Statistica Sinica Preprint No: SS-2025-0221	
Title	Nonparametric Shrinkage Estimation in High Dimensional Glms via Polya Trees
Manuscript ID	SS-2025-0221
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0221
Complete List of Authors	Asaf Weinstein, Jonas Wallin, Daniel Yekutieli and Malgorzata Bogdan
Corresponding Authors	Asaf Weinstein
E-mails	asaf.weinstein@mail.huji.ac.il

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

NONPARAMETRIC SHRINKAGE ESTIMATION IN GENERALIZED LINEAR MODELS VIA POLYA TREES

Asaf Weinstein[1], Jonas Wallin[2], Daniel Yekutieli[3], and Malgorzata Bogdan[4]

[1]*Department of Statistics, Hebrew University of Jerusalem*

ORCID: 0000-0003-2237-2510

[2]*Department of Statistics, Lund University*

ORCID: 0000-0003-0381-6593

[3]*Department of Statistics, Tel Aviv University*

[4]*Institute of Mathematics, University of Wroclaw*

Abstract: Regularization in fitting regression models has been a very active topic of research in the past few decades, but most of the existing methods are designed for particular situations, e.g. for the case of a sparse coefficient vector. We consider the problem of designing *universally* optimal regularized estimators in a given generalized linear model with fixed effects. First, we propose as a contender the Bayes estimator against an *ideal* prior that assigns equal mass to every permutation of the fixed coefficient vector, thus depending on the true coefficients only through their empirical CDF. We prove some optimality properties of this oracle estimator in both the frequentist and Bayesian frameworks. To compete with the oracle estimator, we posit a hierarchical Bayes model where the individual coefficients are modeled as i.i.d. draws from a common distribution π, which is in turn assigned a Polya tree prior that reflects indefiniteness. We demonstrate in examples that the posterior mean of π under the postulated model adapts nonparametrically to the empirical CDF of the true coefficients. Correspondingly, the posterior means of the coefficients themselves are used to mimic the ideal estimator. Numerical experiments show that our method has better estimation and prediction

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

accuracy compared to various parametric and nonparametric alternatives, from relatively standard L_p-regularized estimators to modern penalized-likelihood and Bayesian estimators for high dimensional regression.

Key words and phrases: Hierarchical modeling; empirical Bayes methods; nonparametric inference.

1. Introduction

Supervised learning problems nowadays often entail fitting complex models with thousands or even hundreds of thousands of parameters. Correspondingly, incorporating regularization in the training process is paramount to controlling overfitting and enabling generalization to new examples. Still, *what* exact form of regularization is most adequate for a given problem is generally far from obvious. In this paper we consider the problem of designing an *optimal* regularizer in a setting where the observations $(\boldsymbol{X}_i, Y_i)$, $i = 1, ..., n$, follow a generalized linear model (GLM),

$$Y_i \overset{ind}{\sim} f(y_i; \eta_i, \psi), \quad \mathbb{E}_{\eta_i}[Y_i] = g^{-1}(\eta_i), \quad \eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}, \tag{1.1}$$

where $\boldsymbol{X}_i \in \mathbb{R}^p$ are fixed and known covariate vectors, and f is a given likelihood function corresponding to an exponential family of distributions. The main parameter vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$, and the (optional) nuisance parameter ψ are unknown, and the task is, in general, to estimate $\boldsymbol{\beta}$ under a specified loss function. We will focus here on squared loss in estimating the coefficient vector $\boldsymbol{\beta}$ or the linear predictor $\boldsymbol{\eta} = (\eta_1, ..., \eta_n)^\top = \boldsymbol{X}\boldsymbol{\beta}$, for $\boldsymbol{X} = [\boldsymbol{X}_1, ..., \boldsymbol{X}_n]^\top \in \mathbb{R}^{n \times p}$, but the methods to be described in the sequel apply in principle to any loss function.

A standard approach to obtain a regularized estimator for $\boldsymbol{\beta}$ in (1.1), is to maximize a

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

penalized version of the likelihood,

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_i \log f(Y_i; \eta_i, \psi) + \mathcal{P}_\lambda(\boldsymbol{\beta}), \tag{1.2}$$

where the penalty function $\mathcal{P}_\lambda(\boldsymbol{\beta})$, indexed by the tuning constant λ, is specified in advance (the nuisance parameter ψ is generally treated here as known, and can be estimated subsequently as in a profile-likelihood approach). Penalized-likelihood estimators (1.2) shrink by balancing the log-likelihood against 'suitably disciplined' values of $\boldsymbol{\beta}$, to use the terminology of Ročková and George (2018). There is plenty of modern work proposing and analyzing penalized-likelihood estimators for GLMs, far beyond the relatively standard options of ℓ_p penalties. These include, for example, the convex methods of Bondell and Reich (2008, OSCAR) and Bogdan et al. (2015, SLOPE), which promote both variable selection and parameter tying, or the nonconvex methods of Fan and Li (2001, SCAD) and Zhang (2010, MCP), that simultaneously perform selection and shrinkage estimation with carefully designed nonconcave penalties for attenuating bias.

From a Bayesian perspective, the penalized-likelihood estimator (1.2) can be viewed as a maximum *a posteriori* (MAP) estimator under the (possibly improper) prior $\pi(\boldsymbol{\beta}|\lambda) = \exp(\mathcal{P}_\lambda(\boldsymbol{\beta}))$. The Bayesian viewpoint is often more convenient because *a priori* knowledge about $\boldsymbol{\beta}$ can be incorporated more directly into the model. This connection to Bayes estimators has been exploited in many existing papers that propose different penalty functions by a careful choice of a prior. The majority of modern work concentrates on recovering a sparse coefficient vector, typically by employing various parametric *spike-and-slab* models. George and Foster (2000) consider parametric classes of priors suitable for model selection and sparsity, and propose to estimate the hyperparameters from the data, resulting in empirical Bayes (EB) estimates. The alternative EB methods of Yuan and Lin (2005) offer

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

better computational efficiency. Andersen et al. (2017) generalize the spike-and-slab prior to situations where the coefficient vector has a spatio-temporal structure. The Spike-and-slab Lasso method of Ročková and George (2018) employs a (parametric) modification of the L_1 penalty, that similarly promotes sparsity but is able to reduce the bias of the Lasso. Jiang et al. (2019) extend the methods from Ročková and George (2018) , proposing a Bayesian counterpart of SLOPE. The estimator of Carvalho et al. (2010) uses the horseshoe prior, offering robustness to unknown sparsity level and the handling of large signal components. Many other Bayesian estimators have been proposed over the years, the vast majority of which again ultimately use *parametric* families of priors for regularization.

In the current work we adopt the Bayesian viewpoint discussed above, postulating that β_j are i.i.d. draws from an unknown prior Π, but we make no assumptions at all on this prior, in particular there is no assumption that it is 'sparse'. Instead, we propose a flexible hierarchical modeling scheme to estimate this prior nonparametrically, specifically, we model Π itself as a random realization from some distribution P in order to adapt to the "true" prior (we discuss in Section 3 what this means when β_j are fixed). In this hierarchical Bayes approach, allowing the prior Π to be learned nonparametrically from the observed data calls for endowing it with a 'noninformative' prior. There is more than one option to choose such a prior P on Π, for example specifying P to be the distribution of a Dirichlet process (Ferguson, 1973) is a common choice in existing nonparametric Bayes literature. Here we propose to use for P a *Polya tree* distribution. Polya trees belong to a class of tail-free distributions introduced by Lavine (Lavine, 1992, 1994) as a generalization of Ferguson's Dirichlet process; see also Ghosh and Ramamoorthi (2003) for a textbook account. In particular, the Polya tree allows the random probability measure to be supported on continuous distributions, while

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

maintaining tractability. As detailed in Section 2, Polya trees admit conjugacy (closure) properties, similarly to the Dirichlet process, which make it a convenient choice in practice.

In essence, our approach is a fully Bayes alternative to the nonparametric empirical Bayes approach, in which Π would be regarded as a *fixed* and unknown member of a rich family of distributions specified in advance. The hierarchical Bayes approach in random-effects models has been mentioned by Robbins (1963) in a nonparametric context, but it is more familiar in parametric contexts (Lindley and Smith, 1972; Efron and Morris, 1973). Both the EB and the fully Bayes approaches are ultimately intended to allow Π to be learned ("deconvolved") using the entire set of observations $(\boldsymbol{X}_i, Y_i)$. In Kim et al. (2022) the EB approach is pursued, modeling Π as a mixture of zero-mean normal distributions with different (prespecified) scale parameters. In estimating these hyperparameters, Kim et al. (2022) use a variational approach to handle intractability of the posterior of $\boldsymbol{\beta}$. More precisely, the posterior of $\boldsymbol{\beta}$ is approximated by a product distribution, optimized within a prespecified family to best fit the true posterior in the sense of minimizing the Kullback-Leibler divergence. Avoiding such mean-field approximations of the posterior distribution is what we view as one of the advantages of the MCMC approach proposed in the current article.

Polya tree priors and the related Dirichlet process have some history in application to statistical problems, however existing work mostly focuses on applications to density estimation (Castillo, 2017, and references therein), which under (1.1) would correspond to estimating Π when directly observing $\beta_1, ..., \beta_p$; or to empirical Bayes problems (Antoniak, 1974; Berry and Christensen, 1979) under the usual *sequence* model, to which (1.1) reduces in the very special case where $n = p$ and $\boldsymbol{X} = \boldsymbol{I}_p$ is the identity matrix. The sequence

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

model is simpler than the regression model (1.1), because there are no covariates, and the likelihood of each Y_i depends on its own parameter η_i, with no known relationship between $\eta_1, ..., \eta_p$; for example, $Y_i \sim \mathcal{N}(\eta_i, 1),\ i \leq p$, with $\eta_i \in \mathbb{R}$. By contrast, in (1.1) only the Y_i, not the β_j, are observed, and the likelihood of each Y_i depends on the common vector $\boldsymbol{\beta}$ since $\eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ are all determined by $\boldsymbol{\beta}$. The novelty of the proposed methodology is to extend the use of Polya trees beyond the sequence model to the more general regression model (1.1).

