Biometrika (2019), **106**, 3, *pp.* 533–546
Printed in Great Britain

doi: 10.1093/biomet/asz017
Advance Access publication 24 May 2019

Estimating the error variance in a high-dimensional linear model

BY GUO YU

Department of Statistics, University of Washington, Box 354322, Seattle, Washington 98105, U.S.A

gy63@uw.edu

AND JACOB BIEN

Data Sciences and Operations, Marshall School of Business, University of Southern California, 3670 Trousdale Pkwy, Los Angeles, California 90089, U.S.A.

jbien@usc.edu

SUMMARY

The lasso has been studied extensively as a tool for estimating the coefficient vector in the high-dimensional linear model; however, considerably less is known about estimating the error variance in this context. In this paper, we propose the natural lasso estimator for the error variance, which maximizes a penalized likelihood objective. A key aspect of the natural lasso is that the likelihood is expressed in terms of the natural parameterization of the multi-parameter exponential family of a Gaussian with unknown mean and variance. The result is a remarkably simple estimator of the error variance with provably good performance in terms of mean squared error. These theoretical results do not require placing any assumptions on the design matrix or the true regression coefficients. We also propose a companion estimator, called the organic lasso, which theoretically does not require tuning of the regularization parameter. Both estimators do well empirically compared to pre-existing methods, especially in settings where successful recovery of the true support of the coefficient vector is hard. Finally, we show that existing methods can do well under fewer assumptions than previously known, thus providing a fuller story about the problem of estimating the error variance in high-dimensional linear models.

Some key words: Error variance estimation; Gaussian exponential family; Lasso; Natural parameterization.

1. INTRODUCTION

The linear model

$$y = X\beta^* + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n) \tag{1}$$

is one of the most fundamental models in statistics. It describes the relationship between a response vector $y \in \mathbb{R}^n$ and a fixed design matrix $X \in \mathbb{R}^{n\times p}$. When $p \gg n$, estimating the coefficient vector β^* is a challenging, well-studied problem. Perhaps the most common method in this setting is the lasso (Tibshirani, 1996), which assumes that β^* is sparse and solves the following convex optimization problem:

$$\hat{\beta}_\lambda \in \underset{\beta \in \mathbb{R}^p}{\arg\min} \left(\frac{1}{n} \|y - X\beta\|_2^2 + 2\lambda \|\beta\|_1 \right). \tag{2}$$

Over the past decade, an extensive literature has emerged studying the properties of $\hat{\beta}_\lambda$ from both computational (e.g., Hastie et al., 2015) and theoretical (e.g., Bühlmann & van de Geer, 2011) perspectives.

Compared to the vast amount of work on estimating β^*, relatively little attention has been paid to the problem of estimating σ^2, which captures the noise level or extent to which y cannot be predicted from X. Nonetheless, reliable estimation of σ^2 is important for quantifying the uncertainty in estimating β^*. A series of recent advances in high-dimensional inference (Bühlmann, 2013; Zhang & Zhang, 2014; van de Geer et al., 2014; Lockhart et al., 2014; Javanmard & Montanari, 2014; Lee et al., 2016; Tibshirani et al., 2016; Taylor & Tibshirani, 2017; Ning & Liu, 2017) may very well be the determining factor for the widespread adoption of the lasso and related methods in fields where p-values and confidence intervals are required. Thus, estimating σ^2 reliably in finite samples is crucial.

If β^* were known, then the optimal estimator for σ^2 would of course be $n^{-1}\|y - X\beta^*\|_2^2 = n^{-1}\|\varepsilon\|_2^2$. Thus, a naive estimator for σ^2 based on an estimator $\hat{\beta}$ of β^* would be

$$\hat{\sigma}^2_{\text{naive}} = \frac{1}{n}\|y - X\hat{\beta}\|_2^2. \tag{3}$$

However, a simple calculation in the classical $n > p$ setting shows that such an estimator is biased downward: a least-squares oracle with knowledge of the true support $S = \{j : \beta_j^* \neq 0\}$ scales this to give an unbiased estimator,

$$\hat{\sigma}^2_{\text{oracle}} = \frac{1}{n - |S|}\|y - X_S X_S^+ y\|_2^2, \tag{4}$$

where X_S is a submatrix of X with columns indexed by S and X_S^+ is its pseudoinverse. Many papers in this area discuss the difficulty of estimating σ^2 and warn of the perils of underestimating it: if σ^2 is underestimated then one gets anticonservative confidence intervals, which are highly undesirable (Tibshirani et al., 2018).

Reid et al. (2016) carry out an extensive review and simulation study of several estimators of σ^2 (Fan et al., 2012; Sun & Zhang, 2012; Dicker, 2014), and they devote special attention to studying the estimator

$$\hat{\sigma}^2_R = \frac{1}{n - \hat{s}_\lambda}\|y - X\hat{\beta}_\lambda\|_2^2, \tag{5}$$

where $\hat{\beta}_\lambda$ is as in (2), with λ selected using a crossvalidation procedure, and $\hat{s}_\lambda$ is the number of nonzero elements in $\hat{\beta}_\lambda$. They show that (5) has promising performance in a wide range of simulation settings and provide an asymptotic theoretical understanding of the estimator in the special case where X is an orthogonal matrix.

While intuition from (4) suggests that (5) is a quite reasonable estimator when S can be well recovered, it also points to the question of how well the estimator will perform when S is not well recovered by the lasso. The conditions required for the lasso to recover S are much stricter than the conditions needed for it to do well in prediction (e.g., van de Geer & Bühlmann, 2009). The scale factor $(n - \hat{s}_\lambda)^{-1}$ used in $\hat{\sigma}^2_R$ means that this approach depends not just on the predicted values of the lasso, $X\hat{\beta}_\lambda$, but on the magnitude of the set of nonzero elements in $\hat{\beta}_\lambda$. Indeed, we find that in situations where recovering S is challenging, $\hat{\sigma}^2_R$ tends to yield less favourable empirical performance. The theoretical development in Reid et al. (2016) sidesteps this complication by working in an asymptotic regime in which $\hat{\sigma}^2_R$ behaves like the naive estimator (3). To understand the finite-sample performance of $\hat{\sigma}^2_R$ would require considering the behaviour of the random variable $\hat{s}_\lambda$. Clearly, when $\hat{s}_\lambda \approx n$, even small fluctuations in $\hat{s}_\lambda$ can lead to large fluctuations in $\hat{\sigma}^2_R$. Finally, from a practical standpoint, computing $\hat{s}_\lambda$ is a numerically sensitive operation in that it requires the choice of a threshold size for calling a value numerically zero, and the assurance that one has solved the problem to sufficient precision.

