

Proximal Empirical Bayes for Sparse Regression with Posterior Decision Support

Dimitrios Roxanas

d.roxanas@sheffield.ac.uk

The University of Sheffield

Research Article

Keywords: Quantum game theory, Eisert–Wilkins–Lewenstein protocol, QAOA, Ramanujan mathematics, multi-agent reinforcement learning, supply chain resilience, disruption management. 1

Posted Date: October 8th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11294494/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Proximal Empirical Bayes for Sparse Regression with Posterior Decision Support

Dimitrios Roxanas^{1*}

^{1*}School of Mathematical and Physical Sciences, The University of
Sheffield, Sheffield, S3 7RH, United Kingdom.

Corresponding author(s). E-mail(s): d.roxanas@sheffield.ac.uk;

Abstract

In this work, we revisit estimation and variable selection in sparse Gaussian regression, possibly under affine constraints. We adopt a middle-ground approach between optimisation and fully hierarchical Bayesian methods, preserving much of the computational efficiency of the former while adding posterior uncertainty quantification. In particular, we develop an empirical Bayes framework that separates shrinkage calibration from posterior-informed selection while using the same log-concave model throughout. The resulting workflow is modular: different methods can be used for noise-variance estimation, shrinkage calibration and posterior sampling. At the same time, these tasks are linked, and numerical choices made at one stage can propagate through the workflow and ultimately affect the decision about which variables to retain. We investigate in detail one instance of the framework based on a Laplace prior, with the amount of shrinkage determined by stochastic approximation proximal gradient. The proximal machinery used for this calibration also supports posterior simulation. We further show how the empirical Bayes calibration should be modified when homogeneous constraints reduce the number of free parameters. Nonhomogeneous affine information is instead incorporated softly; when this information is strong, posterior sampling can become poorly conditioned, and we develop a preconditioning strategy to improve mixing and posterior exploration. Synthetic experiments examine how scale calibration, posterior run length and sampler conditioning affect estimation, uncertainty quantification and the final sparse decision. On a diabetes dataset, posterior uncertainty agrees closely with established Bayesian analyses, while the final decision provides a sparser practical summary.

Keywords: sparse regression, variable selection, empirical Bayes, proximal MCMC, uncertainty quantification, affine constraints

1 Introduction

Sparse regression combines two tasks: estimating a coefficient vector and deciding which coordinates are important enough to retain. Penalised estimators such as the Lasso solve both through a sparse optimisation problem (Tibshirani 1996). Under a Gaussian likelihood, the same ℓ_1 penalty corresponds to an independent Laplace prior (Park and Casella 2008), which also provides posterior information about coefficient magnitude, sign and uncertainty. Bayesian variable selection can exploit that information through spike-and-slab models, continuous shrinkage priors, posterior inclusion summaries and decision rules (Bondell and Reich 2012; Hahn and Carvalho 2015). These approaches can be highly effective, but moving from a sparse point estimate to a calibrated posterior decision also introduces additional modelling and computational choices.

Recent work has connected variable selection with optimisation and developed optimisation-based Bayesian methods for sparse and constrained problems (Xu and Duan 2023; Xu et al. 2024; Zhou et al. 2024; Duan 2026). A related line of work in Bayesian inverse problems shows how convex nonsmooth priors can be handled by empirical Bayes calibration, proximal optimisation and proximal Markov chain Monte Carlo (MCMC) (Pereyra 2016; Durmus et al. 2018; Fernandez Vidal et al. 2020; De Bortoli et al. 2020; Pereyra et al. 2020). Moreau–Yosida smoothing is central to this construction because it replaces the nonsmooth prior potential by a differentiable approximation whose gradient is available through the proximal map. The resulting approximation is not tied to one sampler and can also be explored with other gradient-based methods, including Hamiltonian Monte Carlo; see, for example, (Zhou et al. 2024; Xu et al. 2024).

The central contribution of this paper is an end-to-end sparse-regression workflow in which calibration, point estimation, posterior approximation and the final support decision remain distinct statistical objects while sharing the same proximal structure. We first estimate the observation variance, $\hat{\sigma}^2$, then calibrate a global Laplace shrinkage parameter, $\hat{\theta}$, by stochastic approximation proximal gradient (SAPG), compute the nonsmoothed MAP, $\hat{\beta}_{\text{MAP}}$, by FISTA (Beck and Teboulle 2009), and use MYULA to obtain posterior uncertainty from a Moreau-smoothed approximation. The MAP supplies the joint sparse coefficient estimate and candidate configuration, while posterior activation probabilities determine which MAP-supported coordinates remain relevant once uncertainty is taken into account. The resulting pipeline is

$$\hat{\sigma}^2 \longrightarrow \hat{\theta} \longrightarrow \hat{\beta}_{\text{MAP}} \text{ and posterior samples } \longrightarrow \text{selected set } \hat{S}.$$

The separation is deliberate. The observation scale controls the likelihood–prior balance, the empirical Bayes step fixes the global shrinkage level, the nonsmoothed MAP remains the point estimator, and the posterior sample is used for uncertainty and the final decision rather than as a substitute for the sparse optimisation problem. Under a continuous shrinkage prior, posterior draws are almost surely nonzero, so selection requires a practical notion of effect size. We use a posterior-scale threshold and require both $|\hat{\beta}_{\text{MAP},j}| \geq \tau$ and sufficiently high posterior probability that $|\beta_j|$ exceeds the

same threshold. A decision-level Monte Carlo diagnostic then measures whether an eligible coefficient lies close to the probability boundary relative to simulation error.

A second contribution concerns empirical Bayes calibration in the presence of affine information, extending the SAPG methodology of (Fernandez Vidal et al. 2020). Hard homogeneous constraints and nonhomogeneous affine relations look similar at the level of the regression coefficients but behave differently in the marginal-likelihood calibration. For a hard homogeneous restriction $A\beta = 0$, the prior is supported on a lower-dimensional linear space and the SAPG score depends on its intrinsic dimension. For a nonhomogeneous relation, such as $\sum_{j=1}^p \beta_j = 1$, direct restriction

changes the prior normalising constant in a less tractable way. We instead retain the ambient-space homogeneous prior and impose the relation softly through a Gaussian pseudo-observation. This preserves the unconstrained-homogeneous SAPG score while allowing the affine information to enter through the smooth part of the model. Strong soft information also creates a separate numerical issue: low-rank curvature can force a scalar-step Langevin sampler to move slowly in otherwise well-scaled directions. We therefore study a geometry-aware preconditioner that targets this stiffness. Both hard and soft affine structures arise in our concurrent sparse index-tracking application (Roxanas 2025), where portfolio construction used a soft sum-to-one condition and rebalancing used hard self-financing changes with sum zero.

The resulting method is computationally simple. Laplace and elastic net-type penalties have inexpensive proximal maps, so calibration, MAP optimisation and posterior simulation can reuse the same basic numerical operations. Richer hierarchical shrinkage models provide greater prior flexibility, particularly across heterogeneous signal sizes, but require additional modelling, calibration and posterior computation. Our aim here is to determine how far a comparatively simple log-concave model can support posterior analysis and sparse decisions when its computational stages are separated and checked explicitly.

The numerical studies therefore examine the behaviour of the full workflow, not only its final recovery rates. Study 1 evaluates the complete unconstrained workflow in both $n > p$ and $p > n$ regimes and tracks how likelihood-scale calibration and posterior Monte Carlo precision propagate to the final sparse decision. Study 2 examines soft affine information, showing how the same information can improve the fitted model while creating low-rank stiffness in posterior sampling, and quantifies the effect of geometry-aware preconditioning. Study 3 applies the workflow to the diabetes data used in (Efron et al. 2004) and compares its posterior intervals with externally reported Bayesian analyses. The reported 95% Bayesian zero-exclusion pattern agrees with the ProxMCMC, Bayesian lasso and horseshoe analyses reported by (Zhou et al. 2024), while the MAP-plus-posterior rule gives a smaller practical support at its reference operating point. The supplement adds a hard-constraint calibration check and an observation-scale/Moreau sensitivity analysis.

We use the Laplace prior throughout so that the calibration and the role of the final decision remain transparent. The proximal optimisation and sampling components, however, extend to other proper closed convex penalties with tractable proximal maps,

including elastic net penalties. Section 2.3 notes what changes in the empirical Bayes calibration when the simple homogeneous Laplace score is no longer available.

The remainder of the paper is organised as follows. Section 2 gives the model and computational methodology, Section 3 presents the numerical studies, and Section 4 discusses the main findings and extensions. Further derivations and implementation details are provided in the supplement.

2 Model and methodology

We work with n observations and p predictors:

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \quad (1)$$

with $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$. The response and predictors are centred when an intercept is not modelled explicitly, and predictor scaling is stated for each experiment.

Conditional on σ^2 , the likelihood is

$$\pi(y \mid \beta) \propto \exp\{-f_0(\beta)\}, \quad f_0(\beta) = \frac{1}{2\sigma^2} \|y - X\beta\|_2^2. \quad (2)$$

We use the Laplace family

$$\pi(\beta \mid \theta) \propto \exp\{-\theta g(\beta)\}, \quad g(\beta) = \|\beta\|_1, \quad \theta > 0 \quad (3)$$

as the shrinkage prior. Conditional on (σ^2, θ) , the posterior is

$$\pi(\beta \mid y, \sigma^2, \theta) \propto \pi(y \mid \beta, \sigma^2) \pi(\beta \mid \theta).$$

2.1 Likelihood scale calibration

We estimate the observation variance before the empirical Bayes step and treat the result as fixed thereafter. This scale matters because it controls both the likelihood–prior balance and the curvature used in the proximal calculations.

