal distribution given in (7). As in Algorithm 2, a simple Metropolis-Hastings step can be taken to update ϕ.

Since we no longer need to perform importance sampling for each value of ϕ, we no longer constrain ϕ to take values on a discrete set. Yet the need to compute Monte Carlo estimates of the normalizing constant $Z_{\nu,\phi}$ can still slow the sampling algorithm to a debilitating level. However, as the next proposition shows, under arguably most scenarios we have smoothness in $Z_{\nu,\phi}$ which can aid computation significantly.

Theorem 3. *Suppose $L = L(\phi)$ is of full rank and differentiable in ϕ for all $\phi \in \Phi$, and $P(\phi) := L\big(L'L\big)^{-1}L'$. Then for any $\nu > 0$, $Z_{\nu,\phi}$ is continuous in ϕ over Φ.*

The proof is in the Supplementary Material (Sewell, 2023).

We propose, therefore, prior to performing the 2-block Gibbs sampler to perform a two-stage estimation scheme for $Z_{\nu,\phi}$, using a spline regression- or if ϕ is multivariate use tensor product splines or thin-plate splines- for a sequence of values of ϕ predicting $\hat{Z}_{\nu,\phi}$, and then in the Gibbs sampler using estimated values of Z from this spline fit. This approach is given in Algorithm 3.

4 Simulation studies

To evaluate the strengths and weaknesses of the usage of SUBSET priors, we ran two simulation studies. The results shown here correspond to the importance samplers;

Algorithm 2: Two-block MH-within-Gibbs sampler for unknown θ and unknown ϕ (i.e., unknown linear subspace), relying on importance sampling.

Input: Posterior samples $\{\theta_{y,k}\}_{k=1}^{K_y}$ under the base prior; Prior samples $\{\theta_{0,k}\}_{k=1}^{K_0}$; Projection matrix function $P(\cdot)$; Initial values $\theta_{\nu,0}$ and $\phi_{\nu,0}$; Shrinkage weight ν; Prior π_ϕ over the set $\{\phi_q\}_{q=1}^{Q}$; Desired number of posterior samples K_ν.

`/* Precompute key quantites for Gibbs sampler */`
for $q = 1, \ldots, Q$ **do**
  Compute $\widehat{Z}_{\nu,\phi_q} = \frac{1}{K_0}\sum_{k=1}^{K_0} e^{-\frac{\nu}{2}\theta'_{0,k}(I-P(\phi_q))\theta_{0,k}}$
  Compute $w_{(\phi_q)y,k} = e^{-\frac{\nu}{2}\theta'_{y,k}(I-P(\phi_q))\theta_{y,k}}$ for $k = 1, \ldots, K_y$
end

Compute $P^{curr} \leftarrow P(\phi_{\nu,0})$

`/* Perform 2-block Gibbs sampler */`
for $k = 1, \ldots, K_\nu$ **do**
  `/* Draw a new` $\theta|y,\phi$ `*/`
  Set $\theta_{\nu,k} = \theta_{y,k'}$ with probability proportional to $w_{(\phi_{\nu,k-1})y,k'}$

  `/* Draw a new` $\phi|y,\theta$ `*/`
  Draw ϕ^* from the prior over ϕ (π_ϕ)
  Compute $P^* \leftarrow P(\phi^*)$
  Draw $u \sim Unif(0,1)$
  if $u < \exp\left\{-\frac{\nu}{2}\theta'_{\nu,k}(P^{curr} - P^*)\theta_{\nu,k}\right\} \cdot \frac{\widehat{Z}_{\nu,\phi_{\nu,k-1}}}{\widehat{Z}_{\nu,\phi^*}}$ **then**
    $P^{curr} \leftarrow P^*$
    $\phi_{\nu,k} \leftarrow \phi^*$;
  else
    $\phi_{\nu,k} \leftarrow \phi_{\nu,k-1}$
  end
end
return $\{\theta_{\nu,k}, \phi_{\nu,k}\}_{k=1}^{K_\nu}$

using the large sample approximations yielded similar results, which can be found in the Supplementary Material (Sewell, 2023).

