Mr.ASH: A Flexible Empirical Bayes Approach to Multiple Linear Regression with Adaptive Shrinkage Priors

Li Yuekai

City University of Hong Kong

Journal club 2025.02.28

香港城市大學
City University of Hong Kong

Contents

Introduction

Multiple Linear Regression (MLR)

Basic Model

$$
\begin{aligned}
y &= \mathrm{b}_1 x_1 + \mathrm{b}_2 x_2 + \cdots + \mathrm{b}_p x_p + \varepsilon \\
\varepsilon &\sim \mathrm{N}(0, \sigma^2)
\end{aligned}
\tag{1.1}
$$

Challenges

- Rapidly increasing data sizes and high dimensionality.
- How to optimize out-of-sample prediction accuracy, and perform variable selection and inference?

Solutions

- **Penalized Linear Regression**: Incorporate penalty terms to shrink coefficients, improving prediction and interpretability.
- **Bayesian Methods**: Impose parameter priors to shrink coefficients, leveraging empirical Bayes approaches for large-scale regression.

Penalized Linear Regression (PLR) methods

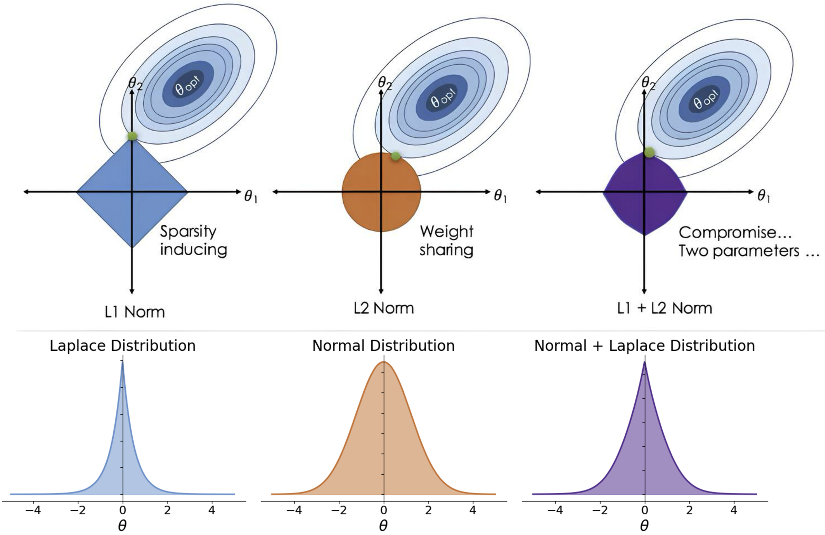
Basic Model

PLR methods estimate the regression coefficients by minimizing a penalized squared-loss function:

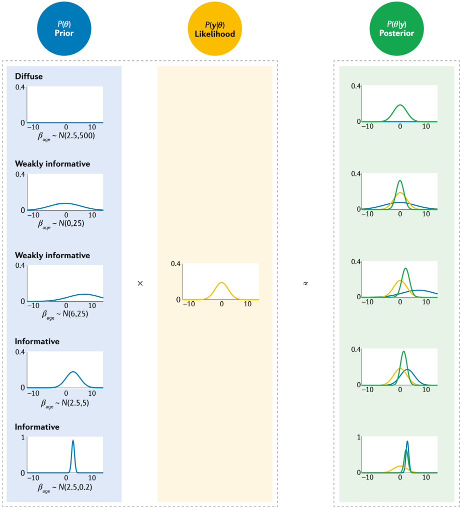
$$\min_{\mathbf{b}\in\mathbb{R}^p} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{b}\|^2 + \sum_{j=1}^{p}\rho(b_j) \tag{1.2}$$

Model Name	**Penalty Term** $\rho(t)$	**Corresponding Prior**
Ridge (L_2 regularization)	$\frac{\lambda t^2}{2}$	Normal
Lasso (L_1 regularization)	$\lambda\|t\|$	Laplace
Elastic Net	$\lambda\cdot(\frac{(1-\eta)t^2}{2}+\eta\|t\|)$	Normal + Laplace
L_0 regularization	$\lambda\cdot\mathbb{I}\{\|t\|>0\}$	
MCP	$\begin{cases}\lambda\|t\| - \frac{t^2}{2\eta}, & \text{if } \|t\|\le\eta\lambda,\\ \frac{\eta\lambda^2}{2}, & \text{otherwise.}\end{cases}$	

Penalty term versus Prior distribution



Bayesian and Empirical Bayes methods



Model Name	Prior ($b_j \sim \cdot$)
LMM	$\mathrm{N}(0, \sigma_b^2)$
BayesA	$\mathrm{t}(0, \nu, \sigma_b^2)$
BVSR,varbvs BayesCπ	$\pi \mathrm{N}(0, \sigma_b^2) + (1-\pi)\delta_0$
BayesB,BayesD BayesDπ	$\pi \mathrm{t}(0, \nu, \sigma_b^2) + (1-\pi)\delta_0$
BSLMM	$\pi \mathrm{N}(0, \sigma_a^2 + \sigma_b^2) +$ $(1-\pi)\mathrm{N}(0, \sigma_b^2)$
Bayesian Lasso	$\mathrm{DE}(0, \theta)$