Based on these observations, we propose in this paper a completely different approach to estimating σ^2. The basic premise of our framework is that when both β^* and σ^2 are unknown, it is convenient to formulate the penalized loglikelihood problem in terms of

$$\phi = \frac{1}{\sigma^2}, \qquad \theta = \frac{\beta}{\sigma^2}, \tag{6}$$

the natural parameters of the Gaussian multi-parameter exponential family with unknown mean and variance. The negative Gaussian loglikelihood is not jointly convex in the (β, σ) parameterization, in fact, even with β fixed, it is nonconvex in σ. However, in the natural parameterization the negative loglikelihood is jointly convex in (ϕ, θ).

We penalize this negative loglikelihood with an ℓ_1-norm on the natural parameter θ and call this new estimator the natural lasso. We show in § 3 that the resulting error variance estimator can in fact be very simply expressed as the minimizing value of the regular lasso problem (2):

$$\hat{\sigma}^2_\lambda = \min_{\beta \in \mathbb{R}^p} \left(\frac{1}{n} \|y - X\beta\|_2^2 + 2\lambda \|\beta\|_1 \right). \tag{7}$$

Observing that the first term is $\hat{\sigma}^2_{\text{naive}}$, we directly see that the natural lasso counters the naive method's downward bias through an additive correction; this is in contrast to $\hat{\sigma}^2_R$'s reliance on a multiplicative correction that sometimes may be unstable. Computing (7) is clearly no harder than solving a lasso and, unlike $\hat{\sigma}^2_R$, does not require determining a threshold for deciding which coefficient estimates are numerically zero. Furthermore, we establish finite-sample bounds on the mean squared error that hold without making any assumptions on the design matrix X. Our theoretical analysis suggests a second approach that is also based on the natural parameterization. The theory that we develop for this method, which we call the organic lasso, relies on weaker assumptions. We find that both methods have competitive empirical performance relative to $\hat{\sigma}^2_R$ and show particular strength in settings in which support recovery is known to be challenging.

Our final contribution is to show that existing methods can also attain high-dimensional consistency under no assumptions on X. In particular, we provide finite-sample bounds for $\hat{\sigma}^2_{\text{naive}}$, with $\hat{\beta}$ in (3) taken to be the standard lasso or the square-root/scaled lasso estimator (Belloni et al., 2011; Sun & Zhang, 2012). Previous results about $\hat{\sigma}^2_{\text{naive}}$ have placed strong assumptions on X. Thus, our work provides a fuller story about the problem of estimating the error variance in high-dimensional linear models.

2. NATURAL PARAMETERIZATION

The negative loglikelihood function in (1) is, up to a constant,

$$L\left(\beta,\sigma^2 \mid X,y\right) = \frac{n}{2}\log\sigma^2 + \frac{\|y - X\beta\|_2^2}{2\sigma^2}.$$

When σ^2 is known, the σ dependence can be ignored, leading to the standard least-squares criterion; however, when σ is unknown, performing a full minimization of the penalized negative loglikelihood amounts to solving a nonconvex optimization problem even with a convex penalty.

The nonconvexity of the Gaussian negative loglikelihood in its variance, or more generally its covariance matrix, is a well-known difficulty (Bien & Tibshirani, 2011). In this context, working instead with the inverse covariance matrix is common (Yuan & Lin, 2007; Banerjee et al., 2008; Friedman et al., 2008). We take an analogous approach when estimation of σ^2 is of interest, considering the natural parameterization (6) of the Gaussian multi-parameter exponential family with unknown variance,

$$L\left(\phi^{-1}\theta,\phi^{-1} \mid X,y\right) = -\frac{n}{2}\log\phi + \frac{1}{2}\phi\left\|y - X\frac{\theta}{\phi}\right\|_2^2 = -\frac{n}{2}\log\phi + \phi\frac{\|y\|_2^2}{2} - y^{\mathrm{T}}X\theta + \frac{\|X\theta\|_2^2}{2\phi}.$$

Observing that attaining sparsity in θ is equivalent to attaining sparsity in β, we propose the following penalized maximum loglikelihood estimator:

$$\left(\hat\theta_\lambda,\hat\phi_\lambda\right) \in \mathop{\arg\min}_{\phi>0,\,\theta}\left\{-\frac{1}{2}\log\phi + \phi\frac{\|y\|_2^2}{2n} - \frac{1}{n}y^{\mathrm{T}}X\theta + \frac{\|X\theta\|_2^2}{2n\phi} + \lambda\Omega(\theta,\phi)\right\} \tag{8}$$

for a convex penalty $\Omega(\theta,\phi)$ that induces sparsity in θ. We will focus on $\Omega(\theta,\phi) = \|\theta\|_1$ in § 3 and $\Omega(\theta,\phi) = \phi^{-1}\|\theta\|_1^2$ in § 4. This problem is jointly convex in (θ,ϕ). While this is a general property of exponential families due to the convexity of the cumulant-generating function, we can see it in this special case because of the convexity of $-\log$ and the convexity of the quadratic-over-linear function (Boyd & Vandenberghe, 2004; Rockafellar, 2015). Given a solution to (8), we can reverse (6) to get estimators for σ^2 and β^*:

$$\tilde\sigma_\lambda^2 = \frac{1}{\hat\phi_\lambda}, \qquad \tilde\beta_\lambda = \frac{\hat\theta_\lambda}{\hat\phi_\lambda}. \tag{9}$$

Before proceeding with an analysis of the estimators (9) with specific choices of $\Omega(\theta,\phi)$, we point out a similarity between our method and that of Städler et al. (2010), who consider a different convexifying reparameterization of the Gaussian loglikelihood, using $\rho = \sigma^{-1}$ and $\gamma = \sigma^{-1}\beta$. They put an ℓ_1-norm penalty on γ, which has the same sparsity pattern as β, and solve

$$\min_{\rho>0,\gamma}\left(-\log\rho + \frac{1}{2n}\|\rho y - X\gamma\|_2^2 + \lambda\|\gamma\|_1\right). \tag{10}$$

Sun & Zhang (2010) give an asymptotic analysis of the solution to (10) under a compatibility condition. A modification of this problem (Antoniadis, 2010) gives the scaled lasso (Sun & Zhang, 2012), which is known to be equivalent to the square-root lasso (Belloni et al., 2011):

$$\tilde\beta_{\mathrm{SQRT}} = \mathop{\arg\min}_{\beta\in\mathbb{R}^p}\left(\frac{1}{\sqrt{n}}\|y - X\beta\|_2 + \lambda\|\beta\|_1\right), \qquad \tilde\sigma_{\mathrm{SQRT}}^2 = \frac{1}{n}\left\|y - X\tilde\beta_{\mathrm{SQRT}}\right\|_2^2. \tag{11}$$

With the same parameterization (ρ, γ), Dalalyan & Chen (2012) propose the scaled Dantzig selector under the assumption of fused sparsity. Under the restricted eigenvalue condition, they establish the same rate of convergence in estimating the error variance as the fast prediction error rate of the standard lasso (Hebiri & Lederer, 2013; Lederer et al., 2018; Dalalyan et al., 2017).