When residual degrees of freedom are available, let $q = \text{rank}(X) < n$ and let $\hat{\beta}_{\text{LS}}$ denote a least-squares solution. The usual Gaussian residual estimator is

$$\hat{\sigma}_{\text{res}}^2 = \frac{\|y - X\hat{\beta}_{\text{LS}}\|_2^2}{n - q}, \quad q = \text{rank}(X) < n. \quad (4)$$

If an intercept is fitted, including implicitly through sample centring, the denominator is reduced by one additional degree of freedom.

When $p \geq n$ and X has full row rank, ordinary least squares leaves no residual degrees of freedom. Under a sparse high-dimensional model we use the scaled Lasso of (Sun and Zhang 2012) as an auxiliary estimator,

$$(\hat{\beta}_{\text{SL}}, \hat{\sigma}_{\text{SL}}) \in \arg \min_{\beta \in \mathbb{R}^p, \sigma > 0} \left\{ \frac{\|y - X\beta\|_2^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_0 \|\beta\|_1 \right\}, \quad \lambda_0 = \sqrt{\frac{2 \log p}{n}}. \quad (5)$$

In finite samples, shrinkage can leave signal in the scaled Lasso residuals and inflate the resulting variance. We therefore use the scaled Lasso only to define the provisional support

$$\widehat{S}_{\text{SL}} = \{j : \widehat{\beta}_{\text{SL},j} \neq 0\}, \quad (6)$$

refit least squares on that support,

$$\widehat{\beta}_{\text{post},\widehat{S}_{\text{SL}}} \in \arg \min_{b \in \mathbb{R}^{|\widehat{S}_{\text{SL}}|}} \|y - X_{\widehat{S}_{\text{SL}}} b\|_2^2, \quad (7)$$

and use

$$\widehat{\sigma}_{\text{postOLS}}^2 = \frac{\|y - X_{\widehat{S}_{\text{SL}}} \widehat{\beta}_{\text{post},\widehat{S}_{\text{SL}}}\|_2^2}{n - \text{rank}(X_{\widehat{S}_{\text{SL}}})}. \quad (8)$$

An intercept adjustment is made when needed. Because the support is data-dependent, (8) is used as a plug-in scale rather than as an unbiased post-selection variance estimator. We emphasise that the scaled Lasso and post-OLS coefficients are not used downstream.

Thus

$$\widehat{\sigma}^2 = \begin{cases} \widehat{\sigma}_{\text{res}}^2, & \text{rank}(X) < n, \\ \widehat{\sigma}_{\text{postOLS}}^2, & \text{rank}(X) = n \text{ under sparse high-dimensional structure.} \end{cases} \quad (9)$$

Section 3.1 shows the finite-sample effect of this choice in the $p > n$ regime. Other scale estimators could be substituted at this stage, including square-root Lasso, robust residual scales and high-dimensional moment estimators under stronger assumptions. The supplementary sensitivity analysis separately examines how alternative supplied scales propagate through the fitted posterior and final selection rule. When the quadratic data-fit term is better viewed as a loss and no physical noise scale is available, the same architecture can instead use a generalised Bayes learning rate or inverse temperature. Our sparse index-tracking application (Roxanas 2025) presents such an example.

2.2 Moreau–Yosida smoothing and MYULA

We write f for the smooth quadratic term in the negative log-posterior: $f = f_0$ in the unconstrained model, and $f = f_{\tau_c}$ when the soft affine term introduced in subsection 2.3 is present. The posterior combines f with a convex but nonsmooth ℓ_1 term. Proximal MCMC methods exploit this structure through Moreau–Yosida smoothing. For a proper closed convex function g , its Moreau–Yosida envelope is

$$g_\lambda(w) = \min_{u \in \mathbb{R}^p} \left\{ g(u) + \frac{1}{2\lambda} \|u - w\|_2^2 \right\}, \quad \lambda > 0, \quad (10)$$

and we write

$$\text{prox}_{\lambda g}(w) = \arg \min_{u \in \mathbb{R}^p} \left\{ g(u) + \frac{1}{2\lambda} \|u - w\|_2^2 \right\}$$

for the corresponding proximal map. The envelope has gradient

$$\nabla g_\lambda(\beta) = \frac{1}{\lambda} \left(\beta - \text{prox}_{\lambda g}(\beta) \right). \quad (11)$$

For the ℓ_1 penalty, the proximal map reduces to coordinatewise soft-thresholding. Moreau smoothing therefore replaces the nonsmooth term by a differentiable approximation with an explicitly available gradient; see, for example, (Durmus et al. 2018) and (Zhou et al. 2024, Section 2.1).

The smoothing parameter λ controls both approximation and computation. As $\lambda \downarrow 0$, the Moreau envelope converges pointwise to the original convex function and, under the conditions used for MYULA, the induced smoothed posterior converges to the nonsmoothed target in total variation (Durmus et al. 2018; Pereyra et al. 2020). Smaller values of λ therefore sharpen the posterior approximation, but also increase the Lipschitz constant of the smoothed gradient and restrict the admissible Langevin timestep. The supplementary sensitivity analysis varies λ on a problem with a synthetic dataset, allowing this approximation–efficiency trade-off to be separated from changes in the supplied observation scale.

Writing $h_\theta(\beta) = \theta g(\beta)$, we use

$$h_\theta^\lambda(\beta) = \min_u \left\{ \theta g(u) + \frac{1}{2\lambda} \|u - \beta\|_2^2 \right\}. \quad (12)$$

Its gradient is

$$\nabla h_\theta^\lambda(\beta) = \frac{1}{\lambda} \left\{ \beta - \text{prox}_{\lambda \theta g}(\beta) \right\}, \quad (13)$$

which is $1/\lambda$ -Lipschitz. We define

$$U_{\theta,\lambda}(\beta) = f(\beta) + h_\theta^\lambda(\beta), \quad \pi_{\theta,\lambda}(\beta \mid y) \propto e^{-U_{\theta,\lambda}(\beta)}. \quad (14)$$

MYULA (Pereyra 2016; Durmus et al. 2018) is built upon an Euler–Maruyama discretisation of the overdamped Langevin diffusion associated with $U_{\theta,\lambda}$:

$$\beta^{(k+1)} = \beta^{(k)} - \delta \nabla U_{\theta,\lambda}(\beta^{(k)}) + \sqrt{2\delta} \xi^{(k+1)}, \quad \xi^{(k+1)} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p), \quad (15)$$

with timestep $\delta > 0$. In our setting, the recursion is

$$\begin{aligned} \beta^{(k+1)} &= \beta^{(k)} - \delta \left[\nabla f(\beta^{(k)}) + \frac{1}{\lambda} \left\{ \beta^{(k)} - \text{prox}_{\lambda \theta g}(\beta^{(k)}) \right\} \right] + \sqrt{2\delta} \xi^{(k+1)}, \\ \xi^{(k+1)} &\stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p). \end{aligned} \quad (16)$$

The continuous-time diffusion has $\pi_{\theta,\lambda}$ as its invariant distribution under the usual regularity conditions. The unadjusted Euler chain introduces a further finite-step approximation, so we distinguish the nonsmoothed posterior π_θ , the smoothed posterior $\pi_{\theta,\lambda}$ and the invariant distribution of the discretised chain, denoted $\pi_{\theta,\lambda,\delta}$. If ∇f

is L_f -Lipschitz, then a global Lipschitz bound for $\nabla U_{\theta,\lambda}$ is

$$L_\lambda = L_f + \frac{1}{\lambda}. \quad (17)$$

The bound

$$0 < \delta \leq \frac{1}{L_\lambda} = \frac{1}{L_f + 1/\lambda}. \quad (18)$$

on the timestep ensures both numerical stability (Pereyra et al. 2020) and the validity of the nonasymptotic MYULA analysis of (Durmus et al. 2018), and the MYULA kernels used within the SAPG methodology of (Fernandez Vidal et al. 2020). In the unconstrained Gaussian model, where $f = f_0$,

$$L_f = \frac{\lambda_{\max}(X^\top X)}{\hat{\sigma}^2}. \quad (19)$$

For computational efficiency, it is necessary to use values of δ that are close to the stability limit (18), so we take

$$\delta = \frac{0.9}{L_f + 1/\lambda}. \quad (20)$$

Our default Moreau choice is

$$r_\lambda := \lambda L_f = 1, \quad \lambda = \frac{1}{L_f}, \quad (21)$$

for which $L_\lambda = 2L_f$ and hence

$$\delta = \frac{0.9}{2L_f}. \quad (22)$$

The three computational objects are consequently kept conceptually separate:

$$\pi_\theta \longrightarrow \pi_{\theta,\lambda} \longrightarrow \pi_{\theta,\lambda,\delta},$$

representing the nonsmooth posterior, the Moreau-smoothed posterior and the finite-step unadjusted chain, respectively. A Metropolis–Hastings correction could instead be added, yielding a Metropolis-Adjusted Langevin Algorithm (MALA)-type kernel that removes the finite-step discretisation bias with respect to the smoothed target $\pi_{\theta,\lambda}$ (though not the Moreau approximation itself), at the cost of an accept–reject step; we retain MYULA here for computational efficiency.