4.1 1-way ANOVA

Our first simulation study used a 1-way ANOVA setting, evaluating the SUBSET prior on the estimation and inference of the group variances. The data we simulated used 6

Algorithm 3: Two-block MH-within-Gibbs sampler for unknown θ and unknown ϕ (i.e., unknown linear subspace), relying on a large sample approximation of the posterior. Note that if Φ is finite, $\{\phi_s\}_{s=1}^S$ should equal Φ, and $\widehat{\widehat{Z}}(\phi) = \widehat{Z}_{\nu,\phi}$ rather than predictions from the spline fit.

Input: Posterior mode m_n and precision matrix Ω_n using the base prior; Prior samples $\{\theta_{0,k}\}_{k=1}^{K_0}$; Projection matrix function $P(\cdot)$; Initial values $\theta_{\nu,0}$ and $\phi_{\nu,0}$; Shrinkage weight ν; Prior π_ϕ; Spline function $f : \Phi \mapsto \Re$; Sequence $\{\phi_s\}_{s=1}^S$; Number of posterior draws K_ν.

`/* Precompute key quantites for Gibbs sampler */`
Compute $P^{curr} \leftarrow P(\phi_{\nu,0})$
for $s = 1, \ldots, S$ **do**
  Compute $\widehat{Z}_{\nu,\phi_s} = \frac{1}{K_0} \sum_{k=1}^{K_0} e^{-\frac{\nu}{2}\theta_{0,k}'(I-P(\phi_s))\theta_{0,k}}$
end
Fit spline model $\widehat{\widehat{Z}}(\cdot)$ from regressing $\widehat{Z}_{\nu,\phi_s}$ on $f(\phi_s)$

`/* Perform 2-block Gibbs sampler */`
for $k = 1, \ldots, K_\nu$ **do**
  `/* Draw a new` $\theta|y,\phi$ `*/`
  Draw $\theta_{\nu,k}$ from $N(\tilde{m}_n, \widetilde{\Omega}_n)$ as defined in (7) using $P = P(\phi_{\nu,k-1})$
  `/* Draw a new` $\phi|y,\theta$ `*/`
  Draw ϕ^* from the prior over ϕ (π_ϕ)
  Compute $P^* \leftarrow P(\phi^*)$
  Draw $u \sim Unif(0,1)$
  if $u < \exp\left\{-\frac{\nu}{2}\theta_{\nu,k}'(P^{curr} - P^*)\theta_{\nu,k}\right\} \cdot \frac{\widehat{\widehat{Z}}(\phi_{\nu,k-1})}{\widehat{\widehat{Z}}(\phi^*)}$ **then**
    $P^{curr} \leftarrow P^*$;
    $\phi_{\nu,k} \leftarrow \phi^*$
  else
    $\phi_{\nu,k} \leftarrow \phi_{\nu,k-1}$
  end
end
return $\{\theta_{\nu,k}, \phi_{\nu,k}\}_{k=1}^{K_\nu}$

groups, each with 20 observations, for a total sample size of 120, i.e.,

$$y_{gi} \overset{iid}{\sim} N(\mu_g, 1/\sigma_g^2), \qquad g = 1, \ldots, 6, \quad i = 1, \ldots, 20.$$

The group means were set to be $(1, 2, \ldots, 6)$, and the group variances were set according to three scenarios:

- *Homoscedasticity.* Each group had a residual variance of 2.

- *Mild heteroscedasticity.* The group variances were $(1, 1.6, 2.2, 2.8, 3.4, 4)$.
- *Strong heteroscedasticity.* The group variances were $(1, 3, 5, 7, 9, 11)$.

Variances, even when not of interest directly, are important for other quantities of interest, e.g., prediction intervals and Exceedance in Pairs Rate[2] (Rosner et al., 2021).

We fit three models to each data set. The first assumed homoscedasticity, using a Normal-Inverse gamma prior on the means and (common) variance, i.e.,

$$\begin{aligned} \mu_g|\sigma^2 &\overset{iid}{\sim} N(0, a/\sigma^2), \\ \sigma^2 &\sim \Gamma^{-1}(b/2, c/2), \end{aligned}$$

We set $a = 1$, $b = 1$, and $c = 2$.