We have focused above on an overview of the proposed methodology, but an important aspect of this work—and something we find to be largely missing in related literature—is providing theoretical justification for the nonparametric Bayes approach even when working in a strictly frequentist setting. Thus, after describing our hierarchical Bayes model more precisely, in Section 3 we motivate our methods as pursuing an *oracle* estimator, defined to be the optimal estimator which knows the true vector $\boldsymbol{\beta}$ but is restricted to using this knowledge *symmetrically*, i.e., without giving *a priori* preference to any of the orderings of $\boldsymbol{\beta}$. This basically means that only information about the unordered vector $\boldsymbol{\beta}$—equivalently, the empirical distribution of the true coefficients $\beta_1, ..., \beta_p$—is available to the oracle. Importantly, the trivial solution $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$ violates this condition, and therefore is eliminated. We provide supporting theoretical analysis showing that our oracle estimator has certain optimality properties in both a frequentist and a Bayesian framework, and explain how this analysis is relevant to the Polya-tree based methodology proposed in this paper.

The rest of the article is organized as follows. Section 2 presents our methodology by describing a hierarchical model that includes a Polya tree prior on univariate distributions. After presenting an analysis of the oracle estimator in Section 3, we provide in Section 4

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

results from a simulation in a logistic regression model. In Section 5 we apply our method to analyze real data of polygenic inheritance. We conclude in Section 6 with some remarks and directions for further research.

2. Methodology

Our approach starts by positing a hierarchical model on the observed data. Thus, we suppose

$$\beta_1, ..., \beta_p \overset{iid}{\sim} \Pi, \tag{2.1}$$

where the distribution Π itself is modeled as random,

$$\Pi \sim P \tag{2.2}$$

for a specified 'noninformative' distribution P. We emphasize that, unless otherwise indicated, (1.1) is the only modeling assumption for the observed data, so (2.1), (2.2), or any of the other suppositions that follow, are merely used to facilitate a shrinkage estimator. To complete the Bayesian model, we further assume

$$\psi \sim h, \tag{2.3}$$

independently of $\boldsymbol{\beta}$, where h is a fully specified vague prior on the nuisance parameter ψ (as usual, h is allowed to depend on the likelihood f in (1.1)). If we fix a loss function and regard h as given, the choice of P determines the Bayes rule for $\boldsymbol{\beta}$, which will be used as a regularized estimator; e.g., for squared loss, this is the posterior mean of $\boldsymbol{\beta}$ under P and h.

Parametric choices of P, e.g. Gaussian, yield Bayes rules that generally resemble parametric EB estimators, some examples of which were mentioned in the introduction. Instead, to allow our hierarchical model to learn a completely unknown Π, we take P to be a *Polya*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

tree distribution. Polya trees, introduced by Ferguson, belong to a class of tail-free distributions on random probability measures that generalize Dirichlet Processes while maintaining tractability (Ferguson, 1973, 1974). Under the hierarchical Bayes model given by (1.1), (2.1), (2.2) and (2.3), we propose to use posterior sampling to provide inference for $\boldsymbol{\beta}$. This is carried out with a Gibbs sampling algorithm, which we construct to take advantage of conjugacy properties of Polya trees when conditioning on certain parts of the unobserved variables.

We proceed with a more precise description of the hierarchical model above, which we will also refer to as the *generative* model, to distinguish it from the frequentist model (1.1), and then explain how we use it to provide (approximate) inference for $\boldsymbol{\beta}$.

The finite Polya tree model. The L-level finite Polya tree (FPT) model generates distributions Π with piecewise constant density functions on a dyadic partition of $\mathcal{I}_0 = (a_{min}, a_{\max}]$, corresponding to a fixed endpoints vector $\boldsymbol{a} := (a_{min} = a_0 \leq a_1 \leq \cdots \leq a_{2^L-1} \leq a_{2^L} = a_{max})$. The dyadic partition consists of subintervals $\mathcal{I}_{l,i} = (a_{(i-1)\cdot 2^{L-l}}, a_{i\cdot 2^{L-l}}]$, for $l = 1 \cdots L$ and $i = 1 \cdots 2^l$. The parameters of the FPT model are the Beta parameters $(\alpha_{l,i}, \beta_{l,i})$, corresponding to subintervals $\mathcal{I}_{l-1,i}$ for $l = 1 \cdots L$ and $i = 1 \cdots 2^{l-1}$. The FPT model has the following components.

I. *Independent Beta random variables.* A vector $\boldsymbol{\phi} = (\phi_{1,1} \cdots \phi_{L,2^{L-1}})$ of independent Beta random variables $\phi_{l,i} \sim Beta(\alpha_{l,i}, \beta_{l,i})$, specifying conditional subinterval probabilities for the dyadic partition. Specifically, $\mathbb{P}(\mathcal{I}_{1,1}|\mathcal{I}_0) = \phi_{1,1}$ and $\mathbb{P}(\mathcal{I}_{1,2}|\mathcal{I}_0) = 1 - \phi_{1,1}$, and, for $l = 2 \cdots L$ and $i = 1 \cdots 2^{l-1}$, $\mathbb{P}(\mathcal{I}_{l,2\cdot i-1}|\mathcal{I}_{l-1,i}) = \phi_{l,i}$, $\mathbb{P}(\mathcal{I}_{l,2\cdot i}|\mathcal{I}_{l-1,i}) = 1 - \phi_{l,i}$.

II. *Subinterval probabilities.* The subinterval probabilities vector $\boldsymbol{\pi}$ has elements $\mathbb{P}(\mathcal{I}_{l,i}) = \pi_{l,i}$, which are products of the conditional subinterval probabilities: $\pi_{1,1} = \phi_{1,1}$, $\pi_{1,2} =$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

$1-\phi_{1,1}$, and, for $l=2,...,L$ and $i=1,...,2^{l-1}$, $\pi_{l,2\cdot i-1}=\phi_{l,i}\cdot\pi_{l-1,i}$ and $\pi_{l,2\cdot i}=(1-\phi_{l,i})\cdot\pi_{l-1,i}$.

III. *Step function PDF.* The step function density Π is specified conditionally on the vector $\boldsymbol{\pi}_L=(\pi_{L,1},...,\pi_{L,2^L})$ of subinterval probabilities at level L,

$$\Pi(\beta|\boldsymbol{\pi}_L)=\pi_{L,1}\cdot\mathbb{1}_{L,1}(\beta)/(a_1-a_0)+\cdots+\pi_{L,2^L}\cdot\mathbb{1}_{L,2^L}(\beta)/(a_{2^L}-a_{2^L-1}), \tag{2.4}$$

for $\beta\in\mathcal{I}_0$, where $\mathbb{1}_{L,i}(\beta)$ is the indicator function corresponding to $\mathcal{I}_{L,i}$.

Taking P to be a level-L FPT, and instantiating $\alpha_{l,i}=\beta_{l,i}\equiv 1$ to reflect noninformativeness, this results in the following generative model:

i) Draw $\psi\sim h(\psi)$

ii) Draw Π from the FPT model with $\phi_{l,i}\sim Beta(1,1)$ for $l=2,...,L$ and $i=1,...,2^{l-1}$

iii) Draw $\beta_j\sim\Pi$, i.i.d., for $j=1,...,p$

iv) Draw $\boldsymbol{Y}$ from (1.1)

The posterior distribution of $\boldsymbol{\Theta}=\{\psi,\boldsymbol{\beta},\boldsymbol{\phi}\}$ is sampled by Gibbs sampling as described in Algorithm 1. The single-site Metropolis-Hastings (MH) algorithm detailed in section S2 of the supplement, is used to sample each component β_j of $\boldsymbol{\beta}$ conditionally on the remaining components of $\boldsymbol{\beta}$, and on ψ, $\boldsymbol{\phi}$ and $\boldsymbol{y}$. To sample ψ conditionally on $(\boldsymbol{\beta},\boldsymbol{y})$, one can use a MH step, unless the prior is (conditionally) conjugate, in which case direct sampling is easy. For instance, for a Gaussian linear model, $\psi=\sigma^2$ and if an inverse-Gamma prior is used then the conditional posterior is also an inverse-Gamma. Finally, for sampling $\boldsymbol{\phi}|\boldsymbol{\beta},\boldsymbol{y}:=\boldsymbol{\phi}|\,\boldsymbol{\beta}$, let

$$N_{l,i}(\boldsymbol{\beta})=\#\{j:\beta_j\in\mathcal{I}_{l,i}\},$$

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

and let $\boldsymbol{N} = (N_{1,1}, ..., N_{L,2^L})$. Ferguson (1974) has already noted the conjugacy of the FPT model, namely, that $\boldsymbol{\phi}|\,\boldsymbol{\beta}$ is a FPT with updated hyper-parameter values,

$$\phi_{l,i}|\boldsymbol{N} \sim Beta(1 + N_{l,2\cdot i-1}, 1 + N_{l,2\cdot i}). \tag{2.5}$$

