A lava attack on the recovery of sums of dense and sparse signals

Victor Chernozhukov
Christian Hansen
Yuan Liao

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP56/15

A LAVA ATTACK ON THE RECOVERY OF SUMS OF DENSE AND SPARSE SIGNALS

By Victor Chernozhukov, Christian Hansen and Yuan Liao

Massachusetts Institute of Technology, University of Chicago and University of Maryland

Common high-dimensional methods for prediction rely on having either a sparse signal model, a model in which most parameters are zero and there are a small number of non-zero parameters that are large in magnitude, or a dense signal model, a model with no large parameters and very many small non-zero parameters. We consider a generalization of these two basic models, termed here a "sparse+dense" model, in which the signal is given by the sum of a sparse signal and a dense signal. Such a structure poses problems for traditional sparse estimators, such as the lasso, and for traditional dense estimation methods, such as ridge estimation. We propose a new penalization-based method, called lava, which is computationally efficient. With suitable choices of penalty parameters, the proposed method strictly dominates both lasso and ridge. We derive analytic expressions for the finite-sample risk function of the lava estimator in the Gaussian sequence model. We also provide a deviation bound for the prediction risk in the Gaussian regression model with fixed design. In both cases, we provide Stein's unbiased estimator for lava's prediction risk. A simulation example compares the performance of lava to lasso, ridge, and elastic net in a regression example using data-dependent penalty parameters and illustrates lava's improved performance relative to these benchmarks.

1. Introduction. Many recently proposed high-dimensional modeling techniques build upon the fundamental assumption of sparsity. Under sparsity, we can approximate a high-dimensional signal or parameter by a sparse vector that has a relatively small number of non-zero components. Various ℓ_1-based penalization methods, such as the lasso and soft-thresholding, have been proposed for signal recovery, prediction, and parameter estimation within a sparse signal framwork. See [22], [15], [36], [19], [18], [44], [43], [41], [6], [9], [21], [5], [30], [39], [7], [42], [29], and others. By virtue of being based on ℓ_1-penalized optimization problems, these methods produce sparse solutions in which many estimated model parameters are set exactly to zero.

Another commonly used shrinkage method is ridge estimation. Ridge estimation differs from the aforementioned ℓ_1-penalized approaches in that it does not produce a sparse solution but instead provides a solution in which all model parameters are estimated to be non-zero. Ridge estimation is thus suitable when the model's parameters or unknown signals contain many very small components, i.e. when the model is dense. See, e.g., [25].

AMS 2000 subject classifications: Primary 62J07; secondary 62J05

Keywords and phrases: high-dimensional models, penalization, shrinkage, non-sparse signal recovery

Ridge estimation tends to work better than sparse methods whenever a signal is dense in such a way that it can not be well-approximated by a sparse signal.

In practice, we may face environments that have signals or parameters which are neither dense nor sparse. The main results of this paper provide a model that is appropriate for this environment and a corresponding estimation method with good estimation and prediction properties. Specifically, we consider models where the signal or parameter, θ, is given by the superposition of sparse and dense signals:

$$\theta = \underbrace{\beta}_{\text{dense part}} + \underbrace{\delta}_{\text{sparse part}}. \tag{1}$$

Here, δ is a sparse vector that has a relatively small number of large entries, and β is a dense vector having possibly many small, non-zero entries. Traditional sparse estimation methods, such as lasso, and traditional dense estimation methods, such as ridge, are tailor-made to handle respectively sparse signals and dense signals. However, the model for θ given above is "sparse+dense" and cannot be well-approximated by either a "dense only" or "sparse only" model. Thus, traditional methods designed for either sparse or dense settings are not optimal within the present context.

Motivated by this signal structure, we propose a new estimation method, called "lava". Let $\ell(\text{data}, \theta)$ be a general statistical loss function that depends on unknown parameter θ, and let p be the dimension of θ. To estimate θ, we propose the "lava" estimator given by

$$\widehat{\theta}_{\text{lava}} = \widehat{\beta} + \widehat{\delta} \tag{2}$$

where $\widehat{\beta}$ and $\widehat{\delta}$ solve the following penalized optimization problem:[1]

$$(\widehat{\beta}, \widehat{\delta}) = \arg \min_{(\beta', \delta')' \in \mathbb{R}^{2p}} \Big\{ \ell(\text{data}, \beta + \delta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\delta\|_1 \Big\}. \tag{3}$$

In the formulation of the problem, λ_2 and λ_1 are tuning parameters corresponding to the ℓ_2- and ℓ_1- penalties which are respectively applied to the dense part of the parameter, β, and the sparse part of the parameter, δ. The resulting estimator is then the sum of a dense and a sparse estimator. Note that the separate identification of β and δ is not required in (1), and the lava estimator is designed to automatically recover the combination $\widehat{\beta} + \widehat{\delta}$ that leads to the optimal prediction of $\beta + \delta$. Moreover, under standard conditions for ℓ_1-optimization, the lava solution exists and is unique. In naming the proposal "lava", we emphasize that it is able, or at least aims, to capture or wipe out both sparse and dense signals.

The lava estimator admits the lasso and ridge shrinkage methods as two extreme cases by respectively setting either $\lambda_2 = \infty$ or $\lambda_1 = \infty$.[2] In fact, it continuously connects the

[1]This combination of ℓ_1 and ℓ_2 penalties for regularization is similar to the robust loss function of [26] for prediction errors, though the two uses are substantially different and are motivated from a very different set of concerns. In addition, we have to choose penalty levels λ_1 and λ_2 to attain optimal performance.

[2]With $\lambda_1 = \infty$ or $\lambda_2 = \infty$, we set $\lambda_1 \|\delta\|_1 = 0$ when $\delta = 0$ or $\lambda_2 \|\beta\|_2^2 = 0$ when $\beta = 0$ so the problem is well-defined.

two shrinkage functions in a way that guarantees it will never produce a sparse solution when $\lambda_2 < \infty$. Of course, sparsity is not a requirement for making good predictions. By construction, lava's prediction risk is less than or equal to the prediction risk of the lasso and ridge methods with oracle choice of penalty levels for ridge, lasso, and lava; see Figure 1. Lava also tends to perform no worse than, and often performs significantly better than, ridge or lasso with penalty levels chosen by data-dependent rules; see Figures 4, 5, and 6.

Note that our proposal is rather different from the elastic net method, which also uses a combination of ℓ_1 and ℓ_2 penalization. The elastic net penalty function is $\theta \mapsto \lambda_2\|\theta\|_2^2 + \lambda_1\|\theta\|_1$, and thus the elastic net also includes lasso and ridge as extreme cases corresponding to $\lambda_2 = 0$ and $\lambda_1 = 0$ respectively. In sharp contrast to the lava method, the elastic net does not split θ into a sparse and a dense part and will produce a sparse solution as long as $\lambda_1 > 0$. Consequently, the elastic net method can be thought of as a sparsity-based method with additional shrinkage by ridge. The elastic net processes data very differently from lava (see Figure 2 below) and consequently has very different prediction risk behavior (see Figure 1 below).

We also consider the post-lava estimator which refits the sparse part of the model:

$$\widehat{\theta}_{\text{post-lava}} = \widehat{\beta} + \widetilde{\delta}, \tag{4}$$

where $\widetilde{\delta}$ solves the following penalized optimization problem:

$$\widetilde{\delta} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \ell(\text{data}, \widehat{\beta} + \delta) : \delta_j = 0 \text{ if } \widehat{\delta}_j = 0 \right\}. \tag{5}$$

This estimator removes the shrinkage bias induced by using the ℓ_1 penalty in estimation of the sparse part of the signal. Removing this bias sometimes results in further improvements of lava's risk properties.

We provide several theoretical and computational results about lava in this paper. First, we provide analytic expressions for the finite-sample risk function of the lava estimator in the Gaussian sequence model and in a fixed design regression model with Gaussian errors. Within this context, we exhibit "sparse+dense" examples where lava significantly outperforms both lasso and ridge. Stein's unbiased risk estimation plays a central role in our theoretical analysis, and we thus derive Stein's unbiased risk estimator (SURE) for lava. We also characterize lava's "Efron's" degrees of freedom ([17]). Second, we give deviation bounds for the prediction risk of the lava estimator in regression models akin to those derived by [5] for lasso. Third, we show that lava estimator can be computed by an application of lasso on suitably transformed data in the regression cases, where transformation is related to ridge. This leads to a computationally efficient, practical algorithm, which we employ in our simulation experiments. Fourth, we illustrate lava's performance relative to lasso, ridge, and elastic net through simulation experiments using penalty levels chosen via either minimizing the SURE or by k-fold cross-validation for all estimators. In our simulations, lava outperforms lasso and ridge in terms of prediction error over a wide range of regression models with coefficients that vary from having a rather sparse structure to having a very dense structure. When the model is very sparse,

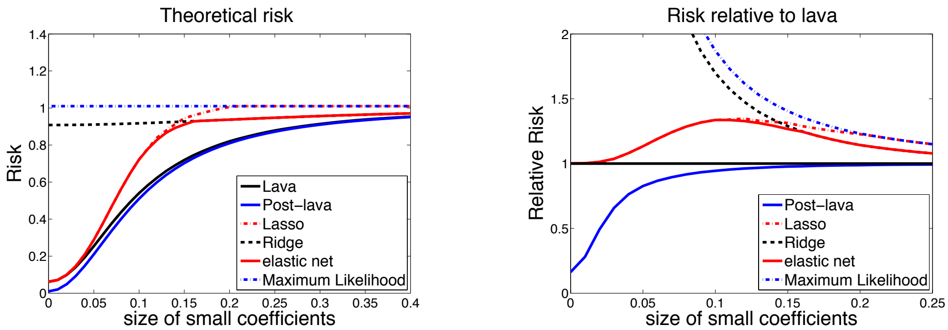

FIG 1. *Exact risk and relative risk functions of lava, post-lava, ridge, lasso, elastic net, and maximum likelihood in the Gaussian sequence model with "sparse+dense" signal structure and* $p = 100$*, using the oracle (risk minimizing) choices of penalty levels. See Section 2.5 for the description of the model. The size of "small coefficients",* q*, is shown on the horizontal axis. The size of these coefficients directly corresponds to the size of the "dense part" of the signal, with zero corresponding to the exactly sparse case. Relative risk plots the ratio of the risk of each estimator to the lava risk. Note that the relative risk plot is over a smaller set of sizes of small coefficients to accentuate comparisons over the region where there are the most interesting differences between the estimators.*

lava performs as well as lasso and outperforms ridge substantially. As the model becomes more dense in the sense of having the size of the "many small coefficients" increase, lava outperforms lasso and performs just as well as ridge. This performance is consistent with our theoretical results.

Our proposed approach complements other recent approaches to structured sparsity problems such as those considered in fused sparsity estimation ([37] and [13]) and structured matrix estimation problems ([10], [12], [20], and [28]). The latter line of research studies estimation of matrices that can be written as low rank plus sparse matrices. Our new results are related to but are sharply different from this latter line of work since our focus is on regression problems. Specifically, our chief objects of interest are regression coefficients along with the associated regression function and predictions of the outcome variable. Thus, the target statistical applications of our developed methods include prediction, classification, curve-fitting, and supervised learning. Another noteworthy point is that it is impossible to recover the "dense" and "sparse" components separately within our framework; instead, we recover the sum of the two components. By contrast, it is possible to recover the low-rank component of the matrix separately from the sparse part in some of the structured matrix estimation problems. This distinction serves to highlight the difference between structured matrix estimation problems and the framework discussed in this paper. Due to these differences, the mathematical development of our analysis needs to address a different set of issues than are addressed in the aforementioned structured matrix estimation problems.

Our work also complements papers that provide high-level general results for penalized estimators such as [32], [11], [40], [8], and [1]. Of these, the most directly related to our work are [40] and [27], which formulate the general problem of estimation in settings where the signal may be split into a superposition of different types $\theta = \sum_{\ell=1}^{L} \theta_\ell$ through the use of penalization of different types for each of the components of the superposition of the form $\sum_{\ell=1}^{L} \text{penalty}_\ell(\theta_\ell)$. Within the general framework, [40] and [27] proceed to focus their study on several leading cases in sparse estimation, emphasizing the interplay between group-wise sparsity and element-wise sparsity and considering problems in multi-task learning. By contrast, we propose and focus on another leading case, which emphasizes the interplay between sparsity and density in the context of regression learning. Our estimators and models are never sparse, and our results on finite-sample risk, SURE, degrees of freedom, sharp deviation bounds, and computation heavily rely on the particular structure brought by the use of ℓ_1 and ℓ_2 norms as penalty functions. In particular, we make use of the fine risk characterizations and tight bounds obtained for ℓ_1-penalized and ℓ_2 penalized estimators by previous of research on lasso and ridge, e.g. [4, 25], in order to derive the above results. The analytic approach we take in our case may be of independent interest and could be useful for other problems with superposition of signals of different types. Overall, the focus of this paper and its results are complementary to the focus and results in the aforementioned research.

Organization. In Section 2, we define the lava shrinkage estimator in a canonical Gaussian sequence model and derive its theoretical risk function. In Section 3, we define and analyze the lava estimator in the regression model. In Section 4 we provide simulation experiments; in Section 5, we collect final remarks; and in the Appendix, we collect proofs.

Notation. The notation $a_n \lesssim b_n$ means that $a_n \le C b_n$ for all n, for some constant C that does not depend on n. The ℓ_2 and ℓ_1 norms are denoted by $\|\cdot\|_2$ (or simply $\|\cdot\|$) and $\|\cdot\|_1$, respectively. The ℓ_0-"norm", $\|\cdot\|_0$, denotes the number of non-zero components of a vector, and the $\|.\|_\infty$ norm denotes a vector's maximum absolute element. When applied to a matrix, $\|\cdot\|$ denotes the operator norm. We use the notation $a \vee b = \max(a,b)$ and $a \wedge b = \min(a,b)$. We use x' to denote the transpose of a column vector x.

2. The lava estimator in a canonical model.

2.1. *The one dimensional case.* Consider the simple problem where a scalar random variable is given by

$$Z = \theta + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

We observe a realization z of Z and wish to estimate θ. Estimation will often involve the use of regularization or shrinkage via penalization to process input z into output $d(z)$, where the map $z \mapsto d(z)$ is commonly referred to as the shrinkage (or decision) function. A generic shrinkage estimator then takes the form $\widehat{\theta} = d(Z)$.

The commonly used lasso method uses ℓ_1-penalization and gives rise to the lasso or soft-thresholding shrinkage function:

$$d_{\text{lasso}}(z) = \arg\min_{\theta \in \mathbb{R}} \left\{ (z-\theta)^2 + \lambda_l |\theta| \right\} = (|z| - \lambda_l/2)_+ \text{sign}(z),$$

where $y_+ := \max(y, 0)$ and $\lambda_l \geq 0$ is a penalty level. The use of the ℓ_2-penalty in place of the ℓ_1 penalty yields the ridge shrinkage function:

$$d_{\text{ridge}}(z) = \arg\min_{\theta \in \mathbb{R}} \left\{ (z-\theta)^2 + \lambda_r |\theta|^2 \right\} = \frac{z}{1+\lambda_r},$$

where $\lambda_r \geq 0$ is a penalty level. The lasso and ridge estimators then take the form

$$\widehat{\theta}_{\text{lasso}} = d_{\text{lasso}}(Z), \quad \widehat{\theta}_{\text{ridge}} = d_{\text{ridge}}(Z).$$

Other commonly used shrinkage methods include the elastic-net ([44]), which uses $\theta \mapsto \lambda_2 |\theta|^2 + \lambda_1 |\theta|$ as the penalty function; hard-thresholding; and the SCAD ([19]), which uses a non-concave penalty function.

Motivated by points made in the introduction, we proceed differently. We decompose the signal into two components $\theta = \beta + \delta$, and use different penalty functions – the ℓ_2 and ℓ_1 – for each component in order to predict θ better. We thus consider the penalty function

$$(\beta, \delta) \mapsto \lambda_2 |\beta|^2 + \lambda_1 |\delta|,$$

and introduce the "lava" shrinkage function $z \mapsto d_{\text{lava}}(z)$ defined by

$$d_{\text{lava}}(z) := d_2(z) + d_1(z), \tag{6}$$

where $d_1(z)$ and $d_2(z)$ solve the following penalized prediction problem:

$$(d_2(z), d_1(z)) := \arg\min_{(\beta,\delta) \in \mathbb{R}^2} \left\{ [z - \beta - \delta]^2 + \lambda_2 |\beta|^2 + \lambda_1 |\delta| \right\}. \tag{7}$$

Although the decomposition $\theta = \beta + \delta$ is not unique, the optimization problem (7) has a unique solution for any given (λ_1, λ_2). The proposal thus defines the lava estimator of θ:

$$\widehat{\theta}_{\text{lava}} = d_{\text{lava}}(Z).$$

For large signals such that $|z| > \lambda_1/(2k)$, lava has the same bias as the lasso. This bias can be removed through the use of the post-lava estimator

$$\widehat{\theta}_{\text{post-lava}} = d_{\text{post-lava}}(Z),$$

where $d_{\text{post-lava}}(z) := d_2(z) + \tilde{d}_1(z)$, and $\tilde{d}_1(z)$ solves the following penalized prediction problem:

$$\tilde{d}_1(z) := \arg\min_{\delta \in \mathbb{R}} \left\{ [z - d_2(z) - \delta]^2 : \delta = 0 \text{ if } d_1(z) = 0 \right\}. \tag{8}$$

The removal of this bias will result in improved risk performance relative to the original estimator in some contexts.