The second model assumed heteroscedasticity, with a prior structure similar to that given above:

$$\begin{aligned} \mu_g|\sigma_g^2 &\overset{ind}{\sim} N(0, a/\sigma_g^2), \\ \sigma_g^2 &\overset{iid}{\sim} \Gamma^{-1}(b/2, c/2). \end{aligned}$$

The third model used the heteroscedastic prior given above as the base prior for a SUBSET prior with ν selected via Bayes factor according to Algorithm 1. We shrunk the group variances towards the subspace spanned by $(1, \ldots, 1)'$, i.e., towards homoscedasticity.

Each model used 50000 posterior draws, and data from each of the three scenarios were generated and analyzed 2000 times.

Table 1 provides the results for the estimation of $\{\sigma_g^2\}_{g=1}^6$ in terms of 95% credible interval (CI) widths, 95% CI coverage rates, and MSE. In the case of homoscedastic data, fitting the homoscedastic model unsurprisingly yields the lowest CI widths and lowest MSE. All approaches yield coverage rates near the nominal level. Importantly, the SUBSET prior improves the performance of the heteroscedastic model in terms of CI width, CI coverage, and MSE. For mildly heteroscedastic data, the homoscedastic model cannot appropriately model the uncertainty due to the hard constraint of homoscedasticity, leading to very low coverage and high MSE. However, compared to the heteroscedastic model, the SUBSET prior yields a 26% reduction in MSE and a 29% reduction in the average CI width, although at a cost of a 0.07 reduction in coverage rate. For the strongly heteroscedastic data, once again the homoscedastic model performs very poorly. Compared to the heteroscedastic model, the SUBSET prior yields a 14% reduction in MSE and a 14% reduction in average CI width, at a cost of a 0.014 reduction in coverage rate.

In short, the SUBSET priors helped improve the fit of the heteroscedastic model when incorrectly specified, and did much better in terms of MSE and CI widths when the true parameters lied outside of the subspace, at the cost of a small reduction in coverage rates.

[2]Roughly, EPR is a measure of how well separated the groups are.

	Homoscedastic	Heteroscedastic	SUBSET
	Homoscedastic data		
CI Width	1.118	3.111	1.926
CI Coverage	0.920	0.943	0.970
MSE	0.103	0.649	0.221
	Mildly heteroscedastic data		
CI Width	1.363	3.768	2.685
CI Coverage	0.378	0.875	0.806
MSE	1.202	1.001	0.738
	Strongly heteroscedastic data		
CI Width	3.066	8.330	7.202
CI Coverage	0.262	0.871	0.857
MSE	12.445	5.193	4.445

Table 1: 1-way ANOVA simulation study results for $\{\sigma_g^2\}_{g=1}^6$. All credible intervals were at 95%.

4.2 Ordinal factor covariate

Our second simulation study used a regression setting with a single ordinal factor covariate with 9 levels. The true group means were $0.005 \times (0, 1, 2^4, 3^4, \ldots, 8^4)$. There were five observations per group, and the residual standard deviation was 1.

We fit three models to each simulated dataset. The first used a Zellner's g prior, setting g to be equal to the sample size (45). The remaining two models were SUBSET priors using the Zellner's g prior as the base prior and the following two linear subspaces:

- *Power:* $\text{span}\left(\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2^\phi & \cdots & 9^\phi \end{pmatrix}'\right)$
- *Geometric:* $\text{span}\left(\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1/\phi & 1/\phi^2 & \cdots & 1/\phi^9 \end{pmatrix}\right)$

Note that the true parameter vector lies in neither of these linear subspaces. We estimated ϕ alongside the true regression coefficients via Algorithm 2. To implement this, for the *power* subspace, we used 15 quantiles coming from a gamma distribution with shape equal to 2 and rate equal to 1; for the *geometric* subspace, we used 15 quantiles coming from a Beta distribution with both shape parameters equal to 2. To determine ν, we temporarily set ϕ to be fixed at its mode (1 for the power subspace, and 1/2 for the geometric subspace) and selected ν by maximizing the Bayes factor via Algorithm 1. As before, each model used 50000 posterior draws, and we generated 2000 datasets.