Algorithm 1 Gibbs Sampler for GLM

Set: number of iterations G, number of levels L, endpoints vector $\boldsymbol{a}$, initial values $\boldsymbol{\Theta}^{(0)}$
Input: $\boldsymbol{y}$, $\boldsymbol{X}$
Output: posterior samples $\boldsymbol{\Theta}^{(1)}, \ldots, \boldsymbol{\Theta}^{(G)}$
for $g = 1$ to G **do**
  for $j = 1$ to p **do**
    Sample $\beta_j^{(g)} \mid \psi^{(g-1)}, \boldsymbol{\phi}^{(g-1)}, \boldsymbol{y}, \beta_1^{(g)}, \ldots, \beta_{j-1}^{(g)}, \beta_{j+1}^{(g-1)}, \ldots, \beta_p^{(g-1)}$ by MH
  end for
  Sample $\psi^{(g)} \mid \boldsymbol{\beta}^{(g)}, \boldsymbol{y}$
  Sample $\boldsymbol{\phi}^{(g)} \mid \boldsymbol{\beta}^{(g)}$ using Eq. (2.5)
end for

3. An oracle shrinkage rule

In this section we provide some theoretical support for the proposed approach. We will need to distinguish, only in this section, between the true, fixed value of the coefficient vector in (1.1), which we denote here $\boldsymbol{\beta}^*$, and a general (potential) value, which we denote $\boldsymbol{\beta}$. While our postulated Bayes model produces a legal estimator, because it is a function of the data only, it is not obvious (i) what object we are actually estimating in terms of the original model (1.1) when $\boldsymbol{\beta}$ is actually fixed, or (ii) why such an object may be a good target to

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

pursue. Addressing (i) first, we now give intuitive reasoning to motivate the Bayes rule under the proposed hierarchical model, as a method for mimicking an *oracle* rule $\widehat{\boldsymbol{\beta}}_{ol}$, which we define to be the Bayes rule in the model specified by the likelihood (1.1) and the prior

$$\boldsymbol{\beta} \sim \widetilde{\Pi}_p^*, \tag{3.1}$$

where we use $\widetilde{\Pi}_p^*$ to denote the (discrete) uniform prior on all $p!$ permutations of $\boldsymbol{\beta}^*$. To see why we expect the Bayes rule under the model given by (1.1), (2.1), (2.2) and (2.3) to be able to adapt to $\widehat{\boldsymbol{\beta}}_{ol}$ when p is large, let us first consider the sequence model mentioned in the introduction. In the sequence model, it is well known from the literature on compound decision (Robbins, 1951, 1956; Zhang, 2003) that the EB approach, which estimates the Bayes rule under (2.1) treating Π as fixed and unknown, generally produces decision rules that asymptotically attain the risk of the oracle Bayes rule against the i.i.d. prior

$$\beta_1, ..., \beta_p \overset{i.i.d.}{\sim} \Pi_p^*, \tag{3.2}$$

where $\Pi_p^* = \frac{1}{p}\sum_{j=1}^p \delta_{\beta_j^*}$ is the *empirical distribution* of the true coefficients $\beta_1^*, ..., \beta_p^*$. In turn, the proposed hierarchical Bayes approach, employing a 'noninformative' prior on Π, is known to produce qualitatively similar estimates to the EB approach (Efron and Morris, 1973; Stein, 1981; Carlin and Louis, 1996). Furthermore, when p is large the Bayes rules under the priors (3.1) and (3.2) are intuitively expected to be similar, in a sense, because the former samples from $\beta_1^*, ..., \beta_p^*$ without replacement, and the latter with replacement; for some formal results we refer the reader to Diaconis and Freedman (1980) and Hannan and Robbins (1955); Greenshtein and Ritov (2009). Taken together, the arguments above give strong justification for regarding the Bayes rule under the proposed hierarchical model, as a competitor (estimator) of the oracle Bayes rule under (3.1). Carrying these ideas

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

over to the more general regression model (1.1), we thus expect that under the proposed hierarchical prior for $\boldsymbol{\beta}$, the corresponding Bayes rule would 'learn' the oracle Bayes rule $\widehat{\boldsymbol{\beta}}_{ol}$. More specifically, we expect the posterior distribution of Π under the Polya tree prior, to concentrate around Π_p^*; this is supported by our simulations in Section 4.

Turning to point (ii) raised at the beginning of this section, we now study properties of $\widehat{\boldsymbol{\beta}}_{ol}$ itself, which, by the argument just presented, is viewed as the population-level target of the estimator produced by the proposed hierarchical Bayes method; note that $\widehat{\boldsymbol{\beta}}_{ol}$ is perfectly well defined in terms of the original frequentist model (1.1). We present two separate results: the first is limited to the Normal linear model, and says that, if $\boldsymbol{X}$ satisfies some conditions, $\widehat{\boldsymbol{\beta}}_{ol}$ minimizes the (point) *risk* $R(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}) := \mathbb{E}_{\boldsymbol{\beta}^*} L(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}})$, among all members in a very natural class of estimators, by appealing to permutation invariance considerations. The second result holds for any GLM and any matrix $\boldsymbol{X}$, but its optimality guarantees are weaker. More specifically, it says that, under *any exchangeable* prior $\widetilde{\Pi}$ on $\boldsymbol{\beta}$, the estimator $\widehat{\boldsymbol{\beta}}_{ol}$ is optimal in terms of the *Bayes risk*,

$$r(\widetilde{\Pi}, \widehat{\boldsymbol{\beta}}) := \int R(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}) \widetilde{\Pi}(d\boldsymbol{\beta}). \tag{3.3}$$

To formalize things, first define the *oracle Bayes rule* to be the minimizer of (3.3) when $\widetilde{\Pi} = \widetilde{\Pi}_p^*$,

$$\widehat{\boldsymbol{\beta}}_{ol}(\boldsymbol{Y}) = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}_{\widetilde{\Pi}_p^*}[L(\boldsymbol{\beta}, \boldsymbol{b}) | \boldsymbol{Y}], \tag{3.4}$$

the subscript on the expectation operator indicating that the (posterior) expectation of $\boldsymbol{\beta}$ is computed under the prior $\widetilde{\Pi}_p^*$. For simplicity, we will assume throughout this section that the nuisance parameter ψ is known; the arguments that follow can be adapted to the case of unknown ψ, but we avoid this so as not to distract from the main ideas.

Our first result is stated under the Gaussian linear model, $\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}^*, \psi^2 \boldsymbol{I})$, assuming

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

$\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has full column rank. In that case, a sufficient statistic is given by

$$\mathbf{Z} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} \sim \mathcal{N}_p(\boldsymbol{\beta}^*, \psi^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1}), \tag{3.5}$$

so that it makes sense to restrict attention to estimators of $\boldsymbol{\beta}^*$ that depend on $\boldsymbol{Y}$ only through $\mathbf{Z}$. Now consider matrices $\boldsymbol{X}$ s.t., for some $-1 < \rho < 1$,

$$\boldsymbol{X}^\top \boldsymbol{X} \propto (1-\rho)\mathbf{I} + \rho \mathbf{1}\mathbf{1}^\top. \tag{3.6}$$

In this case, $\mathbf{Z} - \boldsymbol{\beta}^* \sim \mathcal{N}_p(\mathbf{0}, \psi^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1})$ is an *exchangeable* random vector. The model induced on $\mathbf{Z}$ is therefore *permutation invariant* (PI, see e.g. Berger, 2013), meaning that for any permutation τ, the distribution of $\tau(\mathbf{Z})$ under $\boldsymbol{\beta}^*$ is the same as the distribution of $\mathbf{Z}$ under $\tau(\boldsymbol{\beta}^*)$, where $\tau(\mathbf{Z}) := (Z_{\tau(1)}, ..., Z_{\tau(p)})$ is the reordering of $\mathbf{Z}$ according to τ. Here and throughout, $\boldsymbol{\beta}^*$ indexes the distribution of $\mathbf{Z}$, not of $\tau(\mathbf{Z})$. Restricting attention to estimators $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{Z})$ of $\boldsymbol{\beta}^*$, suppose now that the loss function is itself PI, i.e., $L(\tau(\boldsymbol{\beta}^*), \tau(\widehat{\boldsymbol{\beta}})) = L(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}})$ for any permutation τ; for example, this is satisfied for squared loss $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$ or, under (3.6), for the quadratic loss $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^\top \boldsymbol{X}^\top \boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = \|\widehat{\boldsymbol{\eta}} - \boldsymbol{\eta}^*\|^2$. In that case the entire problem is said to be PI, and the invariance principle calls to limit attention to the class $\mathcal{D}^{PI}$ of all *PI estimators*, i.e., estimators which satisfy

$$\widehat{\boldsymbol{\beta}}(\tau(\mathbf{Z})) = \tau(\widehat{\boldsymbol{\beta}}(\mathbf{Z})) \quad \text{for any permutation } \tau. \tag{3.7}$$

Note that a penalized likelihood estimator (1.2) for *any symmetric* function $\mathcal{P}_\lambda(\boldsymbol{\beta})$, including the plain MLE, will be PI. Also, the oracle Bayes rule $\widehat{\boldsymbol{\beta}}_{ol}$ is PI, since the likelihood (model) is PI and the prior $\widetilde{\Pi}_p^*$ is exchangeable. In fact, the following proposition says that, for any fixed $\boldsymbol{\beta}^*$, the oracle Bayes rule is the *best* PI rule.