From the Karush-Kuhn-Tucker conditions, we obtain the explicit solution to (6).

LEMMA 2.1. *For given penalty levels $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$:*

$$d_{\text{lava}}(z) = (1-k)z + k(|z| - \lambda_1/(2k))_+ \text{sign}(z) \tag{9}$$

$$= \begin{cases} z - \lambda_1/2, & z > \lambda_1/(2k) \\ (1-k)z, & -\lambda_1/(2k) \leq z \leq \lambda_1/(2k) \\ z + \lambda_1/2, & z < -\lambda_1/(2k) \end{cases} \tag{10}$$

where $k := \frac{\lambda_2}{1+\lambda_2}$. The post-lava shrinkage function is given by

$$d_{\text{post-lava}}(z) = \begin{cases} z, & |z| > \lambda_1/(2k), \\ (1-k)z, & |z| \leq \lambda_1/(2k). \end{cases}$$

The left panel of Figure 2 plots the lava shrinkage function along with various alternative shrinkage functions for $z > 0$. The top panel of the figure compares lava shrinkage to ridge, lasso, and elastic net shrinkage. It is clear from the figure that lava shrinkage is different from lasso, ridge, and elastic net shrinkage. The figure also illustrates how lava provides a bridge between lasso and ridge, with the lava shrinkage function coinciding with the ridge shrinkage function for small values of the input z and coinciding with the lasso shrinkage function for larger values of the input. Specifically, we see that the lava shrinkage function is a combination of lasso and ridge shrinkage that corresponds to using whichever of the lasso or ridge shrinkage is closer to the 45 degree line.

It is also useful to consider how lava and post-lava compare with the post-lasso or hard-thresholding shrinkage: $d_{\text{post-lasso}}(z) = z1\{|z| > \lambda_l/2\}$. These different shrinkage functions are illustrated in the right panel of Figure 2.

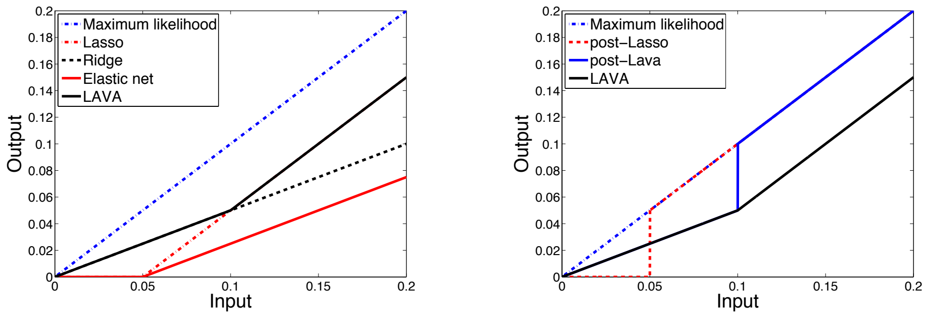

FIG 2. *Shrinkage functions. Here we plot shrinkage functions implied by lava and various commonly used penalized estimators. These shrinkage functions correspond to the case where penalty parameters are set as $\lambda_2 = \lambda_r = 1/2$ and $\lambda_1 = \lambda_l = 1/2$. In each figure, the light blue dashed line provides the 45 degree line coinciding to no shrinkage.*

From (9), we observe some key characteristics of the lava shrinkage function:

1) The lava shrinkage admits the lasso and ridge shrinkages as two extreme cases. The lava and lasso shrinkage functions are the same when $\lambda_2 = \infty$, and the ridge and lava shrinkage functions coincide if $\lambda_1 = \infty$.

2) The lava shrinkage function $d_{\text{lava}}(z)$ is a weighted average of data z and the lasso shrinkage function $d_{\text{lasso}}(z)$ with weights given by $1-k$ and k.

3) The lava never produces a sparse solution when $\lambda_2 < \infty$: If $\lambda_2 < \infty$, $d_{\text{lava}}(z) = 0$ if and only if $z = 0$. This behavior is strongly different from elastic net which always produces a sparse solution once $\lambda_1 > 0$.

4) The lava shrinkage function continuously connects the ridge shrinkage function and the lasso shrinkage function. When $|z| < \lambda_1/(2k)$, lava shrinkage is equal to ridge shrinkage; and when $|z| > \lambda_1/(2k)$, lava shrinkage is equal to lasso shrinkage.

5) The lava shrinkage does *exactly the opposite* of the elastic net shrinkage. When $|z| < \lambda_1/(2k)$, the elastic net shrinkage function coincides with the lasso shrinkage function; and when $|z| > \lambda_1/(2k)$, the elastic net shrinkage is the same as ridge shrinkage.

2.2. *The risk function of the lava estimator in the one dimensional case.* In the one-dimensional case with $Z \sim N(\theta, \sigma^2)$, a natural measure of the risk of a given estimator $\widehat{\theta} = d(Z)$ is given by

$$R(\theta, \widehat{\theta}) \; := \; \mathrm{E}[d(Z) - \theta]^2 = \sigma^2 + \mathrm{E}(Z - d(Z))^2 + 2\mathrm{E}[(Z-\theta)d(Z)]. \tag{11}$$

Let $\mathrm{P}_{\theta,\sigma}$ denote the probability law of Z. Let $\phi_{\theta,\sigma}$ be the density function of Z. We provide the risk functions of lava and post-lava in the following theorem. We also present the risk functions of ridge, elastic net, lasso, and post-lasso for comparison. To the best of our knowledge, the risk function of elastic net is also new in the literature, while that of ridge, lasso and post-lasso are known.

THEOREM 2.1 (Risk Function of Lava and Related Estimators in the Scalar Case). *Suppose $Z \sim N(\theta, \sigma^2)$. Then for $w = \lambda_1/(2k)$, $k = \lambda_2/(1+\lambda_2)$, $h = 1/(1+\lambda_2)$, $d = -\lambda_1/(2(1+\lambda_2)) - \theta$ and $g = \lambda_1/(2(1+\lambda_2)) - \theta$, we have*

$$\begin{aligned}
R(\theta, \widehat{\theta}_{\text{lava}}) &= -k^2(w+\theta)\phi_{\theta,\sigma}(w)\sigma^2 + k^2(\theta - w)\phi_{\theta,\sigma}(-w)\sigma^2 \\
&\quad + (\lambda_1^2/4 + \sigma^2)\mathrm{P}_{\theta,\sigma}(|Z| > w) + (\theta^2 k^2 + (1-k)^2\sigma^2)\mathrm{P}_{\theta,\sigma}(|Z| < w), \\
R(\theta, \widehat{\theta}_{\text{post-lava}}) &= \sigma^2[-k^2 w + 2kw - k^2\theta]\phi_{\theta,\sigma}(w) + \sigma^2[-k^2 w + 2kw + k^2\theta]\phi_{\theta,\sigma}(-w) \\
&\quad + \sigma^2\mathrm{P}_{\theta,\sigma}(|Z| > w) + (k^2\theta^2 + (1-k)^2\sigma^2)\mathrm{P}_{\theta,\sigma}(|Z| < w), \\
R(\theta, \widehat{\theta}_{\text{lasso}}) &= -(\lambda_l/2 + \theta)\phi_{\theta,\sigma}(\lambda_l/2)\sigma^2 + (\theta - \lambda_l/2)\phi_{\theta,\sigma}(-\lambda_l/2)\sigma^2 \\
&\quad + (\lambda_l^2/4 + \sigma^2)\mathrm{P}_{\theta,\sigma}(|Z| > \lambda_l/2) + \theta^2\mathrm{P}_{\theta,\sigma}(|Z| < \lambda_l/2), \\
R(\theta, \widehat{\theta}_{\text{post-lasso}}) &= (\lambda_l/2 - \theta)\phi_{\theta,\sigma}(\lambda_l/2)\sigma^2 + (\lambda_l/2 + \theta)\phi_{\theta,\sigma}(-\lambda_l/2)\sigma^2 \\
&\quad + \sigma^2\mathrm{P}_{\theta,\sigma}(|Z| > \lambda_r/2) + \theta^2\mathrm{P}_{\theta,\sigma}(|Z| < \lambda_r/2), \\
R(\theta, \widehat{\theta}_{\text{ridge}}) &= \theta^2\widetilde{k}^2 + (1-\widetilde{k})^2\sigma^2, \quad \widetilde{k} = \lambda_r/(1+\lambda_r), \\
R(\theta, \widehat{\theta}_{\text{elastic-net}}) &= \sigma^2(h^2\lambda_1/2 + h^2\theta + 2dh)\phi_{\theta,\sigma}(\lambda_1/2) \\
&\quad - \sigma^2(-h^2\lambda_1/2 + h^2\theta + 2gh)\phi_{\theta,\sigma}(-\lambda_1/2) + \theta^2\mathrm{P}_{\theta,\sigma}(|Z| < \lambda_1/2) \\
&\quad + ((h\theta + d)^2 + h^2\sigma^2)\mathrm{P}_{\theta,\sigma}(Z > \lambda_1/2) + ((h\theta + g)^2 + h^2\sigma^2)\mathrm{P}_{\theta,\sigma}(Z < -\lambda_1/2).
\end{aligned}$$

These results for the one-dimensional case are derived through an application of a simple Stein's integration-by-parts trick for Gaussian random variables and provide a key building block for results in the multidimensional case. In particular, we build from these results to show that the lava estimator performs favorably relative to the maximum likelihood estimator, the ridge estimator, and ℓ_1-based estimators (such as lasso and elastic-net) in interesting multidimensional settings. We provide more detailed discussions on the derived risk formulas in the multidimensional case in Section 2.5.

2.3. *Multidimensional case.* We consider now the canonical Gaussian model or the Gaussian sequence model. In this case, we have that

$$Z \sim N_p(\theta, \sigma^2 I_p)$$

is a single observation from a multivariate normal distribution where $\theta = (\theta_1, ..., \theta_p)'$ is a p-dimensional vector. A fundamental result for this model is that the maximum likelihood estimator Z is inadmissible and can be dominated by the ridge estimator and related shrinkage procedures when $p \geq 3$ (e.g. [33]).

In this model, the lava estimator is given by

$$\widehat{\theta}_{\text{lava}} := (\widehat{\theta}_{\text{lava},1}, ..., \widehat{\theta}_{\text{lava},p})' := (d_{\text{lava}}(Z_1), ..., d_{\text{lava}}(Z_p))',$$

where $d_{\text{lava}}(z)$ is the lava shrinkage function as in (10). The estimator is designed to capture the case where

$$\theta = \underbrace{\beta}_{\text{dense part}} + \underbrace{\delta}_{\text{sparse part}}$$

is formed by combining a sparse vector δ that has a relatively small number of non-zero entries which are all large in magnitude and a dense vector β that may contain very many small non-zero entries. This model for θ is "sparse+dense." It includes cases that are not well-approximated by "sparse" models - models in which a very small number of parameters are large and the rest are zero - or by "dense" models - models in which very many coefficients are non-zero but all coefficients are of similar magnitude. This structure thus includes cases that pose challenges for estimators such as the lasso and elastic net that are designed for sparse models and for estimators such as ridge that are designed for dense models.

Remark 2.1. The regression model with Gaussian noise and an orthonormal design is a special case of the multidimensional canonical model. Consider

$$Y = X\theta + U, \quad U \mid X \sim N(0, \sigma_u^2 I_n),$$

where Y and U are $n \times 1$ random vectors and X is an $n \times p$ random or fixed matrix, with n and p respectively denoting the sample size and the dimension of θ. Suppose $\frac{1}{n}X'X = I_p$ a.s.. with $p \leq n$. Then we have the canonical multidimensional model:

$$Z = \theta + \epsilon, \quad Z = \frac{1}{n}X'Y, \quad \epsilon = \frac{1}{n}X'U \sim N(0, \sigma^2 I_p), \quad \sigma^2 = \frac{\sigma_u^2}{n}. \quad \blacksquare$$

All of the shrinkage estimators discussed in Section 2.1 generalize to the multidimensional case. Let $z \mapsto d_e(z)$ be the shrinkage function associated with estimator e in the one dimensional setting where e can take values in the set

$$\mathcal{E} = \{\text{lava}, \text{post-lava}, \text{ridge}, \text{lasso}, \text{post-lasso}, \text{elastic net}\}.$$

We then have a similar estimator in the multidimensional case given by

$$\widehat{\theta}_e := (\widehat{\theta}_{e,1}, ..., \widehat{\theta}_{e,p})' := (d_e(Z_1), ..., d_e(Z_p))'.$$

The following corollary is a trivial consequence of Theorem 2.1 and additivity of the risk function.

Corollary 2.1 (Risk Function of Lava and Related Estimators in the Multi-Dimensional Case). *If $Z \sim N(0, \sigma^2 I_p)$, then for any $e \in \mathcal{E}$ we have that risk function $\mathrm{R}(\theta, \widehat{\theta}_e) := \mathrm{E}\|\theta - \widehat{\theta}_e\|_2^2$ is given by $\mathrm{R}(\theta, \widehat{\theta}_e) = \sum_{j=1}^p R(\theta_j, \widehat{\theta}_{e,j})$, where $R(\cdot,\cdot)$ is the unidimensional risk function characterized in Theorem 2.1.*

These multivariate risk functions are shown in Figure 1 in a prototypical "sparse+dense" model generated according to the model discussed in detail in Section 2.5. The tuning parameters used in this figure are the best possible (risk minimizing or oracle) choices of the penalty levels found by minimizing the risk expression $\mathrm{R}(\theta, \widehat{\theta}_e)$ for each estimator. As guaranteed by the construction of the estimator, the figure illustrates that lava performs no worse than, and often substantially outperforms, ridge and lasso with optimal penalty parameter choices. We also see that lava uniformly outperforms the elastic net.

2.4. *Canonical plug-in choice of penalty levels.* We now discuss simple, rule-of-thumb choices for the penalty levels for lasso (λ_l), ridge (λ_r) and lava (λ_1, λ_2). In the Gaussian model, a canonical choice of λ_l is $\lambda_l = 2\sigma\Phi^{-1}(1 - c/(2p))$, which satisfies

$$\mathrm{P}\left(\max_{j \le p} |Z_j - \theta_j| \le \lambda_l/2\right) \ge 1 - c;$$

see, e.g., [15]. Here $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, and c is a pre-determined significance level which is often set to 0.05. The risk function for ridge is simple, and an analytic solution to the risk minimizing choice of ridge tuning parameter is given by $\lambda_r = \sigma^2(p/\|\theta\|_2^2)$.

As for the tuning parameters for lava, recall that the lava estimator in the Gaussian model is

$$\begin{aligned} \widehat{\theta}_{\text{lava}} &= (\widehat{\theta}_{\text{lava},1}, ..., \widehat{\theta}_{\text{lava},p})', \quad \widehat{\theta}_{\text{lava},j} = \widehat{\beta}_j + \widehat{\delta}_j, \quad j = 1, ..., p, \\ (\widehat{\beta}_j, \widehat{\delta}_j) &= \arg \min_{(\beta_j, \delta_j) \in \mathbb{R}^2} (Z_j - \beta_j - \delta_j)^2 + \lambda_2 |\beta_j|^2 + \lambda_1 |\delta_j|. \end{aligned}$$

If the dense component β were known, then following [15] would suggest setting

$$\lambda_1 = 2\sigma\Phi^{-1}(1 - c/(2p))$$

as a canonical choice of λ_1 for estimating δ. If the sparse component δ were known, one could adopt

$$\lambda_2 = \sigma^2(p/\|\beta\|_2^2)$$

as a choice of λ_2 for estimating β following the logic for the standard ridge estimator.

We refer to these choices as the "canonical plug-in" tuning parameters and use them in constructing the risk comparisons in the following subsection (to complement comparisons given in Figure 1 under oracle tuning parameters). We note that the lasso choice is motivated by a sparse model and does not naturally adapt to or make use of the true structure of θ. The ridge penalty choice is explicitly tied to risk minimization and relies on using knowledge of the true θ. The lava choices for the parameters on the ℓ_1 and ℓ_2 penalties are, as noted immediately above, motivated by the respective choices in lasso and ridge. As such, the motivations and feasibility of these canonical choices are not identical across methods, and the risk comparisons in the following subsection should be interpreted within this light.

2.5. *Some risk comparisons in a canonical Gaussian model with canonical tuning.* To compare the risk functions of lava, lasso, and ridge estimators, we consider a canonical Gaussian model, where

$$\theta_1 = 3, \quad \theta_j = q, \quad j = 2, ..., p,$$

for some index $q \geq 0$. We set the noise level to be $\sigma^2 = 0.1^2$. The parameter θ can be decomposed as $\theta = \beta + \delta$, where the sparse component is $\delta = (3, 0, ..., 0)'$, and the dense component is

$$\beta = (0, q, ..., q)',$$

where q is the "size of small coefficients." The canonical tuning parameters are $\lambda_l = \lambda_1 = 2\sigma\Phi^{-1}(1 - c/(2p))$, $\lambda_r = \sigma^2 p/(3 + q^2(p-1))$ and $\lambda_2 = \sigma^2 p/(q^2(p-1))$.