3. Natural lasso estimator of error variance

We first propose the natural lasso, which is the solution to (8) with $\Omega(\theta, \phi) = \|\theta\|_1$. One might think that solving the natural lasso would involve a specialized algorithm. The following proposition shows, remarkably, that this is not the case.

Proposition 1. *The natural lasso estimator* $(\tilde{\beta}_\lambda, \tilde{\sigma}^2_\lambda)$ *defined in* (9)*, where* $(\hat{\theta}_\lambda, \hat{\phi}_\lambda)$ *is a solution to* (8) *with* $\Omega(\theta, \phi) = \|\theta\|_1$*, satisfies the following properties:*

(i) $\tilde{\beta}_\lambda = \hat{\beta}_\lambda$*, a solution to the standard lasso* (2)*;*
(ii) $\tilde{\sigma}^2_\lambda = \hat{\sigma}^2_\lambda$*, the standard lasso's optimal value* (7)*.*

Furthermore, $\hat{\sigma}^2_\lambda = n^{-1}(\|y\|_2^2 - \|X\hat{\beta}_\lambda\|_2^2)$*.*

The proof of this proposition and all theoretical results that follow can be found in the Supplementary Material. Thus, to get the natural lasso estimator of (β^*, σ^2), one simply solves the standard lasso (2) and returns a solution and the minimal value.

An attractive property of the natural lasso estimator $\hat{\sigma}^2_\lambda$ is the relative ease with which one can prove bounds about its performance. Since $\hat{\sigma}^2_\lambda$ is the optimal value of the lasso problem, the objective value at any vector β provides an upper bound on $\hat{\sigma}^2_\lambda$. Likewise, any dual feasible vector provides a lower bound on $\hat{\sigma}^2_\lambda$. These considerations are used to prove the following lemma, which shows that for a suitably chosen λ, the natural lasso variance estimator gets close to the oracle estimator of σ^2.

Lemma 1. *If* $\lambda \geqslant n^{-1}\|X^T\varepsilon\|_\infty$*, then* $|\hat{\sigma}^2_\lambda - n^{-1}\|\varepsilon\|_2^2| \leqslant 2\lambda\|\beta^*\|_1$.

The result above is deterministic in that it does not rely on any statistical assumptions or arguments. The next result adds such considerations to give a mean squared error bound for the natural lasso.

Theorem 1. *Suppose that each column* X_j *of the matrix* $X \in \mathbb{R}^{n\times p}$ *has been scaled so that* $\|X_j\|_2^2 = n$ *for all* $j = 1, \ldots, p$*, and assume that* $\varepsilon \sim N(0, \sigma^2 I_n)$*. Then, for any constant* $M > 1$*, the natural lasso estimator* (7) *with* $\lambda = \sigma(2Mn^{-1}\log p)^{1/2}$ *satisfies the following relative mean squared error bound:*

$$E\left\{\left(\frac{\hat{\sigma}^2_\lambda}{\sigma^2} - 1\right)^2\right\} \leqslant \left\{\left(8M + 8\frac{p^{1-8M}}{\log p}\right)^{1/2} \frac{\|\beta^*\|_1}{\sigma}\left(\frac{\log p}{n}\right)^{1/2} + \left(\frac{2}{n}\right)^{1/2}\right\}^2.$$

Corollary 1. *The following mean absolute error bound follows from Theorem 1:*

$$E\left|\frac{\hat{\sigma}^2_\lambda}{\sigma^2} - 1\right| = O\left\{\frac{\|\beta^*\|_1}{\sigma}\left(\frac{\log p}{n}\right)^{1/2}\right\}. \tag{12}$$

Proof. This follows from Jensen's inequality. □

Remark 1. Theorem 1 can be easily generalized to the case where the independently and identically distributed zero-mean error ε_i with variance σ^2 is sub-Gaussian or sub-exponential. A high probability bound can be obtained for ε_i with bounded polynomial moments. In particular, for any $m \geqslant 3$, if $E(|\varepsilon_i|^m) \leqslant (m!)^{-1}2K^{m-2}$ for some $K > 0$, and if each column X_j is scaled so that $\sum_{i=1}^n X_{ij}^m = n$ for $j = 1, \ldots, p$, then with $\lambda = 4K\sigma n^{-1/2}(\log p)^{1/2}$ we have that

$$\left| \hat{\sigma}_\lambda^2 - \frac{\|\varepsilon\|_2^2}{n} \right| = O\left\{ \sigma \left\|\beta^*\right\|_1 \left(\frac{\log p}{n}\right)^{1/2} \right\}$$

holds with probability greater than $1 - p^{-1}$.

To put Theorem 1 in context, we devote the remainder of this section to considering what bounds are available for other methods for estimating σ^2. Bayati et al. (2013) propose an estimator of σ^2 based on estimating the mean squared error of the lasso. They show that their estimator of σ^2 is asymptotically consistent with fixed p as $n \to \infty$. In contrast, we provide finite-sample results and these include the $p \gg n$ case. Also, the consistency result in Bayati et al. (2013) is based on the assumption of independent Gaussian features, and in extending this to the case of correlated Gaussian features, the authors invoke a conjecture. In comparison, (12) is essentially free of assumptions on the design matrix.

The natural lasso also compares favourably to the method-of-moments-based estimator of Dicker (2014) in terms of mean squared error bounds. In particular, Dicker (2014) establishes a $O_{\mathrm{P}}[(\sigma^{-2}\tau^2 + 1)\{n^{-2}(p+n)\}^{1/2}]$ relative mean squared error rate, where $\tau^2 = \|\Sigma^{-1/2}\beta^*\|_2^2$ and Σ is the covariance of features X. This rate can be much slower for large p.

Notably, the mean squared error bound in Theorem 1 does not put any assumption on X, β^*, or σ^2. In this sense, the result is analogous to a slow-rate bound (Rigollet & Tsybakov, 2011; Dalalyan et al., 2017), which appears in the lasso prediction consistency context. While it is well known (Sun & Zhang, 2012) or can be easily verified that under stronger conditions, i.e., compatibility or restricted eigenvalue conditions, the naive estimator (3) based on the lasso and $\tilde{\sigma}^2_{\mathrm{SQRT}}$ in (11) attain a faster rate, $O(|S|n^{-1}\log p)$, it is natural to ask whether these two estimators also attain a rate bound as in (12) when the conditions on X are not assumed. The following two results give an affirmative answer to this question.