*Red represents Empirical Bayes methods

Methodology

Unimodal prior distribution

JOURNAL ARTICLE

False discovery rates: a new deal

Matthew Stephens ✉

Biostatistics, Volume 18, Issue 2, April 2017, Pages 275–294,
https://doi.org/10.1093/biostatistics/kxw041
Published: 17 October 2016 **Article history** ▾

PDF | Split View | Cite | Permissions | Share ▾

Summary

We introduce a new Empirical Bayes approach for large-scale hypothesis testing, including estimating false discovery rates (FDRs), and effect sizes. This approach has two key differences from existing approaches to FDR analysis. First, it assumes that the distribution of the actual (unobserved) effects is unimodal, with a mode at 0. This "unimodal assumption" (UA), although natural in many contexts, is not usually incorporated into standard FDR analysis, and we demonstrate how incorporating it brings many benefits. Specifically, the UA facilitates efficient and robust computation—estimating the unimodal distribution involves solving a simple convex optimization problem—and enables more accurate inferences provided that it holds. Second, the method takes as its input two numbers for each test (an effect size estimate and corresponding standard error), rather than the one number usually used (p value or z score). When available, using two numbers instead of one helps account for variation in measurement precision across tests. It also facilitates estimation of effects, and unlike standard FDR methods, our approach provides interval estimates (credible regions) for each effect in addition to measures of significance. To provide a bridge between interval estimates and significance measures, we introduce the term "local false sign rate" to refer to the probability of getting the sign of an effect wrong and argue that it is a superior measure of significance than the local FDR because it is both more generally applicable and can be more robustly estimated. Our methods are implemented in an R package `ashr` available from http://github.com/stephens999/ashr.

Let $\beta = (\beta_1, \ldots, \beta_p)$ denote p "effects" of interest. We test the null hypotheses:

$$H_j : \beta_j = 0 \tag{2.1}$$

- Stephens notes that using a **unimodal prior distribution** can greatly improve FDR and effect size estimation
- Then, he introduces **a mixture of zero-mean normal distributions** as the chosen prior.

Adaptive Shrinkage (ASH) prior

Definition

Stephens considers a family of mixtures of zero-mean normals as prior:

$$\mathcal{G}(\sigma_1^2, \ldots, \sigma_K^2) \triangleq \{g = \sum_{k=1}^{K} \pi_k \mathrm{N}(0, \sigma_k^2) : \pi \in \mathbb{S}^K\} \tag{2.2}$$

- $0 = \sigma_1^2 < \cdots < \sigma_K^2 < \infty$ is a pre-specified grid of component variances.
- π is the unknown mixture proportion, and $\mathbb{S}^K = \{\pi \in \mathbb{R}_+^K : \sum_{i=1}^K \pi_i = 1\}$.

Properties

- Stephens refers to the family of priors (2.2) as **Adaptive SHrinkage** priors due to its flexibility.
- By making the grid of variances sufficiently wide and dense, each component in the mixture contributes a different level of shrinkage. $\mathcal{G}$ includes most distributions used as priors in popular Bayesian regression models.

Normal Means Model with ASH priors (single observation)

Definition

Consider the distribution of a single observation y:

$$
\begin{aligned}
y \mid b, \sigma^2 &\sim \mathrm{N}(b, \sigma^2) \\
b &\sim g \\
g &= \sum_{k=1}^{K} \pi_k \mathrm{N}(0, \sigma^2 \sigma_k^2)
\end{aligned}
\tag{2.3}
$$

- y is a scalar sample; π_k, σ^2 are model parameters to be optimized; σ_k^2 is a pre-specified grid of component variances.
- Here, we use component variances $\sigma^2 \sigma_k^2$, which is equivalent to using ASH prior on the scaled coefficients.

E-Step for Normal Means Model with ASH priors

Posterior Distribution

Due to **Bayes' theorem** and the **conjugacy** of the normal (mixture) prior with the normal likelihood, the posterior distribution can be written as:

$$p(b \mid y, g, \sigma^2) = \sum_{k=1}^{K} \phi_k \mathrm{N}(b; \mu_k, s_k^2) \tag{2.4}$$

where:

$$\text{posterior component means: } \mu_k(y; g, \sigma^2) = \frac{\sigma^2 \sigma_k^2}{\sigma^2 + \sigma^2 \sigma_k^2} \times y \tag{2.5}$$

$$\text{posterior component variances: } s_k^2(y; g, \sigma^2) = \frac{\sigma^2 \sigma_k^2}{\sigma^2 + \sigma^2 \sigma_k^2} \times \sigma^2 \tag{2.6}$$

$$\text{posterior component responsibilities: } \phi_k(y; g, \sigma^2) = \frac{\pi_k L_k}{\sum_{k'=1}^{K} \pi_{k'} L_{k'}} \tag{2.7}$$

$$\text{component marginal likelihoods: } L_k(y; g, \sigma^2) = \int \mathrm{N}(y; b, \sigma^2) \mathrm{N}(b; 0, \sigma^2 \sigma_k^2)\, db \tag{2.8}$$