Figure 1 (given in the introduction) compares risks of lava, lasso, ridge, elastic net, and the maximum likelihood estimators as functions of the size of the small coefficients q, using the ideal (risk minimizing or oracle choices) of the penalty levels. Figure 3 compares risks of lava, lasso, ridge and the maximum likelihood estimators using the "canonical plug-in" penalty levels discussed above. Theoretical risks are plotted as a function of the size of the small coefficients q. We see from these figures that regardless of how we choose the penalty levels – ideally or via the plug-in rules – lava strictly dominates the competing methods in this "sparse+dense" model. Compared to lasso, the proposed lava estimator does about as well as lasso when the signal is sparse and does significantly better than lasso when the signal is non-sparse. Compared to ridge, the lava estimator does about as well as ridge when the signal is dense and does significantly better than ridge when the signal is sparse.

In Section 4, we further explore the use of data-driven choices of penalty levels via cross-validation and SURE minimization; see Figures 4, 5, and 6. We do so in the context of the Gaussian regression model with fixed and random regressors. With either cross-validation or SURE minmization, the ranking of the estimators remains unchanged, with lava consistently dominating lasso, ridge, and the elastic net.

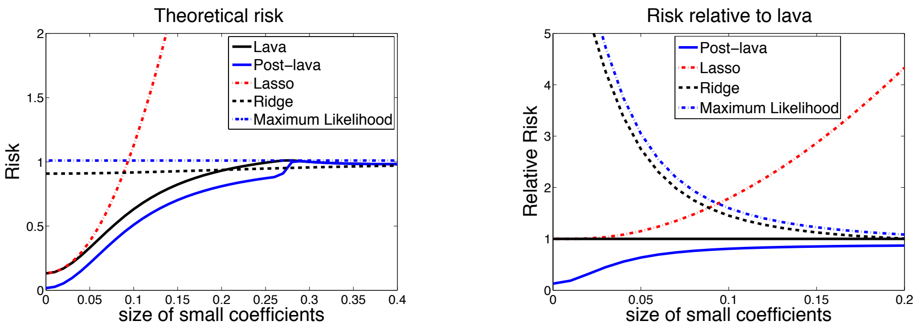

FIG 3. *Exact risk functions of lava, post-lava, ridge, lasso, and maximum likelihood in the Gaussian sequence model with "sparse+dense" signal structure and* $p = 100$ *using the canonical "plug-in" choices of penalty levels given in the text with* $c = .05$*. See Section 2.5 for the description of penalty levels and the model. The size of "small coefficients",* q*, is shown on the horizontal axis. The size of these coefficients directly corresponds to the size of the "dense part" of the signal, with zero corresponding to the exactly sparse case. Relative risk plots the ratio of the risk of each estimator to the lava risk,* $\mathrm{R}(\theta, \widehat{\theta}_e)/\mathrm{R}(\theta, \widehat{\theta}_{\text{lava}})$*. Note that the relative risk plot is over a smaller set of sizes to accentuate comparisons over the region where there are the most interesting differences between the estimators.*

[33] proved that a ridge estimator strictly dominates maximum likelihood in the Gaussian sequence model once $p \geq 3$. In the comparisons above, we also see that the lava estimator strictly dominates the maximum likelihood estimator; and one wonders whether this domination has a theoretical underpinning similar to Stein's result for ridge. The following result provides some (partial) support for this phenomenon for the lava estimator with the plug-in penalty levels. The result shows that, for a sufficiently large p, lava does indeed uniformly dominate the maximum likelihood estimator on the set $\{\theta = \beta + \delta : \|\beta\|_\infty < M, s = \sum_{j=1}^p 1\{\delta_j \neq 0\} \ll p/\log p\}$.

LEMMA 2.2 (Relative Risk of Lava vs. Maximum Likelihood). *Let* $Z \sim N_p(\theta, \sigma^2 I_p)$*, and* (λ_1, λ_2) *be chosen with the plug-in rule given in Section 2.4. Fix a constant* $M > 0$*, and define* $f(M, \sigma) := 15(1 + M^2/\sigma^2)^2 + 9/\sqrt{2\pi\sigma^2}$*. In addition, suppose* $\sigma\sqrt{\log p} > 5M + 33\sigma + 20\sigma^2$*, and* $2p > \pi c^2 \log p \geq 1$*. Then uniformly for* θ *inside the following set:*

$$A(M, s) := \{\theta = \beta + \delta : \|\beta\|_\infty < M, s = \sum_{j=1}^p 1\{\delta_j \neq 0\}\},$$

we have

$$RR := \frac{\mathrm{E}\|\widehat{\theta}_{\text{lava}}(Z) - \theta\|_2^2}{\mathrm{E}\|Z - \theta\|_2^2} \leq \frac{\|\beta\|_2^2}{\sigma^2 p + \|\beta\|_2^2} + \frac{4}{\sqrt{2\pi} p^{1/16}} + 2f(M, \sigma)\frac{s \log p}{p}.$$

REMARK 2.2. Note that this lemma allows unbounded sparse components (i.e., elements of δ are unrestricted), and only requires the dense components to be uniformly bounded $\|\beta\|_\infty < M$, which seems reasonable. We note that

$$R_d^2 = \frac{\|\beta\|_2^2}{\sigma^2 p + \|\beta\|_2^2}$$

measures the proportion of the total variation of $Z - \delta$ around 0 that is explained by the dense part of the signal. If R_d^2 is bounded away from 1 and M and $\sigma^2 > 0$ are fixed, then the risk of lava becomes uniformly smaller than the risk of the maximum likelihood estimator on a compact parameter space as $p \to \infty$ and $s \log p/p \to 0$. Indeed, if R_d^2 is bounded away from 1 and M is fixed,

$$\frac{4}{\sqrt{2\pi} p^{1/16}} + 2f(M,\sigma)\frac{s\log p}{p} \to 0 \implies RR = R_d^2 + o(1) < 1.$$

Moreover, we have $RR \to 0$ if $R_d^2 \to 0$ – namely, in the case that dense component plays a progressively smaller role in explaining the variation of $Z - \delta$, the lava estimator becomes infinitely more asymptotically efficient than the maximum likelihood estimator in terms of relative risk. ∎

Remark 2.3 (Risks of lava and lasso in the sparse case). We mention some implications of Corollary 2.1 in the sparse case. When θ is sparse, we have $\beta = 0$, and hence the canonical choice for the tuning parameter becomes $\lambda_2 = \infty$, and $k = 1$ and $w = \lambda_1/2$. It then immediately follows from Theorem 2.1 and Corollary 2.1 that $\mathrm{R}(\theta, \widehat{\theta}_{\text{lava}}) = \mathrm{R}(\theta, \widehat{\theta}_{\text{lasso}})$. Therefore, in the sparse model, the risk functions of lava and lasso are the same with canonical choices of the tuning parameters. It is useful to analyze the risk function in this case. Applying the same arguments as those in the proof of Lemma 2.2, it can be shown that $R(\theta_j, \widehat{\theta}_{\text{lasso},j})$ in Theorem 2.1 is given by: for some $a > 0$,

$$R(\theta_j, \widehat{\theta}_{\text{lasso},j}) \le \begin{cases} (\frac{\lambda_l^2}{4} + \sigma^2)\mathrm{P}_{\theta_j,\sigma}(|Z_j| > \frac{\lambda_l}{2}) = (\frac{\lambda_l^2}{4} + \sigma^2)\frac{c}{p}, & \theta_j = 0, \\ (\frac{\lambda_l^2}{4} + \sigma^2)\mathrm{P}_{\theta_j,\sigma}(|Z_j| > \frac{\lambda_l}{2}) + \theta_j^2 \mathrm{P}_{\theta_j,\sigma}(|Z_j| < \frac{\lambda_l}{2}) + O(\sigma^2 p^{-a}) & \theta_j \neq 0, \end{cases}$$

where the $O(\cdot)$ is uniform in $j = 1, ..., p$. Hence for $s = \sum_{j=1}^p 1\{\theta_j \neq 0\}$,

$$\mathrm{R}(\theta, \widehat{\theta}_{\text{lava}}) = \mathrm{R}(\theta, \widehat{\theta}_{\text{lasso}}) \lesssim \lambda_l^2\{s + (1 - s/p)\} \lesssim 2^2\sigma^2 (\Phi^{-1}(1 - c/(2p)))^2 \{s + (1 - s/p)\},$$

which is a canonical bound for lasso risk in sparse models. ∎

2.6. *Stein's unbiased risk estimation for lava.* [34] proposed a useful risk estimate based on the integration by parts formula, now commonly referred to as *Stein's unbiased risk estimate* (SURE). This subsection derives SURE for the lava shrinkage in the multivariate Gaussian model. Note that

$$\mathrm{E}\|\widehat{\theta}_{\text{lava}} - \theta\|_2^2 = -p\sigma^2 + \mathrm{E}\|Z - \widehat{\theta}_{\text{lava}}\|_2^2 + 2\mathrm{E}[(Z - \theta)'\widehat{\theta}_{\text{lava}}]. \tag{12}$$

Stein's formula is essential to calculating the key term $\mathrm{E}[(Z - \theta)'\widehat{\theta}_{\text{lava}}]$.

THEOREM 2.2 (SURE for lava). *Suppose* $Z = (Z_1, ..., Z_p)' \sim N_p(\theta, \sigma^2 I_p)$. *Then*

$$\mathrm{E}[(Z-\theta)'\widehat{\theta}_{\text{lava}}] = p(1-k)\sigma^2 + k\sigma^2 \sum_{j=1}^{p} \mathrm{P}_{\theta_j,\sigma}(|Z_j| > \lambda_1/(2k)).$$

In addition, let $\{Z_{ij}\}_{i=1}^n$ *be identically distributed as* Z_j *for each* j. *Then*

$$\widehat{\mathrm{R}}(\theta, \widehat{\theta}_{\text{lava}}) = (1-2k)p\sigma^2 + \frac{1}{n}\sum_{i=1}^{n} \|Z_i - d_{\text{lava}}(Z_i)\|_2^2 + 2k\sigma^2 \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p} 1\{|Z_{ij}| > \lambda_1/(2k)\}$$

is an unbiased estimator of $\mathrm{R}(\theta, \widehat{\theta}_{\text{lava}})$.

3. Lava in the Regression Model.

3.1. *Definition of Lava in the Regression Model.* Consider a fixed design regression model:

$$Y = X\theta_0 + U, \quad U \sim N(0, \sigma_u^2 I_n),$$

where $Y = (y_1, ..., y_n)'$, $X = (X_1, ..., X_n)'$, and θ_0 is the true regression coefficient. Following the previous discussion, we assume that $\theta_0 = \beta_0 + \delta_0$ is "sparse+dense" with sparse component δ_0 and dense component β_0. Again, this coefficient structure includes cases which cannot be well-approximated by traditional sparse models or traditional dense models and will pose challenges for estimation strategies tailored to sparse settings, such as lasso and similar methods, or strategies tailored to dense settings, such as ridge.

In order to define the estimator we shall rely on the normalization condition that

$$n^{-1}[X'X]_{jj} = 1, \quad j = 1, ..., p. \tag{13}$$

Note that without this normalization, the penalty terms below would have to be modified in order to insure equivariance of the estimator to changes of scale in the columns of X.

The lava estimator $\widehat{\theta}_{\text{lava}}$ of θ_0 solves the following optimization problem:

$$\begin{aligned} \widehat{\theta}_{\text{lava}} &:= \widehat{\beta} + \widehat{\delta}, \\ (\widehat{\beta}, \widehat{\delta}) &:= \arg \min_{(\beta',\delta')' \in \mathbb{R}^{2p}} \left\{ \frac{1}{n}\|Y - X(\beta+\delta)\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\delta\|_1 \right\}. \end{aligned} \tag{14}$$

The lava program splits parameter θ into the sum of β and δ and penalizes these two parts using the ℓ_2 and ℓ_1 penalties. Thus, the ℓ_1- penalization regularizes the estimator of the sparse part δ_0 of θ_0 and produces a sparse solution $\widehat{\delta}$. The ℓ_2-penalization regularizes the estimator of the dense part β_0 of θ_0 and produces a dense solution $\widehat{\beta}$. The resulting estimator of θ_0 is the sum of the sparse estimator $\widehat{\delta}$ and the dense estimator $\widehat{\beta}$.

3.2. *A Key Profile Characterization and Some Insights.* The lava estimator can be computed in the following way. For a fixed δ, define

$$\widehat{\beta}(\delta) = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - X(\beta + \delta)\|_2^2 + \lambda_2 \|\beta\|_2^2 \right\}.$$

This ridge program has the solution

$$\widehat{\beta}(\delta) = (X'X + n\lambda_2 I_p)^{-1} X'(Y - X\delta).$$

By substituting $\beta = \widehat{\beta}(\delta)$ into the objective function, we then define an ℓ_1-penalized quadratic program which we can solve for $\widehat{\delta}$:

$$\widehat{\delta} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - X(\widehat{\beta}(\delta) + \delta)\|_2^2 + \lambda_2 \|\widehat{\beta}(\delta)\|_2^2 + \lambda_1 \|\delta\|_1 \right\}. \tag{15}$$

The lava solution is then given by $\widehat{\theta} = \widehat{\beta}(\widehat{\delta}) + \widehat{\delta}$. The following result provides a useful and somewhat unexpected characterization of the profiled lava program.

THEOREM 3.1 (A Key Characterization of the Profiled Lava Program). *Define ridge-projection matrices,*

$$\mathsf{P}_{\lambda_2} = X(X'X + n\lambda_2 I_p)^{-1} X' \text{ and } \mathsf{K}_{\lambda_2} = I_n - \mathsf{P}_{\lambda_2},$$

and transformed data, $\widetilde{Y} = \mathsf{K}_{\lambda_2}^{1/2} Y$ *and* $\widetilde{X} = \mathsf{K}_{\lambda_2}^{1/2} X$. *Then*

$$\widehat{\delta} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|\widetilde{Y} - \widetilde{X}\delta\|_2^2 + \lambda_1 \|\delta\|_1 \right\}, \quad X\widehat{\theta}_{\text{lava}} = \mathsf{P}_{\lambda_2} Y + \mathsf{K}_{\lambda_2} X\widehat{\delta}. \tag{16}$$

The theorem shows that solving for the sparse part $\widehat{\delta}$ of the lava estimator is equivalent to solving a standard lasso problem using transformed data $\widetilde{Y}$ and $\tilde{X}$. This computational characterization is key to both computation and our theoretical analysis of the estimator.

REMARK 3.1 (Insights derived from Theorem 3.1). Suppose δ_0 were known. Let $W = Y - X\delta_0$ be the response vector after removing the sparse signal, and note that we equivalently have $W = X\beta_0 + U$. A natural estimator for β_0 in this setting is then the ridge estimator of W on X: $\widehat{\beta}(\delta_0) = (X'X + n\lambda_2 I_p)^{-1} X'W$. Denote the prediction error based on this ridge estimator as

$$\mathsf{D}_{\text{ridge}}(\lambda_2) = X\widehat{\beta}(\delta_0) - X\beta_0 = -\mathsf{K}_{\lambda_2} X\beta_0 + \mathsf{P}_{\lambda_2} U.$$

Under mild conditions on β_0 and the design matrix, [25] showed that $\frac{1}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|^2 = o_P(1)$.[3] Using Theorem 3.1, the prediction error of lava can be written as

$$X\widehat{\theta}_{\text{lava}} - X\theta_0 = \mathsf{P}_{\lambda_2} Y + \mathsf{K}_{\lambda_2} X\widehat{\delta} - X\beta_0 - X\delta_0 = \mathsf{D}_{\text{ridge}}(\lambda_2) + \mathsf{K}_{\lambda_2} X(\widehat{\delta} - \delta_0). \tag{17}$$

[3]We show in the proof of Theorem 3.3 that $\forall \epsilon > 0$, with probability at least $1 - \epsilon$, $\frac{1}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|^2 \leq B_3 + B_4(\beta_0)$, where B_3 and $B_4(\beta_0)$ are respectively defined in Theorem 3.3 below. Hence a sufficient condition for $\frac{1}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|^2 = o_P(1)$ is that λ_2 and β_0 are such that $B_3 + B_4(\beta_0) = o(1)$.