PROPOSITION 2. *Under the conditions of Theorem 1, the naive estimator* (3) *based on the lasso estimator* $\hat{\beta}_\lambda$ *with* $\lambda = 4\sigma(n^{-1}\log p)^{1/2}$ *has the following bound with probability greater than* $1 - p^{-1}$*:*

$$\left| \hat{\sigma}^2_{\mathrm{naive}} - \frac{\|\varepsilon\|_2^2}{n} \right| \leqslant 16\sigma\|\beta^*\|_1 \left(\frac{\log p}{n}\right)^{1/2}. \tag{13}$$

Relatedly, Chatterjee & Jafarov (2016) also consider a setting with no assumptions on X and derive an error bound $O\{\|\beta^*\|_1^{1/2}(n^{-1}\log p)^{1/4}\}$ for (3) for a lasso estimator $\hat{\beta}_\lambda$ with λ in (2) selected using a crossvalidation procedure.

Lederer et al. (2018) derive a slow-rate bound for the prediction error of the square-root lasso. They show that there exists a value of λ for which $\lambda = 3n^{-1/2}\|X^T\varepsilon\|_\infty\|y - X\tilde{\beta}_{\mathrm{SQRT}}\|_2^{-1}$ and bound $\|X\tilde{\beta}_{\mathrm{SQRT}} - X\beta^*\|_2^2$ at this value. The following result establishes the high-dimensional consistency of $\tilde{\sigma}^2_{\mathrm{SQRT}}$ under no assumptions on X.

Proposition 3. *Under the conditions of Theorem 1, the square-root/scaled lasso estimator $\tilde{\sigma}^2_{\mathrm{SQRT}}$ in (11) based on $\tilde{\beta}_{\mathrm{SQRT}}$ with $\lambda = 3n^{-1/2}\|X^T\varepsilon\|_\infty\|y - X\tilde{\beta}_{\mathrm{SQRT}}\|_2^{-1}$ has the following bound with probability greater than $1 - p^{-1}$:*

$$\left|\tilde{\sigma}^2_{\mathrm{SQRT}} - \frac{\|\varepsilon\|_2^2}{n}\right| \leqslant 12\sigma\|\beta^*\|_1\left(\frac{\log p}{n}\right)^{1/2}. \tag{14}$$

We see that the rate of the natural lasso in (12) matches, up to a constant factor, the rates (13) and (14). The values of λ used in Propositions 2 and 3 are larger than would be necessary for standard prediction error bounds; this technique (Gaynanova, 2018) is key to the proofs of the two propositions. Although the same rate is obtained in Theorem 1, Proposition 2, and Proposition 3, we have not established that this is the best possible rate obtainable in this setting that makes no assumption on X.

4. Organic lasso estimator of error variance

4.1. *Method formulation*

In practice, the value of the regularization parameter λ in (7) may be chosen via crossvalidation; however, Theorem 1 has a regrettable theoretical shortcoming: it requires using a value of λ that itself depends on σ, the very quantity that we are trying to estimate! This is a well-known theoretical limitation of the lasso and related methods that motivated the square-root/scaled lasso. In this section, we propose a second new method, which retains the natural parameterization but remedies the natural lasso's theoretical shortcoming by using a modified penalty. We define the organic lasso as a solution to (8) with $\Omega(\theta, \phi) = \phi^{-1}\|\theta\|_1^2$, i.e.,

$$\left(\check{\theta}_\lambda, \check{\phi}_\lambda\right) = \underset{\phi>0,\ \theta}{\arg\min}\left(-\frac{1}{2}\log\phi + \phi\frac{\|y\|_2^2}{2n} - \frac{1}{n}y^T X\theta + \frac{\|X\theta\|_2^2}{2n\phi} + \lambda\frac{\|\theta\|_1^2}{\phi}\right). \tag{15}$$

We observe that the penalty $\phi^{-1}\|\theta\|_1^2$ is jointly convex in (ϕ, θ) since it can be expressed as $g\{h(\theta), \phi\}$ where $h(\theta) = \|\theta\|_1$ is convex and $g(x, \phi) = \phi^{-1}x^2$ is a jointly convex function that is strictly increasing in x for $x \geqslant 0$ (Boyd & Vandenberghe, 2004; Rockafellar, 2015).

Given a solution to the above problem, we can reverse (6) to give the organic lasso estimators of (β^*, σ^2), i.e., $\check{\beta}_\lambda = \check{\phi}_\lambda^{-1}\check{\theta}_\lambda$ and $\check{\sigma}^2_\lambda = \check{\phi}_\lambda^{-1}$. Furthermore, $\phi^{-1}\|\theta\|_1^2$ still induces sparsity in θ, and thus the final estimate $\check{\beta}_\lambda$ is sparse. In direct analogy to the natural lasso, the following proposition shows that we can find $\check{\sigma}^2_\lambda$ and $\check{\beta}_\lambda$ without actually solving (15).

Proposition 4. *The organic lasso estimators $(\check{\beta}_\lambda, \check{\sigma}^2_\lambda)$ correspond to the solution and minimal value of an ℓ_1^2-penalized least-squares problem*

$$\check{\beta}_\lambda = \underset{\beta\in\mathbb{R}^p}{\arg\min}\left(\frac{1}{n}\|y - X\beta\|_2^2 + 2\lambda\|\beta\|_1^2\right), \tag{16}$$

$$\check{\sigma}^2_\lambda = \min_{\beta\in\mathbb{R}^p}\left(\frac{1}{n}\|y - X\beta\|_2^2 + 2\lambda\|\beta\|_1^2\right). \tag{17}$$

Thus, to compute the organic lasso estimator, one simply solves a penalized least squares problem, where the penalty is the square of the ℓ_1-norm. This can be thought of as the exclusive lasso with a single group (Zhou et al., 2010; Campbell & Allen, 2017). We show in the next section that solving this problem is no harder than solving a standard lasso problem.

One readily sees the connection of the organic lasso to the square-root lasso (11): to get (17), one takes squares of both the loss and the ℓ_1 penalty of (11). However, their origins are actually different in nature: the organic lasso is a maximum of the Gaussian loglikelihood with a scale-equivariant sparsity-inducing penalty under parameterization (6), while (11) minimizes the ℓ_1-penalized Huber concomitant loss function (Antoniadis, 2010; Sun & Zhang, 2012).

4.2. *Algorithm*

Coordinate descent is easy to implement and has steadily maintained its place as a state-of-the-art approach for solving lasso-related problems (Friedman et al., 2007). For coordinate descent to work, one typically verifies separability in the nonsmooth part of the objective function (Tseng, 2001). However, the ℓ_1^2 penalty in (16) is not separable in the coordinates of β. Lorbert et al. (2010) propose a coordinate descent algorithm to solve the pairwise elastic net problem, a generalization of (16), and a proof of the convergence of the algorithm is given in the 2012 PhD thesis of A. Lorbert from Princeton University. In Algorithm 1, we give a coordinate descent algorithm specific to solving (16). The R package `natural` (Yu, 2017; R Development Core Team, 2019) provides a C implementation of Algorithm 1.

Algorithm 1. A coordinate descent algorithm to solve (16).