2.3 Empirical Bayes calibration and affine information

With the observation scale fixed at the preceding stage, we estimate the global shrinkage parameter by maximum marginal likelihood,

$$\hat{\theta} \in \arg \max_{\theta \in \Theta} \pi(y \mid \theta), \quad \pi(y \mid \theta) = \int \pi(y \mid \beta) \pi(\beta \mid \theta) d\beta, \quad (23)$$

over a compact interval $\Theta \subset (0, \infty)$. Once $\hat{\theta}$ has been obtained, it is held fixed for the subsequent MAP and posterior calculations. The computational difficulty in (23) is not the one-dimensional optimisation itself, but evaluating the marginal-likelihood score without access to the integral in closed form.

SAPG (Fernandez Vidal et al. 2020; De Bortoli et al. 2020) addresses this through a Fisher-identity representation of the score and a stochastic-approximation update whose expectation is estimated by a short proximal Markov chain. To make the quantity used in this paper explicit, let C denote the support of the prior, with intrinsic dimension d , and let μ_d denote the induced d -dimensional Lebesgue measure on C . We write

$$Z_C(\theta) = \int_C \exp\{-\theta g(\beta)\} d\mu_d(\beta)$$

for its normalising constant. The marginal-likelihood derivative satisfies

$$\frac{\partial}{\partial \theta} \log \pi(y | \theta) = -\mathbb{E}_\theta[g(\beta) | y] - \frac{\partial}{\partial \theta} \log Z_C(\theta). \quad (24)$$

For a one-homogeneous penalty, $g(t\beta) = tg(\beta)$, on a d -dimensional linear support, the change of variables $u = \theta\beta$ gives

$$Z_C(\theta) = \theta^{-d} Z_C(1).$$

Hence, in logarithmic coordinates $\eta = \log \theta$,

$$\frac{\partial}{\partial \eta} \log \pi(y | e^\eta) = d - \theta \mathbb{E}_\theta[g(\beta) | y]. \quad (25)$$

For the unconstrained Laplace prior, $C = \mathbb{R}^p$, $d = p$, $g(\beta) = \|\beta\|_1$, and $Z_C(\theta) = (2/\theta)^p$.

At iteration k , a short MYULA chain at the current $\theta_k = e^{\eta_k}$ provides states $\{\beta_k^{(\ell)}\}_{\ell=1}^{m_k}$. Writing

$$\bar{g}_k = \frac{1}{m_k} \sum_{\ell=1}^{m_k} g(\beta_k^{(\ell)}), \quad \hat{\Delta}_k = d - \theta_k \bar{g}_k,$$

we use the projected Robbins–Monro recursion

$$\eta_{k+1} = \Pi_{\log \Theta} \left\{ \eta_k + \rho_k \hat{\Delta}_k \right\}, \quad \rho_k = \frac{c}{k + k_0}, \quad (26)$$

followed by a tail/Polyak–Ruppert average of the resulting parameter path. Algorithm 1 summarises the calculation.

Algorithm 1 SAPG calibration of a homogeneous Laplace scale

Require: $y, X, \hat{\sigma}^2$, geometry-specific f , intrinsic dimension d , initial θ_0 , bounds Θ , steps ρ_k .

- 1: Initialise a chain state β_0 and $\eta_0 = \log \theta_0$.
 - 2: **for** $k = 0, \dots, K - 1$ **do**
 - 3: At $\theta_k = e^{\eta_k}$, advance a short MYULA chain and collect m_k states.
 - 4: $\bar{g}_k \leftarrow m_k^{-1} \sum_{\ell=1}^{m_k} g(\beta_k^{(\ell)})$.
 - 5: $\hat{\Delta}_k \leftarrow d - \theta_k \bar{g}_k$.
 - 6: $\eta_{k+1} \leftarrow \Pi_{\log \Theta} \{\eta_k + \rho_k \hat{\Delta}_k\}$.
 - 7: **end for**
 - 8: Return a tail/Polyak–Ruppert average as $\hat{\theta}$.
-

The simple score in (25) is a consequence of one-homogeneity, not a requirement of the broader optimisation–sampling framework. For example, an elastic net prior with fixed mixing parameter still has an explicit proximal map, so the FISTA and proximal MCMC steps remain straightforward. Its empirical Bayes update can use the general Fisher/SAPG identity (24) with the appropriate derivative of the prior normalising constant in place of the term d/θ ; if both elastic net penalty parameters are estimated, the stochastic-approximation update becomes multivariate, but can still be treated by other variants of SAPG already outlined in (Fernandez Vidal et al. 2020).

Hard homogeneous constraints

For a hard linear restriction

$$C_0 = \{\beta : A\beta = 0\}, \quad d = p - \text{rank}(A),$$

the support is a linear subspace, so the same scaling argument applies with its *intrinsic* dimension d . In particular, under the sum-zero condition $\mathbf{1}^\top \beta = 0$, the correct score in (25) uses $d = p - 1$, not p . The same intrinsic dimension enters the natural scale-matching initialisation

$$\theta_0 = \frac{d}{g(\beta_{\text{ref}})}. \quad (27)$$

Here β_{ref} is a preliminary reference fit used only to set the initial shrinkage scale; the experiment-specific choices are given in the supplement. The short MYULA chain used inside SAPG is run within C_0 , using the projected smooth gradient, projected Gaussian innovation and constrained proximal map; the explicit update and implementation details are given in the supplementary material. A paired numerical verification reported there, deliberately replaces $d = p - 1$ by the ambient value p in the sum-zero example.

Nonhomogeneous affine information

The same argument cannot simply be transferred to an affine slice

$$C_b = \{\beta : A\beta = b\}, \quad b \neq 0.$$

Indeed, under the rescaling $u = \theta\beta$, the constraint becomes $Au = \theta b$, so the integration domain itself depends on θ and $Z_{C_b}(\theta)$ does not factor as $\theta^{-d}Z_{C_b}(1)$. A dimension-only contribution to the SAPG score is therefore no longer available. Rather than define the Laplace prior directly on this moving affine slice, we retain the prior on ambient $\mathbb{R}^p$ and impose the nonhomogeneous relation softly through the Gaussian pseudo-observation

$$f_{\tau_c}(\beta) = \frac{1}{2\sigma^2}\|y - X\beta\|_2^2 + \frac{1}{2\tau_c^2}\|A\beta - b\|_2^2. \quad (28)$$

The prior normalising constant then remains proportional to θ^{-p} , so (25) continues to use $d = p$, while τ_c controls both the strength of the affine information and the additional curvature introduced by the quadratic affine term. Section 3.2 studies these two consequences separately.

For the unconstrained and soft affine cases we likewise use $d = p$ in (27); the choices of β_{ref} , projection bounds and averaging windows used in the experiments are recorded in the supplementary material. We note that the exact score (25) belongs to the nonsmoothed model, whereas its posterior expectation is estimated with the inexact MYULA kernel described in Section 2.2. Consequently, $\hat{\theta}$ can inherit some dependence on the Moreau parameter λ . We retain the nonsmoothed marginal likelihood as the empirical Bayes target and assess this numerical dependence separately.

2.4 MAP estimation and role in selection

The nonsmoothed MAP has two roles. It is the primary point estimator under the fitted empirical Bayes model, and its joint sparse configuration provides the candidate set and magnitude anchor for the later selection rule. The posterior gates therefore add uncertainty information without replacing the MAP as the coefficient estimator. Our index-tracking application (Roxanas 2025) is an example where the MAP is more explicitly an intermediate decision vector because portfolio weights or trades must still be converted into an implementable constrained action; here the MAP itself remains the reported point estimate, but the applicability of the method is broader.

Conditional on $(\hat{\sigma}^2, \hat{\theta})$, the primary estimator is

$$\hat{\beta}_{\text{MAP}} \in \arg \min_{\beta} \{f(\beta) + \hat{\theta}\|\beta\|_1\}, \quad (29)$$

where f is the active smooth term defined in subsection 2.2. We compute (29) by FISTA. If ∇f is L_f -Lipschitz, the unconstrained and soft affine iterations use $\gamma = 1/L_f$ and, at the extrapolated point z^k ,

$$\beta^{k+1} = \text{prox}_{\gamma\hat{\theta}g}(z^k - \gamma\nabla f(z^k)). \quad (30)$$

For $g(\beta) = \|\beta\|_1$, this is coordinatewise soft thresholding at level $\gamma\hat{\theta}$.

The empirical Bayes estimate $\hat{\theta}$ is obtained with a MYULA approximation and can therefore depend on the Moreau parameter λ . Once $\hat{\theta}$ is fixed, however, FISTA in (29) optimises the original nonsmoothed objective. For comparison, the smoothed

posterior has mode

$$\hat{\beta}_{\text{MAP},\lambda} \in \arg \min_{\beta} \left\{ f(\beta) + h_{\theta}^{\lambda}(\beta) \right\}. \quad (31)$$

In general $\hat{\beta}_{\text{MAP},\lambda} \neq \hat{\beta}_{\text{MAP}}$. We report the nonsmoothed MAP and use the smoothed mode only to audit the approximation. In particular, we monitor

$$D_2(\lambda) = \frac{\|\hat{\beta}_{\text{MAP},\lambda} - \hat{\beta}_{\text{MAP}}\|_2}{1 + \|\hat{\beta}_{\text{MAP}}\|_2}, \quad D_{\infty}(\lambda) = \|\hat{\beta}_{\text{MAP},\lambda} - \hat{\beta}_{\text{MAP}}\|_{\infty}. \quad (32)$$

For the hard sum-zero constraint $C_0 = \{\beta : \mathbf{1}^{\top} \beta = 0\}$, the constrained proximal map is

$$\text{prox}_{t\theta\|\cdot\|_1}^{C_0}(v) = \arg \min_{u \in C_0} \left\{ \frac{1}{2} \|u - v\|_2^2 + t\theta \|u\|_1 \right\}. \quad (33)$$

Its KKT representation is

$$u_j = S_{t\theta}(v_j - \nu), \quad \sum_{j=1}^p S_{t\theta}(v_j - \nu) = 0, \quad (34)$$

where $S_a(x) = \text{sign}(x)(|x| - a)_+$ is the scalar soft-thresholding operator and the scalar ν is obtained by bisection (Zhou et al. 2024, e.g., Section 5.1). FISTA then operates directly in C_0 :

$$\beta^{k+1} = \text{prox}_{\gamma\hat{\theta}_g}^{C_0}(z^k - \gamma P_{C_0} \nabla f_0(z^k)), \quad (35)$$

with

$$P_{C_0} = I - \frac{1}{p} \mathbf{1}\mathbf{1}^{\top}. \quad (36)$$

2.5 Posterior-informed sparse decisions

A continuous Laplace posterior does not contain inclusion indicators, and MAP nonzero status alone need not imply practical relevance. We therefore separate shrinkage from the final support decision, following the general idea of fitting a continuous model first and sparsifying a posterior summary afterward (Bondell and Reich 2012; Hahn and Carvalho 2015; Piironen et al. 2020). The nonsmoothed MAP supplies the joint magnitude anchor, while MYULA provides posterior-scale information from the smoothed approximation.