Table 2 provides the results for the estimation of the regression coefficients in terms of 95% CI widths, 95% CI coverage rates, and MSE. All methods achieved near 100% coverage rates, although the different methods achieved this with varying CI widths. Compared to posterior inference using the base prior, the SUBSET priors yielded on av-

erage 9% and 3% smaller CI widths for the power and geometric subspaces respectively, and 25% and 8% lower MSE respectively.

		SUBSET	
	Zellner	(power)	(geometric)
CI Width	3.676	3.337	3.571
CI Coverage	0.997	0.997	0.997
MSE	0.398	0.300	0.367

Table 2: Ordinal covariate simulation study results. All credible intervals were at 95%.

5 Pedagogical analyses

The following analyses illustrate the application of SUBSET priors using the R package `SUBSET`[3] developed by the author. Code to replicate these analyses is included in the Supplementary Material (Sewell, 2023).

5.1 Antihypertensive clinical trial (Ordinal covariates)

Sung et al. (2022) conducted a 7 arm randomized clinical trial aimed at discovering the effect of various drug treatments at various doses on reducing the mean sitting systolic blood pressure (MSSBP) over a period of 8 weeks. The efficacy endpoint was the change in MSSBP from baseline to the end of the 8 week period. For the purposes of this analysis, we will focus on the following four treatment arms: placebo, and combination treatments of telmisartan/amlodipine/chlorthalidon at quarter-dose, third-dose, and half-dose.

Table 3 provides the summary statistics from the study, from which it can be seen that there is a failure to observe the expected dose-response relationship between the dose of the combination treatment and the sample mean change in MSSBP (lower is better, reflecting a larger reduction in blood pressure). However, such a non-monotonic relationship is highly implausible. Letting $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ denote the mean change in MSSBP, we can impose our prior beliefs in a monotonic dose-response relationship by shrinking our prior on $\boldsymbol{\mu}$ to favor values on or near the span of

$$\begin{pmatrix} 1 & 0 \\ 1 & \frac{1}{4^\phi} \\ 1 & \frac{1}{3^\phi} \\ 1 & \frac{1}{2^\phi} \end{pmatrix}$$

for some $\phi > 0$. That is, we believe that there may be some placebo effect, and on top of that there is a drug effect that follows some power law.

[3] In R, run `remotes::install_github(''dksewell/SUBSET'')` .

	Mean	SD
Placebo	−5.85	10.74
1/4 dose	−18.87	16.87
1/3 dose	−14.55	14.84
1/2 dose	−19.55	14.75

Table 3: Summary statistics from multi-arm clinical trial on antihypertensive drugs.

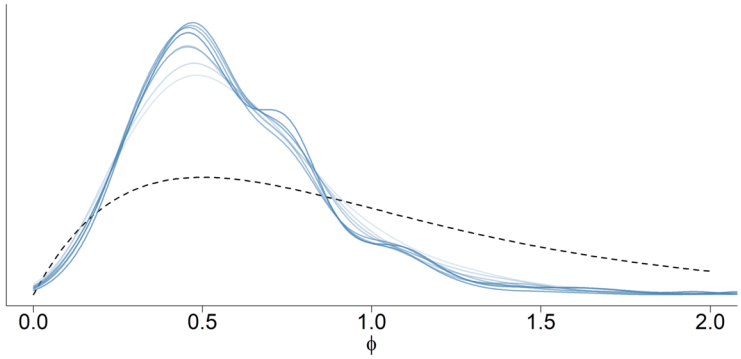


Figure 3: Antihypertensive clinical trial example. Dashed line shows the $\Gamma(2,2)$ prior on ϕ (the parameter dictating the power law of the treatment effect), and the solid lines correspond to the posterior of ϕ using values of ν ranging from 0.25 to 2 graded from lightest to darkest respectively.