Require: Initial estimate $\beta^{(0)} \in \mathbb{R}^p$, $X \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$, and $\lambda > 0$
Set $\beta \leftarrow \beta^{(0)}$ and $r \leftarrow y - X\beta$
For $j = 1, \ldots, p; 1, \ldots, p; \ldots$ until convergence:
  $\beta_j^{\text{new}} \leftarrow (2\lambda + n^{-1}\left\|X_j\right\|_2^2)^{-1}\mathcal{S}(n^{-1}X_j^T r + n^{-1}\left\|X_j\right\|_2^2\beta_j, 2\lambda\|\beta_{-j}\|_1)$
  $r \leftarrow r + X_j\beta_j - X_j\beta_j^{\text{new}}$
  $\beta_j \leftarrow \beta_j^{\text{new}}$
Output β

Each coordinate update requires $O(n)$ operations. In Algorithm 1, $\mathcal{S}(a, b) = \text{sgn}(a)(|a| - b)_+$ is the soft-threshold operator. Empirically Algorithm 1 is as fast as solving a lasso problem. In A. Lorbert's 2012 PhD thesis it is shown that for any initial estimate $\beta^{(0)} \in \mathbb{R}^p$, every limit point of Algorithm 1 is an optimal point of the objective function of (16). This implies that the ℓ_1^2 penalty, although not separable, is well enough behaved that any point that is minimum in every coordinate of the objective function in (16) is indeed a global minimum.

4.3. *Theoretical results*

A first indication that the organic lasso may succeed where the natural lasso falls short is in terms of scale equivariance. As the design X is usually standardized to be unitless, scale equivariance in this context refers to the effect of scaling y.

PROPOSITION 5. *The organic lasso is scale equivariant, i.e., for any $t > 0$,*

$$\check{\beta}_\lambda(ty) = t\check{\beta}_\lambda(y), \qquad \check{\sigma}_\lambda(ty) = t\check{\sigma}_\lambda(y).$$

Scale equivariance is a property associated with the ability to prove results in which the tuning parameter λ does not depend on σ. For example, the square-root/scaled lasso (11) is scale equivariant while the lasso, and thus the natural lasso, is not. In particular, $\hat{\beta}_\lambda(ty) \neq t\hat{\beta}_\lambda(y)$ and $\hat{\sigma}_\lambda(ty) \neq t\hat{\sigma}_\lambda(y)$ for some $t > 0$.

In Lemma 1, we saw how expressing an estimator as the optimal value of a convex optimization problem allows us to take full advantage of convex duality in order to derive bounds on the estimator. We therefore start our analysis of (17) by characterizing its dual problem.

LEMMA 2. *The dual problem of* (17) *is*

$$\max_{u\in\mathbb{R}^n}\left\{\frac{1}{n}\left(\|y\|_2^2-\|y-u\|_2^2\right)-\frac{1}{2\lambda}\left\|\frac{X^Tu}{n}\right\|_\infty^2\right\}.$$

Similar arguments to those in Lemma 1 give a bound expressing $\check{\sigma}_\lambda^2$'s closeness to the oracle estimator of σ^2.

LEMMA 3. *If* $\lambda \geqslant n^{-1}\left\|X^T(\sigma^{-1}\varepsilon)\right\|_\infty$, *then*

$$-2\lambda\sigma^2\left(\frac{\|\beta^*\|_1}{\sigma}+\frac{1}{4}\right)\leqslant \check{\sigma}_\lambda^2-\frac{1}{n}\|\varepsilon\|_2^2\leqslant 2\lambda\left\|\beta^*\right\|_1^2.$$

Comparing this with Lemma 1, we see that the condition on λ depends only on a quantity $\sigma^{-1}\varepsilon\sim N(0,I_n)$ that is independent of σ^2. Indeed, this leads to a mean squared error bound with the desired property of λ not depending on σ.

THEOREM 2. *Suppose that each column X_j of the matrix $X\in\mathbb{R}^{n\times p}$ has been scaled so that $\|X_j\|_2^2=n$ for all $j=1,\ldots,p$, and $\varepsilon\sim N\left(0,\sigma^2 I_n\right)$. Then, for any constant $M>1$, the organic lasso estimator* (17) *with $\lambda=(2Mn^{-1}\log p)^{1/2}$ satisfies the following relative mean squared error bound:*

$$E\left\{\left(\frac{\check{\sigma}_\lambda^2}{\sigma^2}-1\right)^2\right\}\leqslant\left\{\left(8M+8\frac{p^{1-8M}}{\log p}\right)^{1/2}\max\left(\frac{\|\beta^*\|_1^2}{\sigma^2},\frac{\|\beta^*\|_1}{\sigma}+\frac{1}{4}\right)\left(\frac{\log p}{n}\right)^{1/2}+\left(\frac{2}{n}\right)^{1/2}\right\}^2. \tag{18}$$

Compared with Theorem 1, the organic lasso estimator of σ^2 retains the same rate in terms of n and p but has a slower rate in terms of $\sigma^{-1}\|\beta^*\|_1$. Importantly, though, the value of λ attaining (18) does not depend on σ. This tuning-insensitive property is also enjoyed by the square-root/scaled lasso estimate of σ^2, as shown in Proposition 3. As in Remark 1, similar high-probability bounds can be obtained for ε with bounded polynomial moments.

Although not central to our main purpose, the organic lasso estimator (16) of β^* is interesting in its own right. The following theorem gives a slow-rate bound in prediction error.

THEOREM 3. *For any $L>0$, the solution to* (16) *with $\lambda=\{2n^{-1}(\log p+L)\}^{1/2}$ has the following bound on the prediction error with probability greater than $1-\exp(-L)$:*

$$\frac{1}{n}\left\|X\check{\beta}_\lambda-X\beta^*\right\|_2^2\leqslant\left(\sigma^2+4\left\|\beta^*\right\|_1^2\right)\left(\frac{2\log p+2L}{n}\right)^{1/2}.$$

In the Supplementary Material, we provide mappings between the path of the natural lasso, $\{\hat{\beta}_\lambda:\lambda>0\}$, and the path of the organic lasso $\{\check{\beta}_\lambda:\lambda>0\}$. We also include a fast-rate prediction error bound of (16) under a compatibility condition.

5. SIMULATION STUDIES

5.1. *Simulation settings*

Reid et al. (2016) carry out an extensive simulation study to compare many error variance estimators. We have matched their simulation settings fairly closely, so that the performance comparison with various other methods mentioned in Reid et al. (2016) can be inferred. Specifically, all simulations are run with $p = 500$ and $n = 100$. Each row of the design matrix X is generated from a multivariate $N(0, \Sigma)$ distribution, with $\Sigma_{ij} = \rho \in (0, 1)$ for $i \neq j$ and $\Sigma_{ii} = 1$. To generate β^*, we randomly select the indices of $\lceil n^\alpha \rceil$ nonzero elements out of p variables, where $\alpha \in (0, 1)$, and each of the nonzero elements has a value that is randomly drawn from a Laplace distribution with rate 1. The error variance is generated using $\sigma^2 = \tau^{-1}\beta^{*T}\Sigma\beta^*$ for $\tau > 0$. Finally, y is generated following (1).