M-Step for Normal Means Model with ASH priors

Evidence Lower Bound

Given some probability density on $b \in \mathbb{R}$ denoted by q , the evidence lower bound has an analytic expression:

$$\begin{aligned} \mathcal{L}(y; q, g, \sigma^2) &= \log p(y \mid g, \sigma^2) - D_{\mathsf{KL}}(q \mid\mid p_{\mathsf{post}}(b)) \\ &= \mathbb{E}_q[\log p(y \mid b, g, \sigma^2)] - D_{\mathsf{KL}}(q \mid\mid p_{\mathsf{prior}}(b)) \end{aligned} \tag{2.9}$$

where:

$$\mathbb{E}_q[\log p(y \mid b, g, \sigma^2)] = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\bar{b})^2}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{k=1}^{K}[\phi_k(\mu_k^2 + s_k^2 - \bar{b}^2)] \tag{2.10}$$

$$D_{\mathsf{KL}}(q \mid\mid p_{\mathsf{prior}}(b)) = \sum_{k=1}^{K}\phi_k \log\left(\frac{\phi_k}{\pi_k}\right) - \frac{1}{2}\sum_{k=1}^{K}\phi_k\left[1 + \log\left(\frac{s_k^2}{\sigma^2\sigma_k^2}\right) - \frac{\mu_k^2 + s_k^2}{\sigma^2\sigma_k^2}\right] \tag{2.11}$$

and where $\bar{b}$ is the posterior mean of b with respect to q.

M-Step for Normal Means Model with ASH priors

Update for g with Empirical Bayes

Maximizing the ELBO subject to $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$, we have:

$$\pi_k \leftarrow \phi_k, \quad k = 1, \ldots, K. \tag{2.12}$$

Update for σ^2 with Empirical Bayes

Maximizing the ELBO with respect to σ^2 requires setting the derivative to zero, which yields:

$$\sigma^2 \leftarrow \frac{1}{2}\left[(y - \bar{b})^2 \ + \ \sum_{k=1}^{K} \phi_k \left(\mu_k^2 + s_k^2 - \bar{b}^2\right) \ + \ \sum_{k=2}^{K} \phi_k \ \frac{\mu_k^2 + s_k^2}{\sigma_k^2}\right] \tag{2.13}$$

Multiple Regression with ASH priors (Mr.ASH)

Definition

We consider the multiple linear regression model:

$$\mathbf{y} \mid \mathbf{X}, \mathbf{b}, \sigma^2 \sim \mathrm{N}(\mathbf{X}\mathbf{b}, \sigma^2 \mathbf{I}_n) \tag{2.14}$$

where we introduce the latent variable $\gamma_{jk} \in \{0, 1\}$ to rewrite the prior as:

$$\begin{aligned} b_j \mid g, \sigma^2 &\overset{i.i.d.}{\sim} \sum_{k=1}^{K} \gamma_{jk} \mathrm{N}(0, \sigma^2 \sigma_k^2) \\ p(\gamma_{jk} = 1 \mid g) &= \pi_k \end{aligned} \tag{2.15}$$

- $\mathbf{y} \in \mathbb{R}^n$ is a vector of responses; $\mathbf{b} \in \mathbb{R}^p$ is a vector of regression coefficients; $\mathbf{X} \in \mathbb{R}^{n \times p}$ is a matrix whose columns contain predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{R}^n$.
- To simplify presentation, we will assume that $\mathbf{y}$, $\mathbf{X}$ are centralized and the columns of $\mathbf{X}$ are rescaled so that $\|\mathbf{x}_j\| = 1$.

Variational Approximation

Empirical Bayes (EB) Approach

It is convenient to rewrite two steps of EB as solving the optimization problem:

$$(\hat{p}_{\text{post}}, \hat{g}, \hat{\sigma}^2) = \underset{q, g \in \mathcal{G}, \sigma^2 \in \mathbb{R}_+}{\text{argmax}} \; \mathcal{L}(\mathbf{y}; q, g, \sigma^2) \tag{2.16}$$

Variational Empirical Bayes (VEB) Approach

In 2.16, the optimization over q is generally intractable. The VEB approach addresses this by restricting the family of distributions to be optimized:

$$\mathcal{Q} = \left\{ q : q(\mathbf{b}, \boldsymbol{\gamma}) = \prod_{j=1}^{p} q_j(b_j, \boldsymbol{\gamma}_j) \right\} \tag{2.17}$$

Specifically, our VEB approach solves:

$$(\hat{q}, \hat{g}, \hat{\sigma}^2) = \underset{q \in \mathcal{Q}, g \in \mathcal{G}, \sigma^2 \in \mathbb{R}_+}{\text{argmax}} \; \mathcal{L}(\mathbf{y}; q, g, \sigma^2) \tag{2.18}$$

Coordinate Ascent Algorithm for E-Step

Coordinate Ascent in EM Algorithm

The Coordinate Ascent Algorithm is the most commonly used technique for solving factorized variational approximations models.

Algorithm Steps:

1. **Update each q_j:** Compute a posterior distribution under the NM model.
2. **Update g:** Perform a single M-step update for the NM model.
3. **Update σ^2:** Perform a single M-step update for the NM model.

Since we have derived the EM algorithm for the Normal Means model with ASH priors, we only need to use residuals $\bar{\mathbf{r}}_j$ in place of $\mathbf{y}$ and sequentially fit the Normal Means model on different coordinates for $j = 1, \ldots, p$ in E-Step.