As a base prior, we used independent conjugate Normal-Gamma distributions over the mean and precision for each treatment arm. Hence our model for the antihypertensive randomized clinical trial is

$$\begin{aligned} y_i|\text{treatment}_i = k, \boldsymbol{\mu}, \boldsymbol{\tau} &\overset{ind}{\sim} N(\mu_k, \tau_k), \\ \mu_k|\boldsymbol{\tau} &\overset{ind}{\sim} N(a_0, b_0\tau_k), \\ \tau_k &\overset{ind}{\sim} \Gamma(c_0/2, d_0/2), \end{aligned} \tag{10}$$

for $k = 1, \ldots, 4$, where we set $a_0 = -5$, $b_0 = 1$, $c_0 = 3$, and $d_0 = 75$. Additionally, we set a $\Gamma(2,2)$ prior on ϕ, with a prior mean value of ϕ equal to 1 (linear drug effect). We obtained 10000 draws each from the posterior and prior distributions under the base prior and the posterior under the SUBSET prior.

We ran the Gibbs sampler of Algorithm 2 (similar results using Algorithm 3 are provided in the Supplementary Material) using values of ν in $(0.25, 0.5, \ldots, 2)$ and for ϕ considered 50 evenly spaced quantiles from its Gamma prior. Figure 3 shows that for most values of ν, the posterior of ϕ centers near 1/2, implying a prior that shrinks towards a square root relationship between dose and mean change in MSSBP.

Figure 4 shows for the different values of shrinkage, i.e., ν, the posterior mean for the mean change in MSSBP, as well as the posterior probability that the dose response

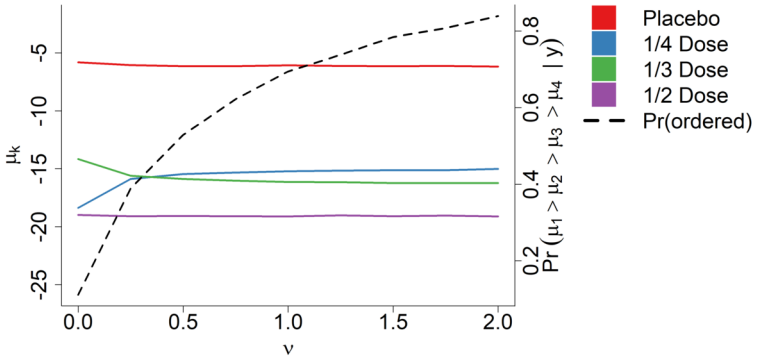


Figure 4: Antihypertensive clinical trial example. The horizontal axis represents differing levels of ν, i.e., shrinkage towards the linear subspace; the left vertical axis corresponding to the solid lines shows the posterior mean estimate of the mean change in MSSBP; the right vertical axis corresponding to the dashed line shows the posterior probability that there is a monotonic relationship between dose and mean change in MSSBP. As ν increases, the expected dose-response relationship emerges.

is monotonic. Although ν increases, the point and interval estimates for the placebo are near constant. Importantly, however, for $\nu \geq 0.5$, the posterior means of $\boldsymbol{\mu}$ reflect a dose-response relationship, i.e., $\mathbb{E}_{\pi_\nu}(\mu_1|y) > \mathbb{E}_{\pi_\nu}(\mu_2|y) > \mathbb{E}_{\pi_\nu}(\mu_3|y) > \mathbb{E}_{\pi_\nu}(\mu_4|y)$, thereby giving us plausible estimates on the effect of the combination treatment on reducing hypertension. While under the base prior there was only a 0.11 posterior probability of a monotonic relationship, this surpassed 0.5 at $\nu = 0.5$ and 0.8 at $\nu = 1.75$.

5.2 Influenza and Pneumonia monthly mortality (Smoothing MA(q) coefficients)

As a second example, we illustrate how to use SUBSET priors to smooth the estimates of a sequence of parameters. We analyzed monthly mortality caused by influenza and pneumonia in the US from 2014-2019 (National Center for Health Statistics, 2022). We detrended the data and based on ACF and PACF plots fit a moving average (MA) model with 15 lags using adaptive MCMC (Scheidegger, 2021). We used as the base prior for the MA coefficients $N(0, 2)$, and for the variance of the residuals a gamma distribution with shape and rate both equal to 2. We obtained 50000 posterior draws under π_0 and removed 10000 as a burn-in period, for a remaining 40000 draws. We obtained 5000 draws from the prior to estimate the Bayes factors.