Hence, lava has vanishing prediction error as long as

$$\frac{1}{n}\| \mathsf{K}_{\lambda_2} X(\widehat{\delta}-\delta_0)\|_2^2 = o_P(1). \tag{18}$$

Condition (18) is related to the performance of the lasso in the transformed problem (16). Examination of (16) shows that it corresponds to a sparse regression model with approximation errors $\mathsf{K}_{\lambda_2}^{1/2} X\beta_0$, akin to those considered in [2, 5]: For $\widetilde{U} = \mathsf{K}_{\lambda_2}^{1/2} U$, decompose $\widetilde{Y} = \widetilde{X}\delta_0 + \widetilde{U} + \mathsf{K}_{\lambda_2}^{1/2} X\beta_0$. Under conditions such as those given in [25], the approximation error obeys

$$\frac{1}{n}\| \mathsf{K}_{\lambda_2}^{1/2} X\beta_0\|_2^2 = o_P(1). \tag{19}$$

It is known that the lasso estimator performs well in sparse models with vanishing approximation errors. The lasso estimator attains rates of convergence in the prediction norm that are the sum of the usual rate of convergence in the case without approximation errors and the rate at which the approximation error vanishes; see, e.g., [2]. Thus, we anticipate that (18) will hold.

To help understand the plausibility of condition (19), consider an orthogonal design where $\frac{1}{n}X'X = I_p$. In this case, it is straightforward to verify that $\mathsf{K}_{\lambda_2}^{1/2} = \mathsf{K}_{\lambda_2^*}$ where $\lambda_2^* = \sqrt{\lambda_2}/(\sqrt{1+\lambda_2}-\sqrt{\lambda_2})$. Hence, $\widetilde{X}\beta_0 = \mathsf{K}_{\lambda_2^*} X\beta_0$ is a component of the prediction bias $\mathsf{D}_{\text{ridge}}(\lambda_2^*) = -\mathsf{K}_{\lambda_2^*} X\beta_0 + \mathsf{P}_{\lambda_2^*} U$ from a ridge estimator with tuning parameter λ_2^* and vanishes under some regularity conditions [25]. We present the formal analysis for the general case in Section 3.5. ■

3.3. *Degrees of Freedom and SURE.* Degrees of freedom is often used to quantify model complexity and to construct adaptive model selection criteria for selecting tuning parameters. In a Gaussian linear regression model $Y \sim N(X\theta_0, \sigma_u^2 I_n)$ with a fixed design, [17] define the degrees of freedom of the mean fit $X\widehat{\theta}$ to be

$$\mathrm{df}(\widehat{\theta}) = \frac{1}{\sigma_u^2}\mathrm{E}[(Y-X\theta_0)'X\widehat{\theta}];$$

and this quantity is also an important component of the mean squared prediction risk:

$$\mathrm{E}\frac{1}{n}\|X\widehat{\theta}-X\theta_0\|_2^2 = -\sigma_u^2 + \mathrm{E}\frac{1}{n}\|X\widehat{\theta}-Y\|_2^2 + \frac{2\sigma_u^2}{n}\mathrm{df}(\widehat{\theta}).$$

Stein ([34])'s SURE theory provides a tractable way of deriving an unbiased estimator of the degrees of freedom, and thus the mean squared prediction risk. Specifically, write $\widehat{\theta} = d(Y,X)$ as a function of Y, conditional on X. Suppose $d(\cdot,X):\mathbb{R}^n \to \mathbb{R}^p$ is almost differentiable; see [31] and [18]. For $f:\mathbb{R}^n\to\mathbb{R}^n$ differentiable at y, define

$$\partial_y f(y) := \Big[\partial f_{ij}(y)\Big], \quad (i,j)\in\{1,...,n\}^2, \quad \partial f_{ij}(y) := \frac{\partial}{\partial y_j} f_i(y),$$

$$\nabla_y \cdot f(y) := \mathrm{tr}(\partial_y f(y)).$$

Let X_i' denote the i-th row of X, $i = 1, ..., n$. Then, from [34], we have that

$$\frac{1}{\sigma_u^2}\mathrm{E}[(Y - X\theta_0)' X d(Y, X)] = \mathrm{E}[\nabla_y \cdot (X d(Y, X))] = \mathrm{E}\mathrm{tr}\big(\partial_y [X d(Y, X)]\big).$$

An unbiased estimator of the term on the right-hand-side of the display may then be constructed using its sample analog.

In this subsection, we derive the degrees of freedom of the lava, and thus a SURE of its mean squared prediction risk. By Theorem 3.1,

$$\nabla_y \cdot (X d_{\mathrm{lava}}(y, X)) = \mathrm{tr}(\mathsf{P}_{\lambda_2}) + \nabla_y \cdot (\mathsf{K}_{\lambda_2} X d_{\mathrm{lasso}}(\mathsf{K}_{\lambda_2}^{1/2} y, \mathsf{K}_{\lambda_2}^{1/2} X)) \tag{20}$$

$$= \mathrm{tr}(\mathsf{P}_{\lambda_2}) + \mathrm{tr}\left(\mathsf{K}_{\lambda_2}\, \partial_y [X d_{\mathrm{lasso}}(\mathsf{K}_{\lambda_2}^{1/2} y, \widetilde{X})]\right), \tag{21}$$

where $d_{\mathrm{lava}}(y, X)$ is the lava estimator on the data (y, X) and $d_{\mathrm{lasso}}(\mathsf{K}_{\lambda_2}^{1/2} y, \mathsf{K}_{\lambda_2}^{1/2} X))$ is the lasso estimator on the data $(\mathsf{K}_{\lambda_2}^{1/2} y, \mathsf{K}_{\lambda_2}^{1/2} X)$ with the penalty level λ_1. The almost differentiability of the map $y \mapsto d_{\mathrm{lasso}}(\mathsf{K}_{\lambda_2}^{1/2} y, \mathsf{K}_{\lambda_2}^{1/2} X)$ follows from the almost differentiability of the map $u \mapsto d_{\mathrm{lasso}}(u, \mathsf{K}_{\lambda_2}^{1/2} X)$, which holds by the results in [16] and [38].

The following theorem presents the degrees of freedom and SURE for lava. Let $\widehat{J} = \{j \le p : \widehat{\delta}_j \ne 0\}$ be the active set of the sparse component estimator $\hat{\delta}$ with cardinality denoted by $|\widehat{J}|$. Recall that $\widetilde{X} = \mathsf{K}_{\lambda_2}^{1/2} X$. Let $\widetilde{X}_{\hat{J}}$ be an $n \times |\widehat{J}|$ submatrix of $\widetilde{X}$ whose columns are those corresponding to the entries in $\widehat{J}$. Let A^- denote the Moore-Penrose pseudo-inverse of a square matrix A.

THEOREM 3.2 (SURE for Lava in Regression). *Suppose $Y \sim N(X\theta_0, \sigma_u^2 I_n)$. Let*

$$\widetilde{\mathsf{K}}_{\hat{J}} = I - \widetilde{X}_{\hat{J}}(\widetilde{X}_{\hat{J}}'\widetilde{X}_{\hat{J}})^- \widetilde{X}_{\hat{J}}'$$

be the projection matrix onto the unselected columns of $\tilde{X}$. We have that

$$\mathrm{df}(\widehat{\theta}_{\mathrm{lava}}) = \mathrm{E}[\mathrm{rank}(\widetilde{X}_{\hat{J}}) + \mathrm{tr}(\widetilde{\mathsf{K}}_{\hat{J}}\, \mathsf{P}_{\lambda_2})].$$

Therefore, the SURE of $\mathrm{E}\frac{1}{n}\|X\widehat{\theta}_{\mathrm{lava}} - X\theta_0\|_2^2$ is given by

$$-\sigma_u^2 + \frac{1}{n}\|X\widehat{\theta}_{\mathrm{lava}} - Y\|_2^2 + \frac{2\sigma_u^2}{n}\mathrm{rank}(\widetilde{X}_{\hat{J}}) + \frac{2\sigma_u^2}{n}\mathrm{tr}(\widetilde{\mathsf{K}}_{\hat{J}}\, \mathsf{P}_{\lambda_2}).$$

3.4. *Post-lava in regression.* We can also remove the shrinkage bias in the sparse component introduced by the ℓ_1-penalization via a *post-selection* procedure. Specifically, let $(\widehat{\beta}, \widehat{\delta})$ respectively denote the lava estimator of the dense and sparse components. Define the *post-lava* estimator as follows:

$$\widehat{\theta}_{\text{post-lava}} = \widehat{\beta} + \widetilde{\delta}, \quad \widetilde{\delta} = \arg\min_{\delta \in \mathbb{R}^p}\left\{\frac{1}{n}\|Y - X\widehat{\beta} - X\delta\|_2^2 : \quad \delta_j = 0 \text{ if } \widehat{\delta}_j = 0\right\}.$$

Let $X_{\hat{J}}$ be an $n \times |\widehat{J}|$ submatrix of X whose columns are selected by $\widehat{J}$. Then we can partition $\widetilde{\delta} = (\widetilde{\delta}_{\hat{J}}, 0)'$, where $\widetilde{\delta}_{\hat{J}} = (X_{\hat{J}}' X_{\hat{J}})^{-} X_{\hat{J}}'(Y - X\widehat{\beta})$. Write $\mathsf{P}_{\hat{J}} = X_{\hat{J}}(X_{\hat{J}}' X_{\hat{J}})^{-} X_{\hat{J}}'$ and $\mathsf{K}_{\hat{J}} = I_n - \mathsf{P}_{\hat{J}}$. The post-lava prediction for $X\theta$ is:

$$X\widehat{\theta}_{\text{post-lava}} = \mathsf{P}_{\hat{J}} Y + \mathsf{K}_{\hat{J}} X\widehat{\beta}.$$

In addition, note that the lava estimator satisfies $X\widehat{\beta} = \mathsf{P}_{\lambda_2}(Y - X\widehat{\delta})$. We then have the following expression of $X\widehat{\theta}_{\text{post-lava}}$.

LEMMA 3.1. *Let $\widehat{U} := Y - X\widehat{\theta}_{\text{lava}}$. Then $X\widehat{\theta}_{\text{post-lava}} = X\widehat{\theta}_{\text{lava}} + \mathsf{P}_{\hat{J}} \widehat{U}$.*

The above lemma reveals that the post-lava corrects the ℓ_1-shrinkage bias of the original lava fit by adding the projection of the lava residual onto the subspace of the selected regressors. This correction is in the same spirit as the post-lasso correction for shrinkage bias in the standard lasso problem; see [2].

REMARK 3.2. We note that SURE for post-lava may not exist, though an estimate of the upper bound of the risk function may be available, because of the impossibility results for constructing unbiased estimators for discontinuous functionals ([23]). ■

3.5. *Deviation Bounds for Prediction Errors.* In the following, we develop deviation bounds for the lava prediction error: $\frac{1}{n}\|X\widehat{\theta}_{\text{lava}} - X\theta_0\|_2^2$. We continue to work with the decomposition $\theta_0 = \beta_0 + \delta_0$ and will show that lava performs well in terms of rates on the prediction error in this setting. According to the discussion in Section 3.2, there are three sources of prediction error: (i) $\mathsf{D}_{\text{ridge}}(\lambda_2)$, (ii) $\widetilde{X}\beta_0$ and (iii) $\mathsf{K}_{\lambda_2} X(\widehat{\delta} - \delta_0)$. The behavior of the first two terms is determined by the behavior of the ridge estimator of the dense component β_0, and the behavior of the third term is determined by the behavior of the lasso estimator on the transformed data.

We assume that $U \sim N(0, \sigma_u^2 I_n)$ and that X is fixed. As in the lasso analysis of [5], a key quantity is the maximal norm of the score:

$$\Lambda = \left\| \frac{2}{n} \widetilde{X}' \widetilde{U} \right\|_\infty = \left\| \frac{2}{n} X' \mathsf{K}_{\lambda_2} U \right\|_\infty.$$

Following [2], we set the penalty level for the lasso part of lava in our theoretical development as

$$\lambda_1 = c\Lambda_{1-\alpha} \text{ with } \Lambda_{1-\alpha} = \inf\{l \in \mathbb{R} : \mathrm{P}(\Lambda \le l) \ge 1 - \alpha\} \tag{22}$$

and $c > 1$ a constant. Note that [2] suggest setting $c = 1.1$ and that $\Lambda_{1-\alpha}$ is easy to approximate by simulation.

Let $S := X'X/n$ and $\bar{V}_{\lambda_2}$ be the maximum diagonal element of

$$V_{\lambda_2} := (S + \lambda_2 I_p)^{-1} S (S + \lambda_2 I_p)^{-1} \lambda_2^2.$$

Then by the union bound and Mill's inequality:

$$\Lambda_{1-\alpha} < \bar{\Lambda}_{1-\alpha} := 2\sigma_u \sqrt{\frac{\bar{V}_{\lambda_2}\log(2p/\alpha)}{n}}. \tag{23}$$

Thus the choice $\Lambda_{1-\alpha}$ is strictly sharper than the union bound-based, classical choice $\bar{\Lambda}_{1-\alpha}$. Indeed, $\Lambda_{1-\alpha}$ is strictly smaller than $\bar{\Lambda}_{1-\alpha}$ even in orthogonal design cases since union bounds are not sharp. In collinear or highly-correlated designs, it is easy to give examples where $\Lambda_{1-\alpha} = o(\bar{\Lambda}_{1-\alpha})$; see [4]. Thus, the gains from using the more refined choice can be substantial.

We define the following design impact factor: For $\widetilde{X} = \mathsf{K}_{\lambda_2}^{1/2} X$,

$$\iota(c,\delta_0,\lambda_1,\lambda_2) := \inf_{\Delta \in \mathcal{R}(c,\delta_0,\lambda_1,\lambda_2)} \frac{\|\widetilde{X}\Delta\|_2/\sqrt{n}}{\|\delta_0\|_1 - \|\delta_0+\Delta\|_1 + c^{-1}\|\Delta\|_1},$$

where $\mathcal{R}(c,\delta_0,\lambda_1,\lambda_2) = \{\Delta \in \mathbb{R}^p \setminus \{0\} : \|\widetilde{X}\Delta\|_2^2/n \le 2\lambda_1(\|\delta_0\|_1 - \|\delta_0+\Delta\|_1 + c^{-1}\|\Delta\|_1)\}$ is the restricted set, and where $\iota(c,\delta_0,\lambda_1,\lambda_2) := \infty$ if $\delta_0 = 0$.

The design impact factor generalizes the restricted eigenvalues of [5] and and is tailored for bounding estimation errors in the prediction norm (cf. [4]). Note that in the best case, when the design is well-behaved and λ_2 is a constant, we have that

$$\iota(c,\delta_0,\lambda_1,\lambda_2) \ge \frac{1}{\sqrt{\|\delta_0\|_0}}\kappa, \tag{24}$$

where $\kappa > 0$ is a constant. Remarks given below provide further discussion.

The following theorem provides the deviation bounds for the lava prediction error.

Theorem 3.3 (Deviation Bounds for Lava in Regression). *We have that with probability* $1-\alpha-\epsilon$

$$\begin{aligned}
\frac{1}{n}\|X\widehat{\theta}_{\text{lava}} - X\theta_0\|_2^2 &\le \frac{2}{n}\|\,\mathsf{K}_{\lambda_2}^{1/2}X(\widehat{\delta}-\delta_0)\|_2^2\|\,\mathsf{K}_{\lambda_2}\,\| + \frac{2}{n}\|\mathsf{D}_{ridge}(\lambda_2)\|_2^2 \\
&\le \inf_{(\delta_0',\beta_0')' \in \mathbb{R}^{2p}:\delta_0+\beta_0=\theta_0} \left\{ \Big(B_1(\delta_0) \vee B_2(\beta_0)\Big)\|\,\mathsf{K}_{\lambda_2}\,\| + B_3 + B_4(\beta_0) \right\},
\end{aligned}$$

where $\|\,\mathsf{K}_{\lambda_2}\,\| \le 1$ *and*

$$\begin{aligned}
B_1(\delta_0) &= \frac{2^3\lambda_1^2}{\iota^2(c,\delta_0,\lambda_1,\lambda_2)} \le \frac{2^5\sigma_u^2 c^2 \bar{V}_{\lambda_2}^2\log(2p/\alpha)}{n\iota^2(c,\delta_0,\lambda_1,\lambda_2)}, \\
B_2(\beta_0) &= \frac{2^5}{n}\|\,\mathsf{K}_{\lambda_2}^{1/2}X\beta_0\|_2^2 = 2^5\lambda_2\beta_0' S(S+\lambda_2 I)^{-1}\beta_0, \\
B_3 &= \frac{2^2\sigma_u^2}{n}\left[\sqrt{\operatorname{tr}(\mathsf{P}_{\lambda_2}^2)} + \sqrt{2}\sqrt{\|\,\mathsf{P}_{\lambda_2}^2\|}\sqrt{\log(1/\epsilon)}\right]^2, \\
B_4(\beta_0) &= \frac{2^2}{n}\|\,\mathsf{K}_{\lambda_2}X\beta_0\|_2^2 = 2^2\beta_0' V_{\lambda_2}\beta_0 \le 2^3 B_2(\beta_0)\|\,\mathsf{K}_{\lambda_2}\,\|.
\end{aligned}$$

REMARK 3.3. As noted before, the "sparse+dense" framework does not require the separate identification of (β_0, δ_0). Consequently, the prediction upper bound is the infimum over all the pairs (β_0, δ_0) such that $\beta_0 + \delta_0 = \theta_0$. The upper bound thus optimizes over the best "split" of θ_0 into sparse and dense parts, δ_0 and β_0. The bound has four components. B_1 is a qualitatively sharp bound on the performance of the lasso for $\mathsf{K}_{\lambda_2}^{1/2}$-transformed data. It involves two important factors: $\bar{V}_{\lambda_2}$ and the design impact factor $\iota(c, \delta_0, \lambda_1, \lambda_2)$. The term B_3 is the size of the impact of the noise on the ridge part of the estimator, and it has a qualitatively sharp form as in [25]. The term B_4 describes the size of the bias for the ridge part of the estimator and appears to be qualitatively sharp as in [25]. We refer the reader to [25] for the in-depth analysis of noise term B_3 and bias term B_4. The term $B_2\| \mathsf{K}_{\lambda_2} \|$ appearing in the bound is also related to the size of the bias resulting from ridge regularization. In examples like the Gaussian sequence model, we have

$$B_4(\beta_0) \lesssim B_2(\beta_0)\| \mathsf{K}_{\lambda_2} \| \lesssim B_4(\beta_0). \tag{25}$$

This result holds more generally whenever $\| \mathsf{K}_{\lambda_2}^{-1} \| \| \mathsf{K}_{\lambda_2} \| \lesssim 1$, which occurs if λ_2 dominates the eigenvalues of S (see Supplementary Material [14] for the proof). ∎

REMARK 3.4 (Comments on Performance in Terms of Rates). It is worth discussing heuristically two key features arising from Theorem 3.3.