Empirical Bayes for M-Step

Evidence Lower Bound

The evidence lower bound of Mr.ASH denoted by:

$$\mathcal{L}(\mathbf{y}; q, g, \sigma^2) = \mathbb{E}_q[\log p(\mathbf{y} \mid \mathbf{X}, \mathbf{b}, g, \sigma^2)] - \sum_{j=1}^{p} D_{\mathsf{KL}}(q_j \mid\mid p_{\mathsf{prior}}) \tag{2.19}$$

Update for g with Empirical Bayes (M-Step)

Maximizing the ELBO subject to $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \geq 0$, we have:

$$\pi_k \leftarrow \frac{1}{p} \sum_{j=1}^{p} \phi_{jk}, \quad k = 1, \dots, K \tag{2.20}$$

Update for σ^2 with Empirical Bayes (M-Step)

Maximizing the ELBO with respect to σ^2 requires setting the derivative to zero:

$$\sigma^2 \leftarrow \frac{\|\mathbf{y} - \mathbf{X}\bar{\mathbf{b}}\|^2 + \sum_{j=1}^{p} \sum_{k=2}^{K} \phi_{jk}(1 + 1/\sigma_k^2)(\mu_{jk}^2 + s_{jk}^2) + \sum_{j=1}^{p} \bar{b}_j^2}{n + p(1 - \pi_1)} \tag{2.21}$$

Algorithm: Coordinate Ascent for fitting Mr.ASH

Require: Data $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^{n}$; number of mixture components, K;
prior variances, $\sigma_1^2 < \cdots < \sigma_K^2$, with $\sigma_1^2 = 0$; initial estimates $\bar{\mathbf{b}}$, $\boldsymbol{\pi}$, σ^2.

$\bar{\mathbf{r}} = \mathbf{y} - \mathbf{X}\bar{\mathbf{b}}$ (compute mean residuals)
repeat
 for $j \leftarrow 1$ **to** p **do**
  $\bar{\mathbf{r}}_j = \bar{\mathbf{r}} + \mathbf{x}_j \bar{b}_j$ (disregard j-th effect in residuals)
  $\hat{b}_j \leftarrow \mathbf{x}_j^T \bar{\mathbf{r}}_j$ (compute OLS estimate)
  for $k \leftarrow 1$ **to** K **do**
   $\phi_{jk} \leftarrow \phi_k(\hat{b}_j; g, \sigma^2)$
   $\mu_{jk} \leftarrow \mu_k(\hat{b}_j; g, \sigma^2)$
  end for (Step 1: update q_j, equations 2.5,2.7)
  $\bar{b}_j \leftarrow \sum_{k=1}^{K} \phi_{jk}\mu_{jk}$ (update posterior mean of b_j)
  $\bar{\mathbf{r}} \leftarrow \bar{\mathbf{r}}_j - \mathbf{x}_j \bar{b}_j$ (update mean residuals)
 end for
 for $k \leftarrow 1$ **to** K **do**
  $\pi_k \leftarrow \sum_{j=1}^{p} \phi_{jk}/p$ (Step 2: update g)
 end for
 $\sigma^2 \leftarrow \frac{\|\bar{\mathbf{r}}\|^2 + \bar{\mathbf{b}}^T(\hat{\mathbf{b}} - \bar{\mathbf{b}}) + \sigma^2 p(1-\pi_1)}{n + p(1-\pi_1)}$ (Step 3: update σ^2)
until termination criterion is met
return $\bar{\mathbf{b}}$, $\boldsymbol{\pi}$, σ^2

Extensions

Variational EB as a Penalized Regression Problem

We define the penalized squared-loss function:

$$
\begin{aligned}
h_{g,\sigma^2}(\bar{\mathbf{b}}) &\triangleq \min_{q_1,\dots,q_p} -\mathcal{L}(\mathbf{y}; q, g, \sigma^2) \\
&= \min_{q_1,\dots,q_p} \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\mathbb{E}_q[\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2] + \sum_{j=1}^{p} D_{\mathsf{KL}}(q_j \mid\mid p_{\mathsf{prior}}) \\
&= \min_{q_1,\dots,q_p} \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{p}\mathsf{Var}_{q_j}(b_j) + \sum_{j=1}^{p} D_{\mathsf{KL}} \\
&= \frac{1}{\sigma^2}\cdot\left\{\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \sum_{j=1}^{p}\rho_{g,\sigma^2}(\bar{b}_j)\right\} + \frac{n}{2}\log(2\pi\sigma^2)
\end{aligned}
\tag{3.1}
$$

where the penalty function with parameters g,σ^2 is:

$$
\rho_{g,\sigma^2} \triangleq \min_{q} \frac{1}{2}\mathsf{Var}_q(b) + \sigma^2 D_{\mathsf{KL}}(q \mid\mid p_{\mathsf{prior}}) \tag{3.2}
$$

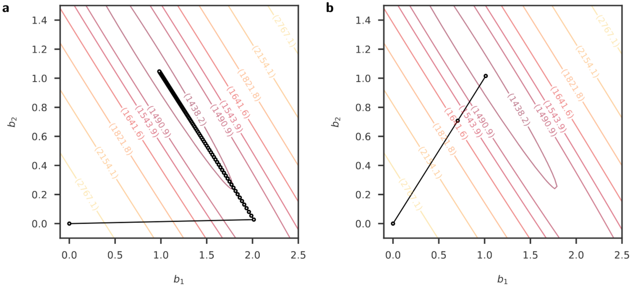
Gradient-Based Optimization Method

GRADIENT-BASED OPTIMIZATION FOR VARIATIONAL EMPIRICAL BAYES MULTIPLE REGRESSION

Saikat Banerjee
Department of Human Genetics
University of Chicago
saikat.banerjee@uchicago.edu