To shrink towards smoothed estimates of the MA coefficients, we used Algorithm 1 using the linear subspace spanned by the natural cubic spline basis functions evaluated at the lags (1-15) using 4 degrees of freedom (implying 3 internal knots). The value of ν which maximized the Bayes factor was 32.7.

Figure 5 shows the posterior mean and 95% credible intervals for the MA coefficients

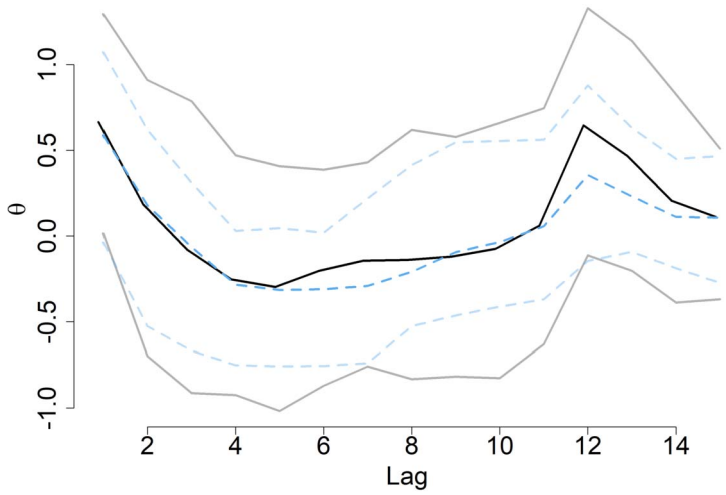


Figure 5: Influenza and pneumonia monthly mortality example. Estimated MA coefficients. Posterior mean and 95% credible interval bounds under the base prior are given by the solid black and gray lines respectively. Under the SUBSET prior with $\nu = 32.7$ (selected via Bayes factor), these are given by the dark blue and light blue dashed lines respectively.

under both the base prior and the SUBSET prior with $\nu = 32.7$. From this figure we can see that both posteriors are telling the same overall story, but that the posterior point and interval estimates under the SUBSET prior are much more smooth.

6 Discussion

The information and beliefs we have about parameters of interest are often relational in nature, which cannot be encoded simply by, e.g., a location shift in the prior distribution. Instead, such prior knowledge leads us to believe that the parameters ought to lie on or near some linear subspace. This type of information is ubiquitous, and yet a comparably small amount of attention has been given it. Previous work has focused almost exclusively on point estimation within a regression setting.

We have proposed a new approach to incorporating relational prior information that can be represented by parameters lying on a linear subspace. Our approach has the following advantages. First, our approach is completely generalizable to any setting, including regression.

Second, we argue that the most logical way to handle prior beliefs described by a linear subspace is to incorporate such information in the prior distribution, which by definition is where our prior beliefs ought to be contained. Doing so allows not only point estimates but all inference to account for this information.

Third, our approach of applying exponential tilting to a base prior does not "overwrite" other prior information that has already been encoded in this base prior such as previous data or scientific domain knowledge. Rather, our approach takes the a priori plausible regions of the parameter space from this prior information and further hones the plausible regions to conform to prior beliefs described by the linear subspace.

Fourth, we have provided methods to obtain posterior inference in a highly computationally efficient manner, allowing for researchers to quickly derive Bayes factors for or conduct sensitivity studies of ν or other facets of the linear subspace. In particular, once K_0 and K_y prior and posterior samples respectively have been obtained under the base prior, computing the Bayes factor in Algorithm 1 for a particular ν costs $\mathcal{O}(K_0 \vee K_y)$. Performing the MH-within-Gibbs sampler of Algorithm 2 has a bottleneck from the multinomial importance resampling of the K_y posterior samples at each step of the Gibbs sampler, leading to a computational cost of $\mathcal{O}(K_\nu K_y \log(K_y))$ (Kronmal and Peterson, 1979), assuming that K_0 grows at the same or slower rate as K_y. Algorithm 3 alleviates this problem entirely by avoiding the resampling of the original K_y posterior draws, and has a computational cost of $\mathcal{O}(K_0 \vee K_\nu)$. Critically, these computational costs are free from the sample size and the computational complexity of evaluating the posterior.