1) In dense models where ridge would work well, lava will work similarly to ridge. Consider any model where there is no sparse component (so $\theta_0 = \beta_0$), where the ridge-type rate $B^* = B_4(\beta_0) + B_3$ is optimal (e.g. [25]), and where (25) holds. In this case, we have $B_1(\delta_0) = 0$ since $\delta_0 = 0$, and the lava performance bound reduces to

$$B_2(\beta_0)\| \mathsf{K}_{\lambda_2} \| + B_3 + B_4(\beta_0) \lesssim B_4(\beta_0) + B_3 = B^*.$$

2) Lava works similarly to lasso in sparse models that have no dense components whenever lasso works well in those models. For this to hold, we need to set $\lambda_2 \gtrsim n \vee \|S\|^2$. Consider any model where $\theta_0 = \delta_0$ and with design such that the restricted eigenvalues κ of [5] are bounded away from zero. In this case, the standard lasso rate

$$B^* = \frac{\|\delta_0\|_0 \log(2p/\alpha)}{n\kappa^2}$$

of [5] is optimal. For the analysis of lava in this setting, we have that $B_2(\beta_0) = B_4(\beta_0) = 0$. Moreover, we can show that $B_3 \lesssim n^{-1}$ and that the design impact factor obeys (24) in this case. To bound $\bar{V}_{\lambda_2}$ in $B_1(\delta_0)$ in this case, by the definition of V_{λ_2} in Section 3.5,

$$S - V_{\lambda_2} := A_{\lambda_2}, \quad A_{\lambda_2} = (S + \lambda_2 I_p)^{-2} S^2 (S + 2\lambda_2 I_p).$$

As long as $\lambda_2 \gtrsim \|S\|^2$, we have $\|A_{\lambda_2}\| \lesssim 1$. Therefore, by the definition of $\bar{V}_{\lambda_2}$,

$$\bar{V}_{\lambda_2} \leq \max_{j \leq p} S_{jj} + \|A_{\lambda_2}\| \lesssim 1,$$

where S_{jj} denotes the jth diagonal element of S. Note that this upper bound of $\bar{V}_{\lambda_2}$ allows a diverging $\|S\|$ as $p \to \infty$, which is not stringent in high-dimensional models. Thus,

$$B_1(\delta_0) \lesssim \frac{\|\delta_0\|_0 \log(2p/\alpha)}{n\kappa^2} = B^*,$$

and $(B_1(\delta_0) \vee B_2(\beta_0))\| \mathsf{K}_{\lambda_2} \| + B_3 + B_4(\beta_0) \lesssim B^*$ follows due to $\| \mathsf{K}_{\lambda_2} \| \leq 1$.

Note that we see lava performing similarly to lasso in sparse models and performing similarly to ridge in dense models in the simulation evidence provided in the next section. This simulation evidence is consistent with the observations made above. ∎

REMARK 3.5 (On the design impact factor). The definition of the design impact factor is motivated by the generalizations of the restricted eigenvalues of [5] proposed in [4] to improve performance bounds for lasso in badly behaved designs. The concepts above are strictly more general than the usual restricted eigenvalues formulated for the transformed data. Let $J(\delta_0) = \{j \leq p : \delta_{0j} \neq 0\}$. For any vector $\Delta \in \mathbb{R}^p$, respectively write $\Delta_{J(\delta_0)} = \{\Delta_j : j \in J(\delta_0)\}$ and $\Delta_{J^c(\delta_0)} = \{\Delta_j : j \notin J(\delta_0)\}$. Define

$$\mathcal{A}(c, \delta_0) = \{v \in \mathbb{R}^p \setminus \{0\} : \|\Delta_{J^c}(\delta_0)\|_1 \leq (c+1)/(c-1)\|\Delta_{J(\delta_0)}\|_1\}.$$

The restricted eigenvalue $\kappa^2(c, \delta_0, \lambda_2)$ is given by

$$\kappa^2(c, \delta_0, \lambda_2) = \inf_{\Delta \in \mathcal{A}(c,\delta_0)} \frac{\|\widetilde{X}\Delta\|_2^2/n}{\|\Delta_{J(\delta_0)}\|_2^2} = \inf_{\Delta \in \mathcal{A}(c,\delta_0)} \frac{X' \mathsf{K}_{\lambda_2} X/n}{\|\Delta_{J(\delta_0)}\|_2^2}.$$

Note that $\mathcal{A}(c, \delta_0) \subset \mathcal{R}(c, \delta_0, \lambda_1, \lambda_2)$ and that

$$\iota(c, \delta_0, \lambda_1, \lambda_2) \geq \inf_{\Delta \in \mathcal{A}(c,\delta_0)} \frac{\|\widetilde{X}\Delta\|_2/\sqrt{n}}{\|\Delta_{J(\delta_0)}\|_1} \geq \frac{1}{\sqrt{\|\delta_0\|_0}}\kappa(c, \delta_0, \lambda_2).$$

Now note that $X' \mathsf{K}_{\lambda_2} X/n = \lambda_2 S(S+\lambda_2 I_p)^{-1}$. When λ_2 is relatively large, $X' \mathsf{K}_{\lambda_2} X/n = \lambda_2 S(S+\lambda_2 I_p)^{-1}$ is approximately equal to S. Hence, $\kappa^2(c, \delta_0, \lambda_2)$ behaves like the usual restricted eigenvalue constant as in [5], and we have a bound on the design impact factor $\iota(c, \delta_0, \lambda_1, \lambda_2)$ as in (24). To understand how $\kappa^2(c, \delta_0, \lambda_2)$ depends on λ_2 more generally, consider the special case of an orthonormal design. In this case, $S = I_p$ and $X' \mathsf{K}_{\lambda_2} X/n = kI_p$ with $k = \lambda_2/(1+\lambda_2)$. Then $\kappa^2(c, \delta_0, \lambda_2) = k$, and the design impact factor becomes $\sqrt{k}/\sqrt{\|\delta_0\|_0}$.

Thus, the design impact factor scales like $1/\sqrt{\|\delta_0\|_0}$ when restricted eigenvalues are well-behaved, e.g. bounded away from zero. This behavior corresponds to the best possible case. Note that design impact factors can be well-behaved even if restricted eigenvalues are not. For example, suppose we have two regressors that are identical. Then $\kappa(c, \delta_0, \lambda_2) = 0$, but $\iota(c, \delta_0, \lambda_1, \lambda_2) > 0$ in this case; see [4]. ∎

4. Simulation Study. In this section, we provide simulation evidence on the performance of the lava estimator. Before proceeding to the simulation settings and results, we review the lava and post-lava estimation procedure. The lava and post-lava algorithm can be summarized as follows:

1) Fix λ_1, λ_2, and define $\mathsf{K}_{\lambda_2} = I_n - X(X'X + n\lambda_2 I_p)^{-1}X'$.

2) For $\widetilde{Y} = \mathsf{K}_{\lambda_2}^{1/2} Y$, and $\widetilde{X} = \mathsf{K}_{\lambda_2}^{1/2} X$, solve for $\widehat{\delta} = \arg\min_{\delta\in\mathbb{R}^p}\{\frac{1}{n}\|\widetilde{Y} - \widetilde{X}\delta\|_2^2 + \lambda_1\|\delta\|_1\}$.

3) Define $\widehat{\beta}(\delta) = (X'X + n\lambda_2 I_p)^{-1}X'(Y - X\delta)$. The lava estimator is $\widehat{\theta}_{\text{lava}} = \widehat{\beta}(\widehat{\delta}) + \widehat{\delta}$.

4) For $W = Y - X\widehat{\beta}(\widehat{\delta})$, solve for $\widetilde{\delta} = \arg\min_{\delta\in\mathbb{R}^p}\{\frac{1}{n}\|W - X\delta\|_2^2, \quad \delta_j = 0 \text{ if } \widehat{\delta}_j = 0\}$.

5) The post-lava estimator is $\widehat{\theta}_{\text{post-lava}} = \widehat{\beta}(\widehat{\delta}) + \widetilde{\delta}$.

The computation of $\mathsf{K}_{\lambda_2}^{1/2}$ is carried by using the eigen-decomposition of XX': Let M be an $n \times n$ matrix whose columns are the eigenvectors of the $n \times n$ matrix XX'. Let $\{v_1, ..., v_n\}$ be the eigenvalues of XX'. Then $\mathsf{K}_{\lambda_2}^{1/2} = MAM'$, where A is an $n\times n$ diagonal matrix with $A_{jj} = (n\lambda_2/(v_j + n\lambda_2))^{1/2}$, $j = 1, ..., n$. Supplementary Material [14] proves this assertion.

We present a Monte-Carlo analysis based on a Gaussian linear regression model: $Y = X\theta + U$, $U \mid X \sim N(0, I_n)$. The parameter θ is a p-vector defined as

$$\theta = (3, 0, \ldots, 0)' + (0, q, \ldots, q)',$$

where $q \geq 0$ is the "size of small coefficients". When q is zero or small, θ can be well-approximated by the sparse vector $(3, 0, \ldots, 0)$. When q is relatively large, θ cannot be approximated well by a sparse vector. We report results for three settings that differ in the design for X. For each setting, we set $n = 100$ and compare the performance of $X\widehat{\theta}$ formed from one of five methods: lasso, ridge, elastic net, lava, and post-lava. For the first two designs for X, we set $p = 2n$ and consider a fixed design matrix for X by generating the design matrix X once and fixing it across simulation replications. We draw our realization of X by drawing rows of X independently from a mean zero multivariate normal with covariance matrix Σ. For the first setting, we set $\Sigma = I_p$, and we use a factor covariance structure with $\Sigma = LL' + I_p$ where the rows of L are independently generated from a $N(0, I_3)$ for the second setting. In the latter case, the columns of X depend on three common factors. In order to verify that the comparisons continue to hold with random design, we also consider the third design, where X is redrawn within each simulation replication by drawing rows of X independently from a mean zero multivariate normal with covariance matrix I_p. In this third setting, we set $p = 5n$.[4] All results are based on $B = 100$ simulation replications.

Within each design, we report two sets of results. For the first set of results, we select the tuning parameters for each method by minimizing SURE. The SURE formula depends on the error variance σ_u^2 which we take as known in all simulations.[5] In the

[4]Results with $p = n/2$, where OLS is also included, are available in supplementary material. The results are qualitatively similar to those given here, with lava and post-lava dominating all other procedures.

[5]In practice, one could obtain an estimator of σ_u^2 using an iterative method as in [2]. We expect that this procedure can be theoretically justified along the lines of [2], and we leave it as a future research direction. In supplementary material, we do provide simulation results where we use SURE with a conservative pre-estimate of σ_u^2 given by taking the estimate of σ_u^2 obtained using coefficients given by

second set of results, all tuning parameters are selected by minimizing 5-fold cross-validation.

To measure performance, we consider the risk measure $\mathrm{R}(\theta, \widehat{\theta}) = \mathrm{E}[\frac{1}{n}\|X\widehat{\theta} - X\theta\|_2^2]$, where the expectation E is conditioned on X for the first two settings. For each estimation procedure, we report the simulation estimate of this risk measure formed by averaging over the simulation replications. We plot the risk measure for the fixed design settings in Figures 4 and 5 and for the random design setting in Figure 6. Each figure shows the simulation estimate of $\mathrm{R}(\theta, \widehat{\theta})$ for each estimation method as a function of q. In Figure 4, we report results for the two fixed-design setting with tuning parameters chosen via minimizing the SURE as defined in Theorem 3.2. We then report results for the two fixed-design settings with tuning parameters chosen by 5-fold cross-validation in Figure 5. Figure 6 contains the results for the random design setting with the two upper panels giving results with tuning parameters selected by minimizing SURE and the lower two panels with tuning parameters selected by minimizing 5-fold cross validation.

The comparisons are similar in all figures with lava and post-lava dominating the other procedures. It is particularly interesting to compare the performance of lava to lasso and ridge. With feasible, data-dependent penalty parameter choices, we still have that the lava and post-lava estimators perform about as well as lasso when the signal is sparse and perform significantly better than lasso when the signal is non-sparse. We also see that the lava and post-lava estimators perform about as well as ridge when the signal is dense and perform much better than ridge when the signal is sparse using data-dependent tuning. When the tuning parameters are selected via cross-validation, the post-lava performs slightly better than the lava when the model is sparse. The gain is somewhat more apparent in the independent design.[6]

5. Discussion.

We propose a new method, called "lava", which is designed specifically to achieve good prediction and estimation performance in "sparse+dense" models. In such models, the high-dimensional parameter is represented as the sum of a sparse vector with a few large non-zero entries and a dense vector with many small entries. This structure renders traditional sparse or dense estimation methods, such as lasso or ridge, sub-optimal for prediction and other estimation purposes. The proposed approach thus complements other approaches to structured sparsity problems such as those considered in fused sparsity estimation ([37] and [13]) and structured matrix decomposition problems ([10], [12], [20], and [28]).

There are a number of interesting research directions that remain to be considered. An immediate extension of the present results would be to consider semi-pivotal estimators

5-fold cross-validated lasso. The results from using this conservative plug-in in calculating SURE are qualitatively similar to those reported with known σ_u^2 in the sense that lava and post-lava dominate the other considered estimators.

[6]We report additional simulation results in the supplementary material ([14]). In the supplement, we also report the percentage of out-of-sample variation explained as an additional risk measure in the random design setting. The qualitative conclusions are similar as those obtained by looking at the risk measure reported here.

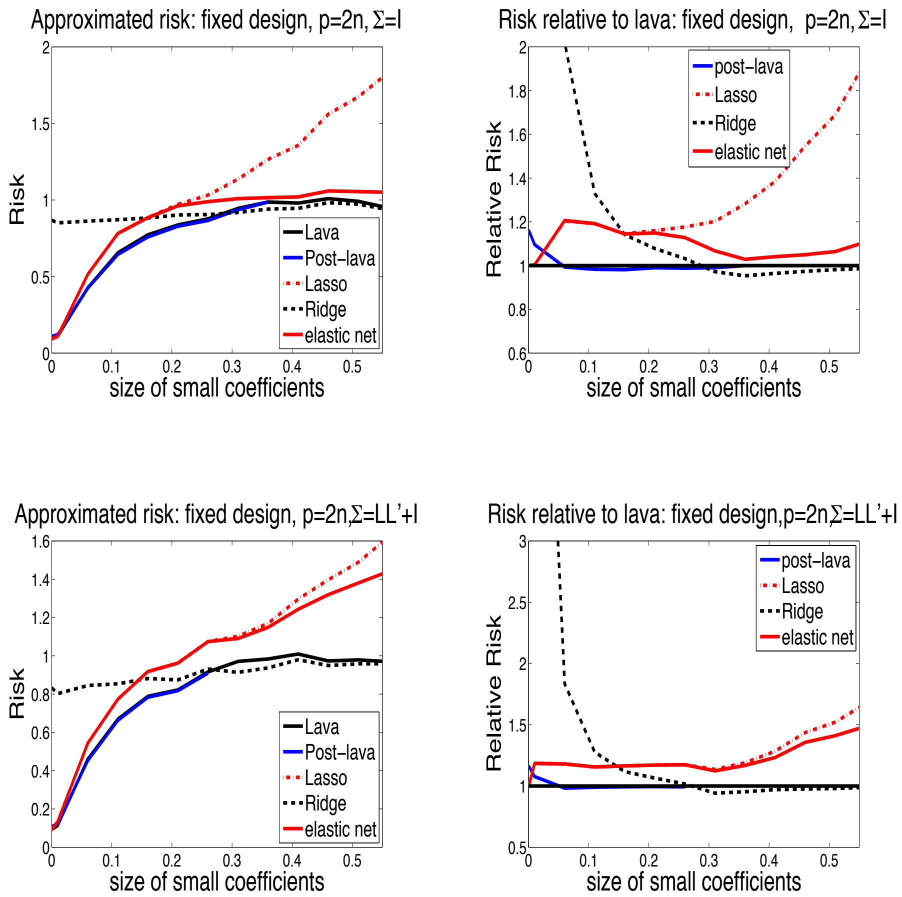

FIG 4. *Risk comparison with tuning done by minimizing SURE in the fixed design simulations. In this figure, we report simulation estimates of risk functions of lava, post-lava, ridge, lasso, and elastic net in a Gaussian regression model with "sparse+dense" signal structure over the regression coefficients. We select tuning parameters for each method by minimizing SURE. The size of "small coefficients" (q) is shown on the horizontal axis. The size of these coefficients directly corresponds to the size of the "dense part" of the signal, with zero corresponding to the exactly sparse case. Relative risk plots the ratio of the risk of each estimator to the lava risk,* $\mathrm{R}(\theta, \widehat{\theta}_{\mathrm{e}})/\mathrm{R}(\theta, \widehat{\theta}_{\mathrm{lava}})$.