Peter Carbonetto
Department of Human Genetics, and
Research Computing Center
University of Chicago
peter.carbonetto@uchicago.edu

Matthew Stephens
Department of Human Genetics, and
Department of Statistics
University of Chicago
mstephens@uchicago.edu



Results

Summary of methods compared in the simulations

method	R package	brief description
PLR methods		
ridge regression	glmnet	convex L_2 penalty
Lasso	glmnet	convex L_1 penalty
Elastic Net	glmnet	linear combination of L_1 and L_2 penalties
SCAD	ncvreg	nonconvex SCAD penalty
MCP	ncvreg	minimax concave penalty
L0Learn	L0Learn	nonconvex L_0, L_0L_1 or L_0L_2 penalty
SSLasso	SSLASSO	adaptive penalty based on Laplace mixture prior
Trimmed Lasso	(MATLAB)	nonseparable "trimmed lasso" penalty
Bayesian and empirical Bayes methods		
BayesB	BGLR	MCMC with spike-and-slab prior ("slab" is t)
Bayesian Lasso	BGLR	MCMC with scaled Laplace prior
varbvs	varbvs	variational inference with spike-and-slab prior
SuSiE	susieR	variational inference for "SuSiE" model

Design of Simulations

- **Experiment 1: Varying sparsity & correlation structure**
  Change the number of non-zero coefficients s, and consider three types of design matrices: (1) *independent variables*; (2) *correlated variables*; (3) *real genotype data* (*Wang et al. 2020*).
- **Experiment 2: Varying proportion of variance explained (PVE)**
  Assess how the PVE affects prediction accuracy and model selection.
- **Experiment 3: Varying signal distributions**
  Normal, Uniform, Laplace, t, and point mass distribution.
- **Experiment 4: Varying the number of predictors p**
  Assess how computational effort change with the number of predictors p.
- **Experiment 5: Varying noise distributions**
  Normal, Uniform, Laplace and t distribution.

Evaluation

RMSE

Each method returns $\hat{\mathbf{b}}$, an estimate of the regression coefficients. We evaluated this estimate using the scaled root-mean-squared error (RMSE) in the test data:

$$\text{scaled RMSE}(\mathbf{y}_{\text{test}}, \hat{\mathbf{b}}) \triangleq \frac{\|\mathbf{y}_{\text{test}} - \mathbf{X}_{\text{test}}\hat{\mathbf{b}}\|/\sqrt{n}}{\sqrt{1-\text{PVE}}/\sigma} \tag{4.1}$$

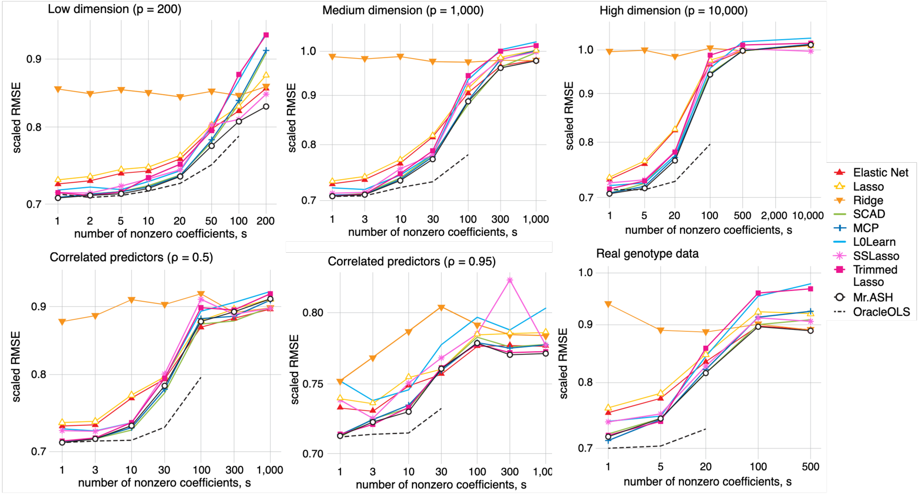
RRMSE

To produce a summary, for each simulation t we calculated the relative prediction error as **the Ratio of the RMSE to the best RMSE** achieved in that simulation among all the methods compared:

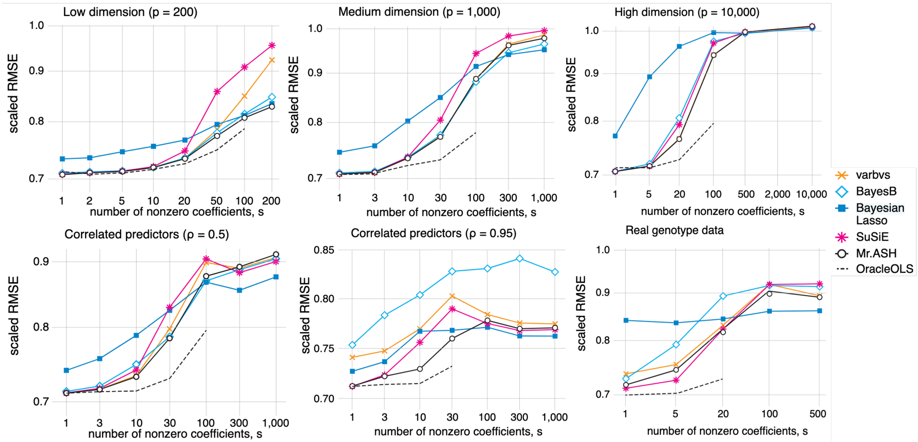
$$\text{RRMSE}_{mt} \triangleq \frac{\text{RMSE}_{mt}}{\min_{m'} \text{RMSE}_{m't}} \tag{4.2}$$

where RMSE_{mt} is the root mean squared error generated by model m for the test set in simulation t.