Our proposed methodology has certain limitations. First, while the use of Bayes factors provides an automated approach to selecting the hyperparameter ν for the case of fixed ϕ, for the case when ϕ is estimated, ν may be determined again using Bayes factors for a user-specified value of ϕ else, as an anonymous reviewer pointed out, one may consider using an alternative approach such as cross-validation or putting a prior on ν and estimating it. Further research into the selection of ν in this context would be worthwhile. Second, with unknown ϕ, the estimation of the normalizing constant $Z_{\nu,\phi}$ may become challenging with higher dimensional ϕ. Third, in the proposed estimation algorithms, the normalizing constant $Z_{\nu,\phi}$ is estimated at least once through a Monte Carlo approach, and in Algorithm 3 a second time through splines. While potentially concerning, in our simulation study (see Supplemental Material for results), this did not seem to have deleterious effects on estimation and inference, and should it appear to be problematic in new scenarios not considered here, work on doubly intractable posteriors (see, e.g., Park and Haran, 2018) may be brought to bear. Fourth, Algorithm 2 relies on using a discrete number of values of ϕ. We anticipate that this will not typically be problematic in practice since ϕ will not be a parameter of primary interest, yet there may be situations where a discretization of Φ is suboptimal.

It is not uncommon for data scientists to obtain unexpected, and perhaps unreasonable, results, such as in the antihypertensive clinical trial of Section 5.1 in which the quarter-dose yielded a greater reduction in MSSBP than the third-dose or the half-dose. A reaction such as "this can't be right!" is indicative that there are in fact prior beliefs about the relations between the model parameters. In the above example, we should feel extremely confident that the reduction from a quarter-dose ought to be less than or equal to that from a third-dose, which in turn ought to be less than or equal to that of a half-dose. This type of prior belief also appears when there is an ordinal covariate in a regression setting. Other examples include when we expect smoothness (of which the

monotonicity above was a special case) across a naturally ordered set of parameters, when we believe that there may be near homoscedasticity, or when we believe there may be equal response rates in a two-population binomial setting. Our proposed sampling algorithms outlined in Section 3 should be reasonably easy to implement for a practicing statistician accustomed to performing Bayesian analyses; still, to further lower the barrier to implementation we have developed the R package `SUBSET` and have illustrated its use in the Supplementary Material where the real data analyses are replicated.

Supplementary Material

Supplementary Material to "Posterior shrinkage towards linear subspaces" (DOI: 10.1214/24-BA1414SUPP; .pdf). This supplementary material provides the proofs for Theorems 2-3, the derivation for the Bayes factors, simulation study results for the large sample approximations, as well as a walkthrough of how to implement the R package SUBSET, including replication of the findings presented in Section 5.

References

An, L., Nkurunziza, S., Fung, K. Y., Krewski, D., and Luginaah, I. (2009). "Shrinkage estimation in general linear models." *Computational Statistics & Data Analysis*, 53(7): 2537–2549. URL https://www.sciencedirect.com/science/article/pii/S0167947308005665 MR2665905. doi: https://doi.org/10.1016/j.csda.2008.11.027. 658

Bernstein, S. (1917). *Theory of Probability*. (In Russian). 666

Blaker, H. (1999). "A class of shrinkage estimators in linear regression." *Canadian Journal of Statistics*, 27(1): 207–220. URL https://onlinelibrary.wiley.com/doi/abs/10.2307/3315502 MR1703631. doi: https://doi.org/10.2307/3315502. 657

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 657

Chen, C.-F. (1985). "On asymptotic normality of limiting density functions with Bayesian implications." *Journal of the Royal Statistical Society: Series B*, 47(3): 540–546. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1985.tb01384.x MR0844485. 666