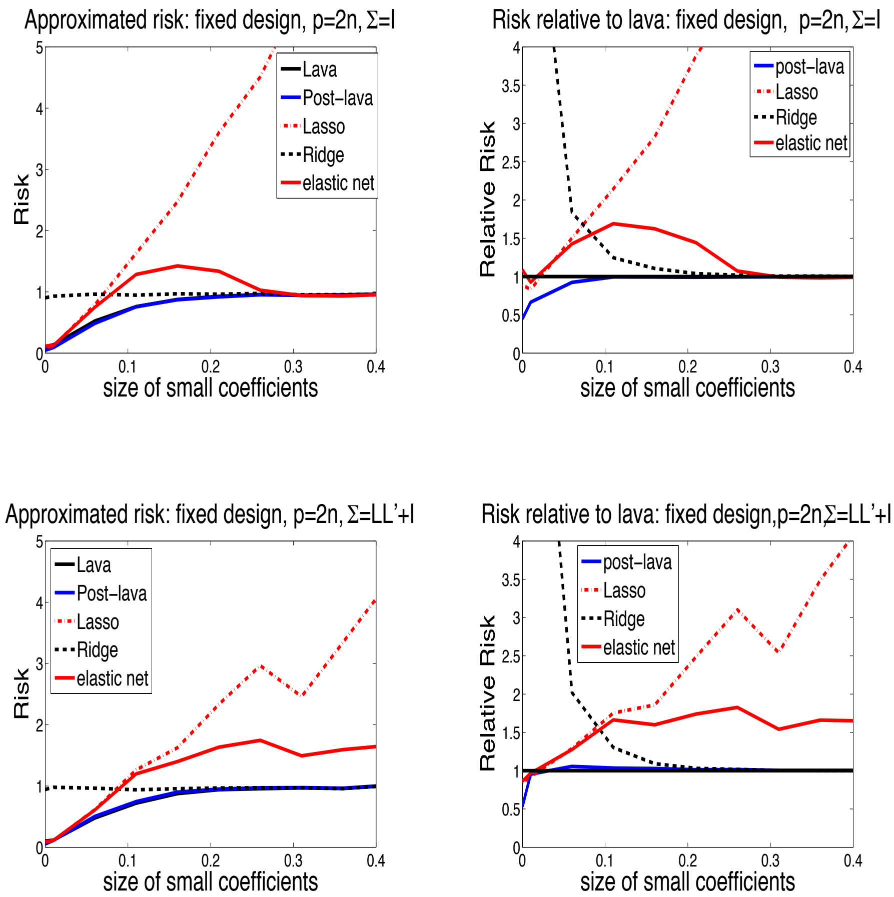

FIG 5. *Risk comparison with tuning done by 5-fold cross-validation in the fixed design simulations. In this figure, we report simulation estimates of risk functions of lava, post-lava, ridge, lasso, and elastic net in a Gaussian regression model with "sparse+dense" signal structure over the regression coefficients. We select tuning parameters for each method by minimizing 5-fold cross-validation. The size of "small coefficients" (q) is shown on the horizontal axis. The size of these coefficients directly corresponds to the size of the "dense part" of the signal, with zero corresponding to the exactly sparse case. Relative risk plots the ratio of the risk of each estimator to the lava risk,* $\mathrm{R}(\theta, \widehat{\theta}_e)/\mathrm{R}(\theta, \widehat{\theta}_{\mathrm{lava}})$.

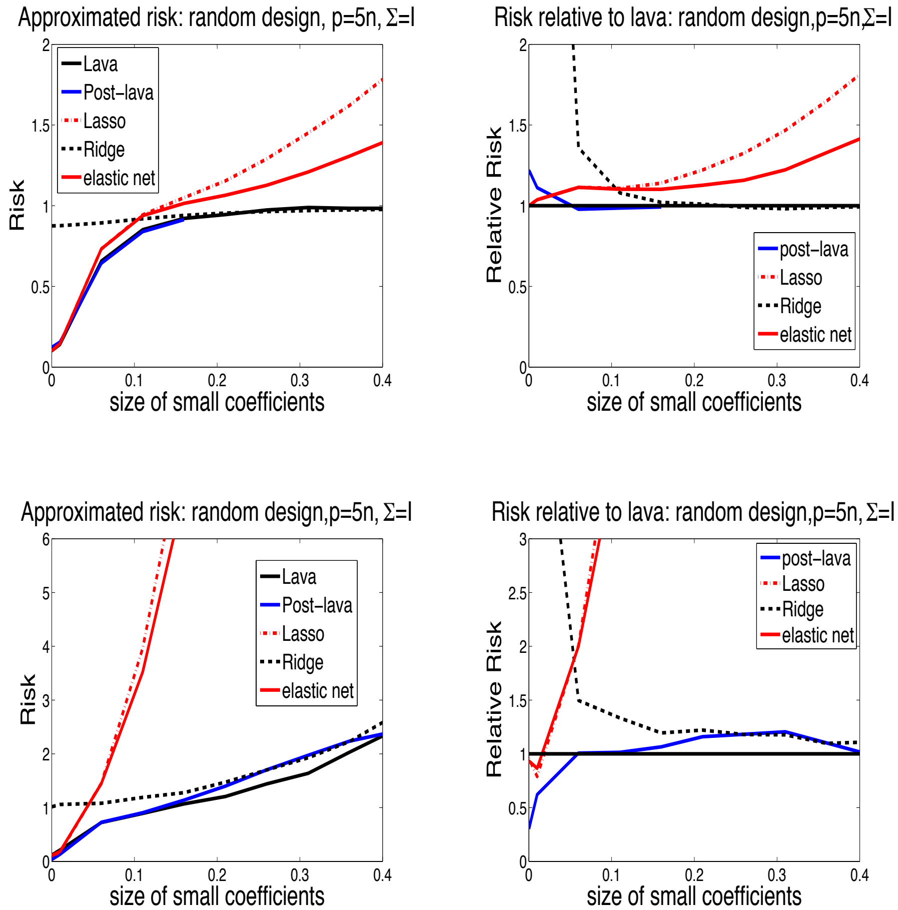

FIG 6. *Risk comparison in the random design simulations. In this figure, we report simulation estimates of risk functions of lava, post-lava, ridge, lasso, and elastic net in a Gaussian regression model with "sparse+dense" signal structure over the regression coefficients. In the upper panels, we report results with tuning parameters for each method selected by minimizing SURE. In the lower panels, we report results with tuning parameters for each method selected by minimizing 5-fold cross-validation. The size of "small coefficients" (q) is shown on the horizontal axis. The size of these coefficients directly corresponds to the size of the "dense part" of the signal, with zero corresponding to the exactly sparse case. Relative risk plots the ratio of the risk of each estimator to the lava risk,* $\mathrm{R}(\theta,\widehat{\theta}_{\mathrm{e}})/\mathrm{R}(\theta,\widehat{\theta}_{\mathrm{lava}})$.

akin to the root-lasso/scaled-lasso of [3] and [35]. For instance, we can define

$$\begin{aligned}\widehat{\theta}_{\text{root-lava}} &:= \widehat{\beta}+\widehat{\delta},\\ (\widehat{\beta},\widehat{\delta}) &:= \arg\min_{\beta,\delta,\sigma}\left\{\frac{1}{2n\sigma^2}\|Y-X(\beta+\delta)\|_2^2+\frac{(1-a)\sigma}{2}+\lambda_2\|\beta\|_2^2+\lambda_1\|\delta\|_1\right\}.\end{aligned}$$

Thanks to the characterization of Theorem 3.1, the method can be implemented by applying root-lasso on appropriately transformed data. The present work could also be extended to accommodate non-Gaussian settings and settings with random designs, and it could also be extended beyond the mean regression problem to more general M- and Z- estimation problems, e.g., along the lines of [32]. There is also a body of work that considers the use of a sum of penalties in penalized estimation, e.g., [10], [12], [28], among others. These methods are designed for models with particular types of structures and targeting different estimation goals than the approach we take in this paper. A more thorough examination of the tradeoffs and structures that may lead one to prefer one approach to the other may be worthwhile. Finally, [40] formulate a general class of problems where the signal may be split into a superposition of different types $\theta=\sum_{\ell=1}^L\theta_\ell$ and use penalization of different types for each of the components of the signal of the form $\sum_{\ell=1}^L \text{penalty}_\ell(\theta_\ell)$. Further analysis of various leading cases, within this class of problems, providing sharp conditions, performance bounds, computational algorithms, and interesting applications, is an important venue for future work.

APPENDIX A: PROOFS FOR SECTION 2

A.1. Proof of Lemma 2.1. Fixing δ, the solution for β is given by $\widehat{\beta}(\delta)=(z-\delta)/(1+\lambda_2)$. Substituting back to the original problem, we obtain

$$\begin{aligned} d_1(z) &= \arg\min_{\delta\in\mathbb{R}}[z-\widehat{\beta}(\delta)-\delta]^2+\lambda_2|\widehat{\beta}(\delta)|^2+\lambda_1|\delta|\\ &= \arg\min_{\delta\in\mathbb{R}} k(z-\delta)^2+\lambda_1|\delta|.\end{aligned}$$

Hence $d_1(z)=(|z|-\lambda_1/(2k))_+\text{sign}(z)$, and $d_2(z)=\widehat{\beta}(d_1(z))=(z_1-d_1(z))(1-k)$. Consequently, $d_{\text{lava}}(z)=d_1(z)+d_2(z)=(1-k)z+kd_1(z)$. ∎

A.2. A Useful Lemma. The proofs rely on the following lemma.

Lemma A.1. *Consider the general piecewise linear function:*

$$F(z)=(hz+d)1\{z>w\}+(ez+m)1\{|z|\le w\}+(fz+g)1\{z<-w\}.$$

Suppose $Z\sim N(\theta,\sigma^2)$. Then

$$\begin{aligned}\mathrm{E}[F(Z)^2] = {}& [\sigma^2(h^2w+h^2\theta+2dh)-\sigma^2(e^2w+e^2\theta+2me)]\phi_{\theta,\sigma}(w)\\ &+[\sigma^2(-e^2w+e^2\theta+2me)-\sigma^2(-f^2w+f^2\theta+2gf)]\phi_{\theta,\sigma}(-w)\\ &+((h\theta+d)^2+h^2\sigma^2)\mathrm{P}_{\theta,\sigma}(Z>w)+((f\theta+g)^2+f^2\sigma^2)\mathrm{P}_{\theta,\sigma}(Z<-w)\\ &+((e\theta+m)^2+e^2\sigma^2)\mathrm{P}_{\theta,\sigma}(|Z|<w).\end{aligned}$$

PROOF. We first consider an expectation of the following form: for any $-\infty \le z_1 < z_2 \le \infty$, and $a, b \in \mathbb{R}$, by integration by part,

$$\begin{aligned} \mathrm{E}(\theta - Z)(aZ+b)1\{z_1 < Z < z_2\} &= \sigma^2 \int_{z_1}^{z_2} \frac{\theta - z}{\sigma^2}(az+b)\phi_{\theta,\sigma}(z)dz \\ &= \sigma^2 (az+b)\phi_{\theta,\sigma}(z)\big|_{z_1}^{z_2} - \sigma^2 a \int_{z_1}^{z_2} \phi_{\theta,\sigma}(z)dz \\ &= \sigma^2[(az_2+b)\phi_{\theta,\sigma}(z_2) - (az_1+b)\phi_{\theta,\sigma}(z_1)] - \sigma^2 a \mathrm{P}_{\theta,\sigma}(z_1 < Z < z_2). \end{aligned} \tag{26}$$

This result will be useful in the following calculations. Setting $a = -1$, $b = \theta$ and $a = 0$, $b = -2(\theta + c)$ respectively yields

$$\mathrm{E}(\theta - Z)^2 1\{z_1 < Z < z_2\} = \sigma^2[(\theta - z_2)\phi_{\theta,\sigma}(z_2) - (\theta - z_1)\phi_{\theta,\sigma}(z_1)] + \sigma^2 \mathrm{P}_{\theta,\sigma}(z_1 < Z < z_2).$$

$$2\mathrm{E}(Z - \theta)(\theta + c)1\{z_1 < Z < z_2\} = \sigma^2[-2(\theta + c)\phi_{\theta,\sigma}(z_2) + 2(\theta + c)\phi_{\theta,\sigma}(z_1)].$$

Therefore, for any constant c,

$$\begin{aligned} \mathrm{E}(Z+c)^2 1\{z_1 < Z < z_2\} &= \mathrm{E}(\theta - Z)^2 1\{z_1 < Z < z_2\} + (\theta + c)^2 \mathrm{P}_{\theta,\sigma}(z_1 < Z < z_2) \\ &\quad + 2\mathrm{E}(Z - \theta)(\theta + c)1\{z_1 < Z < z_2\} \\ &= \sigma^2(z_1 + \theta + 2c)\phi_{\theta,\sigma}(z_1) - \sigma^2(z_2 + \theta + 2c)\phi_{\theta,\sigma}(z_2) \\ &\quad + ((\theta + c)^2 + \sigma^2)\mathrm{P}_{\theta,\sigma}(z_1 < Z < z_2). \end{aligned} \tag{27}$$

If none of h, e, f are zero, by setting $z_1 = w, z_2 = \infty$, $c = d/h$; $z_1 = -\infty, z_2 = -w, c = g/f$ and $z_1 = -w, z_2 = w, c = m/e$ respectively, we have

$$\begin{aligned} \mathrm{E}(hZ+d)^2 1\{Z > w\} &= \sigma^2(h^2 w + h^2\theta + 2dh)\phi_{\theta,\sigma}(w) \\ &\quad + ((\theta h + d)^2 + \sigma^2 h^2)\mathrm{P}_{\theta,\sigma}(Z > w), \\ \mathrm{E}(fZ+g)^2 1\{Z < -w\} &= -\sigma^2(-wf^2 + \theta f^2 + 2gf)\phi_{\theta,\sigma}(-w) \\ &\quad + ((\theta f + g)^2 + \sigma^2 f^2)\mathrm{P}_{\theta,\sigma}(Z < -w), \end{aligned} \tag{28}$$

$$\begin{aligned} \mathrm{E}(eZ+m)^2 1\{|Z| < w\} &= \sigma^2(-we^2 + \theta e^2 + 2me)\phi_{\theta,\sigma}(-w) \\ &\quad - \sigma^2(we^2 + \theta e^2 + 2me)\phi_{\theta,\sigma}(w) \\ &\quad + ((\theta e + m)^2 + \sigma^2 e^2)\mathrm{P}_{\theta,\sigma}(|Z| < w). \end{aligned} \tag{29}$$

If any of h, e, f is zero, for instance, suppose $h = 0$, then $\mathrm{E}(hZ+d)^2 1\{Z > w\} = d^2 \mathrm{P}_{\theta,\sigma}(Z > w)$, which can also be written as the first equality of (28). Similarly, when either $e = 0$ or $f = 0$, (28) still holds.