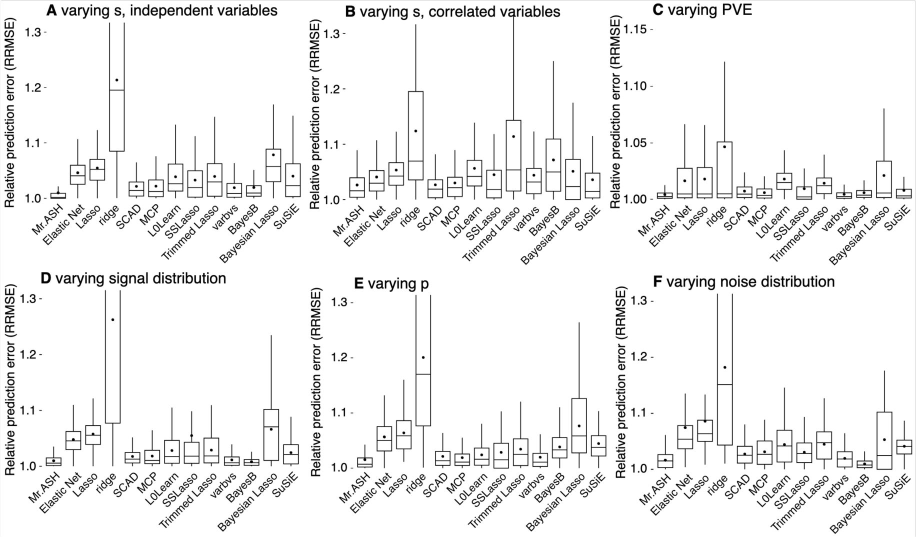
Experiment 1: Varying the Sparsity Level (PLR methods)



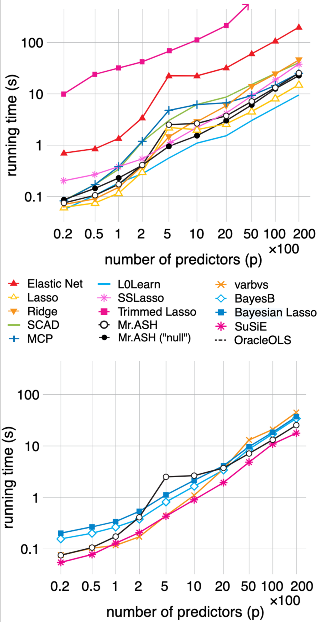
Experiment 1: Varying the Sparsity Level (Bayes methods)



Summary of Results from Experiments 1–5



Contrast of Running Times



method	running time (s)
L0Learn	8.12
Mr.ASH (null)	14.96
SSLasso	16.19
Lasso	17.98
Bayesian Lasso	18.24
Mr.ASH	18.56
MCP	21.19
SuSiE	22.54
BayesB	23.77
ridge regression	32.88
SCAD	33.59
varbvs	52.26
Elastic Net	223.66
Trimmed Lasso	609.54

Thank you!