Floto, G., Kremer, S., and Nica, M. (2022). "The Exponentially tilted Gaussian prior for variational autoencoders." *arXiv preprint arXiv:2111.15646* 658

Hansen, B. E. (2016). "Efficient shrinkage in parametric models." *Journal of Econometrics*, 190(1): 115–132. URL https://www.sciencedirect.com/science/article/pii/S0304407615002365 MR3425275. doi: https://doi.org/10.1016/j.jeconom.2015.09.003. 658

Huber, F. and Koop, G. (2021). "Subspace shrinkage in conjugate Bayesian vector autoregressions." *arXiv preprint arXiv:2107.07804*. MR4596791. 658

James, W. and Stein, C. M. (1961). "Estimation with quadratic loss." *Proceedings of the Fourth Berkely Symposium on Mathematical Statistics and Probability*, 1: 311–319. MR0133191. 657

Kronmal, R. A. and Peterson, A. V. (1979). "On the alias method for generating random variables from a discrete distribution." *The American Statistician*, 33(4): 214–218. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1979.10482697 MR0552783. doi: https://doi.org/10.2307/2683739. 677

Lee, Y. and Birkes, D. (1994). "Shrinking toward submodels in regression." *Journal of Statistical Planning and Inference*, 41(1): 95–111. URL https://www.sciencedirect.com/science/article/pii/0378375894901562 MR1292149. doi: https://doi.org/10.1016/0378-3758(94)90156-2. 658, 664

National Center for Health Statistics (2022). "Monthly counts of deaths by select causes, 2014-2019." URL https://data.cdc.gov/NCHS/Monthly-Counts-of-Deaths-by-Select-Causes-2014-201/bxq8-mugm 675

Oman, S. D. (1982). "Shrinking towards subspaces in multiple linear regression." *Technometrics*, 24(4): 307–311. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1982.10487792 MR0687188. doi: https://doi.org/10.2307/1267825. 658

Park, J. and Haran, M. (2018). "Bayesian Inference in the Presence of Intractable Normalizing Functions." *Journal of the American Statistical Association*, 113(523): 1372–1390. MR3862364. doi: https://doi.org/10.1080/01621459.2018.1448824. 677

Park, T. and Casella, G. (2008). "The Bayesian lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 657

Rosner, G. L., Laud, P. W., and Johnson, W. O. (2021). *Bayesian Thinking in Biostatistics*. Boca Raton, FL, USA: CRC Press. 671

Scheidegger, A. (2021). *adaptMCMC: Implementation of a Generic Adaptive Monte Carlo Markov Chain Sampler*. R package version 1.4. URL https://CRAN.R-project.org/package=adaptMCMC 675

Sewell, D. K. (2023). "Supplement to "Posterior shrinkage towards linear subspaces"." 664, 665, 668, 669, 673

Shen, X. (2002). "Asymptotic normality of semiparametric and nonparametric posterior distributions." *Journal of the American Statistical Association*, 97(457): 222–235. MR1947282. doi: https://doi.org/10.1198/016214502753479365. 666

Shin, M., Bhattacharya, A., and Johnson, V. E. (2020). "Functional Horseshoe Priors for Subspace Shrinkage." *Journal of the American Statistical Association*, 115(532): 1784–1797. PMID: 33716358. MR4189757. doi: https://doi.org/10.1080/01621459.2019.1654875. 657

Sung, K.-C., Sung, J. H., Cho, E. J., Ahn, J. C., Han, S. H., Kim, W., Kim, K. H.,

Sohn, I. S., Shin, J., Kim, S. Y., Kim, K.-i., Kang, S. M., Park, S.-J., Kim, Y.-J., Shin, J.-H., Park, S.-M., and Park, C.-G. (2022). "Efficacy and safety of low-dose antihypertensive combination of amlodipine, telmisartan, and chlorthalidone: A randomized, double-blind, parallel, phase II trial." *The Journal of Clinical Hypertension*, 24(10): 1298–1309. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jch.14570 673

Wiemann, P. and Kneib, T. (2021). "Adaptive shrinkage of smooth functional effects towards a predefined functional subspace." *arXiv preprint arXiv:2101.05630*. 657