From retained MYULA draws, we define

$$\hat{s}_j(\lambda) = \text{sd}_{\pi_{\hat{\theta},\lambda}}(\beta_j \mid y), \quad j = 1, \dots, p. \quad (37)$$

We use the robust global scale

$$\tau_{\text{post}}(\lambda; k) = k \text{ median}_{1 \leq j \leq p} \hat{s}_j(\lambda), \quad (38)$$

which provides a common posterior scale in coefficient units while preventing a small number of coordinates with very large marginal posterior standard deviations from setting the threshold for all variables. We define the activation probability

$$\hat{\pi}_j(\lambda; k) = \Pr_{\pi_{\hat{\theta}, \lambda}}(|\beta_j| \geq \tau_{\text{post}}(\lambda; k) \mid y), \quad (39)$$

estimated by Monte Carlo frequencies.

A marginal activation probability can be large for two different reasons: the posterior may be centred away from zero, or it may simply be very spread out. To make the latter point explicit, suppose for illustration that the marginal posterior of one coefficient is approximately

$$\beta_j \mid y \approx \mathcal{N}(\mu_j, s_j^2).$$

For a fixed practical threshold $\tau > 0$,

$$\Pr(|\beta_j| \geq \tau \mid y) = \Phi\left(\frac{-\tau - \mu_j}{s_j}\right) + 1 - \Phi\left(\frac{\tau - \mu_j}{s_j}\right),$$

where Φ is the standard normal distribution function. If $\mu_j = 0$, this becomes

$$2 \left\{ 1 - \Phi\left(\frac{\tau}{s_j}\right) \right\},$$

which increases with s_j (and hence with the posterior variance s_j^2) for fixed τ . These are “diffuse” in the sense that a weakly identified coordinate can place substantial posterior mass beyond $\pm\tau$ because its marginal variance is large, even when it is centred near zero. The median in (38) limits the influence of a small number of such coordinates on the common threshold, while the MAP magnitude gate below prevents marginal spread alone from triggering selection.

The selected support is

$$\hat{S}_\lambda(k, \pi_\star) = \left\{ j : |\hat{\beta}_{\text{MAP}, j}| \geq \tau_{\text{post}}(\lambda; k), \quad \hat{\pi}_j(\lambda; k) \geq \pi_\star \right\}. \quad (40)$$

We report

$$k \in \{2, 2.5, 3\}, \quad \pi_\star \in \{0.5, 0.75, 0.9\},$$

with (2.5, 0.75) as the reference setting. The MAP gate is tied to the nonsmoothed fitted model, whereas the posterior scale and activation probability depend on the Moreau approximation.

Conditional on passing the MAP magnitude gate, the probability threshold has a simple coordinatewise decision-theoretic interpretation. Let

$$E_j(\lambda; k) = \{|\beta_j| \geq \tau_{\text{post}}(\lambda; k)\} \quad (41)$$

denote the event that coefficient j is practically active, and let $a_j \in \{0, 1\}$ denote the decision to exclude ($a_j = 0$) or select ($a_j = 1$) that coordinate. Let $c_{\text{FP}} > 0$ be the loss

assigned to selecting a practically inactive coefficient and $c_{\text{FN}} > 0$ the loss assigned to excluding a practically active one. With $\mathbb{I}(\cdot)$ denoting the indicator function, consider

$$L(a_j, E_j) = c_{\text{FP}} a_j \mathbb{I}(E_j^c) + c_{\text{FN}} (1 - a_j) \mathbb{I}(E_j). \quad (42)$$

The posterior expected losses of selection and exclusion are therefore

$$c_{\text{FP}} \Pr(E_j^c \mid y) \quad \text{and} \quad c_{\text{FN}} \Pr(E_j \mid y),$$

respectively. The Bayes action selects j when

$$\Pr(E_j \mid y) \geq \frac{c_{\text{FP}}}{c_{\text{FP}} + c_{\text{FN}}}. \quad (43)$$

Thus $\pi_\star$ can be read as a relative false-inclusion/false-exclusion cost threshold. For example, $\pi_\star = 0.75$ corresponds to assigning a false inclusion three times the loss of a false exclusion at this coordinatewise decision stage.

Because the terminal decision depends on Monte Carlo estimates of (38) and (39), we supplement standard autocorrelation and effective-sample-size diagnostics with a decision-level quantity. For retained draw m , define the binary activation sequence

$$I_j^{(m)}(\lambda; k) = \mathbb{I}\left\{|\beta_j^{(m)}| \geq \tau_{\text{post}}(\lambda; k)\right\}.$$

Then $\hat{\pi}_j(\lambda; k)$ is the sample mean of this sequence, and $\text{MCSE}\{\hat{\pi}_j(\lambda; k)\}$ is the Monte Carlo standard error of that mean, accounting for serial dependence through the effective sample size. We define

$$D_j(\lambda; k, \pi_\star) = \frac{|\hat{\pi}_j(\lambda; k) - \pi_\star|}{\text{MCSE}\{\hat{\pi}_j(\lambda; k)\}}. \quad (44)$$

We report the minimum D_j over all coordinates and, more importantly, over coordinates that pass the MAP magnitude gate. If $I_j^{(m)}(\lambda; k)$ is constant over the retained draws, its estimated variance and hence its estimated MCSE are zero. For the interior probability thresholds used here, the empirical activation probability is then either zero or one, so the retained run exhibits no Monte Carlo ambiguity about which side of the probability gate that coordinate lies on. This should not be interpreted as posterior certainty; it is a statement about Monte Carlo resolution conditional on the sampled run and is considered together with the usual mixing diagnostics.

2.6 Geometry-aware preconditioning

The scalar-step MYULA update introduced above is most effective when the posterior curvature is reasonably balanced across directions. This need not hold in regression problems: the likelihood may be poorly scaled across coordinates, while a strong soft affine relation can introduce a small number of directions whose curvature is much

larger than that of the remaining parameter space. Both effects can restrict the admissible timestep and degrade mixing in directions that are not themselves stiff. We therefore use preconditioning as a numerical device matched to the source of the anisotropy.

Let $U_{\theta,\lambda}$ denote the smoothed negative log-posterior and let M be a fixed symmetric positive-definite matrix. The corresponding preconditioned Langevin diffusion is

$$d\beta_t = -M\nabla U_{\theta,\lambda}(\beta_t) dt + \sqrt{2M} dW_t,$$

which has the same invariant distribution as the unpreconditioned diffusion. Its Euler discretisation is

$$\beta^{(k+1)} = \beta^{(k)} - \delta M \nabla U_{\theta,\lambda}(\beta^{(k)}) + \sqrt{2\delta M} \xi^{(k+1)}, \quad \xi^{(k+1)} \sim \mathcal{N}(0, I_p).$$

Thus a constant preconditioner changes the geometry in which the target is explored, without changing the intended smoothed posterior. The timestep must nevertheless be retuned in the transformed geometry. For a quadratic smooth term with Hessian H_f , we denote the corresponding smooth-gradient Lipschitz constant in the preconditioned geometry by

$$L_{f,M} = \lambda_{\max}(M^{1/2} H_f M^{1/2}).$$

Diagonal scaling.

When poor conditioning is distributed across coordinates, one can use a diagonal, or Jacobi, approximation to the Hessian of the smooth term. Writing $H_f = \nabla^2 f$ for the Hessian of the quadratic smooth term, we take

$$D_\varepsilon = \text{diag}(H_f) + \varepsilon I_p, \quad \varepsilon > 0, \quad M_J = D_\varepsilon^{-1},$$

where ε is a small ridge term that prevents zero or very small diagonal entries from producing unstable inverse rescaling. Up to an overall scalar normalisation that can be absorbed into δ , this rescales coordinates according to the diagonal entries of H_f and is inexpensive to construct even when p is large. It is useful when the main numerical difficulty is heterogeneous marginal scaling; being diagonal, it is not intended to remove strong off-diagonal dependence completely.

Low-rank directional correction.