Proposition 1. *In the Gaussian linear model with known ψ, suppose that $\boldsymbol{X}$ has the form* (3.6) *and consider estimating $\boldsymbol{\beta}^*$ under a PI loss function. Then, among all PI estimators*

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

$\widehat{\boldsymbol{\beta}}(\mathbf{Z})$, *the oracle Bayes rule* $\widehat{\boldsymbol{\beta}}_{ol}$ *is optimal, i.e., for any fixed* $\boldsymbol{\beta}^*$,

$$\widehat{\boldsymbol{\beta}}_{ol} = \underset{\widehat{\boldsymbol{\beta}} \in \mathcal{D}^{PI}}{\arg\min}\, R(\boldsymbol{\beta}^*, \widehat{\boldsymbol{\beta}}),$$

We now return to the general case of a GLM with arbitrary $\boldsymbol{X}$. While we are no longer able, in this general setup, to invoke permutation invariance considerations and establish optimality in terms of the frequentist risk as we did in the special case above, we can still show that $\widehat{\boldsymbol{\beta}}_{ol}$ is *Bayes*-optimal simultaneously over a large class of priors. To state our second result, for any fixed p-dimensional prior $\boldsymbol{\beta} \sim \widetilde{\Pi}$, let

$$\widehat{\boldsymbol{\beta}}_{\widetilde{\Pi}}(\boldsymbol{Y}) = \underset{\boldsymbol{b} \in \mathbb{R}^p}{\arg\min}\, \mathbb{E}_{\widetilde{\Pi}}[L(\boldsymbol{\beta}, \boldsymbol{b}) | \boldsymbol{Y}] \tag{3.8}$$

be the minimizer of the posterior expected squared loss under the prior $\widetilde{\Pi}$, i.e., this is the Bayes estimator under $\widetilde{\Pi}$. The following proposition essentially says that, if $\widetilde{\Pi}$ is *any exchangeable* prior, then the oracle Bayes rule $\widehat{\boldsymbol{\beta}}_{ol}$—which does not depend on $\widetilde{\Pi}$—attains smaller (no greater) Bayes risk under $\widetilde{\Pi}$ than the Bayes rule with respect to $\widetilde{\Pi}$. This statement might seem a bit unusual at first, because we know that the Bayes rule under $\widetilde{\Pi}$ is the minimizer of the Bayes risk under $\widetilde{\Pi}$. This apparent discrepancy is reconciled by noting that, whereas $\widehat{\boldsymbol{\beta}}_{\widetilde{\Pi}}$ is a function of $\boldsymbol{Y}$ only, the oracle $\widehat{\boldsymbol{\beta}}_{ol}$ is a function of $\boldsymbol{Y}$ and $\boldsymbol{\beta}$. That is, our oracle gets to see also the *realized* vector $\boldsymbol{\beta}$ up to ordering, and therefore has an advantage even over the Bayes rule that corresponds to the correct prior.

Proposition 2. *Let* $\widetilde{\Pi}$ *be any exchangeable prior on* $\boldsymbol{\beta}$. *Then*

$$r(\widetilde{\Pi}, \widehat{\boldsymbol{\beta}}_{ol}) \leq r(\widetilde{\Pi}, \widehat{\boldsymbol{\beta}}_{\widetilde{\Pi}}),$$

where the oracle Bayes rule $\widehat{\boldsymbol{\beta}}_{ol}$ *is essentially given by* (3.4), *except that, to be precise, we need to condition also on* $\{\boldsymbol{\beta}\}$ *because* $\boldsymbol{\beta}$ *is now random. Thus, formally,* $\widehat{\boldsymbol{\beta}}_{ol}(\boldsymbol{Y}) = \widehat{\boldsymbol{\beta}}_{ol}(\boldsymbol{Y}, \{\boldsymbol{\beta}\}) = \arg\min_{\boldsymbol{b} \in \mathbb{R}^p} \mathbb{E}_{\widetilde{\Pi}_p^*}[L(\boldsymbol{\beta}, \boldsymbol{b}) | \boldsymbol{Y}, \{\boldsymbol{\beta}\}]$.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

The proofs for Propositions 1 and 2 are provided in the Appendix.

Remark 1. There is a slight abuse of notation in the inequality displayed in the statement of Proposition 2 because, formally, the Bayes risk $r(\widetilde{\Pi}, \widehat{\boldsymbol{\beta}})$ is defined only for legal estimators, whereas $\widehat{\boldsymbol{\beta}}_{ol}$ depends also on the true $\boldsymbol{\beta}$. Thus, writing $\mathbb{E}_{\widetilde{\Pi}} L(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{ol})$ is more accurate than $r(\widetilde{\Pi}, \widehat{\boldsymbol{\beta}}_{ol})$, but we chose to use the latter because it makes the statement clearer.

As a simple consequence of Proposition 2 we have

Corollary 1. *For any estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\boldsymbol{Y})$ and any exchangeable prior $\widetilde{\Pi}$,*

$$r(\widetilde{\Pi}, \widehat{\boldsymbol{\beta}}_{ol}) \leq r(\widetilde{\Pi}, \widehat{\boldsymbol{\beta}}).$$

Proof. Recalling that $\mathbb{E}_{\widetilde{\Pi}} L(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}_{\widetilde{\Pi}}) \leq \mathbb{E}_{\widetilde{\Pi}} L(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ by the definition of $\widehat{\boldsymbol{\beta}}_{\widetilde{\Pi}}$, this follows immediately from Proposition 2. □

We now turn to explaining why we intuitively expect the methods proposed in Section 2 to approximate (estimate) the oracle Bayes rule. Our hierarchical Bayes (hBayes) estimates are based on the posterior distribution of ψ and $\boldsymbol{\beta}$ under the generative model, which can be written

$$f(\psi, \boldsymbol{\beta}|\boldsymbol{Y}) = \int f(\psi, \boldsymbol{\beta}, \boldsymbol{\phi}|\boldsymbol{Y}) d\boldsymbol{\phi} = \int f(\psi, \boldsymbol{\beta}|\boldsymbol{\phi}, \boldsymbol{Y}) f(\boldsymbol{\phi}|\boldsymbol{Y}) d\boldsymbol{\phi}. \tag{3.9}$$

Recall that in the generative model, $\boldsymbol{\phi}$ is the random parameter vector that specifies Π, the marginal distribution of the coefficients β_j. Thus, $f(\psi, \boldsymbol{\beta}|\boldsymbol{\phi}, \boldsymbol{Y})$ is the posterior distribution of the parameters $\boldsymbol{\beta}$ and ψ of the model (1.1), under the prior specified by (2.3) and (2.1) for a *fixed* distribution Π. Setting the FPT model hyper-parameters $(\alpha_{l,i}, \beta_{l,i})$ to 1 makes the prior (marginal) distribution of β_j the uniform density on $[a_{min}, a_{max}]$, reflecting a high degree of uncertainty regarding the distribution of the components of $\boldsymbol{\beta}$. In the very special case

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

where $n = p$ and $\boldsymbol{X} = \boldsymbol{I}_p$, (1.1) reduces to a sequence model (with "free", i.e., unrelated, parameters $\eta_j = \beta_j$), and our method resembles the hierarchical constructs in Antoniak (1974). In the sequence model, where the parameters η_i are unrelated, unlike in (1.1), putting a 'nonparametric' noninformative prior on Π is a fully Bayes alternative to (nonparametric) empirical Bayes strategies, which, in turn, pursue consistent estimation of the empirical distribution $\widetilde{\Pi}_p^*$ of the β_j's (see, e.g. Zhang, 2003; Brown and Greenshtein, 2009), in a model where $\boldsymbol{\beta}$ is fixed. Our hierarchical model in Section 2 extends these ideas to regression models (GLMs). More specifically, by extending the Polya tree prior to the more general model (1.1) we still expect the posterior of Π to approximate $\widetilde{\Pi}_p^*$, the empirical distribution of the (fixed) β_j's. It is worth remarking here that the particular choice of a Polya tree prior is just one option we found to work well in our experiments, but other 'noninformative' choices such as distributions based on Dirichlet priors, could also be considered. In our simulation studies we will show that the posterior mean of Π under the generative model does indeed provide a good approximation of $\widetilde{\Pi}_p^*$. Referring to the representation (3.9), this implies that $f(\boldsymbol{\phi}|\boldsymbol{Y})$ assigns large weights to $\boldsymbol{\phi}$ corresponding to distributions Π which are similar to $\widetilde{\Pi}_p^*$. In turn, according to (3.9), the posterior distribution of ψ and $\boldsymbol{\beta}$, and the corresponding Bayes estimates of the β_j's under the generative model, will be similar to the posterior distributions and Bayes rules under the prior (2.1) with Π replaced by $\widetilde{\Pi}_p^*$.

4. Simulations

We turn to a simulation study for demonstrating the utility of our methods. We focus on a logistic regression model and use the simulation setup of Sur and Candès (2019), where $\boldsymbol{X}$ consists of $n = 4000$ rows and $p = 800$ columns of i.i.d. $\mathcal{N}(0, 1/n)$ entries. Three different

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

configurations for the coefficient vector $\boldsymbol{\beta}$ are considered: (I) $\boldsymbol{\beta}$ has 100 replicates of -10, 100 replicates of $+10$, and 600 zeros; (II) $\boldsymbol{\beta}$ consists of 800 i.i.d. $\mathcal{N}(3, 16)$ realizations; (III) $\boldsymbol{\beta}$ consists of 400 i.i.d. $\mathcal{N}(7, 1)$ realizations, and 400 zeros. For each of the experiments (I)-(III) we ran 30 Monte Carlo rounds, generating

$$Y_i \sim Bernoulli(q_i), \qquad q_i = 1/(1 + \exp(-\boldsymbol{X}_i^\top \boldsymbol{\beta})), \tag{4.1}$$

and holding $\boldsymbol{X}$ and $\boldsymbol{\beta}$ fixed through the 30 simulation runs. We calculated the root mean square error (RMSE), the average over the 30 runs of $\sqrt{\sum_{j=1}^{p}(\beta_j - \hat{\beta}_j)^2/p}$, for six estimators of the vector $\boldsymbol{\beta}$: the maximum likelihood estimator ("MLE"), $\hat{\boldsymbol{\beta}}_{MLE}$; the bias-corrected maximum likelihood estimator ("adj-MLE") of Sur and Candès (2019), which, by the calculations they report in this example, results in $\hat{\boldsymbol{\beta}}_{MLE}/1.499$ for all cases (I)-(III); an L_2-penalized estimator ("Ridge") and an L_1-penalized estimator ("LASSO"), implemented using the `cv.glmnet` function from the `glmnet` package Friedman et al. (2010) with default specifications; a Bayesian logistic regression with an R2-D2 prior (Zhang et al., 2022; Aguilar and Bürkner, 2023), implemented using the `brms` package (Bürkner, 2021) with default settings and a total of 4000 samples drawn from two chains ("R2-D2"); and the proposed hierarchical Bayes estimator ("hBayes") using a Polya tree of level $L = 6$, with $a_{min} = \min(-24, \hat{\boldsymbol{\beta}}_{MLE} - 0.5)$ and $a_{max} = \max(24, \hat{\boldsymbol{\beta}}_{MLE} + 0.5)$, divided evenly into 64 subintervals, with 500 iterations of the Gibbs sampler using the first 100 iterations as burn-in. For each implementation of the hBayes approach we generally set $[a_{min}, a_{max}]$ to be slightly larger than range of the components of $\hat{\boldsymbol{\beta}}_{MLE}$, to ensure large overlap of the support of $\Pi(\beta|\boldsymbol{\pi}_L)$ for each simulation run in each experiment (I)-(III). As a reference, we also computed the oracle Bayes estimator ("oBayes") given in Equation (3.4). Performance of the Oracle Bayes estimates was evaluated by running 500 iterations of a permutation Gibbs sampler, described

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

in Section S2.1 of the supplement, using the first 100 iterations as burn-in.