Each model is indexed by a triplet (ρ, α, τ), where ρ captures the correlation among features, α determines the sparsity of β^*, and τ characterizes the signal-to-noise ratio. We vary $\rho, \alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ and $\tau \in \{0.3, 1, 3\}$. We compute a Monte Carlo estimate of both the mean squared error $E\{(\sigma^{-1}\hat{\sigma} - 1)^2\}$ and $E(\sigma^{-1}\hat{\sigma})$ as the measure of performance. The methods under comparison include (a) the naive estimator (3) with $\hat{\beta}_\lambda$ in (2), (b) the degrees of freedom adjusted estimator $\hat{\sigma}_R^2$ in (5) (Reid et al., 2016), (c) the square-root/scaled lasso (Belloni et al., 2011; Sun & Zhang, 2013), (d) the natural lasso (7), and (e) the organic lasso (17). As a benchmark, we also include the oracle $n^{-1}\|\varepsilon\|_2^2$. The R simulator package (Bien, 2016; R Development Core Team, 2019) was used for all simulations.

5.2. *Methods with the regularization parameter selected by crossvalidation*

We carry out two sets of simulations. In the first set, we compare the performance of the aforementioned methods with regularization parameter selected in a data-adaptive way. In particular, five-fold crossvalidation is used to select the tuning parameter for each method.

Due to space constraints, we present a subset of the results in Fig. 1. Additional results are presented in the Supplementary Material. The result for the square-root/scaled lasso is averaged over 100 repetitions due to the large computational time. For all other methods, the results are averaged over 1000 repetitions. Overall, the natural lasso does well in adjusting the downward bias of the naive estimator, while other methods tend to produce underestimates. In each panel, we fix the signal-to-noise ratio, τ, and correlations among features, ρ, and vary the model sparsity, α. All estimates get worse with growing α, except for the natural lasso, which improves as the true β^* gets denser. In particular, both the natural lasso and the organic lasso gain performance advantage over other methods when the underlying models do not satisfy conditions for the support recovery of the lasso solution. From left to right, Fig. 1 illustrates the effect of increasing ρ. As observed in Reid et al. (2016), high correlations can be helpful: all curves approach the oracle as ρ increases. Finally, we find that the organic lasso is uniformly better or equivalent to $\hat{\sigma}_R^2$.

Paired t-tests and Wilcoxon signed-rank tests show that the differences in mean squared errors of the different methods are significant at the 5% level for almost all points shown in Fig. 1.

Results in the Supplementary Material also show the natural lasso estimator doing well when the signal-to-noise ratio is low: the performances of all methods degrade as τ gets large. This is expected from Theorems 1 and 2, and is also observed in Reid et al. (2016).

5.3. *Methods with fixed choice of regularization parameter*

Although solving (17) is fast enough for one to use crossvalidation with the organic lasso, Theorem 2 implies that $\lambda_0 = (2n^{-1}\log p)^{1/2}$ is a theoretically sound choice of regularization

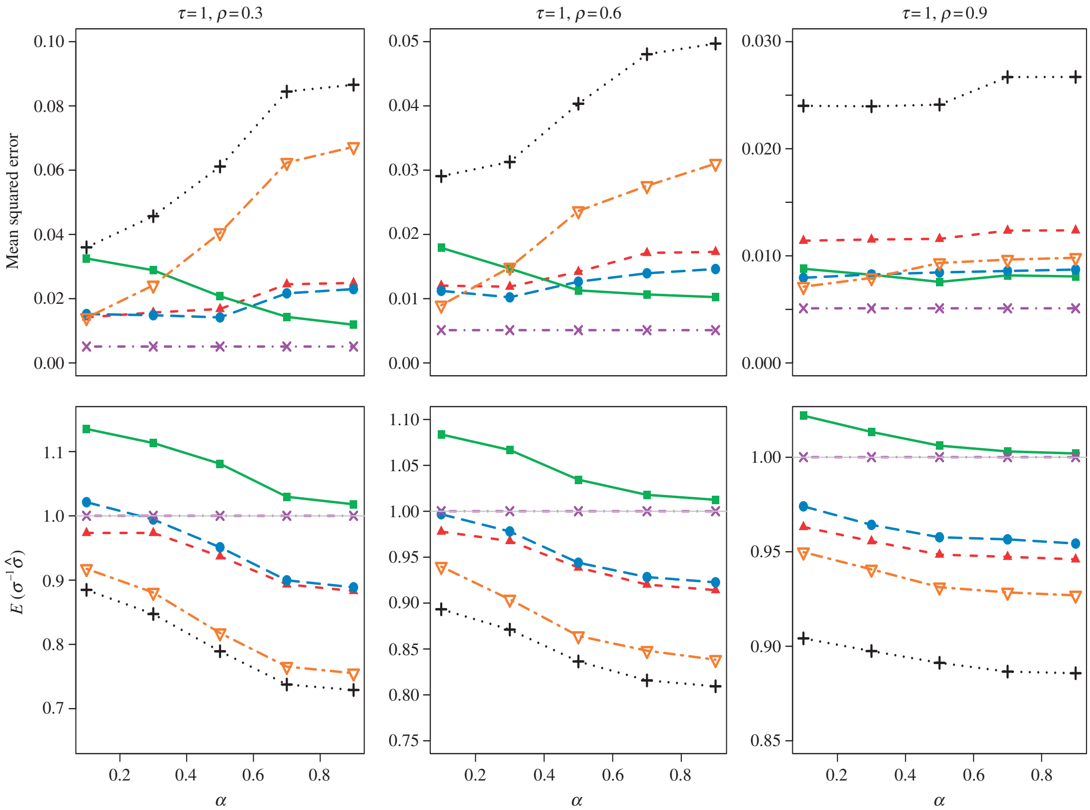

Fig. 1. Simulation results of methods using crossvalidation. From left to right, columns show Monte Carlo estimates of the mean squared error in the top panels and of $E(\sigma^{-1}\hat{\sigma})$ in the bottom panels, for various methods in three simulation settings. Line styles and their corresponding methods: naive; $\hat{\sigma}_R^2$; the square-root/scaled lasso; the natural lasso; the organic lasso; the oracle.

parameter. We also conjecture that a sharper rate may be obtainable at $\lambda_1 \geqslant n^{-2}\|X^T\epsilon\|_\infty^2$, where $\epsilon \sim N(0,1)$. With high probability, $n^{-2}\|X^T\epsilon\|_\infty^2 \approx n^{-1}\log(p)$. Thus, we also show the performance of the organic lasso with tuning parameter values $\lambda_2 = n^{-1}\log(p)$ and λ_3 taken to be a Monte Carlo estimate of $E(n^{-2}\|X^T\epsilon\|_\infty^2)$, where the expectation is with respect to $\epsilon \sim N(0,1)$.