A different geometry arises from a soft affine relation. For a single condition $a^\top \beta = b$, the pseudo-observation adds

$$\frac{1}{2\tau_c^2} (a^\top \beta - b)^2$$

to the smooth potential and hence contributes the rank-one Hessian term $\tau_c^{-2} a a^\top$. Let

$$u = \frac{a}{\|a\|_2}, \quad P_{\parallel} = u u^\top, \quad P_{\perp} = I - P_{\parallel}.$$

When the largest eigenvalue of the smooth Hessian is driven primarily by the soft affine term, it is useful to distinguish the overall smooth-gradient Lipschitz constant from the corresponding constant within the tangent space. For this quadratic problem, define

$$L_f = \lambda_{\max}(H_f), \quad L_{\perp} = \sup_{\substack{v \perp u \\ \|v\|_2=1}} v^{\top} H_f v.$$

Thus L_f is the largest eigenvalue of the smooth Hessian, while $L_{\perp}$ is its largest Rayleigh quotient over directions tangent to the affine hyperplane. We then use

$$M_{\phi} = P_{\perp} + \phi P_{\parallel}, \quad \phi = \min \left\{ 1, \frac{L_{\perp}}{L_f} \right\}.$$

The update therefore leaves tangent directions unchanged while damping movement in the single stiff direction enough to bring its numerical scale closer to the rest of the problem. The same idea extends to several affine relations by replacing $P_{\parallel}$ with the projector onto the relevant low-dimensional constraint subspace.

There is one additional interaction with Moreau smoothing. If the soft affine Hessian term materially increases L_f , the default choice $\lambda = 1/L_f$ also makes the Moreau approximation unnecessarily sharp in the better-scaled tangent directions. In that case we decouple the two scales and use the tangent-space constant $L_{\perp}$ to set the Moreau parameter, while the timestep is chosen from the Lipschitz bound in the preconditioned geometry.

Hence the two preconditioners address different mechanisms: diagonal scaling treats broadly distributed coordinatewise imbalance, whereas the projector construction treats identifiable low-rank stiffness. We assess the effect of preconditioning through mixing, effective sample size and the Monte Carlo precision of the posterior quantities entering the final decision rule.

3 Numerical studies

Study 1 tests the full unconstrained workflow and, in particular, whether calibration and Monte Carlo choices alter the final sparse decision in the $n > p$ and $p > n$ regimes. Study 2 isolates the statistical and numerical consequences of soft affine information. Study 3 asks whether the resulting posterior uncertainty and sparse decision remain credible on a real regression problem with correlated predictors. The supplement contains the hard-constraint calibration check and the observation-scale/Moreau sensitivity study. Unless stated otherwise, selection results use $(k, \pi_{\star}) = (2.5, 0.75)$ as the reference setting and report the full grid as a sensitivity analysis.

3.1 Study 1: End-to-end sparse regression

Study 1 evaluates the complete workflow under a known sparse ground truth, with emphasis on how upstream calibration and posterior Monte Carlo precision propagate to the final support decision. We use the same signal pattern in an overdetermined and an underdetermined regime and report likelihood-scale calibration, MAP estimation and prediction, sparse selection, and decision-level Monte Carlo precision.

Rows of the Gaussian design are generated with AR(1) covariance $\Sigma_{jk} = 0.4^{|j-k|}$, after which each realised predictor column is rescaled to $\|X_j\|_2 = \sqrt{n}$. The response follows $y = X\beta^* + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, I_n)$, so that $\sigma = 1$. There are ten nonzero coefficients, with alternating signs and decreasing magnitudes, rescaled so that the population signal-to-noise ratio satisfies

$$\frac{\beta^{*\top} \Sigma \beta^*}{\sigma^2} = 3.$$

We consider

$$(n, p) = (500, 200) \quad \text{and} \quad (n, p) = (200, 500),$$

with $R = 100$ independently generated datasets in each regime; the additional predictors in the high-dimensional regime are null. The decision rule is evaluated over $k \in \{2, 2.5, 3\}$ and $\pi_* \in \{0.50, 0.75, 0.90\}$, with $(2.5, 0.75)$ used only as a reference operating point. SAPG uses 4000 iterations with tail averaging from iteration 2000. The initial calculation retained 4000 MYULA draws after burn-in and was subsequently refined to 40,000 retained draws with all upstream quantities fixed. The full settings and the complete decision grid are given in the supplementary material. Throughout the numerical studies, TPR denotes true-positive rate, FDR false-discovery rate, and support RMSE the root mean squared coefficient error restricted to the true support.

The two regimes use the two scale-estimation branches of subsection 2.1. When $n > p$, the classical residual estimator is centred close to the truth: across the $R = 100$ datasets,

$$\text{mean}(\hat{\sigma}/\sigma) = 0.9979, \quad \text{sd}(\hat{\sigma}/\sigma) = 0.0415.$$

The high-dimensional branch required more care. In an initial five-dataset pilot, the raw scaled Lasso estimate was materially inflated, with mean $\hat{\sigma}_{\text{SL}} = 1.514$, and the initially narrow SAPG projection interval also produced boundary hits. Widening the projection interval removed this numerical artefact, but selection remained weak when the inflated scale was retained. Replacing the estimated scale by the true $\sigma = 1$ recovered much of the lost sensitivity and improved prediction performance. This separated the two effects and motivated a feasible correction to the scale estimate: we use the scaled Lasso only to identify a provisional support, refit OLS on that support, and estimate σ^2 from the refitted residuals as in (8). In the same five datasets, this reduced the mean estimated scale from 1.514 to 1.138, with behaviour much closer to the true-noise comparison.

Accordingly, the reported $R = 100$ experiment uses the post-selection OLS scale, while the scaled Lasso coefficients themselves play no role beyond defining the provisional support. In the $p > n$ datasets, the raw scaled Lasso scale has mean 1.4820, whereas the post-OLS estimate has mean 1.1173 and standard deviation 0.0776. The remaining upward displacement is consistent with provisional supports that omit weak signals; the supplementary analysis records their composition explicitly. This sequence is important for the interpretation of the later studies as it clarifies the role of the observation-scale estimate in the full workflow.

Table 1 summarises the output at the reference decision setting, using the final 40,000-draw posterior calculations for all selection quantities. Relative L_2 error and

true-support RMSE refer to the nonsmoothed MAP, while prediction risk is evaluated under the population design covariance.

Table 1 Study 1 results at the reference setting $(k, \pi_\star) = (2.5, 0.75)$. Parentheses give standard errors over $R = 100$ datasets; selection uses the 40,000-draw MYULA run.

Regime	$\hat{\sigma}/\sigma$	$\hat{\theta}$	Rel. L_2	Supp. RMSE	Pred. risk	TPR	FDR	Size	Exact
$n = 500, p = 200$	0.998 (.004)	15.711 (.041)	0.215 (.003)	0.065 (.001)	0.130 (.003)	0.989 (.003)	0.006 (.002)	9.96 (.04)	0.830 (.038)
$n = 200, p = 500$	1.117 (.008)	17.334 (.067)	0.343 (.004)	0.165 (.003)	0.347 (.009)	0.429 (.012)	0.000 (.000)	4.29 (.12)	0.000 (.000)

The low-dimensional regime, $n > p$, behaves as a well-informed sparse regression problem. The noise scale is accurately calibrated, estimation and prediction errors remain moderate, and the reference rule recovers on average 98.9% of the true variables with FDR 0.0065; the exact ten-variable support is recovered in 83% of datasets. The high-dimensional regime, $p \geq n$, is qualitatively different. Estimation and prediction remain meaningful, but the final decision at the reference setting, $(k, \pi_\star) = (2.5, 0.75)$, is conservative: it selects 4.29 variables on average with TPR 0.429, but no false positives are observed in the 100 replications. Signal-wise selection frequencies decrease almost monotonically with true coefficient magnitude, from 0.96 for the largest signal to zero for the smallest. The loss of exact support recovery is therefore driven by omissions of progressively weaker signals, rather than by indiscriminate inclusion of zero coordinates.

Figure 1 makes the two-stage mechanism explicit in the high-dimensional regime: MAP eligibility remains high for moderate signals, while the posterior activation gate progressively removes weaker effects.

The full 3×3 decision grid is obtained by post-processing the same posterior sample; neither the model nor the empirical Bayes calibration is refitted when k or $\pi_\star$ changes. In the low-dimensional regime, the grid displays the expected sensitivity-specificity tradeoff: the permissive setting $(2, 0.50)$ has TPR 1 and FDR 0.105, while more stringent settings almost eliminate false discoveries with only a modest loss of sensitivity. In the high-dimensional regime, FDR is zero over almost the whole grid; TPR ranges from 0.675 at $(2, 0.50)$ to 0.229 at $(3, 0.90)$. Thus, the two decision parameters mainly control how much weak-signal sensitivity is retained once the problem becomes information-limited.

Increasing the retained MYULA sample from 4000 to 40,000 draws separates the statistical behaviour above from finite Monte Carlo precision. Median coordinate ESS increases from 212 to 2093 in the $n > p$ regime and from 134 to 1269 in the $p > n$ regime, while aggregate TPR and FDR move by less than one percentage point. Let $\hat{S}_{4k}$ and $\hat{S}_{40k}$ denote the selected sets obtained from the shorter and longer posterior

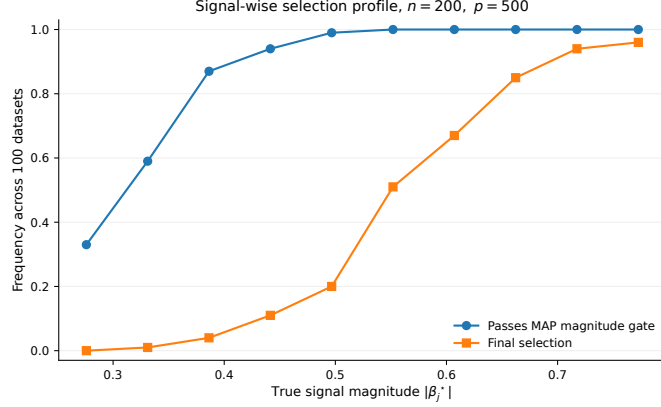


Fig. 1 Signal-wise behaviour at the reference decision rule in the high-dimensional regime. For each true signal, the figure shows the frequency with which it passes the MAP magnitude gate and the frequency with which it is retained by the complete MAP-plus-posterior rule.

runs. These sets are identical in 96% of low-dimensional replicates and 74% of high-dimensional replicates; the mean values of $|\hat{S}_{4k} \triangle \hat{S}_{40k}|$, the number of coordinates selected in exactly one of the two runs, are only 0.04 and 0.33, respectively.