Table 1 reports the RMSE for the five estimators and the oracle in each of the three experiments (I)-(III). In all three experiments the RMSE of hBayes was considerably smaller compared to Lasso, Ridge and R2-D2, and only slightly larger than that of the oracle. The approximation is particularly good in Experiment (II), where the distribution of the parameter vector is relatively easy to estimate. Among the other methods, R2-D2 performed best in all three experiments.

	MLE	oBayes	hBayes	R2-D2	LASSO	Ridge	adj-MLE
(I)	5.11 (0.06)	1.86 (0.03)	**1.97** (0.02)	2.33 (0.01)	2.53 (0.04)	2.85 (0.06)	3.05 (0.03)
(II)	5.21 (0.06)	2.36 (0.01)	**2.38** (0.01)	2.69 (0.01)	3.09 (0.10)	2.91 (0.05)	3.12 (0.03)
(III)	5.25 (0.07)	2.02 (0.01)	**2.10** (0.01)	2.63 (0.01)	3.06 (0.12)	2.90 (0.07)	3.11 (0.03)

Table 1: Root mean square error (RMSE) in the three simulated examples. The first column indicates the experiment setting. Numbers are averages of the RMSE from 30 simulation runs of each simulated example. Parentheses show standard errors for these averages

Figure 1 presents, for each experiment (I)-(III), the empirical cumulative distribution function (CDF) of the coordinates β_j of the *true* vector $\boldsymbol{\beta}$, along with estimates of this (true) empirical CDF corresponding to the MLE and the proposed method, see caption for details. In all three scenarios our hierarchical Bayes method was able to recover the overall shape of the true empirical CDF, whereas this shape is undetectable from the empirical CDF of the noisy maximum likelihood estimates. Our approach produces smoothed distribution estimates that are shrunk toward the uniform distribution. This can be seen more clearly in Scenarios 1 and 3, where the distribution of the β_j's has point mass.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

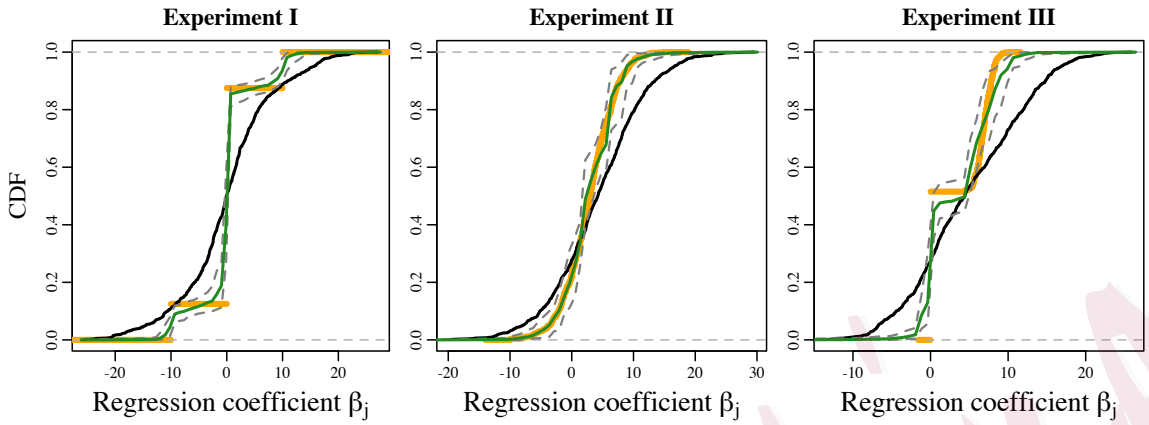

Figure 1: Empirical CDF of the true coordinates β_j of $\boldsymbol{\beta}$ (orange), and estimates thereof, for the simulation experiments (I)-(III). Black curve is empirical CDF of the p components of $\hat{\boldsymbol{\beta}}_{MLE}$. Solid green curve is the (estimated) posterior median of the CDF of Π under the generative model for the proposed method. Dashed green lines mark 0.025 and 0.975 (estimated) quantiles of this posterior.

5. Unraveling polygenic inheritance

We now apply our hierarchical Bayes approach to gain insight into the genetic architecture of polygenically inherited traits. In this context the explanatory variables are appropriately coded genotypes of genetic markers and the vector of regression coefficients represents the influence of specific genomic regions on the trait.

Many genetic studies point out that genetically inherited traits are often influenced by many genes with small effects, distributed over the whole genome (Price et al., 2008; Fraser et al., 2010, 2011; Turchin et al., 2012; Visscher and Haley, 1996; Vilhjálmsson and Nordborg, 2013). As discussed in, e.g., Wallin et al. (2021), analyzing the respective genetic data with classical "sparse" regression models leads to highly unsatisfactory results. Instead, geneticists often use mixed linear models, where the polygenic effect is represented by one

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

random effect, meant to capture the effect of all polygenes (see e.g., Kang et al. (2010)), or by many small random effects at all markers (Piepho (2009); Endelman (2011)). In the latter case the estimation of individual effects is often performed using ridge regression, which yields an empirical version of the BLUP when the genetic effects arise from the Normal distribution and the tuning parameter is adjusted according to the ratio of the variance of this distribution and the variance of the noise term. In Wallin et al. (2021) the classical mixed model is further extended by allowing a nonzero mean in the random effect, which is suitable when the investigated population is an admixture of populations subject to different selection pressures.

All methods mentioned above assume that the polygenic effects come from a Normal distribution. As shown in the real data analysis below, this assumption can be grossly inadequate. Thus, unraveling the inheritance of polygenic traits is an interesting case for our nonparametric Bayes approach. In the following section we report the results of a simulation study and a real data analysis to illustrate the advantages of our method in analyzing such genetic data.

5.1 Simulation study

Our study follows the design of the simulation study for the experimental backcross design from Wallin et al. (2021). Thus, we simulated data for $n = 400$ individuals from the backcross, where the marker genotypes can take only two values, $X_{ij} \in \{1, -1\}$, which coincide with the ancestry indictors (i.e. parental line indicators) at given loci. We simulated 10 chromosomes, each of the length of 150cM (centiMorgan), with markers spaced every 1cM. This means that for the consecutive markers on the same chromosome $P(X_{ij} = X_{i(j+1)}) \approx 0.99$,

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

while the markers on different chromosomes are independent. Following Wallin et al. (2021), the trait values are generated as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{X}_m\tilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \tag{5.1}$$

where $\boldsymbol{X}$ is the 400×1500 incidence matrix with all marker genotypes, $\boldsymbol{\epsilon} \sim \mathcal{N}_{400}(\boldsymbol{0}, 0.1\boldsymbol{I})$, and $\boldsymbol{X}_m$ is the 400×4 matrix, whose first column consists of all ones (to model the intercept term) and the remaining three columns form a subset of $\boldsymbol{X}$ containing genotypes of markers strongly associated with the trait, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_0, \ldots, \tilde{\beta}_3) \in \mathbb{R}^4$. The elements of the polygenic random effects vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_{1500})$ are i.i.d. random variables from a generalized Laplace distribution, where the Normal mean-variance mixture is of the form

$$\gamma_j = \mu + \tau\left(\xi(V_i - 1) + \sqrt{V_i}Z_i\right),$$

with $V_i \sim \Gamma(\nu, \nu)$ and $Z_i \sim \mathcal{N}(0, 1)$. The parameter μ represents the expected value of γ_j, ξ controls the asymmetry of the distribution, and ν the shape of the distribution. In our simulation we set $(\mu, \xi, \tau, \nu) = (-0.01, -2, 0.05, 0.75)$, which generates a spiked, strongly asymmetric distribution with exponential tails. We set $\boldsymbol{X}_m = [1_{400}, X_{300}, X_{750}, X_{1200}]$ and $\boldsymbol{\beta} = (0, 0.2, 0.2, -0.2)$, s.t. the first two signals are in the opposite direction of the polygenic effect, and the third is in the same direction.

We analyze our simulated data sets with the proposed hierarchical Bayes approach using a FPT of level $L = 8$, and with some classical and modern methods for the analysis of high-dimensional regression models. Among methods targeting sparse signals, we included LASSO (Tibshirani, 1996) with the tuning parameter selected by cross-validation, and four Bayesian variable selection methods: the Sum-of-single-effects method (SuSiE) by Wang et al. (2020); the *varbvsmix* algorithm of Carbonetto et al. (2017), as implemented in the

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

varbvs package in R (VARBVSMIX); Spike-and-Slab LASSO (SSL) of Ročková and George (2018); and the Expectation-Maximization Variable Selection (EMVS) method of Ročková and George (2014). Among methods targeting dense signals, we included Ridge Regression (Hoerl and Kennard, 1970) with the tuning parameter selected by cross-validation; Bayesian Multiple Regression with Adaptive Shrinkage (Mr. ASH) by Kim et al. (2022); and the method of Wallin et al. (2021) based on the mixed regression model (5.1).