We compare the organic lasso at these three fixed values of tuning parameter to the square-root/scaled lasso estimator (11) of error variance, which is another method whose theoretical choice of λ does not depend on σ. Sun & Zhang (2012) find that λ_0 works very well for (11), which we denote by scaled(1), and Sun & Zhang (2013) propose a refined choice of λ, which is proved to attain a sharper rate, denoted by scaled(2). The results of all the methods are averaged over 1000 repetitions.

Figure 2 shows similar patterns to Fig. 1. Specifically, a large value of ρ helps all methods, while performance generally degrades for denser β^*. Although not shown here, all methods struggle as τ increases. The theoretically justified tuning parameter λ_0 for the organic lasso appears in practice to overshrink the estimate of β^* and thus to overestimate σ^2, leading to poor performance; however, the organic lasso with the smaller tuning parameter values λ_2 and λ_3 does quite well, generally outperforming the square-root/scaled lasso-based methods.

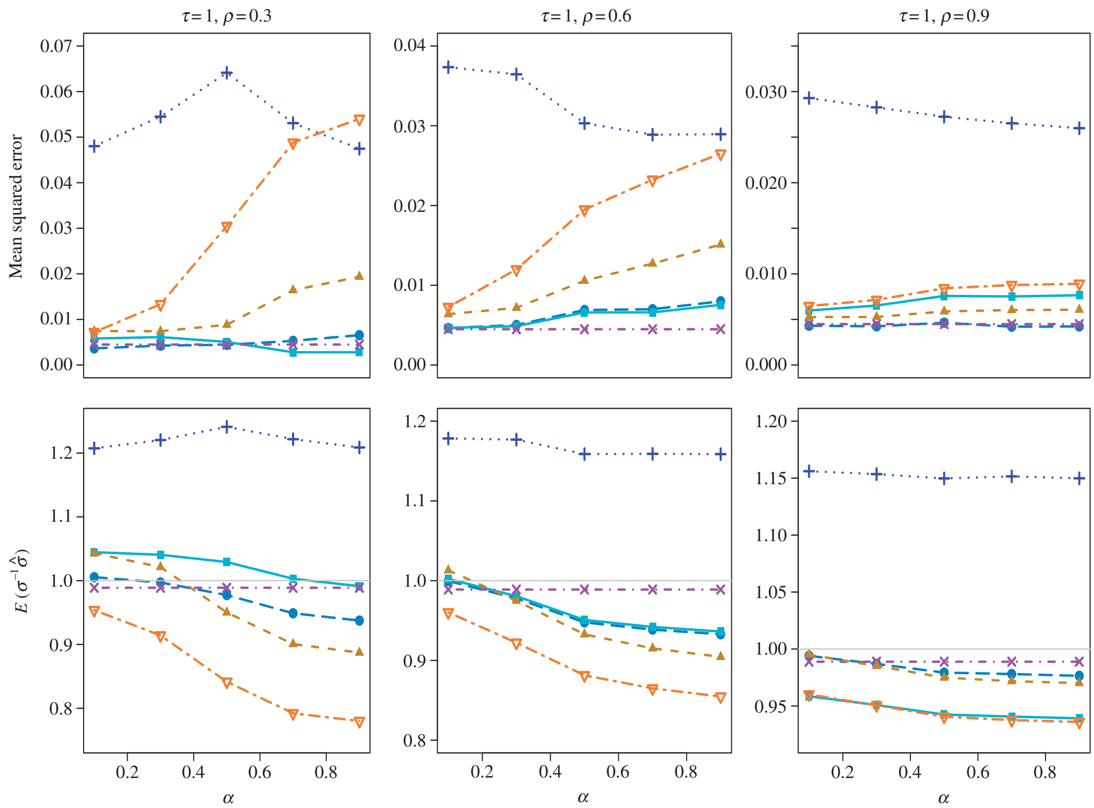

Fig. 2. Simulation results of methods using prespecified regularization parameter values. From left to right, columns show Monte Carlo estimates of the mean squared error in the top panels and of $E(\sigma^{-1}\hat{\sigma})$ in the bottom panels for various methods in three simulation settings. Line styles and their corresponding methods: +, organic (λ_0); ■, organic (λ_2); ●, organic (λ_3); ▲, scaled(1); ▽, scaled (2); ×, the oracle.

6. Error estimation for the Million Song dataset

We apply our error variance estimators to a subset of the Million Song dataset available at `https://archive.ics.uci.edu/ml/datasets/yearpredictionmsd`. The data consist of information about 463 715 songs, and the primary goal is to model the release year of a song using $p = 90$ of its timbre features. The dataset has a very large sample size so that we can reliably estimate the ground truth of the target of estimation on a very large set of held-out data. In particular, we randomly select half of the songs for this purpose and use $\bar{\sigma}^2 = (n-p)^{-1}\|y - X\hat{\beta}_{\rm LS}\|_2^2$ to form our ground truth, where $\hat{\beta}_{\rm LS}$ is the least-squares estimator of β^*. In practice, model (1) may rarely hold, which alters the interpretation of error variance estimation. Suppose the response vector y has mean μ and covariance matrix Σ. Then $\bar{\sigma}^2$ can be thought of as an estimator of the population quantity

$$\min_\beta \frac{1}{n} E\left(\|y - X\beta\|_2^2\right) = \frac{1}{n}\,\mathrm{tr}(\Sigma) + \frac{1}{n}\left\|\left(I - XX^+\right)\mu\right\|_2^2 .$$

In the special case where $\Sigma = \sigma^2 I_n$ and $\mu = X\beta^*$, as in (1), $\bar{\sigma}^2$ reduces to the linear model noise variance σ^2.