The decision-aware diagnostic D_j , defined in (44), measures the distance between the estimated activation probability for coefficient j and the probability threshold π_* , in Monte Carlo standard-error units. For each replicate, we restrict attention to coordinates that pass the MAP magnitude gate and record

$$D_{\min, \text{elig}} = \min_{j: |\hat{\beta}_{\text{MAP}, j}| \geq \tau_{\text{post}}} D_j.$$

Thus, $D_{\min, \text{elig}}$ measures how close the most decision-sensitive MAP-eligible coefficient is to the posterior-probability boundary: a small value means that at least one eligible coefficient has an activation probability close to π_* relative to its Monte Carlo uncertainty. The median of this quantity increases from 7.03 to 27.87 in the $n > p$ regime and from 0.81 to 2.16 in the $p > n$ regime. Even after 40,000 draws, 46% of the high-dimensional replicates contain at least one MAP-eligible coordinate within two MCSE units of the activation threshold. Thus, the aggregate recovery rates are stable, while individual high-dimensional inclusion decisions can remain close enough to the probability gate for Monte Carlo precision to matter. Increasing the retained sample tenfold was computationally inexpensive with MYULA in these experiments; environment-specific timings are reported in the supplement.

Figure 2 visualises the same distinction: the longer posterior run moves the most decision-sensitive eligible coordinates farther from the probability boundary in Monte Carlo standard-error units.

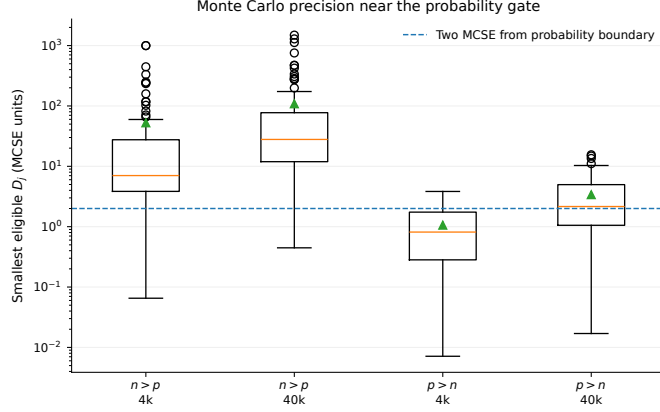


Fig. 2 Monte Carlo precision of the posterior-probability gate under the 4000- and 40,000-draw calculations. The plotted quantity is the smallest D_j among MAP-eligible coordinates in each replicate; the dashed line marks two Monte Carlo standard errors from the probability threshold.

In conclusion, Study 1 establishes the baseline behaviour used by the remaining experiments. It shows that the complete pipeline is well behaved in the $n > p$ regime, that the high-dimensional difficulty is primarily conservative loss of weak signals, and that nuisance-scale calibration and posterior Monte Carlo error can be diagnosed separately.

3.2 Study 2: Soft affine information and numerical geometry

Study 2 isolates the second part of the methodology: how affine information affects both empirical Bayes calibration and posterior computation. In our index-tracking application (Roxanas 2025), construction uses a soft sum-to-one condition and rebalancing uses hard self-financing changes with $\mathbf{1}^\top \Delta w = 0$. Here we isolate the soft nonhomogeneous case in a controlled regression experiment. We examine how stronger sum-to-one information changes calibration, estimation and selection, and how the same term changes the geometry seen by the Langevin sampler. The hard homogeneous calibration is checked separately in the supplement.

We use $R = 100$ independently generated datasets with $n = 200$, $p = 100$, known $\sigma = 1$, and eight nonzero coefficients satisfying $\mathbf{1}^\top \beta^* = 1$. The smooth potential is augmented by the Gaussian pseudo-observation term

$$\frac{1}{2\tau_c^2} (\mathbf{1}^\top \beta - 1)^2$$

that encodes the constraint. Writing $L_0 = \lambda_{\max}(X^\top X)$, we index the strength of the affine information by

$$\kappa_c = \frac{p}{\tau_c^2 L_0}, \quad \kappa_c \in \{0, 0.1, 1, 10\}.$$

This parameter has a direct geometric interpretation: the rank-one affine Hessian $\tau_c^{-2} \mathbf{1}\mathbf{1}^\top$ has nonzero eigenvalue $p/\tau_c^2 = \kappa_c L_0$. Thus κ_c compares the curvature introduced in the sum direction with the baseline likelihood-curvature scale; $\kappa_c = 0$ corresponds to omitting the pseudo-observation. Because the relation is imposed softly rather than by restricting the prior support, the prior remains on the ambient space and the SAPG score continues to use $d = p$ throughout. The full simulation settings and decision grids are reported in the supplement.

The empirical Bayes scale changes only moderately as the affine information strengthens, from mean $\hat{\theta} = 9.772$ at $\kappa_c = 0$ to 10.346 at $\kappa_c = 10$. The effect on adherence to the supplied relation is much larger: the mean posterior value of $|\mathbf{1}^\top \beta - 1|$ decreases from 0.685 to 0.112. Point estimation and selection improve more modestly. MAP relative L_2 error falls from 0.2557 to 0.2482, while at the reference decision rule, $(k, \pi_\star) = (2.5, 0.75)$, mean TPR increases from 0.8025 to 0.8200; FDR remains below 0.003 at both endpoints. Table 2 summarises the final posterior calculation used for each value of κ_c . Here Med. ESS is the ensemble median of the coordinatewise median effective sample size, Weak-tangent ESS is the ensemble median of the effective sample size after projection onto the least-curved tangent-space eigendirection, and D_{elig} is the median, across datasets, of the smallest D_j among MAP-eligible coordinates.

Table 2 Study 2 results at $(k, \pi_\star) = (2.5, 0.75)$. The $\kappa_c = 10$ row uses the geometry-aware calculation; the baseline scalar-step calculation is reported in the supplement.

κ_c	$\hat{\theta}$	MAP rel. L_2	$E \mathbf{1}^\top \beta - 1 $	Med. ESS	Weak- tangent ESS	TPR	FDR	D_{elig}
0	9.772	0.2557	0.6851	301.6	146.5	0.8025	0.00268	5.49
0.1	9.785	0.2549	0.5955	302.9	146.9	0.8063	0.00268	5.49
1	9.976	0.2518	0.3418	212.7	106.5	0.8150	0.00125	4.31
10	10.346	0.2482	0.1118	302.7	160.8	0.8200	0.00250	5.03

As discussed in Sections 1 and 2, the same affine information that improves adherence also changes the numerical problem. When MYULA is used without geometry-aware preconditioning, a single Euclidean timestep is chosen from the global smoothness constant L_f . The effect is already visible at $\kappa_c = 1$ as mixing deterioration: median coordinate ESS falls from about 302 to 213. At $\kappa_c = 10$, it becomes pronounced. Across the ensemble,

$$\frac{L_f}{L_\perp} = 10.3977$$

on average, where $L_\perp$ is the smooth curvature restricted to the tangent space orthogonal to $\mathbf{1}$. More explicitly, if v_1 is the unit leading eigenvector of the smooth Hessian and $u = \mathbf{1}/\sqrt{p}$ is the unit sum direction, then the ensemble mean of $|v_1^\top u|^2$ is 0.999366. Thus the direction of largest curvature is, to numerical accuracy, almost parallel to the sum direction. The scalar timestep is therefore governed almost entirely by this single

stiff direction. Median coordinate ESS falls to 33.8, the median ESS in the weakest tangent direction to 20.1, and the median minimum decision distance among MAP-eligible variables to 1.25. In contrast, the sum direction itself has median ESS above 1100. Therefore, the difficulty is not failing to explore the affine direction; it is that curvature in that direction forces the rest of the sampler to evolve much more slowly.

For $\kappa_c = 10$ we consequently use the rank-one preconditioning matrix from Section 2.6,

$$M_\phi = P_\perp + \phi P_\parallel, \quad \phi = \min\{1, L_\perp/L_f\},$$

and set the Moreau scale from the tangent curvature $L_\perp$ rather than from the affine-dominated L_f . The mean ϕ is 0.0962, and in the transformed geometry the mean smooth-curvature ratio is reduced to $L_{f,M}/L_\perp = 1.022$. Using M_ϕ with λ fixed changes the geometry of the Langevin dynamics while leaving the intended smoothed posterior unchanged. Here we additionally replace the default $\lambda = 1/L_f$ by $\lambda = 1/L_\perp$. Since $L_\perp < L_f$, this uses a larger Moreau parameter and therefore changes the smoothed approximation itself, not only how that approximation is explored. For this reason, we designed this branch primarily as a geometry-aware posterior calculation rather than as a pure preconditioner-only comparison; the timestep choice is audited separately in the supplement.