Therefore, summing up the three terms of (28) yields the desired result. ∎

A.3. Proof of Theorem 2.1. Recall that $\widehat{\theta}_{\text{lava}} = (1-k)Z + k d_{\text{lasso}}(Z)$ is a weighted average of Z and the soft-thresholded estimator with shrinkage parameters $\lambda_1/(2k)$ and $k = \lambda_2/(1+\lambda_2)$. Since $d_{\text{lasso}}(Z)$ is a soft-thresholding estimator, results from [15] give that $\mathrm{E}[(Z-\theta)d_{\text{lasso}}(Z)] = \sigma^2 \mathrm{P}_{\theta,\sigma}(|Z| > \lambda_1/(2k))$. Therefore, for $w = \lambda_1/(2k)$,

$$2\mathrm{E}[(Z-\theta)d_{\text{lava}}(Z)] = 2(1-k)\sigma^2 + 2k\sigma^2 \mathrm{P}_{\theta,\sigma}(|Z| > w). \tag{30}$$

Next we verify that

$$\text{(31)} \qquad \begin{aligned} \mathrm{E}(Z - d_{\text{lava}}(Z))^2 &= -k^2(w+\theta)\phi_{\theta,\sigma}(w)\sigma^2 + k^2(\theta - w)\phi_{\theta,\sigma}(-w)\sigma^2 \\ &\quad +(\lambda_1^2/4)\mathrm{P}_{\theta,\sigma}(|Z| > w) + k^2(\theta^2+\sigma^2)\mathrm{P}_{\theta,\sigma}(|Z| < w). \end{aligned}$$

By definition,

$$\text{(32)} \qquad d_{\text{lava}}(z) - z \;=\; \begin{cases} -\lambda_1/2, & z > \lambda_1/(2k), \\ -kz, & -\lambda_1/(2k) < z \le \lambda_1/(2k), \\ \lambda_1/2, & z < -\lambda_1/(2k). \end{cases}$$

Let $F(z) = d_{\text{lava}}(z) - z$. The claim then follows from applying Lemma A.1 by setting $h = f = m = 0$, $d = -\lambda_1/2$, $e = -k$, $g = \lambda_1/2$, and $w = \lambda_1/(2k)$. Hence

$$\begin{aligned} \mathrm{E}(Z - d_{\text{lava}}(Z))^2 \;&=\; -k^2(w+\theta)\phi_{\theta,\sigma}(w)\sigma^2 + k^2(\theta - w)\phi_{\theta,\sigma}(-w)\sigma^2 \\ &\quad +(\lambda_1^2/4)\mathrm{P}_{\theta,\sigma}(|Z| > w) + k^2(\theta^2+\sigma^2)\mathrm{P}_{\theta,\sigma}(|Z| < w). \end{aligned}$$

The risk of lasso is obtained from setting $\lambda_2 = \infty$ and $\lambda_1 = \lambda_l$ in the lava risk. The risk of ridge is obtained from setting $\lambda_1 = \infty$ and $\lambda_2 = \lambda_r$ in the lava risk.

As for the risk of post-lava, note that

$$d_{\text{post-lava}}(z) - \theta = \begin{cases} z - \theta, & |z| > \lambda_1/(2k) \\ (1-k)z - \theta, & |z| \le \lambda_1/(2k). \end{cases}$$

Hence applying Lemma A.1 to $F(z) = d_{\text{post-lava}}(z) - \theta$, i.e. by setting $h = f = 1, e = 1-k$ and $d = m = g = -\theta$, we obtain:

$$\begin{aligned} R(\theta, \widehat{\theta}_{\text{post-lava}}) \;&=\; \sigma^2[-k^2w + 2kw - k^2\theta]\phi_{\theta,\sigma}(w) + \sigma^2[-k^2w + 2kw + k^2\theta]\phi_{\theta,\sigma}(-w) \\ &\quad +\sigma^2\mathrm{P}_{\theta,\sigma}(|Z| > w) + (k^2\theta^2 + (1-k)^2\sigma^2)\mathrm{P}_{\theta,\sigma}(|Z| < w). \end{aligned}$$

Finally, the elastic net shrinkage is given by $d_{\text{enet}}(z) = \frac{1}{1+\lambda_2}(|z| - \lambda_1/2)_+\mathrm{sgn}(z)$. The risk of elastic net then follows from Lemma A.1 by setting $F(z) = d_{\text{enet}}(z) - \theta$, $w = \lambda_1/2, h = f = 1/(1+\lambda_2), e = 0$, $d = -\lambda_1/(2(1+\lambda_2)) - \theta$ and $g = \lambda_1/(2(1+\lambda_2)) - \theta$. ■

A.4. Proof of Lemma 2.2. For $\widehat{\theta}_{\text{lava}} = (\widehat{\theta}_{\text{lava},j})_{j=1}^p$, we have $\mathrm{E}\|\widehat{\theta}_{\text{lava}} - \theta\|_2^2 = \sum_{j=1}^p R(\theta_j, \widehat{\theta}_{\text{lava},j})$. We bound $R(\theta_j, \widehat{\theta}_{\text{lava},j})$ uniformly over $j = 1, ..., p$. Since $Z_j \sim N(\theta_j, \sigma^2)$, by Theorem 2.1,

$$\begin{aligned} R(\theta_j, \widehat{\theta}_{\text{lava},j}) \;&=\; \mathrm{E}(\widehat{\theta}_{\text{lava},j} - \theta_j)^2 = I + II + III, \\ I \;&:=\; -k^2(w+\theta_j)\phi_{\theta_j,\sigma}(w)\sigma^2 + k^2(\theta_j - w)\phi_{\theta_j,\sigma}(-w)\sigma^2 \\ II \;&:=\; (\lambda_1^2/4 + \sigma^2)\mathrm{P}_{\theta_j,\sigma}(|Z_j| > w), \\ III \;&:=\; (\theta_j^2 k^2 + (1-k)^2\sigma^2)\mathrm{P}_{\theta_j,\sigma}(|Z_j| < w). \end{aligned}$$

Let $L := (1 + M^2/\sigma^2)/2$. By Mill's ratio inequality, if $\frac{2p}{\pi c^2} \ge \log p \ge \frac{1}{\pi c^2}$, then $2\sqrt{2\log p} > \lambda_1/\sigma > 2\sqrt{\log p}$. By definition, $w = \lambda_1/(2k)$, $k = \lambda_2/(1+\lambda_2)$, $\lambda_2 = \sigma^2 p/\|\beta\|_2^2$, and the condition $\|\beta\|_\infty \le M$, we have $\sigma\sqrt{\log p} < \lambda_1/2 \le w \le \lambda_1 L \le 2\sigma\sqrt{2\log p}L$.

We divide our discussions into case: **1)** $\delta_j = 0$, **2)** $0 < |\delta_j| \leq 2L\lambda_1$, and **3)** $|\delta_j| > 2L\lambda_1$.

1): $\delta_j = 0$. In this case, as long as $\sigma\sqrt{\log p} > 2M$, $|\theta_j| = |\beta_j| \leq M < w/2$.

To bound I, we note since $|\theta_j| < w$, we have $I < 0$. To bound II,

$$\begin{aligned} II &\leq_{(1)} (w^2+\sigma^2)\frac{\sigma}{\sqrt{2\pi}(w-\theta_j)}e^{-(w-\theta_j)^2/2\sigma^2} + (w^2+\sigma^2)\frac{\sigma}{\sqrt{2\pi}(w+\theta_j)}e^{-(w+\theta_j)^2/2\sigma^2} \\ &\leq_{(2)} 2\times(w^2+\sigma^2)\frac{2\sigma}{\sqrt{2\pi}w}e^{-w^2/(8\sigma^2)} \\ &\leq_{(3)} \frac{8w\sigma^2}{\sqrt{2\pi}\sigma}e^{-w^2/(8\sigma^2)} \leq_{(4)} \frac{4\sigma^2}{\sqrt{2\pi}}e^{-w^2/(16\sigma^2)} \leq \frac{4\sigma^2}{\sqrt{2\pi}p^{1/16}}. \end{aligned}$$

In the above, (1) follows from the Mill's ratio inequality: $\int_x^\infty e^{-t^2/2}dt \leq x^{-1}e^{-x^2/2}$ for $x \geq 0$. Also note that $w \pm \theta_j > 0$. Hence we can apply the Mill's ratio inequality respectively on $\mathrm{P}_{\theta_j,\sigma}(Z_j > w)$ and $\mathrm{P}_{\theta_j,\sigma}(Z_j < -w)$. (2) is due to $w \pm \theta_j > w/2$. (3) follows since $\sigma^2 \leq w^2$ when $p \geq e$. Finally, for any $a > 0$, and any $x > 1 + a^{-1}$, $ax^2 > \log x$. Set $a = 64^{-1}$; when $\sqrt{\log p} > 33$, $\log(2w/\sigma) < w^2/(16\sigma^2)$. Hence $\frac{2w}{\sigma}e^{-w^2/(8\sigma^2)} \leq e^{-w^2/(16\sigma^2)}$, which gives (4).

To bound III, we note $III \leq \beta_j^2 k^2 + (1-k)^2\sigma^2$. Therefore,

$$\sum_{\delta_j=0} R(\theta_j, \widehat{\theta}_{\text{lava},j}) \leq \|\beta\|_2^2 k^2 + (1-k)^2\sigma^2 p + \frac{4\sigma^2 p}{\sqrt{2\pi}p^{1/16}} = (1-k)p\sigma^2 + \frac{4\sigma^2 p}{\sqrt{2\pi}p^{1/16}},$$

where we used the equality $(1-k)^2\sigma^2 p + k^2\|\beta\|_2^2 = (1-k)p\sigma^2$ for $k = \frac{\sigma^2 p}{\|\beta\|_2^2 + \sigma^2 p}$.

2): $0 < |\delta_j| \leq 2L\lambda_1$. To bound I, since $|\theta_j| \leq M + 4\sigma\sqrt{2\log p}L < 5\sigma\sqrt{2\log p}L$,

$$I \leq |\theta_j|\sigma^2(\phi_{\theta_j,\sigma}(w) + \phi_{\theta_j,\sigma}(-w)) \leq 5\sigma\sqrt{2\log p}L\sigma^2\frac{2}{\sqrt{2\pi\sigma^2}} = \frac{10\sigma^2 L}{\sqrt{\pi}}\sqrt{\log p}.$$

To bound $II + III$, we note $II + III \leq \lambda_1^2/4 + 2\sigma^2 + \theta_j^2$. Hence

$$\sum_{0<|\delta_j|\leq 2L\lambda_1} R(\theta_j, \widehat{\theta}_{\text{lava},j}) \leq (\lambda_1^2/4 + 2\sigma^2 + \theta_j^2)s \leq (2\sigma^2 + (5L^2 + 0.25)\lambda_1^2)s,$$

where we used $\theta_j^2 \leq M^2 + 2M|\delta_j| + \delta_j^2 \leq 5L^2\lambda_1^2$.

3) $|\delta_j| > 2L\lambda_1$. In this case, $|w \pm \theta_j| \geq |\theta_j| - w \geq |\delta_j| - M - w > |\delta_j|/3$, and $|w \pm \theta_j| \leq w + M + |\delta_j| < 5|\delta_j|/3$. Also, $|\delta_j| > 2\sigma\sqrt{\log p}$. To bound I, when $\sqrt{\log p} > 20\sigma$, $|\delta_j| > 1 + 20\sigma^2$, and $|\delta_j| < \exp(|\delta_j|^2/(20\sigma^2))$. So

$$\begin{aligned} I &\leq 5|\delta_j|\sigma^2(\phi_{\theta_j,\sigma}(w) + \phi_{\theta_j,\sigma}(-w))/3 \leq \frac{10\sigma^2}{3\sqrt{2\pi\sigma^2}}|\delta_j|e^{-|\delta_j|^2/(18\sigma^2)} \\ &\leq \frac{10\sigma^2}{3\sqrt{2\pi\sigma^2}}e^{-|\delta_j|^2/(180\sigma^2)} \leq \frac{10\sigma}{3\sqrt{2\pi}}p^{-1/45}. \end{aligned}$$

To bound II, we have $II \leq \lambda_1^2/4 + \sigma^2$. To bound III, since $|\theta_j|^2 + \sigma^2 < 5|\delta_j|^2/4$, we have $III \leq 5|\delta_j|^2\mathrm{P}_{\theta_j,\sigma}(|Z_j| < w)/4$. Supppspse $\theta_j > 0$, then $w - \theta_j \leq -|\delta_j|/4$, and

$$III \leq 1.25|\delta_j|^2\mathrm{P}(N(0,1) < (w-\theta_j)/\sigma) \leq 1.25|\delta_j|^2\mathrm{P}(N(0,1) < -|\delta_j|/(4\sigma))$$

$$\leq \quad 1.25|\delta_j|^2 \frac{1}{\sqrt{2\pi}} \frac{4\sigma}{|\delta_j|} e^{-|\delta_j|^2/(32\sigma^2)} \leq \frac{5\sigma}{\sqrt{2\pi}} p^{-1/45}.$$

Therefore,

$$\sum_{|\delta_j|>2L\lambda_1} R(\theta_j, \widehat{\theta}_{\text{lava},j}) \leq s(\lambda_1^2/4 + \sigma^2 + \frac{9\sigma}{\sqrt{2\pi}} p^{-1/45}).$$

Combining the three cases yields:

$$\begin{aligned}
\mathrm{E}\|\widehat{\theta}_{\text{lava}} - \theta\|_2^2 &\leq (1-k)p\sigma^2 + \frac{4\sigma^2 p}{\sqrt{2\pi}p^{1/16}} + (3\sigma^2 + (5L^2+0.5)\lambda_1^2 + \frac{9\sigma}{\sqrt{2\pi}} p^{-1/45})s \\
&\leq (1-k)p\sigma^2 + \frac{4\sigma^2 p}{\sqrt{2\pi}p^{1/16}} + (57L^2\sigma^2 \log p + \frac{9\sigma}{\sqrt{2\pi}} p^{-1/45})s
\end{aligned}$$

where we used $0.5 < 2L^2$ and $\lambda_1^2 \leq 8\sigma^2 \log p$. Since $\mathrm{E}\|Z-\theta\|_2^2 = p\sigma^2$, we have

$$\begin{aligned}
\frac{\mathrm{E}\|\widehat{\theta}_{\text{lava}}(Z) - \theta\|_2^2}{\mathrm{E}\|Z-\theta\|_2^2} &\leq 1 - k + \frac{4}{\sqrt{2\pi}p^{1/16}} + 15(1+M^2/\sigma^2)^2 \frac{s\log p}{p} + \frac{9}{\sqrt{2\pi}\sigma} \frac{s}{p^{1+1/45}} \\
&\leq 1 - k + \frac{4}{\sqrt{2\pi}p^{1/16}} + f(M,\sigma)(\log p + p^{-1/45})\frac{s}{p} \\
&\leq 1 - k + \frac{4}{\sqrt{2\pi}p^{1/16}} + 2f(M,\sigma)\frac{s\log p}{p}. \quad \blacksquare
\end{aligned}$$

A.5. Proof of Theorem 2.2. The first result follows from equation (30) in the proof of Theorem 2.1; the second result follows directly from (12). ■

APPENDIX B: PROOFS FOR SECTION 3

B.1. Proof of Theorem 3.1. Let $\mathsf{Q}_{\lambda_2} = [X'X + n\lambda_2 I_p]$. Then for any $\delta \in \mathbb{R}^p$

$$\begin{aligned}
X\{\widehat{\beta}(\delta) + \delta\} &= X\{\mathsf{Q}_{\lambda_2}^{-1} X'(Y - X\delta) + \delta\} \\
&= \mathsf{P}_{\lambda_2} Y + (I_p - \mathsf{P}_{\lambda_2})X\delta = \mathsf{P}_{\lambda_2} Y + \mathsf{K}_{\lambda_2} X\delta.
\end{aligned}$$

The second claim of the theorem immediately follows from this.