Table 2: Estimates of MRNE obtained by averaging RNE over 200 simulation runs

	lasso	susie	vbsmix	ssl	emvs	horseshoe	ridge	mr.ash	mix	hBayes	oracle
$\hat{\boldsymbol{\beta}}$	0.89	3.10	1.10	3.81	0.94	0.95	0.96	0.95	0.77	0.76	0.74
$\hat{\boldsymbol{\beta}}_s$	0.75	4.07	0.89	3.40	0.61	0.63	0.76	0.72	0.48	0.46	0.45
$\hat{\boldsymbol{\eta}}$	0.07	0.31	0.11	0.24	0.03	0.03	0.02	0.07	0.02	0.02	0.02

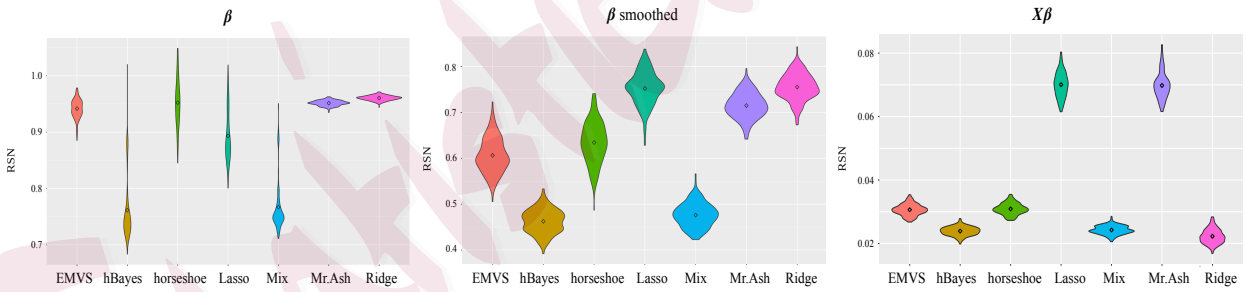

Figure 2: Empirical distribution of relative norm of the error (RNE) on 200 simulation runs

In Tables 2 and 3, we provide the mean relative norm of the error (MRNE) in estimation,

$$MRNE(\hat{\beta}) = \mathbb{E}\left(\frac{||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||}{||\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}||}\right),$$

where $\bar{\boldsymbol{\beta}} = \frac{1}{p}\sum_{j=1}^{p}\beta_j$. Empirical distributions of the relative norm of the error (RNE) are visualized in Figure 2. Due to the strong correlation between neighboring markers, it is quite

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

Table 3: Estimated MRNE for the large fixed effects

	β_1	β_2	β_3	smoothed	β_1	β_2	β_3
lasso	0.11	0.11	0.09	lasso	0.07	0.07	0.05
ridge	0.19	0.19	0.18	ridge	0.10	0.10	0.12
mix	0.03	0.06	0.03	mix	0.02	0.02	0.02
hBayes	0.05	0.07	0.04	hBayes	0.02	0.02	0.02
mr ash	0.19	0.19	0.17	mr ash	0.18	0.18	0.02
EMVS	0.12	0.13	0.11	EMVS	0.10	0.11	0.10
SSlasso	0.21	0.22	0.20	SSlasso	0.21	0.22	0.19
varbvsmix	0.20	0.20	0.14	varbvsmix	0.21	0.21	0.07
oracle	0	0	0	oracle	0.02	0.02	0.02

difficult to appropriately estimate individual genetic effects. As an alternative approach, we propose to estimate a smoothed version of $\boldsymbol{\beta}$, denoted $\boldsymbol{\beta}_s$, by a correspondingly smoothed version of $\hat{\boldsymbol{\beta}}_s$. Thus, we define

$$MRNE(\hat{\boldsymbol{\beta}}_s) = \mathbb{E}\left(\frac{\|\hat{\boldsymbol{\beta}}_s - \boldsymbol{\beta}_s\|}{\|\boldsymbol{\beta}_s - \bar{\boldsymbol{\beta}}_s\|}\right),$$

where $\boldsymbol{\beta}_s$ and $\hat{\boldsymbol{\beta}}_s$ are obtained by averaging $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$, respectively, over ± 5 cM windows within the chromosome boundaries. Table 2 and Figure 2 display the results for the entire vector $\boldsymbol{\beta}$, while Table 3 shows the accuracy in estimating the three larger fixed effects, β_1, β_2 and β_3. Table 2 and Figure 2 include also MRNE of the error in prediction, i.e., in estimating $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$. All MRNEs are estimated based on 200 independent replicates of the entire experiment, where each uses an independent draw of the design matrix $\boldsymbol{X}$, the vector

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

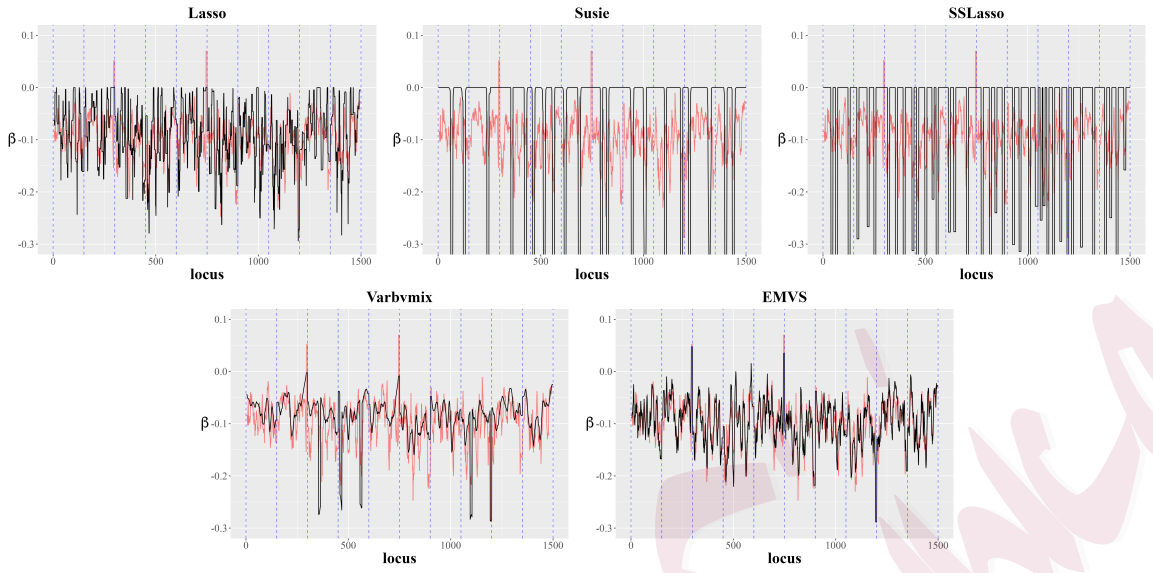

Figure 3: Results of the analysis of one simulated data by the methods aimed for the selection of important variables under the sparsity assumption. Going left to right, up to down we have: first plot is the cross-validated LASSO, second is the *SuSiE* algorithm of Wang et al. (2020), followed by the spike and slab lasso of Ročková and George (2018), and *varbvsmix* algorithm of Carbonetto et al. (2017) and *EMVS* algorithm of Ročková and George (2014). Red lines mark the true genetic effects and the black lines their estimates

of polygenic effects $\boldsymbol{\gamma}$ and the vector $\boldsymbol{\epsilon}$ of error terms. As shown in Table 2, two methods targeting sparse signals, SuSiE and SSL, perform poorly on our example.

Figure 3 shows that these methods result in many large false discoveries in the direction of the summary polygenic effect, and fail to locate two signals of the opposite sign. These observations align with the findings of Wallin et al. (2021), which highlight the shortcomings of "sparse" signal methods when applied to traits with a polygenic component. Two other spike and slab methods, *varbsmix* and *EMVS*, perform substantially better, with *EMVS* adapting particularly well and performing similarly to methods targeted at dense signals. It

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

is also interesting to note the relatively good performance of cross-validation LASSO.

Figure 2 presents violin plots of RNE for our hierarchical Bayes method and some of the competitors mentioned above. The hierarchical Bayes method produces the best results in terms of estimation error, slightly outperforming the specialized method of Wallin et al. (2021). Both of these methods yield the smallest MRNE while exhibiting appreciable variability when estimating regression coefficients at each of the very dense locations. This "variability" effect disappears after smoothing the estimate over a 10 cM window. Interestingly, in terms of estimation error, ridge regression and *Mr.Ash* are outperformed by *EMVS* and cross-validation Lasso. As illustrated in Figure 4, this is a result of oversmoothing of the signal in *ridge* and *Mr. Ash*. However, in terms of prediction accuracy, ridge regression does well, while cross-validation Lasso and *Mr. Ash* are substantially worse, compared to other methods included in Figure 2.