The resulting gain in Monte Carlo reliability is large. Relative to the baseline scalar-step $\kappa_c = 10$ calculation, median coordinate ESS increases from 33.8 to 302.7, the median minimum coordinate ESS from 9.1 to 127.4, and weak-tangent ESS from 20.1 to 160.8. The median minimum D_j among MAP-eligible variables increases from 1.25 to 5.03. Taking, in each dataset, the MAP-eligible coordinate whose activation probability lies closest to the posterior-probability threshold, the median standardised separation from that threshold increases from about 1.25 to about 5 Monte Carlo standard errors. Thus the probability-gate decisions are much better resolved relative to Monte Carlo error. In contrast, the reported selection summaries move very little: mean selected-set size changes from 6.65 to 6.58, mean TPR from 0.825 to 0.820, and mean FDR from 0.0063 to 0.0025.

Figure 3 makes the geometry of the $\kappa_c = 10$ calculation more explicit across the $R = 100$ datasets. The weak- and strong-tangent summaries are the ESS values along the least- and most-curved eigendirections of the smooth Hessian restricted to the tangent space, while the median-coordinate ESS gives a robust summary of typical coordinatewise mixing. The sum-direction ESS isolates the affine normal direction $u = \mathbf{1}/\sqrt{p}$. Without preconditioning, the sum direction itself is already explored efficiently, whereas mixing deteriorates substantially in the tangent directions and across the original coordinates. Preconditioning largely restores that exploration while leaving the sum-direction ESS at the same high order.

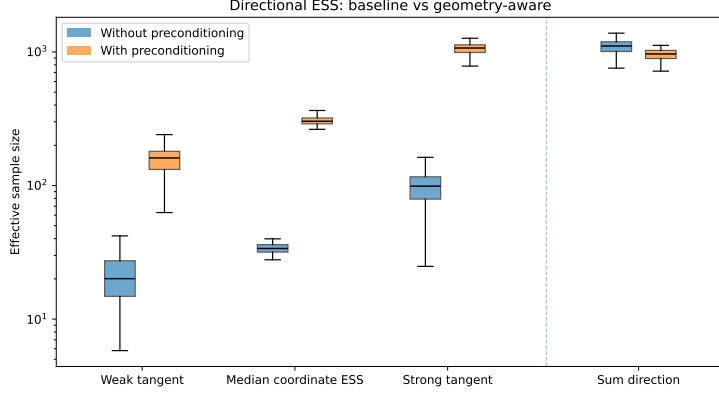


Fig. 3 Directional effective sample sizes at $\kappa_c = 10$, across $R = 100$ datasets, with and without preconditioning. The four summaries are the weak-tangent, median-coordinate, strong-tangent and sum-direction ESS.

Study 2 therefore separates two consequences of softly specifying affine information. Strengthening the soft relation genuinely changes the fitted posterior and substantially improves adherence to the relation, with only modest changes in estimation and support recovery. At the same time, its rank-one curvature can make a scalar-step posterior calculation unnecessarily inefficient. Matching the numerical geometry to that low-rank structure restores decision-level Monte Carlo precision without materially changing the sparse conclusion.

3.3 Study 3: Diabetes data and external benchmarking

Study 3 provides an external check on the complete workflow in a real regression problem with correlated predictors. We examine calibration, posterior computation, stability of the sparse decision across the prespecified grid, and agreement of the resulting posterior uncertainty with established Bayesian analyses of the same data.

We worked with the diabetes data used by (Efron et al. 2004). The outcome there is a quantitative measure of disease progression over a year, and the covariates are Age, Sex, Body Mass Index (BMI), average Blood Pressure (BP), and six blood serum measurements (labelled S1-S6). There are $n = 442$ observations and $p = 10$ predictors; the response is centred and divided by its sample standard deviation, and the predictors are used as supplied. The design is full rank, while the Gram matrix has condition number about 470. The largest absolute pairwise predictor correlation is approximately 0.897. The model is therefore small enough to remain fully interpretable, but contains weak coefficient-attribution directions that make posterior uncertainty nontrivial.

Accounting for the fitted intercept through the residual degrees of freedom, the classical residual estimator gives $\hat{\sigma}^2 = 0.493440$. The extended empirical Bayes calculation gave $\hat{\theta} = 0.337289$, which was then fixed for all downstream calculations; the SAPG settings and convergence diagnostics are reported in the supplement. We retained the default Moreau choice $\lambda = 1/L_f$. In this example, the nonsmoothed FISTA MAP had eight nonzero coordinates, while the corresponding smoothed mode

was numerically very close; the detailed mode comparison is also reported in the supplement.

The initial scalar-step MYULA run showed slow exploration in the weakest likelihood directions, but the sampler is sufficiently computationally cheap here that extending the run was preferable to introducing an additional preconditioner or retuning λ . The final calculation therefore retained the same fitted model, Moreau scale and sampler geometry, with a longer simulation horizon. Median coordinate ESS was 1589.6, the minimum coordinate ESS was 198.8, and the weakest-likelihood-direction ESS was 192.2. The full short-versus-long audit is given in the supplement.

The long-chain 95% marginal credible intervals exclude zero for Sex, BMI, BP, and S5. The final sparse action is stricter than interval zero-exclusion because it also applies the practical magnitude threshold and posterior activation gate from subsection 2.5. At the reference operating point $(k, \pi_\star) = (2.5, 0.75)$, the selected support is

$$\{\text{BMI, BP, S5}\}.$$

Thus, the posterior-informed decision reduces the eight-coordinate nonsmoothed MAP support to three coordinates at the reference operating point. The full prespecified grid in Table 3 gives a more informative stability summary: BMI and S5 are selected in all nine settings, blood pressure in eight, Sex in two, and every other variable in none.

Table 3 Diabetes selection grid. The reference setting $(k, \pi_\star) = (2.5, 0.75)$ is shown in bold.

	$\pi_\star = .50$	$\pi_\star = .75$	$\pi_\star = .90$
$k = 2.0$	Sex, BMI, BP, S5	Sex, BMI, BP, S5	BMI, BP, S5
$k = 2.5$	BMI, BP, S5	BMI, BP, S5	BMI, BP, S5
$k = 3.0$	BMI, BP, S5	BMI, BP, S5	BMI, S5

Two coordinates show why the posterior stage adds information beyond the sparse MAP. For Sex, the 95% credible interval lies entirely below zero, but the MAP magnitude lies just below the reference practical threshold τ_{post} . Its reference activation probability is only 0.503, so Sex behaves as a practical-magnitude boundary case, and enters the selected support only at the less stringent $k = 2$ settings with $\pi_\star \leq 0.75$. Variable S3 shows the complementary mechanism: at $k = 2$ its MAP magnitude exceeds the practical threshold, but its posterior activation probability remains below 0.5, so it is selected in none of the nine settings. The distinction between MAP eligibility and posterior-supported magnitude is therefore visible in the real-data example itself.

The same operational data scaling is used in the public diabetes demonstration of (Zhou et al. 2024). Their accompanying supplementary notebook stores the numerical interval output underlying its published comparison, so we reconstructed that comparison from the authors’ recorded numerical values rather than rerunning their

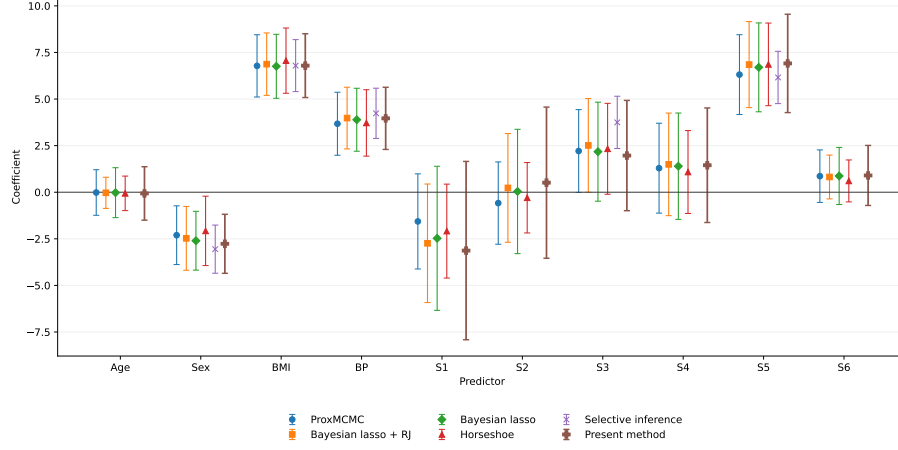


Fig. 4 95% coefficient intervals for the diabetes data. Benchmark intervals are reconstructed from the public notebook of Zhou et al. (2024); the present intervals use the long MYULA run. Symbols mark interval midpoints only for visual alignment.

mixed Julia/R analysis. Figure 4 appends the interval from our long MYULA calculation (EB-Laplace + MYULA analysis) to the recorded 95% intervals for ProxMCMC, Bayesian lasso with and without reversible-jump selection, and horseshoe intervals, together with the selective-inference intervals reported in the same notebook across Age, Sex, BMI, BP and S1–S6. For the reconstructed benchmark methods, centre markers are interval midpoints used only for visual alignment. The benchmark reconstruction and all interval endpoints are documented in the supplement.