From the remaining data that were not previously used to yield $\bar{\sigma}^2$, we randomly form training datasets of size n and compare the performance of various error variance estimators. We vary

Table 1. *Mean squared error of noise variance estimation for the Million Song dataset: mean and standard errors, over* 1000 *replications, of the squared error of various methods; each entry is multiplied by* 100

n	20	40	60	80	100	120
naive	17.02 (0.68)	8.48 (0.41)	5.28 (0.26)	3.80 (0.17)	3.03 (0.13)	2.43 (0.10)
$\hat{\sigma}^2_R$	10.74 (0.45)	5.92 (0.29)	3.57 (0.17)	2.57 (0.11)	2.23 (0.10)	1.75 (0.08)
natural(cv)	8.82 (0.38)	5.23 (0.27)	3.47 (0.16)	2.61 (0.12)	2.39 (0.11)	2.01 (0.09)
organic(cv)	8.08 (0.32)	4.23 (0.20)	2.59 (0.12)	2.00 (0.08)	1.72 (0.08)	1.54 (0.07)
scaled(1)	7.43 (0.37)	4.92 (0.25)	3.84 (0.17)	3.08 (0.13)	2.94 (0.12)	2.75 (0.11)
scaled(2)	7.11 (0.28)	3.36 (0.15)	2.23 (0.10)	2.57 (0.83)	1.61 (0.07)	1.46 (0.07)
organic(λ_2)	5.87 (0.24)	3.17 (0.14)	1.93 (0.09)	1.40 (0.06)	1.20 (0.05)	1.02 (0.05)
organic(λ_3)	5.72 (0.24)	3.15 (0.14)	1.99 (0.09)	1.45 (0.07)	1.28 (0.05)	1.12 (0.05)

n in $\{20, 40, 60, 80, 100, 120\}$ to gauge the performance of these methods in situations in which $n < p$ and $n \approx p$. For each n, we repeat the data selection and error variance estimation on 1000 disjoint training sets, and report estimates of the mean squared error $E\{(\bar{\sigma}^{-1}\hat{\sigma} - 1)^2\}$ in Table 1 and estimates of $E(\bar{\sigma}^{-1}\hat{\sigma})$ in the Supplementary Material.

All methods produce a substantial performance improvement over the naive estimator for a wide range of values of n. The natural and organic lassos with crossvalidation perform either better or comparably to $\hat{\sigma}^2_R$ and are in some, but not all, cases outperformed by scaled(2). When n gets large, the natural lasso shows some upward bias, which is less problematic than downward bias. The organic lasso with the fixed choice of λ_2 or λ_3 performs extremely well for all n.

Future research directions include the analysis of the proposed methods with smaller values of λ, and extending the natural parameterization to penalized nonparametric regression.

Acknowledgement

We thank Irina Gaynanova for a useful conversation that helped us prove Propositions 2 and 3. Jacob Bien was supported by the U.S. National Science Foundation.

Supplementary material

Supplementary Material available at *Biometrika* online includes the detailed proofs of all theoretical results and extended results of the numerical studies in § 5 and § 6.

References

Antoniadis, A. (2010). Comments on: ℓ_1-penalization for mixture regression models. *Test* **19**, 257–8.

Banerjee, O., El Ghaoui, L. & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Mach. Learn. Res.* **9**, 485–516.

Bayati, M., Erdogdu, M. A. & Montanari, A. (2013). Estimating lasso risk and noise level. In *Advances in Neural Information Processing Systems 26 (NIPS'13)*. New York: Curran Associates, pp. 944–52.

Belloni, A., Chernozhukov, V. & Wang, L. (2011). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98**, 791–806.

Bien, J. (2016). The simulator: an engine to streamline simulations. *arXiv:* 1607.00021.

Bien, J. & Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika* **98**, 807–20.

Boyd, S. & Vandenberghe, L. (2004). *Convex Optimization*. Cambridge: Cambridge University Press.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**, 1212–42.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer.

Campbell, F. & Allen, G. I. (2017). Within group variable selection through the exclusive lasso. *Electron. J. Statist.* **11**, 4220–57.

CHATTERJEE, S. & JAFAROV, J. (2016). Prediction error of cross-validated lasso. *arXiv:* 1502.06291v2.
DALALYAN, A. & CHEN, Y. (2012). Fused sparsity and robust estimation for linear models with unknown variance. In *Advances in Neural Information Processing Systems 25 (NIPS'12)*. New York: Curran Associates, pp. 1259–67.
DALALYAN, A. S., HEBIRI, M. & LEDERER, J. (2017). On the prediction performance of the lasso. *Bernoulli* **23**, 552–81.
DICKER, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika* **101**, 269–84.
FAN, J., GUO, S. & HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc.* B **74**, 37–65.
FRIEDMAN, J., HASTIE, T., HÖFLING, H. & TIBSHIRANI, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302–32.
FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**, 432–41.
GAYNANOVA, I. (2018). Prediction and estimation consistency of sparse multi-class penalized optimal scoring. *arXiv:* 1809.04669.
HASTIE, T., TIBSHIRANI, R. & WAINWRIGHT, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, Florida: CRC Press.
HEBIRI, M. & LEDERER, J. (2013). How correlations influence lasso prediction. *IEEE Trans. Info. Theory* **59**, 1846–54.
JAVANMARD, A. & MONTANARI, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–909.
LEDERER, J., YU, L. & GAYNANOVA, I. (2018). Oracle inequalities for high-dimensional prediction. *arXiv:* 1608.00624v2.
LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.
LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. & TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42**, 413–68.
LORBERT, A., EIS, D. J., KOSTINA, V., BLEI, D. M. & RAMADGE, P. J. (2010). Exploiting covariate similarity in sparse regression via the pairwise elastic net. *Proc. Mach. Learn. Res*. **9**, 477–84.
NING, Y. & LIU, H. (2017). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Ann. Statist.* **45**, 158–95.
R DEVELOPMENT CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
REID, S., TIBSHIRANI, R. & FRIEDMAN, J. (2016). A study of error variance estimation in lasso regression. *Statist. Sinica* **26**, 35–67.
RIGOLLET, P. & TSYBAKOV, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731–71.
ROCKAFELLAR, R. T. (2015). *Convex Analysis*. Princeton: Princeton University Press.
STÄDLER, N., BÜHLMANN, P. & VAN DE GEER, S. (2010). ℓ_1-penalization for mixture regression models (with discussion). *Test* **19**, 209–85.
SUN, T. & ZHANG, C.-H. (2010). Comments on: ℓ_1-penalization for mixture regression models. *Test* **19**, 270–75.
SUN, T. & ZHANG, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–98.
SUN, T. & ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *J. Mach. Learn. Res.* **14**, 3385–418.
TAYLOR, J. & TIBSHIRANI, R. (2017). Post-selection inference for ℓ_1-penalized likelihood models. *Can. J. Statist.* **46**, 41–61.
TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.
TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. & WASSERMAN, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.* **46**, 1255–87.
TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. & TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Statist. Assoc.* **111**, 600–20.
TSENG, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *J. Optimiz. Theory Appl.* **109**, 475–94.
VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.
VAN DE GEER, S. A. & BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.* **3**, 1360–92.
YU, G. (2017). *natural: Estimating the Error Variance in a High-Dimensional Linear Model*. R package version 0.9.0.
YUAN, M. & LIN, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94**, 19–35.
ZHANG, C.-H. & ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc.* B **76**, 217–42.
ZHOU, Y., JIN, R. & HOI, S. (2010). Exclusive lasso for multi-task feature selection. *Proc. Mach. Learn. Res*. **9**, 988–95.

[*Received on* 5 *December* 2017. *Editorial decision on* 13 *November* 2018]