5.2 Real Data Analysis

We use our hierarchical Bayes approach ($L = 8$ as in the simulation in the previous subsection) to analyze the popular *Drosophila* data (Zeng et al., 2000). The purpose of the analysis is to identify genes influencing the shape of the posterior lobe of the male genital arch in *Drosophila*. The size and shape variation of the males' posterior lobes (which are highly correlated) are quantified by averaging over both sides of the morphometric descriptor (PC1) based on elliptical Fourier and principal components analyses. These data were extensively analyzed in Zeng et al. (2000), Bogdan et al. (2008) and Wallin et al. (2021), using different approaches based on the different multiple regression models. Zeng et al. (2000) and Bogdan et al. (2008) use a standard fixed effects model. Zeng et al. (2000) report 17 Quantitative

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

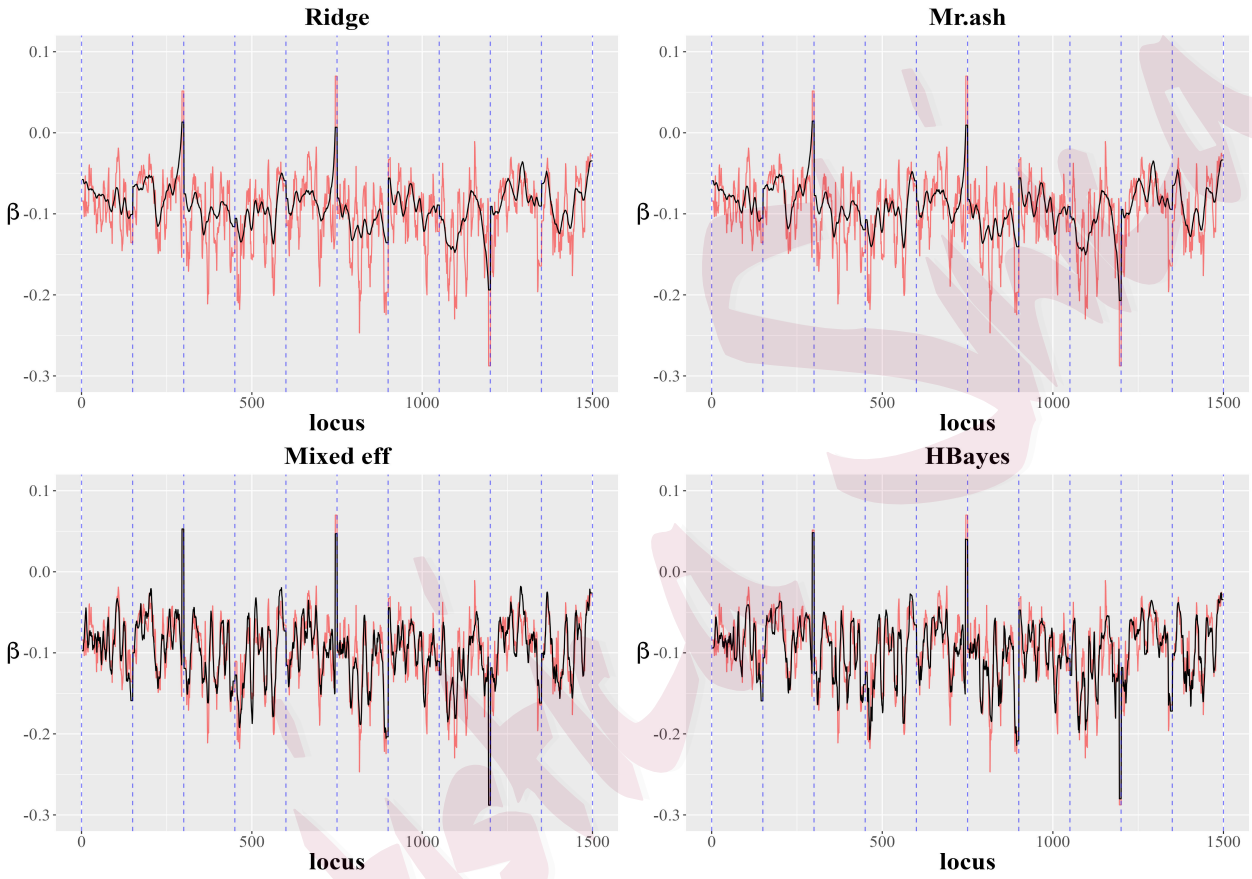

Figure 4: Results of the analysis of one simulated data set. Going left to right, up to down we have: first plot is ridge, second is Mr.Ash of Kim et al. (2022), third is the Wallin et al. (2021) method based on the mixed effect model, and fourth is our nonparametric Bayes approach. Red lines mark the true genetic effects and the black lines their estimates.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

Trait Loci (QTL), approximately uniformly distributed over the two chromosomes, with two of the strongest QTL located close to the centers of these chromosomes.

The dataset includes genotypes of 39 markers on 2 autosomes for $n = 491$ individuals. Following Bogdan et al. (2008) and Wallin et al. (2021), we used $m = 161$ pseudo-marker explanatory variables spaced every 2cM. The values of these pseudo-markers are calculated as the conditional expectations of the corresponding genotypes, given the genotypes of observed flanking markers, as in the regression interval mapping of Haley and Knott (1992). Such pseudo-marker explanatory variables are more strongly correlated than the markers spaced every 2cM. Thus, to control the variance of locus specific estimates of regression coefficients, we performed our analysis using only every third of the pseudo-markers, i.e. using pseudo-markers spaced every 6cM. Figure 5 shows 95% and 50% Bayesian credible intervals for the loci-specific genetic effects and the estimates of the cumulative distribution of $\boldsymbol{\beta}$ for our nonparametric approach and the random effects model in Wallin et al. (2021).

Our analysis indicates a systematic negative polygenic effect, i.e. many relatively weak QTL effects of the negative sign on both chromosomes. The nonparametric Bayes estimates are more "peaky" than the estimates from a random effects model, especially in the direction of negative values. This difference is most visible in the plot of the 50% credible intervals, which for the proposed method are entirely contained on the side of negative values. Also, the median posterior distribution from the nonparametric Bayes approach is strongly asymmetric: it is almost truncated at zero, and has a relatively heavy negative tail. Heavy tails of the true polygenic distribution are also reflected in the 95% credible intervals, which are wider for the nonparametric Bayes approach than for the mixed model of Wallin et al. (2021). The Normal random effects model seems to over-smooth, and it also places more weight on

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

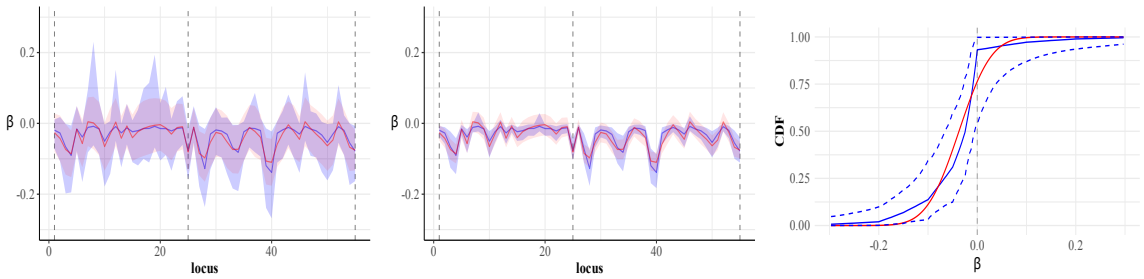

Figure 5: Two left panels are the posterior estimates of $\boldsymbol{\beta}$ ordered on the chromosomes. Blue and red areas represent pointwise credible intervals for our nonparametric Bayes method and the random effect model of Wallin et al. (2021). The lines are corresponding posterior means. The plot to the left presents intervals at $5\% - 95\%$ coverage level, and in the center, for level $25\% - 75\%$. Right panel shows estimate of the CDF of the β_j's for the random effects model (red) and the posterior median from the nonparametric Bayes method (blue). Dashed lines are pointwise 95% credible intervals.

the positive values. Despite these differences, the summary results from both methods are quite similar. Both methods estimate heritability (% of trait variability explained by genetic causes, $h^2 = (\hat{\mathrm{V}}[Y] - \mathbb{E}[\sigma^2|Y])/\hat{\mathrm{V}}[Y])$ of 72%, and the 95% credible intervals for the posterior CDF of the polygenic effects cover the estimated Normal prior for the random effects model. Our approach gives more refined estimates for locations with strong polygenic effects, and is able to capture asymmetry and heavy tails in the distribution.

6. Discussion

In this work we have proposed a new regularization method for estimation in GLMs. The method is inspired by an oracle shrinkage rule, defined as Bayes against the uniform prior on all permutations of the *true* vector $\boldsymbol{\beta}$. The theory in Section 3 presents two different

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

notions of optimality for the aforementioned oracle, a frequentist notion and a more general Bayesian notion, providing a formal justification for pursuing that oracle and motivating our nonparametric Bayes method (the relationship between the oracle and the data-driven hierarchical Bayes estimator is explained in Section 3). The simulation results provide proof of concept that the empirical distribution of the true coefficients β_j can be estimated using the proposed hierarchical Bayes method, and that, on the computational side, both the oracle and hierarchical Bayes procedures can be approximated effectively via Gibbs sampling. The simulations in Section 4 consider fixed coefficient vectors and asymptomatically independent $\boldsymbol{X}$ matrix, which, informally speaking, yield a model that is approximately permutation invariant. In this type of models the oracle Bayes rule actually minimizes *frequentist* risk (the result in Proposition 1 under the Normal linear model being a special case). The simulation in Section 5.1 which includes a non-exchangeable $\boldsymbol{X}$ matrix and a coefficient vector in which all but 3 of the 1500 components are i.i.d. random samples, is in essence a Bayesian simulation that generates repeated joint realizations of the parameter and the data, and thus the average RNE entries reported in Table 2 are practically evaluating *average* risks and demonstrate the result in Corollary 1. It is important to note that while our theoretical results apply to the *expected* loss of the oracle Bayes estimator, in our simulations the oracle Bayes estimators consistently achieved the smallest *realized* loss, and the loss of the proposed hierarchical Bayes estimates was close to that. On the computational side, the Gibbs sampling implementation explored in this paper may become impractical in moderately high dimensions. To extend the proposed hierarchical Bayes approach to higher dimensions, one will likely need to forego MCMC schemes and consider combining the Polya tree prior with Variational Bayes methods.