The Bayesian interval comparisons show the same zero-exclusion pattern. ProxMCMC, Bayesian lasso with and without reversible-jump variable selection, the horseshoe, and the present analysis all give exactly the same 95% Bayesian zero-exclusion pattern:

$$\{\text{Sex}, \text{BMI}, \text{BP}, \text{S5}\}.$$

For the strongly identified coefficients BMI, BP and S5, the interval endpoints are also close across the Bayesian procedures. The largest differences occur in the correlated serum block, where the present analysis yields wider marginal credible intervals without changing the common zero-exclusion pattern. Selective inference additionally reports an interval excluding zero for S3, but those intervals are conditional frequentist objects and the associated lasso support solves a different selection problem.

The interval differences in the correlated serum block are also consistent with the geometry of the regression problem. Strong predictor correlation creates weakly identified directions of the likelihood: some linear combinations of coefficients are well determined by the data, while movements along other combinations change the fitted response only slightly. Posterior mass can therefore extend substantially along these weak directions, and this spread is inherited by the marginal intervals of the individual coefficients even when prediction, or better-identified coefficient combinations, remain comparatively stable. Different prior and shrinkage constructions regularise these weak directions differently, so larger differences between Bayesian procedures

are most naturally expected precisely in this part of the parameter space; in our calculation, extending MYULA greatly increased the effective sample size in the weakest likelihood directions, but the marginal posterior standard deviations did not decrease and, in fact, increased slightly on average.

Study 3 therefore provides two complementary real-data checks. The posterior uncertainty from the calibrated proximal workflow is consistent with established Bayesian analyses of the same dataset, while the MAP-plus-posterior rule adds a separate practical-magnitude decision layer. Reporting the full decision grid, rather than only its reference operating point, makes that distinction explicit.

4 Discussion

The main contribution of this paper is a sparse Bayesian regression workflow in which calibration, optimisation, posterior simulation and selection are treated as separate tasks but implemented through a common proximal structure. The global shrinkage parameter is calibrated by empirical Bayes, the nonsmoothed MAP remains the reported point estimator, the smoothed posterior supplies uncertainty, and the final rule converts that uncertainty into a sparse decision. This separation makes it possible to diagnose where a change in the final support comes from: likelihood-scale calibration, shrinkage calibration, the fitted sparse mode, posterior uncertainty, or Monte Carlo error. It also retains the numerical advantages of convex regularisation without requiring discrete model-space exploration or a high-dimensional hierarchy of local shrinkage parameters.

The numerical studies show why this separation matters. In the $n > p$ regime, scale calibration, MAP recovery and the final decision are all stable, and increasing the posterior run length has little effect on aggregate recovery. In the $p > n$ regime, by contrast, weak-signal losses arise primarily from the statistical calibration of the problem rather than from an excess of false discoveries, while the decision-aware Monte Carlo diagnostic identifies individual coefficients whose inclusion remains close to the probability boundary. Study 2 exposes a different failure mode: strong soft affine information improves adherence to the supplied relation but creates low-rank stiffness that can make a scalar-step Langevin calculation unnecessarily slow. Geometry-aware scaling restores effective posterior exploration with little change to the sparse conclusion. On the diabetes data, the posterior intervals agree closely with several established Bayesian analyses while the additional magnitude-and-probability rule gives a smaller practical support.

The affine calculations make a further distinction between statistical specification and numerical geometry. A hard homogeneous constraint changes the support of the prior and hence the dimension entering the empirical Bayes score. A nonhomogeneous affine relation does not admit the same scaling argument, but can be introduced softly without changing the homogeneous prior normalisation. That soft formulation preserves the basic SAPG machinery, while its induced curvature can be handled separately at the sampling stage. Statistical specification and numerical geometry are therefore related but need not be treated as the same problem.

The framework is modular by construction. Observation-scale calibration, empirical Bayes shrinkage, the nonsmoothed MAP, posterior approximation and the final decision can each be changed without redefining the entire procedure. The proximal optimisation and sampling components extend directly to other proper closed convex penalties with tractable proximal maps, including elastic net penalties, and Moreau smoothing can be paired with gradient-based samplers other than MYULA. When a Gaussian observation-noise interpretation is unavailable, the same architecture can instead use a generalised Bayes learning rate.

There are some limits to the present formulation. Plug-in estimates of the observation and shrinkage scales do not propagate hyperparameter uncertainty, and a single global Laplace scale cannot adapt to heterogeneous signals as flexibly as global–local or richer hierarchical priors. The final gate is coordinatewise once the joint MAP has supplied the candidate configuration, so strongly correlated problems may benefit from decisions based more directly on joint posterior geometry. These limitations do not alter the computational point of the paper: a simple convex sparse model can support calibrated shrinkage, a nonsmoothed sparse estimator, posterior uncertainty, affine information and an auditable terminal decision within one reusable proximal architecture. The resulting workflow provides a computationally efficient middle ground between reporting a penalised point estimate alone and fitting a substantially richer hierarchical selection model.

Supplementary information. The supplement contains full decision grids, additional SAPG and MYULA diagnostics, MAP-smoothing and preconditioning audits, the complete diabetes benchmark, the hard-constraint calibration check, and the observation-scale/Moreau sensitivity study. It is currently accessible from the repository given under Code availability.

Declarations

Funding. The author received no specific funding for this work.

Competing interests. The author declares no competing interests.

Data availability. The diabetes data are publicly available through the sources associated with (Efron et al. 2004) and the public notebook accompanying (Zhou et al. 2024).

Code availability. Code for the reported synthetic experiments, the diabetes analysis and benchmark reconstruction is available at <https://github.com/droxanas/ProxEBS-SparseRegression>. The repository also contains the analysis scripts, fixed random seeds, and the diabetes and benchmark input files needed to reproduce the reported numerical results and figures.

References

- Bondell, H.D., Reich, B.J.: Consistent High-Dimensional Bayesian Variable Selection via Penalized Credible Regions. *Journal of the American Statistical Association* **107**(500), 1610–1624 (2012) <https://doi.org/10.1080/01621459.2012.716344>
- Beck, A., Teboulle, M.: A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM Journal on Imaging Sciences* **2**(1), 183–202 (2009) <https://doi.org/10.1137/080716542>
- De Bortoli, V., Durmus, A., Pereyra, M., Fernandez Vidal, A.: Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach. Part II: Theoretical Analysis. *SIAM Journal on Imaging Sciences* **13**(4), 1990–2028 (2020) <https://doi.org/10.1137/20M1339842>
- Durmus, A., Moulines, É., Pereyra, M.: Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. *SIAM Journal on Imaging Sciences* **11**(1), 473–506 (2018) <https://doi.org/10.1137/16M1108340>
- Duan, L.L.: Revisiting Bayesian Variable Selection via Optimization. arXiv preprint arXiv:2604.21009 (2026) <https://doi.org/10.48550/arXiv.2604.21009>
- Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least Angle Regression. *The Annals of Statistics* **32**(2), 407–499 (2004) <https://doi.org/10.1214/009053604000000067>
- Fernandez Vidal, A., De Bortoli, V., Pereyra, M., Durmus, A.: Maximum Likelihood Estimation of Regularization Parameters in High-Dimensional Inverse Problems: An Empirical Bayesian Approach. Part I: Methodology and Experiments. *SIAM Journal on Imaging Sciences* **13**(4), 1945–1989 (2020) <https://doi.org/10.1137/20M1339829>
- Hahn, P.R., Carvalho, C.M.: Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective. *Journal of the American Statistical Association* **110**(509), 435–448 (2015) <https://doi.org/10.1080/01621459.2014.993077>
- Park, T., Casella, G.: The Bayesian Lasso. *Journal of the American Statistical Association* **103**(482), 681–686 (2008) <https://doi.org/10.1198/016214508000000337>
- Pereyra, M.: Proximal Markov Chain Monte Carlo Algorithms. *Statistics and Computing* **26**, 745–760 (2016) <https://doi.org/10.1007/s11222-015-9567-4>
- Pereyra, M., Mieses, L.V., Zygalakis, K.C.: Accelerating proximal Markov chain Monte Carlo by using an explicit stabilized method. *SIAM Journal on Imaging Sciences* **13**(2), 905–935 (2020) <https://doi.org/10.1137/19M1283719>
- Piironen, J., Paasiniemi, M., Vehtari, A.: Projective Inference in High-Dimensional Problems: Prediction and Feature Selection. *Electronic Journal of Statistics* **14**(1),

- 2155–2197 (2020) <https://doi.org/10.1214/20-EJS1711>
- Roxanas, D.: Low-Turnover Rebalancing for Sparse Index Tracking. arXiv preprint arXiv:2512.22109v2 (2025) <https://doi.org/10.48550/arXiv.2512.22109>
- Sun, T., Zhang, C.-H.: Scaled Sparse Linear Regression. *Biometrika* **99**(4), 879–898 (2012) <https://doi.org/10.1093/biomet/ass043>
- Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(1), 267–288 (1996) <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- Xu, M., Duan, L.L.: Bayesian inference with the l1-ball prior: solving combinatorial problems with exact zeros. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **85**(5), 1538–1560 (2023) <https://doi.org/10.1093/jrsssb/qqad076>
- Xu, M., Zhou, H., Hu, Y., Duan, L.L.: Bayesian inference using the proximal mapping: Uncertainty quantification under varying dimensionality. *Journal of the American Statistical Association* **119**(547), 1847–1858 (2024) <https://doi.org/10.1080/01621459.2023.2220170>
- Zhou, X., Heng, Q., Chi, E.C., Zhou, H.: Proximal MCMC for Bayesian Inference of Constrained and Regularized Estimation. *The American Statistician* **78**(4), 379–390 (2024) <https://doi.org/10.1080/00031305.2024.2308821>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementSC.pdf](#)