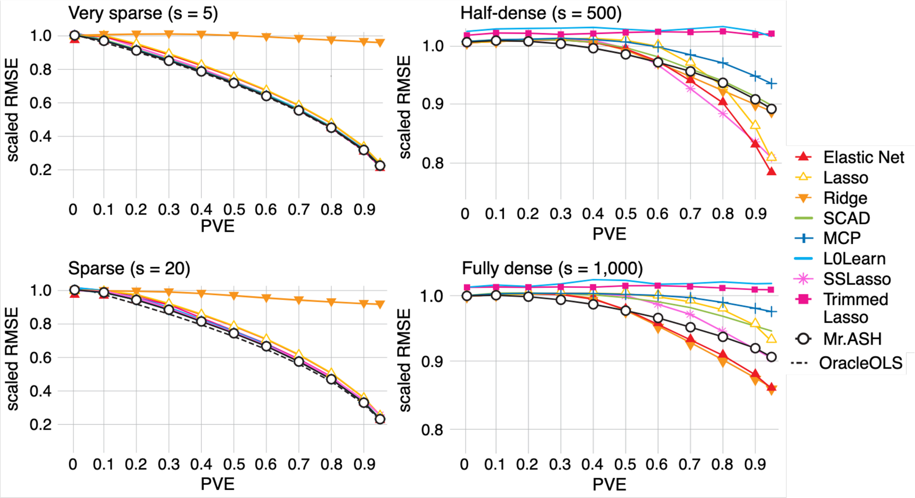
Appendix

Default Parameter Setting of ASH Prior

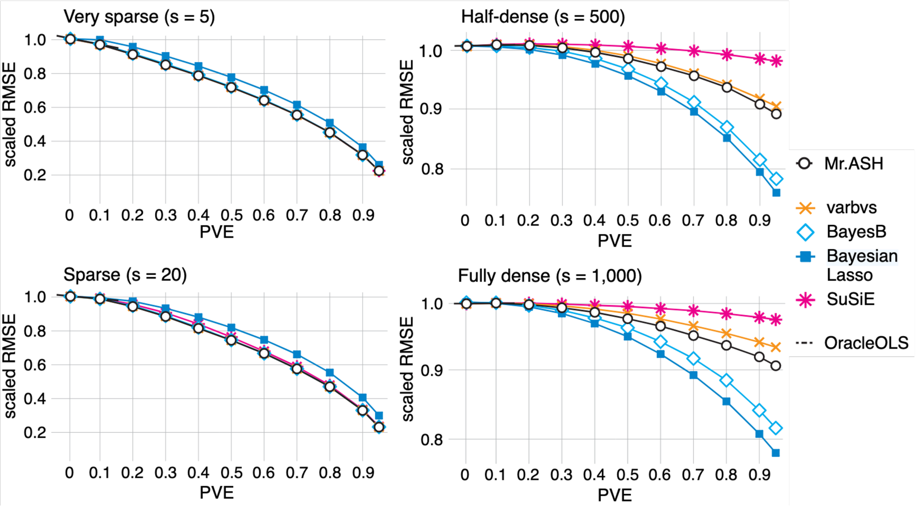
```
# precompute x_j^T x_j
w                    = colSums(data$X^2)
data$w               = w

# set sa2 if missing
if ( is.null(sa2) ) {
  sa2                = (2^((0:19) / 20) - 1)^2
  sa2                = sa2 / median(data$w) * n
}
K                    = length(sa2)
data$sa2             = sa2
```

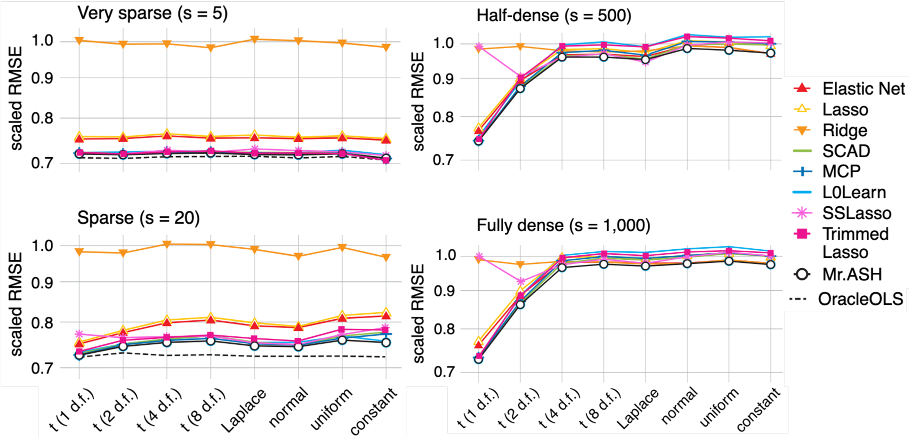
Experiment 2: Varying PVE (PLR methods)



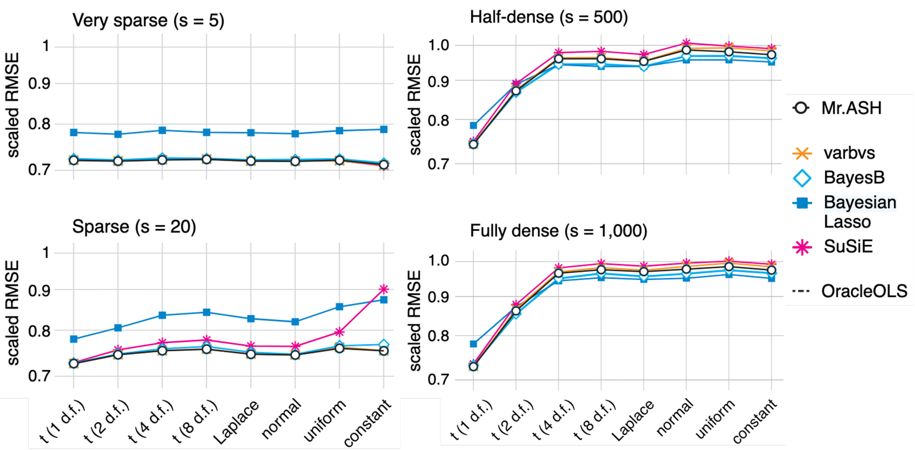
Experiment 2: Varying PVE (Bayes methods)