We conclude with some further comments on the relevance of the results in Section 3

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

to an analysis of the proposed method itself, which is an important direction for future work. To formally analyze the proposed Polya tree-based method, one could start, as in Antoniak (1974), with a genuinely Bayesian setup where $\beta_j \sim \Pi_0$, i.i.d., for some fixed and unknown distribution Π_0. In this case the oracle Bayes rule would be naturally defined as the minimizer of the Bayes risk under Π_0. If $P(\cdot|\boldsymbol{Y})$ denotes the posterior of Π in our generative model, which postulates *a priori* that $\Pi \sim P$ for a Polya tree distribution P, then a primary goal would be to show that $P(\cdot \mid \boldsymbol{Y})$ converges to Π_0 under suitable conditions. If such a result can be obtained, it would generalize existing consistency results on density estimation with Polya tree priors (e.g., Castillo, 2017). If $P(\cdot \mid \boldsymbol{Y})$ converges to Π_0, which in terms of the representation in (3.9) means that the posterior of $\boldsymbol{\phi}$ converges to some $\boldsymbol{\phi}_0$ (corresponding to Π_0), then the Bayes rule for $\boldsymbol{\beta}$ under our generative model should in turn be consistent for the Bayes rule under the true prior Π_0. Existing consistency results in the literature, for example those in Castillo (2017), correspond to the situation where the β_j's are observed. In our case, where only the Y_i's are observed, and the likelihood of each Y_i depends on the common vector parameter $(\beta_1, \ldots, \beta_p)$, is considerably more challenging and deserves separate consideration.

Code and data availability. Code required to reproduce the numerical experiments in this paper is available at the GitHub repository `https://github.com/JonasWallin/hBayes` under `experiment/article`.

Supplementary Material

The online Supplementary Material contains proof, and details on the Gibbs sampling algorithm.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

Acknowledgments

A.W. was supported by the Israeli Science Foundation (ISF) under grant no. 2679/24. M.B. and J.W. were supported by the Swedish Research Council under grant no. 2020-05081.

References

Aguilar, J. E. and P.-C. Bürkner (2023). Intuitive joint priors for bayesian linear multilevel models: The r2d2m2 prior. *Electronic Journal of Statistics 17*(1), 1711–1767.

Andersen, M. R., A. Vehtari, O. Winther, and L. K. Hansen (2017). Bayesian inference for spatio-temporal spike-and-slab priors. *Journal of Machine Learning Research 18*(139), 1–58.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174.

Berger, J. O. (2013). *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media.

Berry, D. A. and R. Christensen (1979). Empirical bayes estimation of a binomial parameter via mixtures of dirichlet processes. *The Annals of Statistics 7*(3), 558–568.

Bogdan, M., F. Frommlet, P. Biecek, R. Cheng, J. K. Ghosh, and R. W. Doerge (2008). Extending the modified bayesian information criterion (mbic) to dense markers and multiple interval mapping. *Biometrics 64*(4), 1162–1169.

Bogdan, M., E. van den Berg, C. Sabatti, W. Su, and E. J. Candès (2015). Slope—adaptive variable selection via convex optimization. *The Annals of Applied Statistics 9*(3), 1103–1140.

Bondell, H. and B. Reich (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics 64*(1), 115–123.

Brown, L. D. and E. Greenshtein (2009). Nonparametric empirical bayes and compound decision approaches to estimation of a high-dimensional vector of normal means. *The Annals of Statistics 37*(4), 1685–1704.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

Bürkner, P.-C. (2021). Bayesian item response modeling in R with brms and Stan. *Journal of Statistical Software 100*(5), 1–54.

Carbonetto, P., X. Zhou, and M. Stephens (2017). varbvs: Fast variable selection for large-scale regression. *arXiv preprint arXiv:1709.06597*.

Carlin, B. P. and T. A. Louis (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. London: Chapman & Hall.

Carvalho, C. M., N. G. Polson, and J. G. Scott (2010). The horseshoe estimator for sparse signals. *Biometrika 97*(2), 465–480.

Castillo, I. (2017). Pólya tree posterior distributions on densities. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, Volume 53, pp. 2074–2102. Institut Henri Poincaré.

Diaconis, P. and D. Freedman (1980). Finite exchangeable sequences. *The Annals of Probability 8*(4), 745–764.

Efron, B. and C. Morris (1973). Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association 68*(341), 117–130.

Endelman, J. B. (2011). Ridge regression and other kernels for genomic selection with r package rrblup. *The Plant Genome 4*, 250–255.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The Annals of Statistics 1*(2), 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The annals of statistics 2*(4), 615–629.

Fraser, H. B., T. Babak, J. Tsang, Y. Zhou, B. Zhang, M. Mehrabian, and E. E. Schadt (2011). Systematic detection of polygenic cis-regulatory evolution. *PLOS Genetics 7*(3), e1002023.

Fraser, H. B., A. Moses, and E. E. Schadt (2010). Evidence for widespread adaptive evolution of gene expression in

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

budding yeast. *Proceedings of the National Academy of Sciences of the United States of America 107*, 2977–2982.

Friedman, J., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

George, E. and D. P. Foster (2000). Calibration and empirical bayes variable selection. *Biometrika 87*(4), 731–747.

Ghosh, J. K. and R. Ramamoorthi (2003). *Bayesian nonparametrics*. Springer.

Greenshtein, E. and Y. Ritov (2009). Asymptotic efficiency of simple decisions for the compound decision problem. In J. Rojo (Ed.), *Optimality: The Third Erich L. Lehmann Symposium*, Volume 49 of *IMS Lecture Notes–Monograph Series*, pp. 266–275. Institute of Mathematical Statistics.

Haley, C. and S. Knott (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*, 315–324.

Hannan, J. F. and H. Robbins (1955). Asymptotic solutions of the compound decision problem for two completely specified distributions. *The Annals of Mathematical Statistics 26*(1), 37–51.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Jiang, W., M. Bogdan, J. Josse, B. Miasojedow, V. Rockova, and T. Group (2019). Adaptive bayesian slope–high-dimensional model selection with missing values. *arXiv preprint arXiv:1909.06631*.

Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin (2010). Variance component model to account for sample structure in genome- wide association studies. *Nature Genetics 42*, 348–354.

Kim, Y., W. Wang, P. Carbonetto, and M. Stephens (2022). A flexible empirical bayes approach to multiple linear regression and connections with penalized regression. *arXiv preprint arXiv:2208.10910*.

Lavine, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *The Annals of Statistics 20*(3), 1222–1235.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

Lavine, M. (1994). More aspects of polya tree distributions for statistical modelling. *The Annals of Statistics 22*(3), 1161–1176.

Lindley, D. V. and A. F. Smith (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society: Series B (Methodological) 34*(1), 1–18.

Piepho, H.-P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Science 49*, 1165–1176.

Price, A. L., N. J. Patterson, D. Hancks, S. Myers, D. Reich, V. G. Cheung, and R. S. Spielman (2008). Effects of *cis* and *trans* genetic ancestry on gene expression in african americans. *PLOS Genetics 4*(12), e1000294.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, pp. 131–149. University of California Press.

Robbins, H. (1956). An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, Berkeley, California, pp. 157–163. University of California Press.

Robbins, H. (1963). The empirical bayes approach to testing statistical hypotheses. *Revue de l'Institut International de Statistique 31*(3), 195–208.

Ročková, V. and E. I. George (2014). Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association 109*(506), 828–846.

Ročková, V. and E. I. George (2018). The spike-and-slab lasso. *Journal of the American Statistical Association 113*(521), 431–444.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*(6), 1135–1151.

Sur, P. and E. J. Candès (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences 116*(29), 14516–14525.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Turchin, M., C. Chiang, C. Palmer, S. Sankararaman, D. Reich, C. GIANT, and J. N. Hirschhorn (2012). Evidence of widespread selection on standing variation in europe at height-associated snps. *Nature Genetics 44*, 1015–1019.

Vilhjálmsson, B. and M. Nordborg (2013). The nature of confounding in genome-wide association studies. *Nature Reviews Genetics 14*, 1–2.

Visscher, P. M. and C. S. Haley (1996). Detection of quantitative trait loci in line crosses under infinitesimal genetic models. *Theoretical and Applied Genetics 93*, 691–702.

Wallin, J., M. Bogdan, P. A. Szulc, R. W. Doerge, and D. O. Siegmund (2021, 01). Ghost QTL and hotspots in experimental crosses: novel approach for modeling polygenic effects. *Genetics 217*(3). iyaa041.

Wang, G., A. Sarkar, P. Carbonetto, and M. Stephens (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology 82*(5), 1273–1300.

Yuan, M. and Y. Lin (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association 100*(472), 1215–1225.

Zeng, Z., J. Liu, L. Stam, C. Kao, J. Mercer, and C. Laurie (2000). Genetic architecture of a morphological shape difference between two drosophila species. *Genetics 154*(1), 299–310.

Zhang, C.-H. (2003). Compound decision theory and empirical bayes methods. *The Annals of Statistics 31*(2), 379–390.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics 38*(2), 894 – 942.

Zhang, Y. D., B. P. Naughton, H. D. Bondell, and B. J. Reich (2022). Bayesian regression using a prior on the model fit: The r2-d2 shrinkage prior. *Journal of the American Statistical Association 117*(538), 862–874.

Statistica Sinica: Preprint
doi:10.5705/ss.202025.0221

Department of Statistics, Hebrew University of Jerusalem

E-mail: asaf.weinstein@mail.huji.ac.il

Department of Statistics, Lund University

E-mail: jonas.wallin@stat.lu.se

Department of Statistics and OR, Tel Aviv University

E-mail: yekutiel@tauex.tau.ac.il

Institute of Mathematics, University of Wroclaw

E-mail: malgorzata.bogdan@uwr.edu.pl