Further, to show the first claim, we can write for any $\delta \in \mathbb{R}^p$,

$$\|Y - X\widehat{\beta}(\delta) - X\delta\|_2^2 = \|(I_n - \mathsf{P}_{\lambda_2})(Y - X\delta)\|_2^2 = \|\mathsf{K}_{\lambda_2}(Y - X\delta)\|_2^2,$$
$$n\lambda_2\|\widehat{\beta}(\delta)\|_2^2 = n\lambda_2\|\mathsf{Q}_{\lambda_2}^{-1} X'(Y - X\delta)\|_2^2.$$

The sum of these terms is equal to

$$(Y - X\delta)'[\mathsf{K}_{\lambda_2}^2 + n\lambda_2 X \mathsf{Q}_{\lambda_2}^{-1} \mathsf{Q}_{\lambda_2}^{-1} X'](Y - X\delta) = \|\mathsf{K}_{\lambda_2}^{1/2}(Y - X\delta)\|_2^2,$$

where the equality follows from the observation that, since $\mathsf{K}_{\lambda_2}^2 = I_n - 2X\mathsf{Q}_{\lambda_2}^{-1} X' + X\mathsf{Q}_{\lambda_2}^{-1} X'X\mathsf{Q}_{\lambda_2}^{-1} X'$ and $[X'X + n\lambda_2 I_p]\mathsf{Q}_{\lambda_2}^{-1} = I_p$, we have

$$\begin{aligned}
\mathsf{K}_{\lambda_2}^2 + n\lambda_2 X \mathsf{Q}_{\lambda_2}^{-1} \mathsf{Q}_{\lambda_2}^{-1} X' &= I_n - 2X\mathsf{Q}_{\lambda_2}^{-1} X' + X\mathsf{Q}_{\lambda_2}^{-1}[X'X + n\lambda_2 I_p]\mathsf{Q}_{\lambda_2}^{-1} X' \\
&= I_n - X\mathsf{Q}_{\lambda_2}^{-1} X' = I - \mathsf{P}_{\lambda_2} = \mathsf{K}_{\lambda_2}.
\end{aligned}$$

Therefore, after multiplying by n, the profiled objective function in (15) can be expressed as: $\|\mathsf{K}_{\lambda_2}^{1/2}(Y - X\delta)\|_2^2 + n\lambda_1\|\delta\|_1$. This establishes the first claim. ■

B.2. Proof of Theorem 3.2. Consider the following lasso problem:

$$h_\lambda(\widetilde{y}) := \arg\min_{\delta\in\mathbb{R}^p}\left\{\frac{1}{n}\|\widetilde{y} - \mathsf{K}_{\lambda_2}^{1/2} X\delta\|_2^2 + \lambda\|\delta\|_1\right\}.$$

Let $g_\lambda(\widetilde{y}, X) := \widetilde{X}h_\lambda(\widetilde{y})$, where $\widetilde{X} := K_{\lambda_2}^{1/2}X$. By Lemmas 1, 3 and 6 of [38], $y \mapsto g_{\lambda_1}(y, X)$ is continuous and almost differentiable, and $\partial g_{\lambda_1}(\widetilde{y}, X)/\partial\widetilde{y} = \widetilde{X}_{\hat{J}}(\widetilde{X}'_{\hat{J}}\widetilde{X}_{\hat{J}})^{-}\widetilde{X}'_{\hat{J}}$. Then by Theorem 3.1, $\widetilde{X}d_{\text{lava}}(y, X) = \widetilde{X}h_{\lambda_1}(K_{\lambda_2}^{1/2}y) = g_{\lambda_1}(K_{\lambda_2}^{1/2}y, X)$. Therefore,

$$\nabla_y\cdot(\mathsf{K}_{\lambda_2} Xd_{\text{lava}}(y, X)) \;=\; \operatorname{tr}\left(\mathsf{K}_{\lambda_2}^{1/2}\frac{\partial g_{\lambda_1}(\mathsf{K}_{\lambda_2}^{1/2} y, X)}{\partial y}\right) = \operatorname{tr}(\mathsf{K}_{\lambda_2}^{1/2}\widetilde{X}_{\hat{J}}(\widetilde{X}'_{\hat{J}}\widetilde{X}_{\hat{J}})^{-}\widetilde{X}'_{\hat{J}}\mathsf{K}_{\lambda_2}^{1/2}).$$

It follows from (20) that

$$\begin{aligned}
\mathrm{df}(\widehat{\theta}) &= \operatorname{tr}(\mathsf{P}_{\lambda_2}) + \mathrm{E}\operatorname{tr}(\mathsf{K}_{\lambda_2}^{1/2}\widetilde{X}_{\hat{J}}(\widetilde{X}'_{\hat{J}}\widetilde{X}_{\hat{J}})^{-}\widetilde{X}'_{\hat{J}}\mathsf{K}_{\lambda_2}^{1/2})\\
&= \operatorname{tr}(\mathsf{P}_{\lambda_2}) + \mathrm{E}\operatorname{tr}(\widetilde{X}_{\hat{J}}(\widetilde{X}'_{\hat{J}}\widetilde{X}_{\hat{J}})^{-}\widetilde{X}'_{\hat{J}}(I-\mathsf{P}_{\lambda_2}))\\
&= \operatorname{tr}(\mathsf{P}_{\lambda_2}) + \mathrm{E}\operatorname{tr}(\widetilde{X}_{\hat{J}}(\widetilde{X}'_{\hat{J}}\widetilde{X}_{\hat{J}})^{-}\widetilde{X}'_{\hat{J}}) - \mathrm{E}\operatorname{tr}(\widetilde{X}_{\hat{J}}(\widetilde{X}'_{\hat{J}}\widetilde{X}_{\hat{J}})^{-}\widetilde{X}'_{\hat{J}}\mathsf{P}_{\lambda_2})\\
&= \mathrm{E}\operatorname{rank}(\widetilde{X}_{\hat{J}}) + \mathrm{E}\operatorname{tr}(\widetilde{\mathsf{K}}_{\hat{J}}\mathsf{P}_{\lambda_2}). \quad\blacksquare
\end{aligned}$$

B.3. Proof of Lemma 3.1. Note that $X\widehat{\theta}_{\text{lava}} + \mathsf{P}_{\hat{J}}\widehat{U} = \mathsf{P}_{\hat{J}}Y + \mathsf{K}_{\hat{J}}X\widehat{\theta}_{\text{lava}} = \mathsf{P}_{\hat{J}}Y + \mathsf{K}_{\hat{J}}X\widehat{\beta} + \mathsf{K}_{\hat{J}}X\widehat{\delta}$ and $X\widetilde{\theta}_{\text{post-lava}} = \mathsf{P}_{\hat{J}}Y + \mathsf{K}_{\hat{J}}X\widehat{\beta}$. Hence it suffices to show that $\mathsf{K}_{\hat{J}}X\widehat{\delta} = 0$. In fact, let $\widehat{\delta}_{\hat{J}}$ be the vector of zero components of $\widehat{\delta}$, then $X\widehat{\delta} = X_{\hat{J}}\widehat{\delta}_{\hat{J}}$. So $\mathsf{K}_{\hat{J}}X\widehat{\delta} = \mathsf{K}_{\hat{J}}X_{\hat{J}}\widehat{\delta}_{\hat{J}} = 0$ since $\mathsf{K}_{\hat{J}}X_{\hat{J}} = 0$. ∎

B.4. Proof of Theorem 3.3. **Step 1.** By (17),

$$\begin{aligned}
\frac{1}{n}\|X\widehat{\theta}_{\text{lava}} - X\theta_0\|_2^2 &\le \frac{2}{n}\|\mathsf{K}_{\lambda_2}X(\widehat{\delta}-\delta_0)\|_2^2 + \frac{2}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|_2^2\\
&\le \frac{2}{n}\|\mathsf{K}_{\lambda_2}^{1/2}X(\widehat{\delta}-\delta_0)\|_2^2\|\mathsf{K}_{\lambda_2}\| + \frac{2}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|_2^2,
\end{aligned}$$

since $\|\mathsf{K}_{\lambda_2}\| \le 1$ as shown below. Step 2 provides the bound $(B_1(\delta_0)\vee B_2(\beta_0))\|\mathsf{K}_{\lambda_2}\|$ for the first term, and Step 3 provides the bound $B_3 + B_4(\beta_0)$ on the second term.

Furthermore, since $X'\mathsf{K}_{\lambda_2}X = n\lambda_2 S(S+\lambda_2 I)^{-1}$, we have

$$B_2(\beta_0) = \frac{8}{n}\|\widetilde{X}\beta_0\|_2^2 = \frac{8}{n}\beta_0'X'\mathsf{K}_{\lambda_2}X\beta_0 = 8\lambda_2\beta_0'S(S+\lambda_2 I)^{-1}\beta_0.$$

Also, to show that $\|\mathsf{K}_{\lambda_2}\|_2 \le 1$, we let $\mathsf{P}_{\lambda_2} = U_1D_1U_1'$ be the eigen-decomposition of P_{λ_2}, then $\|\mathsf{K}_{\lambda_2}\| = \|U_1(I-D_1)U_1'\| = \|I-D_1\|$. Note that all the nonzero eigenvalues of D_1 are the same as those of $(X'X+n\lambda_2 I)^{-1/2}X'X(X'X+n\lambda_2 I)^{-1/2}$, and are $\{d_j/(d_j+n\lambda_2), j\le\min\{n,p\}\}$, where d_j is the jth largest eigenvalue of $X'X$. Thus $\|I-D_1\| = \max\{\max_j n\lambda_2/(d_j+n\lambda_2), 1\} \le 1$.

Combining these bounds yields the result.

Step 2. Here we claim that on the event $\|\frac{2}{n}\widetilde{X}'\widetilde{U}\|_\infty \le c^{-1}\lambda_1$, which holds with probability $1-\alpha$, we have

$$\frac{1}{n}\|\widetilde{X}(\widehat{\delta}-\delta_0)\|_2^2 \le \frac{4\lambda_1^2}{\iota^2(c,\delta_0,\lambda_1,\lambda_2)} \vee \frac{4^2\|\widetilde{X}\beta_0\|_2^2}{n} = B_1(\delta_0)\vee B_2(\beta_0).$$

By (16), for any $\delta\in\mathbb{R}^p$,

$$\frac{1}{n}\|\widetilde{Y}-\widetilde{X}\widehat{\delta}\|_2^2 + \lambda_1\|\widehat{\delta}\|_1 \le \frac{1}{n}\|\widetilde{Y}-\widetilde{X}\delta_0\|_2^2 + \lambda_1\|\delta_0\|_1.$$

Note that $\widetilde{Y}=\widetilde{X}\delta_0+\widetilde{U}+\widetilde{X}\beta_0$, which implies the following basic inequality: for $\Delta=\widehat{\delta}-\delta_0$, on the event $\|\frac{2}{n}\widetilde{X}'\widetilde{U}\|_\infty \le c^{-1}\lambda_1$,

$$\begin{aligned}\frac{1}{n}\|\widetilde{X}\Delta\|_2^2 &\le \lambda_1\left(\|\delta_0\|_1-\|\delta_0+\Delta\|_1+\left|\frac{2}{n}\Delta\widetilde{X}'\widetilde{U}\right|\right)+2\left|\frac{1}{n}(\widetilde{X}\Delta)'(\widetilde{X}\beta_0)\right|\\ &\le \lambda_1\left(\|\delta_0\|_1-\|\delta_0+\Delta\|_1+c^{-1}\lambda\|\Delta\|_1\right)+2\left\|\frac{1}{\sqrt{n}}\widetilde{X}\Delta\right\|_2\left\|\frac{1}{\sqrt{n}}\widetilde{X}\beta_0\right\|_2,\end{aligned}$$

or, equivalently,

$$\frac{1}{n}\|\widetilde{X}\Delta\|_2^2\left(1-\frac{2\left\|\frac{1}{\sqrt{n}}\widetilde{X}\beta_0\right\|_2}{\left\|\frac{1}{\sqrt{n}}\widetilde{X}\Delta\right\|_2}\right)\le\lambda_1\Big(\|\delta_0\|_1-\|\delta_0+\Delta\|_1+c^{-1}\lambda\|\Delta\|_1\Big).$$

If $\left\|\frac{1}{\sqrt{n}}\widetilde{X}\Delta\right\|_2 \le 4\left\|\frac{1}{\sqrt{n}}\widetilde{X}\beta_0\right\|_2$, then we are done. Otherwise we have that

$$\frac{1}{n}\|\widetilde{X}\Delta\|_2^2 \le 2\lambda_1\Big(\|\delta_0\|_1-\|\delta_0+\Delta\|_1+c^{-1}\|\Delta\|_1\Big).$$

Thus $\Delta\in\mathcal{R}(c,\delta_0,\lambda_1,\lambda_2)$ and hence by the definition of the design-impact factor

$$\frac{1}{n}\|\widetilde{X}\Delta\|_2^2 \le 2\lambda_1\frac{\frac{1}{\sqrt{n}}\|\widetilde{X}\Delta\|_2}{\iota(c,\delta_0,\lambda_1,\lambda_2)} \implies \frac{1}{\sqrt{n}}\|\widetilde{X}\Delta\|_2 \le \frac{2\lambda_1}{\iota(c,\delta_0,\lambda_1,\lambda_2)}.$$

Combining the two cases yields the claim.

Step 3. Here we bound $\frac{2}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|_2^2$. We have

$$\frac{2}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|_2^2 \le \frac{4}{n}\|\,\mathsf{K}_{\lambda_2}X\beta_0\|_2^2+\frac{4}{n}\|\,\mathsf{P}_{\lambda_2}U\|_2^2.$$

By [25]'s exponential inequality for deviation of quadratic form of sub-Gaussian vectors the following bound applies with probability $1-\epsilon$:

$$\begin{aligned}\frac{4}{n}\|\,\mathsf{P}_{\lambda_2}U\|_2^2 &\le \frac{4\sigma_u^2}{n}[\mathrm{tr}(\mathsf{P}_{\lambda_2}^2)+2\sqrt{\mathrm{tr}(\mathsf{P}_{\lambda_2}^4)\log(1/\epsilon)}+2\|\,\mathsf{P}_{\lambda_2}^2\,\|\log(1/\epsilon)],\\ &\le \frac{4\sigma_u^2}{n}[\mathrm{tr}(\mathsf{P}_{\lambda_2}^2)+2\sqrt{\mathrm{tr}(\mathsf{P}_{\lambda_2}^2)\|\,\mathsf{P}_{\lambda_2}^2\,\|\log(1/\epsilon)}+2\|\,\mathsf{P}_{\lambda_2}^2\,\|\log(1/\epsilon)],\\ &\le \frac{4\sigma_u^2}{n}\left[\sqrt{\mathrm{tr}(\mathsf{P}_{\lambda_2}^2)}+\sqrt{2}\sqrt{\|\,\mathsf{P}_{\lambda_2}^2\|}\sqrt{\log(1/\epsilon)}\right]^2 = B_3,\end{aligned}$$

where the second inequality holds by Von Neumann's theorem ([24]), and the last inequality is elementary.

Furthermore, note that $\mathsf{K}_{\lambda_2} X = \lambda_2 X(S + \lambda_2 I)^{-1}$. Hence

$$B_4(\beta_0) = \frac{4}{n} \| \mathsf{K}_{\lambda_2} X\beta_0\|_2^2 = 4\lambda_2^2\beta_0'(S + \lambda_2 I)^{-1}S(S + \lambda_2 I)^{-1}\beta_0 = 4\beta_0' V_{\lambda_2}\beta_0. \quad \blacksquare$$

REFERENCES

[1] Banerjee, A., Chen, S., Fazayeli, F. and Sivakumar, V. (2015). Estimation with norm regularization. Tech. rep.
[2] Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19** 521–547.
[3] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806.
[4] Belloni, A., Chernozhukov, V. and Wang, L. (2014). Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics* **42** 757–788.
[5] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37** 1705–1732.
[6] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics* **1** 169–194.
[7] Bunea, F., Tsybakov, A. B., Wegkamp, M. H. and Barbu, A. (2010). Spades and mixture models. *The Annals of Statistics* **38** 2525–2558.
[8] Cai, T. T., Liang, T. and Rakhlin, A. (2015). Geometric inference for general high-dimensional linear inverse problems. Tech. rep.
[9] Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35** 2313–2351.
[10] Candès, E. J., Li, X., Ma, Y. and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)* **58** 11.
[11] Chandrasekaran, V., Recht, B., Parrilo, P. A. and Willsky, A. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics* **12** 805–849.
[12] Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization* **21** 572–596.
[13] Chen, Y. and Dalalyan, A. (2012). Fused sparsity and robust estimation for linear models with unknown variance. *Advances in Neural Information Processing Systems* 1259–1267.
[14] Chernozhukov, V., Hansen, C. and Liao, Y. (2015). A lava attack on the recovery of sums of dense and sparse signals. Tech. rep., MIT.
[15] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90** 1200–1224.
[16] Dossal, C., Kachour, M., Fadili, J. M., Peyré, G. and Chesneau, C. (2011). The degrees of freedom of the lasso for general design matrix. *arXiv preprint:1111.1162* .
[17] Efron, B. (2004). The estimation of prediction error. *Journal of the American Statistical Association* **99** 619–642.
[18] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics* **32** 407–499.
[19] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.
[20] Fan, J., Liao, Y. and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements (with discussion). *Journal of the Royal Statistical Society, Series B* **75** 603–680.
[21] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (with discussion)* **70** 849–911.
[22] Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.
[23] Hirano, K. and Porter, J. R. (2012). Impossibility results for nondifferentiable functionals. *Econometrica* **80** 1769–1790.
[24] Horn, R. A. and Johnson, C. R. (2012). *Matrix analysis.* Cambridge university press.
[25] Hsu, D., Kakade, S. M. and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics* **14** 569–600.

[26] Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Statistics* **53** 73101.
[27] Jalali, A., Ravikumar, P. and Sanghavi, S. (2013). A dirty model for multiple sparse regression. *IEEE Transactions on Information Theory* **59** 7947–7968.
[28] Klopp, O., Lounici, K. and Tsybakov, A. B. (2014). Robust matrix completion. *arXiv preprint arXiv:1412.8132* .
[29] Loh, P.-L. and Wainwright, M. J. (2013). Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*.
[30] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics* **37** 246–270.
[31] Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *The Annals of Statistics* **28** 1083–1104.
[32] Negahban, S., Yu, B., Wainwright, M. J. and Ravikumar, P. K. (2009). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*.
[33] Stein, C. M. (1956). Inadmissibility of the usual estimator for the mean of a multivariate distribution. *Proc. Third Berkeley Symp. Math. Statist. Prob.* 197–206.
[34] Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* **9** 1135–1151.
[35] Sun, T. and Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99** 879–898.
[36] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288.
[37] Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108.
[38] Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40** 1198–1232.
[39] Wainwright, M. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory* **55** 2183–2202.
[40] Yang, E. and Ravikumar, P. K. (2013). Dirty statistical models. In *Advances in Neural Information Processing Systems 26* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Weinberger, eds.). Curran Associates, Inc., 611–619.
URL http://papers.nips.cc/paper/5092-dirty-statistical-models.pdf
[41] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68** 49–67.
[42] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894–942.
[43] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7** 2541–2563.
[44] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 301–320.

Department of Economics,
Massachusetts Institute of Technology,
Cambridge, MA 02139, USA.
E-mail: vchern@mit.edu

Booth School of Business,
University of Chicago,
Chicago, IL 60637, USA.
E-mail: Christian.Hansen@chicagobooth.edu

Department of Mathematics,
University of Maryland.
College Park, MD 20742, USA.
E-mail: yuanliao@umd.edu