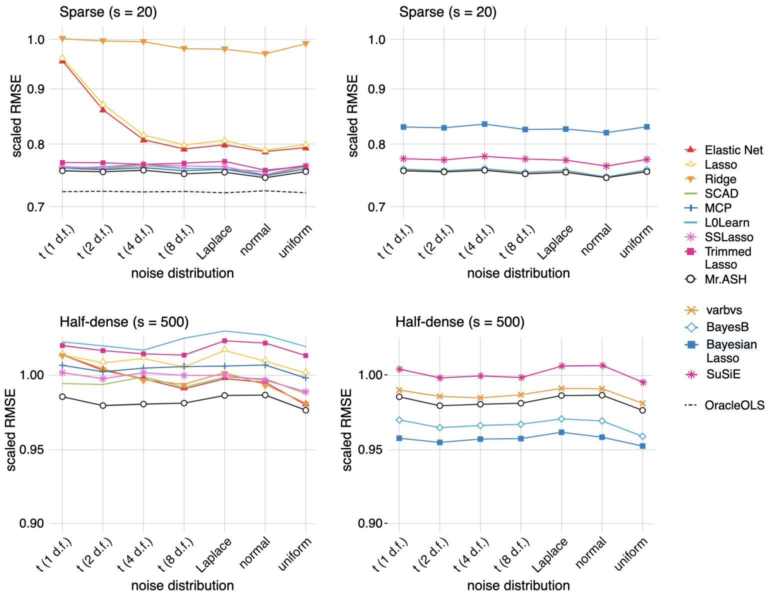
Experiment 3: Varying Signal Distributions (PLR methods)



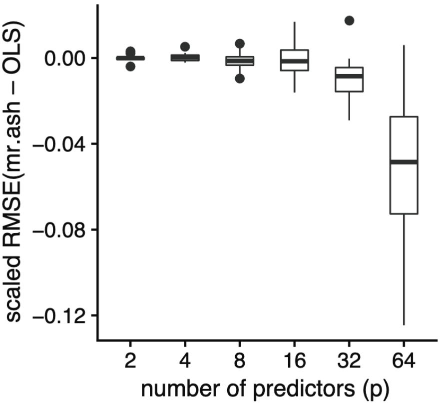
Experiment 3: Varying Signal Distributions (Bayes methods)



Experiment 5: Varying Noise Distributions



Additional Experiment: Simulations with $p < n$

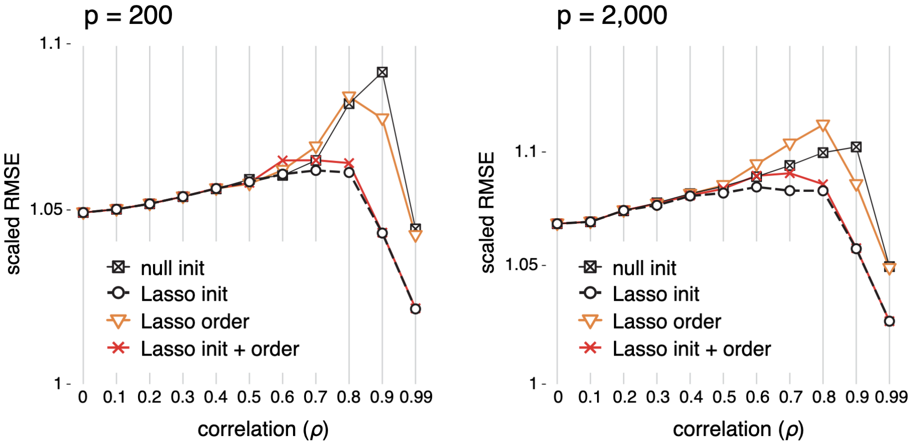


Setting:

- PVE $= 0.5$
- $n = 200$
- $p \in \{2, 4, 8, 16, 32, 64\}$
- $s = p$ (all coefficients are non-zero)
- simulations times: 20

Mr.ASH performs similarly to OLS when p is very small and outperforms the OLS estimate as p increases.

Comparison of Different Initializations



Expectation Maximization (EM) Algorithm

Goal Function

Consider a probabilistic model in which we collectively denote all of the observed variables by $\mathbf{X}$ and all of the hidden variables by $\mathbf{Z}$. Our goal is to maximize the **marginal likelihood** that is given by:

$$p(\mathbf{X} \mid \boldsymbol{\theta}) = \int_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \, d\mathbf{Z} \tag{6.1}$$

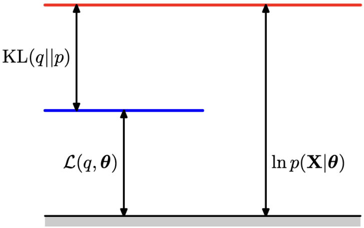
Take the following decomposition:

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathsf{KL}(q||p) \tag{6.2}$$

where we have defined:

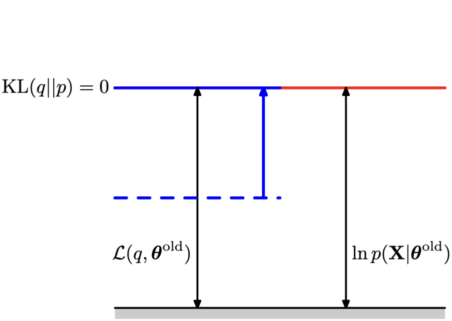
$$\mathcal{L}(q, \boldsymbol{\theta}) = \int_{\mathbf{Z}} q(\mathbf{Z}) \ln\{\frac{p(\mathbf{X}, \mathbf{Z} \mid \theta)}{q(\mathbf{Z})}\} \, d\mathbf{Z} \tag{6.3}$$

$$\mathsf{KL}(q||p) = -\int_{\mathbf{Z}} q(\mathbf{Z}) \ln\{\frac{p(\mathbf{Z} \mid \mathbf{X}, \theta)}{q(\mathbf{Z})}\} \, d\mathbf{Z} \tag{6.4}$$



Expectation Maximization (EM) Algorithm

- E-Step:



- M-Step: