Econometric Theory, **39**, 2023, 189–218.
doi:10.1017/S0266466621000347

STRETCHING THE NET: MULTIDIMENSIONAL REGULARIZATION

JAUME VIVES-I-BASTIDA
Massachusetts Institute of Technology

This paper derives asymptotic risk (expected loss) results for shrinkage estimators with multidimensional regularization in high-dimensional settings. We introduce a class of multidimensional shrinkage estimators (MuSEs), which includes the elastic net, and show that—as the number of parameters to estimate grows—the empirical loss converges to the oracle-optimal risk. This result holds when the regularization parameters are estimated empirically via cross-validation or Stein's unbiased risk estimate. To help guide applied researchers in their choice of estimator, we compare the empirical Bayes risk of the lasso, ridge, and elastic net in a spike and normal setting. Of the three estimators, we find that the elastic net performs best when the data are moderately sparse and the lasso performs best when the data are highly sparse. Our analysis suggests that applied researchers who are unsure about the level of sparsity in their data might benefit from using MuSEs such as the elastic net. We exploit these insights to propose a new estimator, the *cubic net*, and demonstrate through simulations that it outperforms the three other estimators for any sparsity level.

1. INTRODUCTION

Estimation problems that involve a large number of unknown parameters have become increasingly common in economics. Applied researchers with large datasets at their disposal are focusing on empirical questions that require the estimation of many treatment effects—often with many subgroups—and prediction problems with many covariates. A typical approach to dealing with these high-dimensional problems is to use methods from the machine learning literature. Yet given the wide range of available methods, choosing the right one is not always straightforward. In the economics literature, for example, considerable attention has been given to regularization-based methods with a single regularization parameter; examples include the lasso and ridge methods.

I thank Rachael Meager for invaluable guidance provided throughout the supervision of this paper at the London School of Economics, and Alberto Abadie, Maximilian Kasy, and Matthew Levy for giving insightful comments and feedback. I also thank David Card for starting me on this line of research. A co-editor of this journal and three anonymous referees provided very useful comments that contributed to improve this paper. Address correspondence to Jaume Vives-i-Bastida, MIT Department of Economics, 50 Memorial Dr, Cambridge, MA 02142, USA; e-mail: vives@mit.edu.

Abadie and Kasy (2019; hereafter AK19) provide some guidance on how to choose among such regularization methods. Our study offers guidance for methods, such as the elastic net (Zou and Hastie, 2005), with *multiple* regularization parameters.

Regularization through loss function penalization is a straightforward way of improving the mean squared error (MSE) of an estimator and thereby helping to prevent overfitting. The success of these methods relies on the data-driven choice of the regularization parameters that control the amount of shrinkage on the estimated parameters. Therefore, the choice of loss function penalization and the procedure used to choose the regularization parameters are key for the success of regularization-based methods. The chief takeaways from AK19 are that, among methods that use only one type of regularization (e.g., lasso, ridge, and pretest): (a) no method universally dominates the others, in terms of risk; and (b) under mild conditions, data-driven choices of the regularization parameter that use Stein's unbiased risk estimator or cross-validation (CV) are guaranteed to work well in high-dimensional settings. More precisely, AK19's main theoretical result ensures that the risk function of an estimator evaluated at a data-driven choice of the regularization parameter is uniformly close to the risk function of the infeasible estimator using the oracle-optimal regularization parameter, the regularization parameter that minimizes the true risk.

1.1. Contribution

There are many applied settings in which the researcher is unsure whether the parameters being estimated come from a sparse or a dense data distribution. In these cases, the researcher might want to minimize the mean squared error by using a more flexible regularization-based method with multiple types of regularization. This paper presents guidelines concerning the choice of regularization-based method for high-dimensional estimation problems. First, we show that the AK19 result on the uniform validity of a data-driven approach to selecting the regularization parameter extends, under mild conditions, to estimators with multiple regularization parameters. Second, we compare the elastic net, lasso, and ridge methods in both a spike and normal setting and establish that—in terms of risk—the elastic net performs well for many levels of sparsity. Finally, we propose a new and better-performing estimator, the *cubic net*, and use Monte Carlo simulations to demonstrate that it outperforms the lasso, ridge, and elastic net methods.

1.2. Setup

As in AK19, we focus on the problem of estimating the unknown means $\mu_1, \dots, \mu_p$ of observed random variables $X_1, \dots, X_p$, where the number p of parameters can be large. This setting is quite flexible and is applicable also to more general prediction and estimation problems, including multivariate regression under a suitable transformation. In these extended settings, our results apply only to

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

those cases in which the sample size exceeds the number of parameters.[1] We consider componentwise estimators of the form $\hat{\mu}_i = m(X_i, \boldsymbol{\lambda})$, with the difference that we allow the regularization parameter vector $\boldsymbol{\lambda}$ to be multidimensional. It is intuitive that $\boldsymbol{\lambda}$ controls the amount of shrinkage, with higher values of the components shrinking the estimates. Such shrinkage estimators are important in high-dimensional settings, because they have been shown to outperform the maximum likelihood estimator when $p \geq 3$ (Stein, 1956). Two reasonable assumptions are that (i) if $\boldsymbol{\lambda} = \mathbf{0}$, then the estimates remain unregularized, $m(x, \mathbf{0}) = x$, and (ii) at the limit $\boldsymbol{\lambda} = \boldsymbol{\infty}$, where $\boldsymbol{\infty}$ is a vector where all elements are ∞, the estimates fully shrink to zero. Our results apply to componentwise estimators that satisfy these assumptions along with a monotonicity assumption. We refer to these estimators as *multidimensional shrinkage estimators* (MuSEs). Many commonly used shrinkage estimators are MuSEs—not only the aforementioned lasso, ridge, and elastic net, but also the adaptive lasso (Zou, 2006), scad (Fan and Li, 2001), and lava (Chernozhukov et al. 2017) estimators.

1.3. Theoretical Results

The primary object of interest in this study is the risk function for componentwise estimators. To define that function, we consider a setting in which (X_i, μ_i) are realizations of the random variables (X, μ) with a joint distribution π. Our notion of risk in this setup is the integrated (or empirical Bayes) risk, defined as the average componentwise loss of the estimator integrated over π. In the case of the squared loss, this notion of risk can be understood as the expected MSE. We cannot compute such risk directly, in practice, because the underlying distribution π may be unknown; hence, we rely on the empirical counterparts—namely the compound loss and the compound risk (i.e., average loss and risk across components). According to our results, as the number p of parameters grows: (i) the differences between the compound loss, compound risk, and integrated risk vanish uniformly; (ii) the integrated risk function is uniformly close to the integrated risk function evaluated at the oracle-optimal regularization parameters; and (iii) we can use data-driven methods such as CV and Stein's unbiased risk estimator (a.k.a. *Sure*) to estimate the risk function consistently. These results for MuSEs are analogous to the AK19 results for estimators with scalar regularization.

1.4. Spike and Normal Risk

To study the performance of regularization-based methods with multiple regularization parameters, we derive and compare the integrated risk functions of the elastic net, ridge, and lasso when π follows a spike and normal distribution. This parametric assumption allows us to evaluate how the various methods fare under

[1] In Section A.1 in the Appendix, we discuss one of AK19's examples that shows how to transform this model for multivariate regression and why we require the sample size to be larger than the number of parameters.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

different sparsity settings. Our analysis is in line with AK19 and the literature: the ridge estimator performs best in nonsparse settings and the lasso performs best in very sparse settings. The elastic net is better for “in-between” scenarios, prompting our recommendation that it should be used when the researcher is unsure about the data’s sparsity. It is interesting that, according to our results, the ridge estimator should rarely be used, because that method is nearly always outperformed by the elastic net. We complement this analysis by comparing the cross-validated MSEs of each method for a prediction problem with a varying number of parameters. For estimation problems with a small number of parameters, we observe that the lasso and ridge methods perform better; but as the problem’s dimensionality increases, the elastic net matches their performance. Finally, we introduce the cubic net and use simulations to show that it outperforms lasso, ridge and, elastic net methods irrespective of the setting’s sparsity.

1.5. Overview

Section 2 sets up the theoretical framework and defines the MuSEs. Our paper’s main theoretical results are presented in Section 3. In Section 4, we derive the risk formulas for the normal, spike, and slab distributions and then compare the estimators by way of simulations. Section 5 introduces the cubic net estimator and evaluates its performance via Monte Carlo simulations. We conclude in Section 6 with a brief summary and suggestions for further research. All proofs are relegated to the Appendix.

2. SETUP

We consider the problem of estimating a p-dimensional vector $\boldsymbol{\mu}$ from a random vector $\boldsymbol{X}$. Each component of $\boldsymbol{\mu}$ can be understood as a parameter to be estimated in a high-dimensional problem. This setting can be extended to more general prediction and estimation problems as shown in Section A.1 in the Appendix. In particular, we consider

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix},$$

where the components of $\boldsymbol{X}$ are mutually independent and $X_i \sim P_i$. Each distribution P_i has finite mean μ_i, and we denote the vector of distributions by $\boldsymbol{P} = \left[P_1, \ldots, P_p\right]$.

In this paper, we focus on componentwise estimators of μ_i with multidimensional regularization. We define a componentwise estimator as a function $m\colon \mathbb{R} \times [0,\infty]^k \to \mathbb{R}$ that depends on X_i and on k nonnegative regularization parameters denoted by the vector $\boldsymbol{\lambda}$; hence, our estimate of each parameter will be given by $\hat{\mu}_i = m(X_i, \boldsymbol{\lambda})$. A crucial observation is that our estimate $\hat{\mu}_i$ will depend not just

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

directly on X_i, but also indirectly on $\boldsymbol{X}$—through the common regularization vector $\boldsymbol{\lambda}$ when that vector is estimated from $\boldsymbol{X}$ by a data-driven procedure.

2.1. MuSE Estimators

Many of the regularization-based estimators used, in practice, can be formulated as component-wise estimators. In this paper, we focus on a large class of componentwise estimators that we refer to as MuSEs.

Definition 1 (MuSE)**.** A componentwise estimator $m\colon \mathbb{R}\times[0,\infty]^k \to \mathbb{R}$ is an MuSE if it satisfies the following conditions:

(i) $m(x,\boldsymbol{\lambda})$ is monotonic in all components of $\boldsymbol{\lambda}$ for all $x \in \mathbb{R}$; and
(ii) $m(x,\mathbf{0}) = x$ and $\lim_{\boldsymbol{\lambda}\to\boldsymbol{\infty}} m(x,\boldsymbol{\lambda}) = 0$ for all $x \in \mathbb{R}$.

This definition ensures that any MuSE has two desirable properties. First, increasing any of the regularization parameters must lead to more shrinkage. Second, maximal shrinkage (to zero) can be achieved as we increase the regularization parameters without bound; in the absence of regularization, there is no shrinkage and the estimate is simply x (i.e., the maximum likelihood estimate).

Some of the most commonly used regularization-based methods satisfy the MuSE conditions. In this paper, we focus on the lasso, ridge, elastic net, and lava methods, which are characterized by the following componentwise estimator functions:

$$\begin{aligned}
m_{\text{ridge}}(x,\lambda) &= \operatorname*{argmin}_{m\in\mathbb{R}}\{(x-m)^2+\lambda m^2\} && \text{(ridge)}\\
&= \frac{1}{1+\lambda}x;\\
m_{\text{lasso}}(x,\lambda) &= \operatorname*{argmin}_{m\in\mathbb{R}}\{(x-m)^2+2\lambda|m|\} && \text{(lasso)}\\
&= \mathbb{1}(x<-\lambda)(x+\lambda)+\mathbb{1}(x>\lambda)(x-\lambda);\\
m_{\text{EN}}(x,\boldsymbol{\lambda}) &= \operatorname*{argmin}_{m\in\mathbb{R}}\{(x-m)^2+2\lambda_1|m|+\lambda_2 m^2\} && \text{(elastic net)}\\
&= \mathbb{1}(x<-\lambda_1)\frac{x+\lambda_1}{1+\lambda_2}+\mathbb{1}(x>\lambda_1)\frac{x-\lambda_1}{1+\lambda_2};\\
m_{\text{lava}}(x,\boldsymbol{\lambda}) &= \operatorname*{argmin}_{(\delta,\beta)\in\mathbb{R}^2}\{(x-\beta-\delta)^2+\lambda_1|\delta|+\lambda_2\beta^2\} && \text{(lava)}\\
&= \mathbb{1}\left(x>\frac{\lambda_1}{2k}\right)\left(x-\frac{\lambda_1}{2}\right)+\mathbb{1}\left(-\frac{\lambda_1}{2k}\le x\le\frac{\lambda_1}{2k}\right)(1-k)x\\
&\quad+\mathbb{1}\left(x<-\frac{\lambda_1}{2k}\right)\left(x+\frac{\lambda_1}{2}\right).
\end{aligned}$$

In the last equality, $k=\lambda_2/(1+\lambda_2)$.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

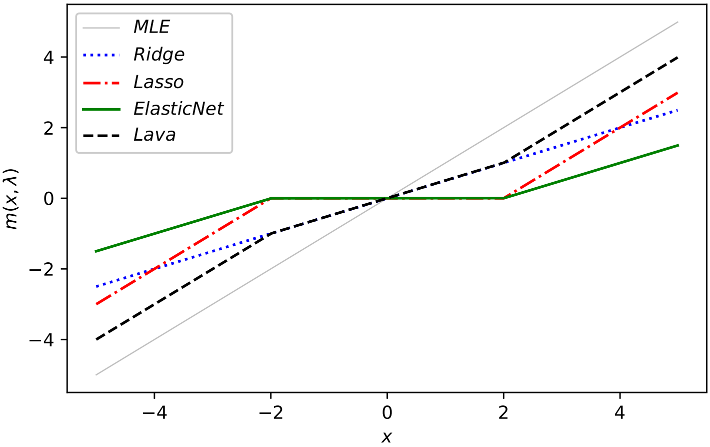

Figure 1. Componentwise estimator functions for the MLE, ridge, lasso, elastic net, and lava. For ridge, lasso, elastic net, and lava, the regularization parameters are set to $\lambda_1 = 1$ and $\lambda_2 = 2$. A color version of this figure can be found in the online Appendix.

To develop a full appreciation of the different shrinkage types induced by these estimators, in Figure 1, we plot the functions for regularization parameters $\lambda_1 = 1$ and $\lambda_2 = 2$. In this figure, the 45° line represents the unregularized estimate, which coincides with the maximum likelihood estimate. The regularized estimators either shrink the estimates toward the maximum likelihood estimate, as does the ridge estimator (dotted line), or—like the lasso estimator (dot-dashed line)—induce sparsity by setting the estimate to zero in a region. The appeal of the elastic net (solid line) is that it applies both types of shrinkage; thus, it sets the estimate to zero in the same region as the lasso but otherwise behaves like the ridge. In contrast, the lava estimator (dashed line) applies ridge-like shrinkage only in the area where the lasso shrinks estimates to zero. In Section 4, we focus on the ridge, lasso, and elastic net methods in order to assess the trade-off between shrinking all estimates uniformly and inducing sparseness.

2.2. Measuring Risk

We are interested in the efficiency of componentwise estimators as we increase the number p of parameters. Given a realization of the vector $\boldsymbol{X}$, a standard measure of efficiency is the compound squared loss:

$$\begin{aligned} L_p(\boldsymbol{X}, m(\cdot,\boldsymbol{\lambda}), \boldsymbol{P}) &= \|m(\boldsymbol{X},\boldsymbol{\lambda}) - \boldsymbol{\mu}\|^2 \\ &= \frac{1}{p}\sum_{i=1}^{p}(m(X_i,\boldsymbol{\lambda}) - \mu_i)^2. \end{aligned}$$

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

Given the component distribution $\boldsymbol{P}$, we define the compound risk R_p as the average expected loss over the components:[2]

$$\begin{aligned} R_p(m(\cdot,\boldsymbol{\lambda}),\boldsymbol{P}) &= \mathbb{E}[L_p(\boldsymbol{X},m(\cdot,\boldsymbol{\lambda}),\boldsymbol{P})|\boldsymbol{P}] \\ &= \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}[(m(X_i,\boldsymbol{\lambda})-\mu_i)^2|P_i] \\ &= \frac{1}{p}\sum_{i=1}^{p}\int (m(X_i,\boldsymbol{\lambda})-\mu_i)^2\,dP_i. \end{aligned}$$

One of our goals is to show that, as the number of parameters grows, the expected loss of our estimator converges to the risk function of the estimator when the underlying distribution $\boldsymbol{P}$ is known. A notion of risk that captures this idea is the empirical Bayes risk (Robbins, 1956), otherwise known as integrated risk. We follow AK19 and define integrated risk by considering $P_1,\ldots,P_p$ to be draws from an underlying distribution $\Pi \in Q$, for a set of distributions Q with bounded fourth moments, that induces a joint distribution π for (X_i,μ_i). In this setting, the integrated risk $\bar{r}_\pi$ refers to the expected risk over π:

$$\begin{aligned} \bar{r}_\pi(m(\cdot,\boldsymbol{\lambda})) &= \mathbb{E}_\pi[L_p(\boldsymbol{X},m(\cdot,\boldsymbol{\lambda}),\boldsymbol{P})] \\ &= \iint (m(X_i,\boldsymbol{\lambda})-\mu_i)^2\,dP_i\,d\Pi(P_i). \end{aligned}$$

It is worth noting that the setting we consider differs from the one adopted in other analyses of shrinkage estimators (e.g., Leeb and Potscher, 2006; Jia and Yu, 2010), which deliver negative consistency results. First of all, we are interested in the integrated risk of our estimator and not in the consistency of our parameter estimation. In the second place, as the number of components increases, we get a larger sample of the generating distribution Π that is common to all tuples (X_i,μ_i); this allows for better estimates of the integrated risk.

2.3. Optimal Regularization

The risk function depends on the regularization vector $\boldsymbol{\lambda}$. The choice of regularization is extremely important in minimizing risk. As shown by James and Stein (1961) and Morris (1983), for settings with at least three parameters ($p \geq 3$), the estimator that minimizes risk is a shrinkage estimator. With this in mind and given $\boldsymbol{P}$, we define the so-called *oracle-selector* regularization parameter as

$$\boldsymbol{\lambda}^*(\boldsymbol{P}) = \underset{\boldsymbol{\lambda}\in[0,\infty]^k}{\operatorname{argmin}}\, R_p(m(\cdot,\boldsymbol{\lambda}),\boldsymbol{P}).$$

[2] Section A.2 in the Appendix explains why we use the compound risk rather than the minimax criterion.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

The population-level oracle-selector is similarly given by

$$\bar{\boldsymbol{\lambda}}^*(\pi) = \underset{\boldsymbol{\lambda}\in[0,\infty]^k}{\text{argmin}}\, \bar{r}_\pi(m(\cdot,\boldsymbol{\lambda})).$$

3. THEORETICAL RESULTS

Now, we present our principal theoretical results. We show that, under some regularity conditions and when the regularization parameter $\boldsymbol{\lambda}$ is chosen using a data-driven procedure, the risk function of MuSE estimators is uniformly close to the oracle-optimal risk with respect to the joint distribution π of (X_i,μ_i). The oracle-optimal risk is the integrated risk for a given shrinkage estimator evaluated at the oracle selector.[3] This result is the multidimensional regularization extension of AK19, who establish it for estimators with scalar regularization. The result is useful, because: (a) it confirms that oracle-optimal risk can be achieved using data-driven procedures to hyper-tune the regularization parameter; and (b) it ensures that comparing the oracle-optimal risk is a valid method for evaluating the performance of different MuSEs. In Section 4, we focus on this comparison in a spike and normal setting.

We organize our findings as follows. Theorem 1 states that, as the number of parameters increases, the compound loss converges in L^2 to the integrated risk under the MuSE assumptions, when the fourth moments are bounded, and assuming a technical condition on the componentwise function and the data. These assumptions are satisfied by the MuSEs considered in this paper: the lasso, ridge, and elastic net. Theorems 2 and 3 characterize the uniform convergence of (i) the compound loss to the infeasible minimum loss and (ii) the integrated risk to the oracle-optimal integrated risk when the regularization parameter is chosen to minimize the empirical loss. These two theorems are critical, because, in practice, the regularization parameter is chosen by way of a data-driven procedure that relies on empirical loss minimization. Thus Theorems 2 and 3 ensure that we can, in fact, achieve the oracle-optimal integrated risk as p increases. Finally, in Theorem 4, we establish that CV can be used as a data-driven method to estimate the regularization parameter.

In what follows, we use shorthand notation for the main objects of interest: (i) $L_p(\boldsymbol{\lambda}) \equiv L_p(\boldsymbol{X}, m(\cdot,\boldsymbol{\lambda}),\boldsymbol{P})$, and (ii) $\bar{r}_\pi(\boldsymbol{\lambda}) \equiv \bar{r}_\pi(m(\cdot,\boldsymbol{\lambda}))$.

3.1. Uniform Risk Convergence

Our first result (Theorem 1) states that, as the number p of parameters grows, the compound loss converges (in L^2) to the integrated risk for any MuSE. The intuition behind this result is that, as p increases, we have more observations from

[3] The oracle-optimal risk should not be confused with the optimal risk over a class of alternative feasible shrinkage estimators; the former is, rather, optimal within the class of estimators indexed by the regularization parameter.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

the underlying joint distribution π of (X_i, μ_i); hence, the compound loss becomes a better approximation of the integrated risk over π for any regularization parameter. The practical relevance of this result is to ensure that, in high-dimensional settings, the compound loss of MuSEs will be close to the infeasible integrated risk for a broad set of underlying distributions. In particular, Theorem 1 holds for any underlying distribution with bounded fourth moments and for any MuSE that satisfies a technical restriction on the estimating function m.

THEOREM 1 (General uniform risk convergence). *For an MuSE estimator, suppose that the following assumptions hold.*

(i) $\sup_{\pi\in Q}\mathbb{E}_\pi\left[X^4\right] < \infty$.
(ii) *For all* $i \in \{1,\ldots,k\}$ *and for all* $\varepsilon_i > 0$, *there exists* $0 = \lambda_1^i < \cdots < \lambda_T^i = \infty$ *subject to* $\mathbb{E}_\pi\left[(|X-\mu|+|\mu|)\left|m(X,\boldsymbol{\lambda}_j^i)-m(X,\boldsymbol{\lambda}_{j-1}^i)\right|\right] \leq \varepsilon_i$, *for all* $j \in \{1,\ldots,T\}$ *and for all* $\pi \in Q$.

Then,

$$\sup_{\pi\in Q}\mathbb{E}_\pi\left[\sup_{\boldsymbol{\lambda}\in[0,\infty]^k}(L_p(\boldsymbol{\lambda})-\bar{r}_\pi(\boldsymbol{\lambda}))^2\right]\to 0.$$

The proof of Theorem 1 relies on the two intermediate lemmas given in Sections A.3 and A.4 in the Appendix. The conditions for Theorem 1 apply to the lasso, ridge, and elastic net estimators. For condition (ii), observe that, if we treat one regularization parameter as a constant, the condition is satisfied for the other parameter because the elastic net then behaves like the ridge or the lasso. It follows that condition (ii) holds also for the elastic net.

3.2. Uniform Risk and Loss Consistency

Next, we show that the convergence results hold when we estimate the regularization parameter. Let $\hat{\boldsymbol{\lambda}}_p$ be an estimator of the oracle selector $\bar{\boldsymbol{\lambda}}^*(\pi)$. Theorem 2 stipulates conditions under which the compound loss function converges uniformly over π to the empirical Bayes risk as p increases—provided that the regularization parameter $\hat{\boldsymbol{\lambda}}_p$ is chosen so as to minimize a uniformly consistent estimate of the integrated risk.

THEOREM 2 (General uniform loss consistency). *For all* $\varepsilon > 0$, *assume that*

$$\sup_{\pi\in Q}P_\pi\left(\sup_{\boldsymbol{\lambda}\in[0,\infty]^k}|L_p(\boldsymbol{\lambda})-\bar{r}_\pi(\boldsymbol{\lambda})|>\varepsilon\right)\to 0,$$

and that there exist functions $r^*_\pi(\boldsymbol{\lambda}), c_\pi$, *and* $\hat{r}_p(\boldsymbol{\lambda})$ *such that* $\bar{r}_\pi(\boldsymbol{\lambda}) = r^*_\pi(\boldsymbol{\lambda})+c_\pi$ *and*

$$\sup_{\pi\in Q}P_\pi\left(\sup_{\boldsymbol{\lambda}\in[0,\infty]^k}|\hat{r}_p(\boldsymbol{\lambda})-\bar{r}_\pi(\boldsymbol{\lambda})|>\varepsilon\right)\to 0.$$

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

Then,

$$\sup_{\pi \in Q} P_\pi \left(\left| L_p(\hat{\boldsymbol{\lambda}}_p) - \inf_{\boldsymbol{\lambda} \in [0,\infty]^k} L_p(\boldsymbol{\lambda}) \right| > \varepsilon \right) \to 0,$$

where $\hat{\boldsymbol{\lambda}}_p = \operatorname{argmin}_{\boldsymbol{\lambda} \in [0,\infty]^k} \hat{r}_p(\boldsymbol{\lambda})$.

The proof follows directly from AK19 and Theorem 1, which also gives conditions under which the first assumption in Theorem 2 is satisfied (since L^2 convergence implies convergence in probability). Theorem 3 states that, under the conditions of Theorem 1, we also obtain uniform risk consistency in this sense: the integrated risk evaluated at $\hat{\boldsymbol{\lambda}}_p$ is uniformly close to the oracle-optimal risk as the number of components grows.

THEOREM 3 (General uniform risk consistency). *Under the assumptions of Theorem 1,*

$$\sup_{\pi \in Q} \left| \bar{r}_\pi(m(X, \hat{\boldsymbol{\lambda}}_p), \pi) - \inf_{\boldsymbol{\lambda} \in [0,\infty]^k} \bar{r}_\pi(\boldsymbol{\lambda}) \right| \to 0.$$

3.3. Data-Driven Risk Estimation

Theorems 2 and 3 lead naturally to proposing a data-driven method to estimate the oracle-optimal regularization parameter and thereby obtain a consistent integrated risk estimator. Abadie and Kasy (2019) propose two methods: Stein's unbiased risk estimator, or *Sure*; and CV. The result for *Sure* follows directly from AK19, and, in Section A.7 in the Appendix, we describe it in more detail for the case of elastic net. Next, we focus on a result to choose the tuning parameter using CV.

3.4. Cross-Validation

To derive the result for CV, we consider (as in AK19) a setting in which n observations are available for each parameter. For each $i \in \{1,\dots,p\}$, we draw an independent and identically distributed (i.i.d.) sample $(x_{1i},\dots,x_{ni},\mu_i,\sigma_i)$ from π; here, π is the distribution of the random variable $(x_1,\dots,x_k,\mu,\sigma)$. We assume that, conditional on (X,μ), the samples $(x_1,\dots,x_k)$ are i.i.d. and assume also that, for all $j \in \{1,\dots,n\}$,

$$\mathbb{E}[x_j|\mu,\sigma] = \mu, \qquad \text{var}[x_j|\mu,\sigma] = \sigma^2.$$

We define the compound loss from using $n-1$ observations to estimate the nth observation as follows:

$$L_{p,n}(\boldsymbol{\lambda}) = \frac{1}{p}\sum_{i=1}^{p}(m(X_{n-1i},\boldsymbol{\lambda}) - \mu_i)^2;$$

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

here, $X_{ni} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$ is the data used to estimate the nth term x_{ni} with mean μ_i. Then, the "leave one out" CV estimator is given by

$$r_{p,n}(\boldsymbol{\lambda}) = \frac{1}{p}\sum_{i=1}^{p}(m(X_{n-1i},\boldsymbol{\lambda}) - x_{ni})^2.$$

Theorem 4 shows that this CV estimator converges to the integrated risk as the number of components p increases and n is fixed.

THEOREM 4 (General uniform CV consistency). *For MuSEs and for* $\sup_\pi \mathbb{E}_\pi[x_j^4] < \infty, j \in \{1,\dots,n\}$, *where* $n \geq 2$, *let*

$$\bar{r}_{\pi,k}(\boldsymbol{\lambda}) = \mathbb{E}_\pi[r_{p,n}(\boldsymbol{\lambda})] = \mathbb{E}_\pi[m(X_{n-1i},\boldsymbol{\lambda}) - x_{ni})^2] + \mathbb{E}_\pi[\sigma^2],$$

and let $\hat{\boldsymbol{\lambda}}_p = \operatorname{argmin}_{\boldsymbol{\lambda}\in[0,\infty]^k} r_{p,n}(\boldsymbol{\lambda})$. *Then,*

$$\sup_{\pi\in Q} \mathbb{E}_\pi\left[\sup_{\boldsymbol{\lambda}\in[0,\infty]^k} (r_{p,n}(\boldsymbol{\lambda}) - \bar{r}_{\pi,n}(\boldsymbol{\lambda}))^2\right] \to 0,$$

and

$$\sup_{\pi\in Q} P_\pi\left[\left|L_{p,n}(\hat{\boldsymbol{\lambda}}_p) - \inf_{\boldsymbol{\lambda}\in[0,\infty]^k} L_{p,n}(\boldsymbol{\lambda})\right| > \varepsilon\right] \to 0 \quad \forall\varepsilon > 0.$$

When combined with our previous results, Theorem 4 implies that using an MuSE with data-driven regularization is a valid approach to minimizing expected risk in high-dimensional settings. Following this insight, in the next section, we derive analytical results for the purpose of comparing the integrated risk of MuSEs with the expected MSE obtained through CV—for a spike and normal DGP.

4. COMPARING RIDGE, LASSO, AND ELASTIC NET

In this section, we compare the oracle-optimal integrated risk for the lasso, risk, and elastic net estimators in the spike and normal setting. Doing so allows us to evaluate the performance of the three estimators for data generating processes (DGPs) characterized by varying degrees of sparsity. By describing the scenarios in which each estimator performs best, we provide guidance for applied researchers looking to choose among MuSEs. We complement this analysis by simulating MSEs for a prediction task while changing the number of parameters.

4.1. Spike and Normal Process

To study the performance of regularization-based methods with multiple regularization parameters, we derive and compare the integrated risk functions of the elastic net, ridge, and lasso methods when π follows a spike and normal distribution. This parametric assumption allows us to evaluate how the methods fare under settings that differ in terms of sparsity. Our findings suggest that the ridge method yields only a marginal improvement over the elastic net in nonsparse settings and that the elastic net performs better in the in-between scenarios. We therefore recommend that the elastic net should be used by researchers who are not sure about their data's sparsity—and might always be preferred to the ridge method unless the researcher is confident that the data are not sparse. We complement this analysis by comparing the cross-validated MSEs of each method for a prediction problem with a varying number of parameters. Our results indicate that the lasso performs better for problems with a small number of parameters, but that, as we increase the dimension, the elastic net matches the lasso's performance.

Recall that, in our setup, (X_i, μ_i) is assumed to have joint distribution π. We shall consider the parametric family where $X_i \sim N(\mu_i, \sigma^2)$ with probability p, $\mu_i = 0$, and $\mu_i \sim N(\mu_0, \sigma_0^2)$ with probability $(1 - \mathrm{p})$. This framework allows us to generate data with different sparse and dense components. For example, increasing the parameter p makes the spike of zeros "taller" and therefore increases the *sparsity* of the DGP; in contrast, a larger σ_0 makes the *dense* component "wider."

In Propositions 1 and 2 (see Sections A.9 and A.10 in the Appendix), we derive the componentwise and integrated risk functions for this setting. When evaluated at the optimal selector, these functions are the measure of interest for comparing the performance of different estimators. The oracle-optimal selector can be found analytically for the ridge estimator, but it must be found numerically for the lasso and the elastic net. For a plot of the integrated risk surfaces, see Figure A.1.

An intuitive way of comparing the estimators is by finding the estimator with lowest integrated risk. In Figure 2, we project the estimator with minimum risk to the (μ_0, σ_0) plane. As expected, ridge and lasso dominate the cases with no sparsity $(\mathrm{p} = 0)$ and significant sparsity $(\mathrm{p} = 0.75)$. Elastic net only has minimum integrated risk for a small region in cases with moderate sparsity.

Even though Figure 2 seems to reflect poorly on the elastic net, measures of *relative* integrated risk between the estimators tell a different story; namely, the ridge estimator should almost never be used. In support of this claim, Figure 3 plots the ratio of the elastic net's integrated risk to that of the lasso and the ridge. We can see from the figure's lower panel that the ridge and the elastic net have nearly the same integrated risk function for low-sparsity settings, while the latter has lower risk in highly sparse settings. In other words: the elastic net always performs better—even if just weakly—than does the ridge.

The upper panel of Figure 3 suggests another interesting fact: the elastic net performs better than the lasso for some values of (μ_0, σ_0) regardless of how sparse the data are. When $\mu_0 = 1$ and σ_0 is small, the conditional distributions $X_i | \mu_0 \neq 0$ and $X_i | \mu_0 = 0$ do not overlap and are close enough that inducing too much sparsity might lead to increased integrated risk. In that case, the lasso performs worse, because it sets some positive values belonging to $X_i | \mu_0 = 0$ to zero, whereas the elastic net shrinks them. As μ_0 increases, the lasso improves, because the region where it shrinks values to zero is farther from the dense component. Elsewhere, the behavior is as expected: the elastic net has lower risk only in lower sparsity data settings.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

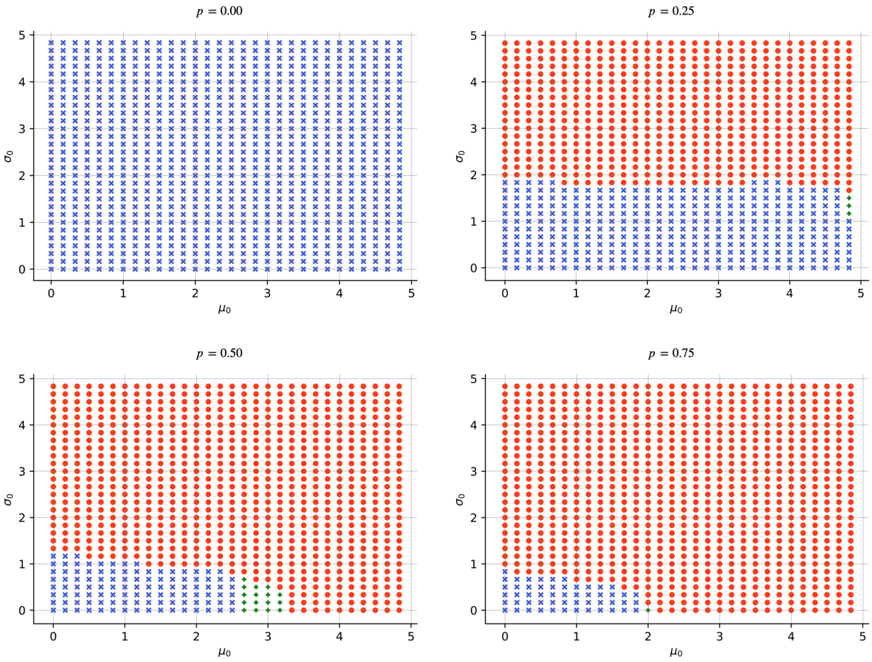

FIGURE 2. Minimum integrated risk estimators for a 900-point (μ_0, σ_0) grid. The "x", filled dot, and "+" markers correspond to (respectively) the ridge, lasso, and elastic net having the minimum integrated risk in the (μ_0, σ_0) plane. A color version of this figure can be found in the online Appendix.

4.2. Expected Mean Squared Errors

Another way of evaluating the different methods is by comparing their MSE in a prediction task. The process we follow is first to train the models on a dataset of size N that is generated from the spike and normal distribution and then to hypertune the regularization parameters using 10-fold CV. Next, we compute the MSE for an independent test set drawn from the same distribution and repeat the process 200 times to approximate the expected MSE. We do this for $N \in \{50, 250, 1,000\}$ in order to investigate how the estimators differ as we increase the number of parameters. The results are tabulated in Section A.11 in the Appendix for $p \in \{0, 0.25, 0.5, 0.75\}$, $\mu_0 \in \{0, 2, 4\}$, and $\sigma_0 \in \{1, 3, 5\}$. Three conclusions follow from this exercise.

First, an increase in N reduces the expected MSE overall, and so the difference between the models becomes less significant. Thus, when $N = 50$, the lasso has minimum MSE in 16 cases, the ridge in 12 cases, and the elastic net in 8 cases. For $N = 250$, the respective numbers are 15, 6, and 15; for $N = 1,000$, we obtain 11, 12, and 13. It is intuitive that, for larger values of N, all models perform closer to the oracle-optimal integrated risk. *Second*, for low-sparsity settings, the three methods behave similarly; yet for high-sparsity settings, the lasso and the elastic net have

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

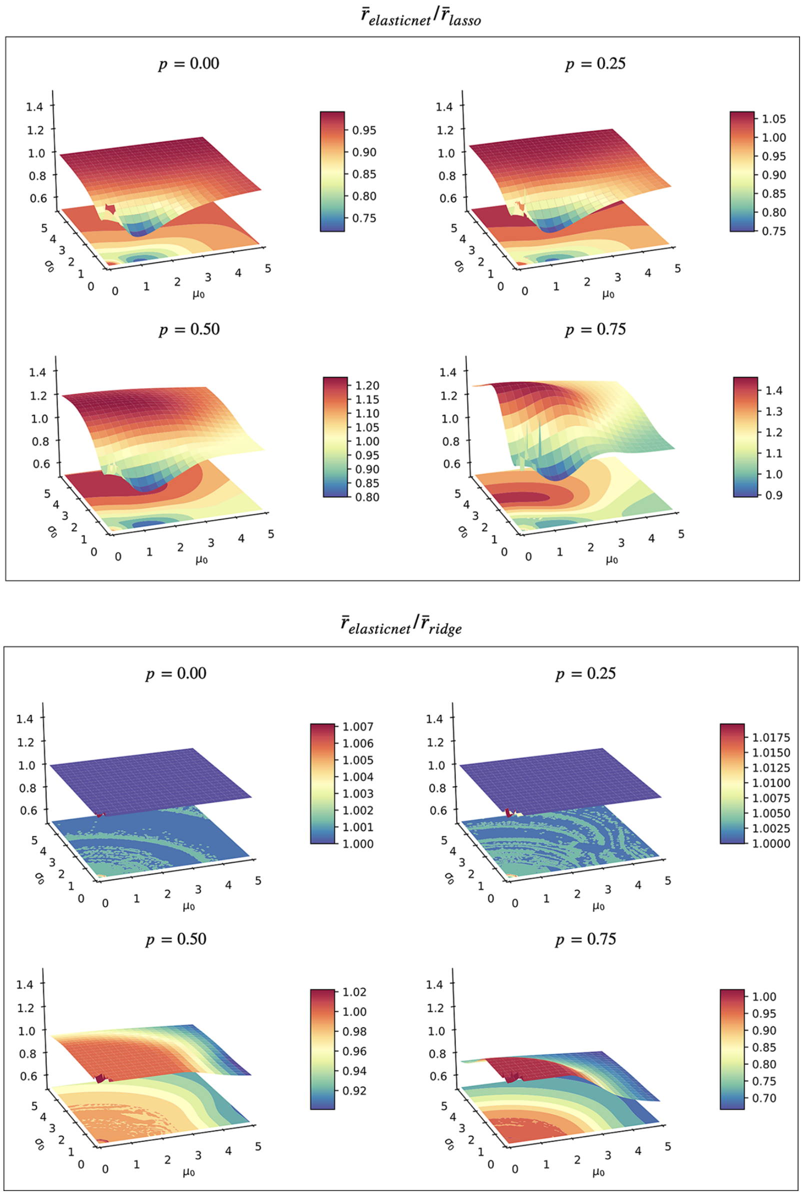

FIGURE 3. Elastic net's integrated risk relative to lasso and ridge for the spike and normal distribution. Values that exceed unity indicate that the elastic net has higher integrated risk. The discontinuities in the neighborhood of zero occur, because (i) we use a numerical solver to compute the regularization parameter for the lasso and the elastic net and (ii) in the neighborhood of zero, the risk functions are so close to zero that small perturbations lead to larger changes in the ratio. Contours of the surface are projected in the (μ_0, σ_0) plane. A color version of this figure can be found in the online Appendix.

lower MSE. *Third*, when N is smaller, the elastic net's expected MSE is higher. The reason is that approximating the oracle selector for two regularization parameters is more difficult, and so the elastic net converges more slowly (than do the lasso and the ridge) to the oracle-optimal integrated risk. Furthermore, shrinkage estimators with multidimensional regularization are more flexible and, hence, are more likely to overfit in small samples. It follows that applied researchers dealing with a low-dimensional problem might well be better served by a shrinkage estimator (e.g., lasso or ridge) with just one type of regularization.

5. THE CUBIC NET

Now, we propose a new MuSE: the *cubic net*, which is similar to the elastic net but with a cubic feature. Instead of inducing sparsity, the cubic net smoothly shrinks the parameters toward zero following a cubic polynomial. The idea is that this "softer" regularization might be preferable for cases in which the dense and sparse components of the data are separated—for example, in the spike and normal setting with $\mu_0 = 1$ and small σ_0 (see Figure 3). The componentwise function of the cubic net is given by

$$m_{\mathrm{CN}}(x, \boldsymbol{\lambda}) = \mathbb{1}(x < -\lambda_1)\frac{x}{1+\lambda_2} + \mathbb{1}(x > \lambda_1)\frac{x}{1+\lambda_2} + \mathbb{1}(|x| < \lambda_1)\frac{x^3}{\lambda_1^2(1+\lambda_2)}.$$

The cubic net has the same inflection points as the lava and coincides with the ridge estimator outside the cubic shrinkage region (see Figure A.2). Furthermore, the cubic net satisfies the MuSE assumptions and also the conditions of Theorem 1, so we can use oracle-optimal integrated risk to compare the cubic net with the lasso and the ridge. In this case, we estimate the integrated risk by sampling 10,000 times (with replacement) from the spike and normal generating distribution and then compute the average MSE of the cubic net for a 900-point (μ_0, σ_0) grid.

Figure 4 plots the minimum risk estimator for each parameter value. Overall, the cubic net seems to have lower risk than the lasso and the ridge regardless of the sparsity level. Perhaps surprisingly, it performs like ridge for p = 0 and better than the lasso for high-sparsity settings. It also universally outperforms the elastic net in a similar exercise. The main feature that drives these results is that the cubic net has lower integrated risk for larger (μ_0, σ_0); the reason is that it mitigates the error from setting a parameter to zero when it is not a true zero.

Although this analysis may not apply to other settings, it does suggest that there is a value in constructing shrinkage estimators with different types of regularization—especially for researchers who are unaware of their data's sparsity structure.

6. CONCLUSION

Applied researchers using regularization-based estimators might be interested in estimators with multiple types of regularization. In this paper, we characterize

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

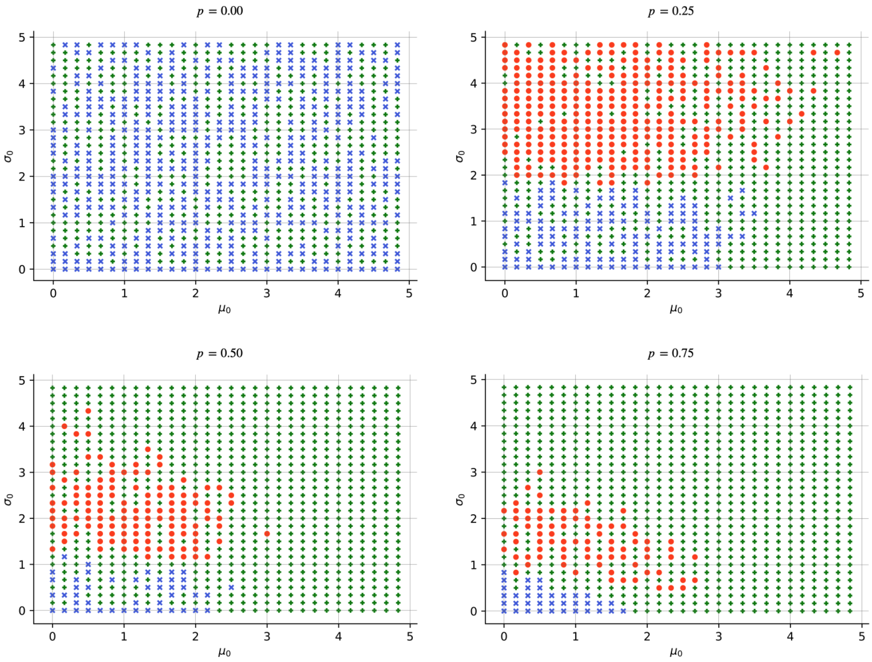

FIGURE 4. Minimum integrated risk estimators, including the cubic net, for a 900-point (μ_0, σ_0) grid. The "x", filled dot, and "+" markers correspond to (respectively) the ridge, lasso, and cubic net having the minimum risk. The risk surfaces are nonmonotonic, because the cubic net integrated risk is simulated. A color version of this figure can be found in the online Appendix.

a large class of such estimators, the MuSEs, and provide theoretical guarantees for the uniform convergence to the oracle-optimal risk when the regularization parameters are chosen by a data-driven procedure (e.g., CV or *Sure*). To provide guidance regarding the choice among the various MuSEs, we compare the lasso, ridge, and elastic net methods in spike and normal settings. Our analysis suggests that the elastic net should be used in settings with medium and low sparsity or when the researcher is uncertain about the structure of the DGP. The ridge estimator should be favored only when the researcher is confident that the setting is not sparse—and even then, its superiority vis-à-vis the elastic net appears to be marginal. In contrast, the lasso estimator performs best when the data are extremely sparse. The applied researcher should also consider the problem's dimension, since MuSEs with multiple regularization parameters (e.g., the elastic net) might exhibit slower convergence rates when the risk estimate is data-driven; so for lower-dimensional problems, the ridge and lasso methods might perform better. Finally, we introduce the cubic net estimator and show that it can outperform the lasso, ridge, and elastic net; thus, we demonstrate that other MuSEs could be of practical relevance. Hence, a promising avenue for future research is the analysis of other

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

MuSEs, such as the adaptive lasso and scad, and their behavior under different DGPs.

APPENDIX

A.1. From Multiple Means to Multivariate Regression

This example is discussed also—with the same setting—in Abadie and Kasy (2019).

Risk minimization in a multivariate regression setting can be approached as a "many means" estimation problem similar to the one described in this paper. Consider the standard multivariable linear problem

$$Y = X'\beta + \varepsilon;$$

here (for simplicity), Y and ε are scalar, X is a $p \times 1$ vector of features, and $\varepsilon | X \sim N(0,\sigma^2)$. The risk minimization problem for a sample $(Y_i, X_i)_{i=1}^n$ and loss function L requires that we minimize

$$R = \mathbb{E}[L(Y,\hat{Y})].$$

Suppose that X is drawn from the empirical distribution of $(X_1,\dots,X_n)$ for $n > p$, and assume linearity of conditional expectations and also normality. Note that these assumptions can be relaxed; we make them for ease of explanation and to avoid asymptotic arguments. Then, we can derive the form of the risk for the square loss as

$$R = \mathrm{tr}\big(\Omega \cdot \mathbb{E}[(\hat{\beta}-\beta)(\hat{\beta}-\beta)']\big) + \mathbb{E}[\varepsilon^2],$$

where we assume that $\Omega = \frac{1}{n}\sum_{i=1} X_i X_i'$ has full rank. Consider the orthogonalization $\tilde{X} = \Omega^{1/2}X$ and the ordinary least-squares regression of y on $\tilde{X}$; then, conditional on $(X_1,\dots,X_n)$, we have

$$\tilde{X} \sim N\left(\Omega^{1/2}\beta, \frac{\sigma^2}{n} I_p\right).$$

Under this change of variables, it is easy to see that (i) the problem is equivalent to the many-means problem with $\mu = \Omega^{1/2}\beta$ and (ii) we can construct regularized estimators $\hat{\mu}$ for μ by componentwise shrinkage of $\hat{X}$—as discussed previously and in Abadie and Kasy (2019).

It is noteworthy that our results and those of AK19 do *not* apply when the sample size is smaller than the number of parameters of interest. The reason is that if $n > p$, then Ω has rank of at most n. If Ω has rank n, then the least-squares predictor is $\hat{Y}_j = Y_j$, and so we can directly set both $\hat{X} = X$ and $\mu = \mathbb{E}[Y \mid X_1,\dots,X_n]$. Then, the risk of prediction is the MSE plus $\mathbb{E}[\varepsilon^2]$.

A.2. Motivation for Using the Compound Risk

Consider the means estimation setup described in Section 2. Given a loss function l and a limiting distribution L_{μ_i}, we want to minimize the asymptotic risk:

$$r(m(X_i,\boldsymbol{\lambda}), P_{\mu_i}) \equiv \int l(m(X_i,\boldsymbol{\lambda}),\mu_i)\, dL_{\mu_i}.$$

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

To do so, we could use the minimax criterion (van der Vaart, 1998) that defines the best estimator relative to a loss function l as the estimator that minimizes the maximum asymptotic risk over all randomized estimators:

$$r_{\max}(m(X_i,\boldsymbol{\lambda}),P_{\mu_i})) = \sup_{\mu_i} \mathbb{E}_{\mu_i} l(m(X_i,\boldsymbol{\lambda})-\mu_i).$$

It follows from Anderson's lemma (see van der Vaart, 1998) that, for a bowl-shaped (convex and symmetric about the origin) loss function, $m(X_i,\boldsymbol{\lambda}) = X_i$ is a lower bound for asymptotic risk and is therefore minimax (i.e., it minimizes the maximum risk). Furthermore, this occurs when the limit distribution L_{μ_i} is normal. In fact, X_i is the best regular estimator. A regular estimator is defined as asymptotically equivariant-in-law; in other words, the distribution of the limiting distribution does not depend on local parameters. Yet as we have seen in the motivating example of Stein (1956), when we estimate multiple means and if $p \geq 3$, then a shrinkage estimator with a non-normal limiting distribution can have a lower asymptotic risk than the best regular estimator. So instead of using the minimax risk, we are interested in defining the compound loss and risk over all components.

A.3. Lemma 1

The proof is similar to that for the scalar case in AK19 (but with minor adaptations).

LEMMA. *For a finite partition of regularization parameters* $0 = \lambda_0^i < \cdots < \lambda_T^i$ (*where* λ^i *is the* ith *component of* $\boldsymbol{\lambda}$) *and for some* $c \in \mathbb{R}$, *define*:

$$u_j^i = \sup_{\lambda^i \in [\lambda_{j-1}^i, \lambda_j^i]} L(\boldsymbol{\lambda}_i);$$

$$l_j^i = \inf_{\lambda^i \in [\lambda_{j-1}^i, \lambda_j^i]} L(\boldsymbol{\lambda}_i).$$

Here, $\boldsymbol{\lambda}_i = [c_1,\ldots,c_{i-1},\lambda^i,c_{i+1},\ldots,c_k]'$ *for some constants* $c_l \in [0,\infty]$, *and L is the vector-valued loss function. Assume that, for all* $\varepsilon_i > 0$, *there exists a partition* $0 = \lambda_0^i < \cdots < \lambda_T^i$ *with* $T(\varepsilon_i)$ *such that*

$$\sup_{\pi \in Q} \max_{1 \leq j \leq T} \mathbb{E}_\pi [u_j^i - l_j^i] \leq \varepsilon_i,$$

$$\sup_{\pi \in Q} \max_{1 \leq j \leq T} \max\{\mathrm{var}_\pi(l_j^i), \mathrm{var}_\pi(u_j^i)\} < \infty.$$

Then,

$$\sup_{\pi \in Q} \mathbb{E}_\pi \left[\sup_{\lambda^i \in [0,\infty]} (L_p(\boldsymbol{\lambda}_i) - \bar{r}_\pi(\boldsymbol{\lambda}_i))^2 \right] = o(1).$$

It follows that

$$\sup_{\lambda^i \in [0,\infty]} (L_p(\boldsymbol{\lambda}_i) - \bar{r}_\pi(\boldsymbol{\lambda}_i))^2 = o_{\mathrm{p}}(1)$$

with respect to any measure $\pi \in Q$.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

Proof. Fix a $\boldsymbol{\lambda}_i$ that depends on a set of constants and a parameter λ^i, and consider the partition $0 = \lambda^i_0 < \cdots < \lambda^i_T = \infty$. Then, by construction, we have:

$$\mathbb{E}_p[L(\boldsymbol{\lambda}_i)] - \mathbb{E}_\pi[L(\boldsymbol{\lambda}_i)] \leq \mathbb{E}_p[u^i_j] - \mathbb{E}_\pi[u^i_j] + \mathbb{E}_\pi[u^i_j - l^i_j];$$
$$\mathbb{E}_p[L(\boldsymbol{\lambda}_i)] - \mathbb{E}_\pi[L(\boldsymbol{\lambda}_i)] \geq \mathbb{E}_p[l^i_j] - \mathbb{E}_\pi[l^i_j] + \mathbb{E}_\pi[u^i_j - l^i_j].$$

Therefore,

$$\sup_{\boldsymbol{\lambda}_i \in [0,\infty]} (\mathbb{E}_n[L(\boldsymbol{\lambda}_i)] - \mathbb{E}_\pi[L(\boldsymbol{\lambda}_i)])^2 \leq \sum_{j=1}^{T}((\mathbb{E}_n[u^i_j] - \mathbb{E}_\pi[u^i_j])^2 + (\mathbb{E}_n[l^i_j] - \mathbb{E}_\pi[l^i_j])^2 + \varepsilon_i^2$$
$$+ 2\varepsilon \sum_{j=1}^{T}(|\mathbb{E}_n[u^i_j] - \mathbb{E}_\pi[u^i_j]| + |\mathbb{E}_n[l^i_j] - \mathbb{E}_\pi[l^i_j]|).$$

Now, taking expectations over π yields the desired result:

$$\mathbb{E}_\pi\Big[\sup_{\boldsymbol{\lambda}_i \in [0,\infty]} (\mathbb{E}_p[L(\boldsymbol{\lambda}_i)] - \mathbb{E}_\pi[L(\boldsymbol{\lambda}_i)])^2\Big]$$
$$\leq \sum_{j=1}^{T}\left(\frac{\mathrm{var}_\pi(u^i_j)}{p} + \frac{\mathrm{var}_\pi(l^i_j)}{p}\right) + \varepsilon_i^2 + 2\varepsilon \sum_{j=1}^{T}\left(\sqrt{\frac{\mathrm{var}_\pi(u^i_j)}{p} + \frac{\mathrm{var}_\pi(l^i_j)}{p}}\right).$$

Hence, as $p \to \infty$, we obtain $\mathbb{E}_\pi[\sup_{\boldsymbol{\lambda}_i \in [0,\infty]}(\mathbb{E}_p[L(\boldsymbol{\lambda}_i)] - \mathbb{E}_\pi[L(\boldsymbol{\lambda}_i)])^2] \to 0$ as required. □

A.4. Lemma 2

LEMMA. *Let the assumptions of Lemma A.3 be satisfied for all* $\lambda^i (i \in \{1, \ldots, k\})$, *and let* $m(x, \boldsymbol{\lambda})$ *be monotonic in all components of* $\boldsymbol{\lambda}$ *for all* $x \in \mathbb{R}^*$. *In addition, let* $m(x, \mathbf{0}) = x$ *and* $\lim_{\boldsymbol{\lambda} \to \boldsymbol{\infty}} m(x, \boldsymbol{\lambda}) = 0$. *Then,*

$$\sup_{\pi \in Q} \mathbb{E}_\pi\left[\sup_{\boldsymbol{\lambda} \in [0,\infty]^k} (L_p(\boldsymbol{\lambda}) - \bar{r}_\pi(\boldsymbol{\lambda}))^2\right] = o(1).$$

Proof . We show that if Lemma A.3 holds, then we can find a regularization path for the components of $\boldsymbol{\lambda}$ that asymptotically bounds $\sup_{\boldsymbol{\lambda} \in [0,\infty]^k}(L_p(\boldsymbol{\lambda}) - \bar{r}_\pi(\boldsymbol{\lambda}))^2$.

Without loss of generality, consider Lemma A.3 for λ^1, $c \in \mathbb{R}$, and let $A^1 = \{\lambda^1 \in \mathbb{R}^*_+ | (\mathbb{E}_p[L(\boldsymbol{\lambda}^1)] - \mathbb{E}_\pi[L(\boldsymbol{\lambda}^1)])^2\}$. It follows that $\sup_{\boldsymbol{\lambda}^1} A^1 = o_{\mathrm{p}}(1)$ with respect to π. Since m is monotonic for all λ components and $m(x, \mathbf{0}) = x$ and $\lim_{\boldsymbol{\lambda} \to \boldsymbol{\infty}} m(x, \boldsymbol{\lambda}) = 0$, A^1 is bounded below by the oracle lambda (the optimal lambda) and above by $\tilde{\boldsymbol{\lambda}}^1$ which happens when $\lambda^1 = \tilde{\lambda}^1$ (the rest of the components are equal to c). Note that $\tilde{\lambda}^1$ exists, because $\lambda^i \in \mathbb{R}^*$, the extended reals are compact, and the loss function is continuous in lambda, so by Weierstrass Theorem, the extrema must exist. It is straightforward to show that $(\mathbb{E}_p[L(\tilde{\boldsymbol{\lambda}}^1)] - \mathbb{E}_\pi[L(\tilde{\boldsymbol{\lambda}}^1)])^2 \in A^1$.

Hence, let:

$$\tilde{\lambda}^1 = \sup_{\lambda^1} A^1,$$

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

and define the $\boldsymbol{\lambda}^2 = [\tilde{\lambda}^1, \lambda^2, c, \dots, c]$ and repeat the same procedure as before fixing $\tilde{\lambda}^1$ as a constant. It follows that:

$$\sup(A^1) \leq \sup(A^2) \leq \cdots \leq \sup(A^k).$$

Furthermore, we can repeat this for any initial choice of $c \in \mathbb{R}_+$ and so:

$$\sup_{\boldsymbol{\lambda}} A \leq \sup_{c \in \mathbb{R}_+} A^k$$

by the limiting conditions on the choice of c, and knowing from Lemma A.3 that $A^k = o_{\mathrm{p}}(1)$, the result follows. It is sufficient to ensure that the choice of c would not be made better by ∞. This is the case since we know that it is never optimal for $\tilde{\lambda}^i$ to be ∞ because of the limiting conditions on m. If $\boldsymbol{\lambda} = \boldsymbol{\infty}$, that would mean that the estimator $m(x,.) = 0$ is the maximizer of A. However, it can be shown that since m is monotonic in $\boldsymbol{\lambda}$, we would increase the value of the empirical loss by setting $m > 0$ for components with measure centered on a positive value and $m < 0$ for components with measure centered on a negative value. Furthermore, we can achieve this by increasing/decreasing the λ parameters. Hence, it is never strictly dominant to have $c = \infty$. □

A.5. Proof of Theorem 1

We need to show that the assumptions of Theorem 1 imply the assumptions of Lemma A.4.

Proof. We follow the AK19 proof closely and show that the conditions for Lemma A.4 are satisfied. The only conditions that we need to check are those of Lemma A.3 (i.e., since the others are assumed by Theorem 1). First, note that the convexity of fourth powers yields bounded variances. Second, the square loss is strictly convex, and so any extrema are achieved at the boundary. It follows that, for u_j^i and l_j^i as defined in Lemma A.3,

$$\mathrm{var}_{\pi}(u_j^i) \leq \mathbb{E}_{\pi}[(u_j^i)^2] \leq \mathbb{E}_{\pi}[\max\{(X-\mu)^4, \mu^4\}] \leq \mathbb{E}_{\pi}[(X-\mu)^4]\mathbb{E}_{\pi}[\mu^4].$$

Similarly, the variance of l_j^i is also bounded as $\mathrm{var}_{\pi}(l_j^i) \leq \mathbb{E}_{\pi}[(u_j^i)^2]$. Hence, the condition of Lemma A.3 on the variances is satisfied. For the condition on the expectations, we use the strong monotonicity of m with respect to $\boldsymbol{\lambda}$. We need to find a partition for each i such that $\mathbb{E}_{\pi}[u_j^i - l_j^i] < \varepsilon_i$. Fix $\boldsymbol{\lambda}^i = [c_1, \dots, c_{i-1}, \lambda^i, c_{i+1}, \dots, c_k]$ for some constants c_j, in which case $m(\cdot, \boldsymbol{\lambda}_i)$ will be monotonic in λ^i. Consider a partition $0 = \lambda_0^i < \cdots < \lambda_T^i$ with $T(\varepsilon_i)$. Then, for any λ_j^i: since the compound loss L is convex (because the square loss is convex and the support is bounded), it follows that the supremum lies at the boundary. So for the interval $[\lambda_{j-1}^i, \lambda_j^i]$, we have

$$u_j^i = \max\{L(\boldsymbol{\lambda}_{j-1}^i), L(\boldsymbol{\lambda}_j^i)\},$$
$$l_j^i = \min\{L(\boldsymbol{\lambda}_{j-1}^i), L(\boldsymbol{\lambda}_j^i)\}.$$

Note that if μ is predicted correctly for some point in the interval, then $l_j^i = 0$. However, in both cases, we obtain

$$u_j^i - l_j^i = |L(\boldsymbol{\lambda}_{j-1}^i) - L(\boldsymbol{\lambda}_j^i)|,$$

since if $l_j^i = 0$ then $u_j^i - l_j^i = \max\{L(\boldsymbol{\lambda}_{j-1}^i), L(\boldsymbol{\lambda}_j^i)\}$. Yet because one of the maxima's two terms is zero, the right-hand side is equivalent to $|L(\boldsymbol{\lambda}_{j-1}^i) - L(\boldsymbol{\lambda}_j^i)|$, since the compound loss function is nonnegative. Let $m_j = m(X, \boldsymbol{\lambda}_j^i)$; then,

$$u_j^i - l_j^i = |(m_j - \mu)^2 - (m_{j-1} - \mu)^2| \leq (|m_j - \mu| + |m_{j-1} - \mu|)|m_j - m_{j-1}|.$$

By condition (ii) in Theorem 1, the expectation of this result is bounded; hence, the condition in Lemma A.3 is satisfied, and so the result follows. □

A.6. Proof of Theorem 3

This proof is similar to AK19's proof for the scalar case but is adapted for MuSEs.

Proof. By the assumptions of Theorem 1, convergence in L^2 implies convergence in probability, and so

$$\sup_{\pi \in Q} P_\pi \left(\sup_{\boldsymbol{\lambda} \in [0,\infty]^k} |L_p(\hat{\boldsymbol{\lambda}}_p) - \inf_{\boldsymbol{\lambda} \in [0,\infty]^k} L_p(\boldsymbol{\lambda})| > \varepsilon \right) \to 0,$$

since $\bar{r}(m(X, \hat{\boldsymbol{\lambda}}_p), \pi) = \mathbb{E}_\pi [L_p(\hat{\boldsymbol{\lambda}}_p)]$ by definition. We need to show that uniform convergence in probability implies L^1 convergence. By van der Vaart (1998, Theorem 2.20), this is the case if and only if the estimator sequence $L_p(\hat{\boldsymbol{\lambda}}_p)$ is asymptotically uniform integrable. Hence, we must check that

$$\lim_{M \to \infty} \lim \sup_{p \to \infty} \mathbb{E}_\pi |L_p(\hat{\boldsymbol{\lambda}}_p)| 1(|L_p(\hat{\boldsymbol{\lambda}}_p)| > M) = 0.$$

By the conditions of Theorem 1, we know that if $\hat{\boldsymbol{\lambda}} = 0$, then $m(x, \mathbf{0}) = x$ and $L_p(\mathbf{0}) = \frac{1}{p} \sum_i^p (X_i - \mu_i)^2$. We choose $\hat{\boldsymbol{\lambda}}_p$ to minimize the empirical compound risk $\bar{r}$. Given the convexity of fourth powers of X, the bowl shape of L, and the boundary conditions in Theorem 1, we conclude that $|L_p(\hat{\boldsymbol{\lambda}}_p)|$ is bounded from above; otherwise, Theorem 1 would not hold—a contradiction. It follows that, as $M \to \infty$, we have uniform integrability, and so the theorem holds. □

A.7. *Sure* Derivation

The main takeaway is that, under some assumptions, the risk estimator solves a penalized loss function minimization problem of the following form:

$$\hat{r}_p = \underset{\boldsymbol{\lambda}}{\operatorname{argmin}} \frac{1}{p} \sum_{i=1}^{p} (m(X_i, \boldsymbol{\lambda}) - X_i)^2 + \mathit{Penalty}.$$

The penalty term, which should not be confused with the penalty term built in $m(X_i, \boldsymbol{\lambda})$, depends on the gradient of the componentwise estimator function and captures the type

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

of shrinkage that the estimator induces. For the estimators, we consider, in this paper, the penalties are

$$
\begin{aligned}
\text{Ridge}: &\ \frac{2}{1+\lambda}.\\
\text{Lasso}: &\ \frac{2}{p}\sum_{i=1}^{p}\mathbb{1}(|X_i|>\lambda).\\
\text{Elastic net}: &\ \frac{2}{p}\sum_{i=1}^{p}\frac{1}{1+\lambda_2}\mathbb{1}(|X_i|>\lambda_1).
\end{aligned}
$$

As expected, the penalty for the elastic net combines both types of regularization, both shrinking the estimates toward zero and inducing sparsity.

To derive the formula for *Sure*, we consider the setting where $\mu\sim\Pi$ and $X|\mu\sim N(\mu,1)$ as in Abadie and Kasy 2019. We impose the additional assumptions that X has a density and m is differentiable with respect to x. Let $f_\pi=\Pi*\phi$ be the marginal density of X, where ϕ is the standard normal pdf. Consider an estimator $m(X,\boldsymbol{\lambda})$ of μ differentiable with respect to x everywhere except for $\{x_1,\dots,x_J\}$ where m might be discontinuous. Then, let $\nabla_x m$ be the gradient with respect to X and $\Delta m_j(\boldsymbol{\lambda})=\lim_{x\downarrow x_j}m(x,\boldsymbol{\lambda})-\lim_{x\uparrow x_j}m(x,\boldsymbol{\lambda})$, for $j\in\{1,\dots,J\}$. Under the assumption that $\mathbb{E}_\pi[(m(X,\boldsymbol{\lambda})-X)^2]<\infty$, $\mathbb{E}_\pi[\nabla_x(m(X,\boldsymbol{\lambda})]<\infty$, and $m(x,\boldsymbol{\lambda})-x)\phi(x-\mu)\to 0$ as $|x|\to 0$, it can be shown that the *Sure* risk estimator is given by:

$$
\hat{r}_p=\underset{\boldsymbol{\lambda}\in[0,\infty]^k}{\operatorname{argmin}}\ \frac{1}{p}\sum_{i=1}^{p}(m(X_i,\boldsymbol{\lambda})-X_i)^2+2\left(\frac{1}{p}\sum_{i=1}^{p}\nabla_x m(X_i,\boldsymbol{\lambda})+\sum_{j=1}^{J}\Delta m_j(\boldsymbol{\lambda})\hat{f}(x_i)\right),
$$

where $\hat{f}$ is the estimator for f_π.

A.8. Proof of Theorem 4

Proof. Consider the following decomposition as in AK19:

$$
\begin{aligned}
r_{p,n}(\boldsymbol{\lambda})&=\frac{1}{p}\sum_{i=1}^{p}\left[(m(X_{n-1i},\boldsymbol{\lambda})-\mu_i)^2+(x_{ni}-\mu_i)^2+2((m(X_{n-1i},\boldsymbol{\lambda})-\mu_i)(x_{ni}-\mu_i)\right]\\
&=L_{n,p}(\boldsymbol{\lambda})+\frac{1}{p}\sum_{i=1}^{p}(x_{ni}-\mu_i)^2+\frac{2}{p}\sum_{i=1}^{p}(m(X_{n-1i},\boldsymbol{\lambda})-\mu_i)(x_{ni}-\mu_i).
\end{aligned}
$$

Note that $\mathbb{E}_\pi\left[\sup_{\boldsymbol{\lambda}\in[0,\infty]^k}(r_{p,n}(\boldsymbol{\lambda})-\bar{r}_{\pi,n}(\boldsymbol{\lambda}))^2\right]$ is bounded above by the sum of the squares for each term in $r_{p,n}$ minus its population counterpart. Now, observe that the general loss consistency theorem holds for the first term (the CV loss), and so the first term converges uniformly to $\mathbb{E}_\pi[m(X_{n-1i},\boldsymbol{\lambda})-x_{ni})^2]$. Given the convexity of fourth powers and the equality $\operatorname{var}_\pi(x_j)=\sigma^2$, it follows that $\frac{1}{p}\sum_{i=1}^{p}(x_{ni}-\mu_i)^2$ converges uniformly in probability to $\mathbb{E}_\pi[\sigma^2]$ by a suitable law of large numbers. Because $\bar{r}_{\pi,k}(\boldsymbol{\lambda})=$

$\mathbb{E}_\pi[m(X_{n-1i},\boldsymbol{\lambda})-x_{ni})^2]+\mathbb{E}_\pi[\sigma^2]$ asymptotically, the $\mathbb{E}_\pi[\sigma^2]$ cancels out. For the last term, note that the linearity of the expectations operator and Jensen's inequality together imply that

$$\begin{aligned}
&\mathbb{E}_\pi\left[\left(\frac{1}{p}\sum_{i=1}^{p}(m(X_{n-1i},\boldsymbol{\lambda})-\mu_i)(x_{ni}-\mu_i)\right)^4\right]\\
&\quad\le \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}_\pi[(m(X_{n-1},\boldsymbol{\lambda})-\mu)^4]\mathbb{E}_\pi[(x_n-\mu)^4]\\
&\quad\le \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}_\pi[(X_{n-1}-\mu)^4+\mu^4]\mathbb{E}_\pi[(x_n-\mu)^4].
\end{aligned}$$

It follows from the assumptions of Theorem 1 and the inequality $\sup_\pi \mathbb{E}_\pi[x_j^4]<\infty$ that the two expectations are uniformly tight and so $o_{\text{p}}(1)$. Therefore, $(\hat{r}_{p,n}(\boldsymbol{\lambda})-\bar{r}_{\pi,n}(\boldsymbol{\lambda}))$ tends to zero in mean squared for all $\boldsymbol{\lambda}$ and the first result holds. Given this first result, the second result follows from Theorem 2, because all conditions are satisfied. □

A.9. Proposition 1

PROPOSITION 1. *Given $X_i\sim N(\mu_i,\sigma_i)$, for $i\in\{1,\dots,p\}$, the componentwise risks are given by*

$$\begin{aligned}
r_{m_{\text{ridge}}} &= \left(\frac{1}{1+\lambda}\right)^2\sigma_i^2+\left(1-\frac{1}{1+\lambda}\right)^2\mu_i^2,\\
r_{m_{\text{lasso}}} &= \left(1+\Phi\left(\frac{-\lambda-\mu_i}{\sigma_i}\right)-\Phi\left(\frac{\lambda-\mu_i}{\sigma_i}\right)\right)(\sigma_i^2+\lambda^2)\\
&\quad+\left(\left(\frac{-\lambda-\mu_i}{\sigma_i}\right)\phi\left(\frac{\lambda-\mu_i}{\sigma_i}\right)+\left(\frac{\lambda+\mu_i}{\sigma_i}\right)\phi\left(\frac{-\lambda-\mu_i}{\sigma_i}\right)\right)\sigma_i^2\\
&\quad+\left(\Phi\left(\frac{\lambda-\mu_i}{\sigma_i}\right)-\Phi\left(\frac{-\lambda-\mu_i}{\sigma_i}\right)\right)\mu_i^2,\\
r_{m_{\text{EN}}} &= l\big(-2\mu_i^2-2(l-1)\lambda_1\mu_i+l(\mu^2+\sigma_i^2)+l\lambda_1^2\big)\left(1-\Phi\left(\frac{\lambda_1-\mu_i}{\sigma_i}\right)\right)\\
&\quad+l\sigma(l(\mu_i+\lambda_1)-2\mu-2l\lambda_1)\phi\left(\frac{\lambda_1-\mu_i}{\sigma_i}\right)\\
&\quad+l\big(-2\mu_i^2+2(l-1)\lambda_1\mu_i+l(\mu_i^2+\sigma_i^2)+l\lambda_1^2\big)\left(1-\Phi\left(\frac{-\lambda_1-\mu_i}{\sigma_i}\right)\right)\\
&\quad-l\sigma_i(l(\mu_i-\lambda_1)-2\mu_i+2l\lambda_1)\phi\left(\frac{-\lambda_1-\mu_i}{\sigma_i}\right);
\end{aligned}$$

here, $l=1/(1+\lambda_2)$, and ϕ and Φ are the standard normal pdf and cumulative distribution function (cdf), respectively.

The results for lasso and ridge are derived in AK19. Our goal here is to derive the form of $r(m_{\text{EN}}(X_i,\boldsymbol{\lambda}),P_{\mu_i})=\mathbb{E}[(m(X_i,\boldsymbol{\lambda}),\mu_i]$ for a normal limiting distribution $L_{\mu_i}=N(\mu_i,\sigma_i^2)$. We drop the subscript i in order to simplify the notation.

We have that

$$m_{\mathrm{EN}} = 1(x > \lambda_1)\left(\frac{x-\lambda_1}{1+\lambda_2}\right) + 1(x < -\lambda_1)\left(\frac{x+\lambda_1}{1+\lambda 2}\right).$$

Then, for $l = 1/(1+\lambda_2)$,

$$\begin{aligned} m_{\mathrm{EN}} - \mu = {} & l(x-\mu)1(|x| > \lambda_1) + l\lambda_1(1(x < -\lambda_1) - 1(x > \lambda_1)) \\ & - l\lambda_1\mu 1(|x| > \lambda_1) - \mu 1(|x| < \lambda_1). \end{aligned}$$

By squaring, taking expectations, and simplifying, we obtain

$$\begin{aligned} \mathbb{E}[(m_{\mathrm{EN}} - \mu)^2] = {} & l^2\mathbb{E}[(x-\mu)1(|x| > \lambda_1] \\ & + 2l^2(\lambda_1 - \lambda_2\mu)\mathbb{E}[1(x < -\lambda_1)(x-\mu)] \\ & - 2l^2(\lambda_1 + \lambda_2\mu)\mathbb{E}[1(x > \lambda_1)(x-\mu)] \\ & + l^2\lambda_1^2(1+\mu^2)1(|x| > \lambda_1) \\ & + 2l^2\lambda_1\lambda_2\mu\mathbb{E}[1(x > \lambda_1) - 1(x < -\lambda_1)] \\ & + \mu^2 E(1(|x| < \lambda_1). \end{aligned}$$

Next, we use the following standard normal integration results to derive the necessary integrals:

$$\begin{aligned} & \textstyle\int_a^b v^2\phi(v)\,dv = \Phi(b) - \Phi(a) - [b\phi(b) - a\phi(a)]; \\ & \textstyle\int_a^b v\phi(v)\,dv = \phi(a) - \phi(b). \end{aligned}$$

These results allow us to derive the following expressions:

$$\begin{aligned} & E\big[\big(\tfrac{x-\mu}{\sigma}\big)^2 1(|x| > \lambda_1)\big] = 1 + \Phi\big(\tfrac{-\lambda_1-\mu}{\sigma}\big) - Phi\big(\tfrac{\lambda_1-\mu}{\sigma}\big) + \big(\tfrac{\lambda_1-\mu}{\sigma}\big)\phi\big(\tfrac{\lambda_1-\mu}{\sigma}\big) \\ & \quad - \big(\tfrac{-\lambda_1-\mu}{\sigma}\big)\phi\big(\tfrac{-\lambda_1-\mu}{\sigma}\big); \\ & E\big[\big(\tfrac{x-\mu}{\sigma}\big)1(x > \lambda_1)\big] = \phi\big(\tfrac{\lambda_1-\mu}{\sigma}\big); \\ & \mathbb{E}[1(|x| > \lambda_1)] = 1 - \Phi\big(\tfrac{\lambda_1-\mu}{\sigma}\big) + \Phi\big(\tfrac{-\lambda_1-\mu}{\sigma}\big); \\ & \mathbb{E}[1(|x| < \lambda_1)] = \Phi\big(\tfrac{\lambda_1-\mu}{\sigma}\big) - \Phi\big(\tfrac{-\lambda_1-\mu}{\sigma}\big). \end{aligned}$$

By applying these results to the expectations derived previously and then simplifying, we obtain the desired result. The calculations are not shown, because they are both tedious and uninteresting.

A.10. Proposition 2

PROPOSITION 2. *Let π be such that $\mu_i \sim_{\mathrm{iid}} N(\mu_0, \sigma_0)$ with probability p and* 0 *otherwise, and let $X_i \sim N(\mu_i, \sigma^2)$. Then, the integrated risk for the ridge estimator may be written as*

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

$$\bar{r}_{m_{\text{ridge}}} = \left(\frac{1}{1+\lambda}\right)^2 \sigma^2 + \left(\frac{\lambda}{1+\lambda}\right)^2 (\mu_0^2+\sigma_0^2),$$

where $\bar{\lambda}^*(\pi) = \frac{\sigma^2}{(1-\mathrm{p})(\mu_0^2+\sigma_0^2)}$.

For the lasso and the elastic net, the normal part is given by

$$\begin{aligned}
\bar{r}_1(m_{\text{lasso}}) = {} & \left(1+\Phi\left(\frac{-\lambda-\mu_0}{s}\right)-\Phi\left(\frac{\lambda-\mu_0}{s}\right)\right)(\sigma^2+\lambda^2) \\
& +\left(\Phi\left(\frac{\lambda-\mu_0}{s}\right)-\Phi\left(\frac{-\lambda-\mu_0}{s}\right)\right)(\mu_0^2+\sigma_0^2) \\
& -\frac{1}{\sqrt{s}}\phi\left(\frac{\lambda-\mu_0}{s}\right)(\lambda+\mu_0)s \\
& -\frac{1}{\sqrt{s}}\phi\left(\frac{-\lambda-\mu_0}{s}\right)(\lambda-\mu_0)s; \\
\bar{r}_1(m_{\text{EN}}) = {} & l(-2(\sigma_0^2+\mu_0^2)-2(l-1)\lambda_1\mu_0+l(\mu_0^2+s^2)+l\lambda_1^2) \\
& -l(l-2)\left[\Phi\left(\frac{\lambda_1-\mu_0}{\sqrt{s}}\right)(\mu_0^2+\sigma_0^2)-\frac{1}{\sqrt{s}}\sigma_0^2\phi\left(\frac{\lambda_1-\mu_0}{\sqrt{s}}\right)\right. \\
& \left.\times\left(\lambda_1+\mu_0+\sigma^2\frac{\lambda_1-\mu_0}{s}\right)\right] \\
& +2l(l-1)\left[\Phi\left(\frac{\lambda_1-\mu_0}{\sqrt{s}}\right)\mu_0-\frac{\sigma_0^2}{\sqrt{s}}\phi\left(\frac{\lambda_1-\mu_0}{\sqrt{s}}\right)\right]-l^2(\sigma^2+\lambda_1^2)\Phi\left(\frac{\lambda_1-\mu_0}{\sqrt{s}}\right) \\
& +\frac{l\sigma^2}{\sqrt{s}}\phi\left(\frac{\lambda_1-\mu_0}{\sqrt{s}}\right)\left[\left(-\lambda_1+\frac{\sigma^2(\mu_0-\lambda_1)}{s}\right)(l-2)+l\lambda_1\right] \\
& +l(l-2)\left[\Phi\left(\frac{-\lambda_1-\mu_0}{\sqrt{s}}\right)(\mu_0^2+\sigma_0^2)-\frac{1}{\sqrt{s}}\sigma_0^2\phi\left(\frac{-\lambda_1-\mu_0}{\sqrt{s}}\right)\right. \\
& \left.\times\left(-\lambda_1+\mu_0+\sigma^2\frac{-\lambda_1-\mu_0}{s}\right)\right] \\
& +2l(l-1)\left[\Phi\left(\frac{-\lambda_1-\mu_0}{\sqrt{s}}\right)\mu_0-\frac{\sigma_0^2}{\sqrt{s}}\phi\left(\frac{-\lambda_1-\mu_0}{\sqrt{s}}\right)\right] \\
& +l^2(\sigma^2+\lambda_1^2)\Phi\left(-\frac{\lambda_1-\mu_0}{\sqrt{s}}\right) \\
& -\frac{l\sigma^2}{\sqrt{s}}\phi\left(\frac{-\lambda_1-\mu_0}{\sqrt{s}}\right)\left[\left(\lambda_1+\frac{\sigma^2(\mu_0+\lambda_1)}{s}\right)(l-2)+l\lambda_1\right] \\
& +\sigma_0^2+\mu_0^2.
\end{aligned}$$

In addition, when $\mu = 0$, *we have*

$$\begin{aligned}
\bar{r}_0(m_{\text{lasso}}) &= 2\Phi\left(\frac{-\lambda}{\sqrt{s}}\right)(\sigma^2+\lambda^2)-2\frac{\lambda}{\sigma}\phi\left(\frac{\lambda}{\sqrt{s}}\right)\sigma^2, \\
\bar{r}_0(m_{\text{EN}}) &= 2\Phi\left(\frac{-\lambda_1}{\sqrt{s}}\right)l^2(\sigma^2+\lambda_1^2)-2l^2\sigma\phi\left(\frac{\lambda_1}{\sqrt{s}}\right);
\end{aligned}$$

here, $l = 1/(1+\lambda_2)$ *and* $s = \sigma^2 + sigma_0^2$*. Then, the integrated risk is given by*

$$\begin{aligned}
\bar{r}_{\text{lasso}} &= \text{p}\bar{r}_0(m_{\text{lasso}}) + (1-\text{p})\bar{r}_1(m_{\text{lasso}});\\
\bar{r}_{\text{EN}} &= \text{p}\bar{r}_0(m_{\text{EN}}) + (1-\text{p})\bar{r}_1(m_{\text{EN}}).
\end{aligned}$$

The lasso and ridge results are derived in AK19. In order to obtain the integrated risk, we must integrate the elastic net risk with respect to the cdf of $N(\mu_0, \sigma_0^2)$ for the case where $\mu \neq 0$. Let $N(\mu) = \frac{1}{\sigma_0}\phi\big(\frac{\mu_0-\mu}{\sigma_0}\big)$. Then, it follows that $\mu_0 = \int_{-\infty}^{\infty} \mu N(\mu)\,d\mu$ and that $\sigma_0^2 + \mu_0^2 = \int_{-\infty}^{\infty} \mu^2 N(\mu)\,d\mu$. Furthermore, we derive 12 other Gaussian integrals from first principles with $s = \sigma^2 + \sigma_0^2$. Consider the integral limits to be from $-\infty$ to ∞. Then, the following equalities hold:

$$\begin{aligned}
&\int \Phi\big(\tfrac{\pm\lambda_1-\mu}{\sigma}\big)N(\mu)\,d\mu = \Phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big);\\
&\int \phi\big(\tfrac{\pm\lambda_1-\mu}{\sigma}\big)N(\mu)\,d\mu = \tfrac{-\sigma}{\sqrt{s}}\phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big);\\
&\int \tfrac{\pm\lambda_1-\mu}{\sigma}\phi\big(\tfrac{\pm\lambda_1-\mu}{\sigma}\big)N(\mu)\,d\mu = \tfrac{\sigma^2(\pm\lambda_1-\mu_0)}{s}\tfrac{1}{\sqrt{s}}\phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big);\\
&\int \mu\phi\big(\tfrac{\pm\lambda_1-\mu}{\sigma}\big)N(\mu)\,d\mu = \tfrac{1}{\sqrt{s}}\phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big)\sigma\big(\pm\lambda_1 + \tfrac{\sigma^2(\mu_0\mp\lambda_1)}{s}\big);\\
&\int \mu^2\Phi\big(\tfrac{\pm\lambda_1-\mu}{\sigma}\big)N(\mu)\,d\mu = \Phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big)(\mu_0^2 + sigma_0^2) - \tfrac{1}{\sqrt{s}}\phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big)(\pm\lambda_1+\mu_0)\sigma_0^2 +\\
&\qquad \sigma^2\sigma_0^2\tfrac{1}{\sqrt{s}}\phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big)\big(\tfrac{\pm\lambda_1-\mu_0}{s}\big);\\
&\int \mu\Phi\big(\tfrac{\pm\lambda_1-\mu}{\sigma}\big)N(\mu)\,d\mu = \mu_0\Phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big) - \tfrac{\sigma_0^2}{\sqrt{s}}\phi\big(\tfrac{\pm\lambda_1-\mu_0}{\sqrt{s}}\big).
\end{aligned}$$

We use these integrals to compute the required expectations with respect to μ for the normal risk of the elastic net. (The derivations are again tedious and uninteresting, so they are not reproduced here.) The result follows from applying these 12 integrals and then simplifying.

A.11. Integrated Risk (Figure A.1)

Figure A.1 illustrates the integrated risk for each estimator when $p \in \{0.0, 0.25, 0.5, 0.75\}$ for a grid of (μ_0, σ_0).

Figure A.1 highlights some relevant features of the estimators. As expected, integrated risk improves as p increases, because the spike and dense components become more similar. The ridge's integrated risk surface is symmetric and does not plateau as μ_0 increases. At the same time, the lasso induces sparsity and therefore achieves lower integrated risk for smaller values of μ_0, when the dense component's mean is closer to zero. The elastic net trades off both behaviors and manages integrated risk well even as μ_0 increases.

A.12. Mean Squared Error Comparison

Table A.1 presents the MSE comparison for all p and (μ_0, σ_0) values that we consider. In bold are the minimum values of the MSE for each DGP.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

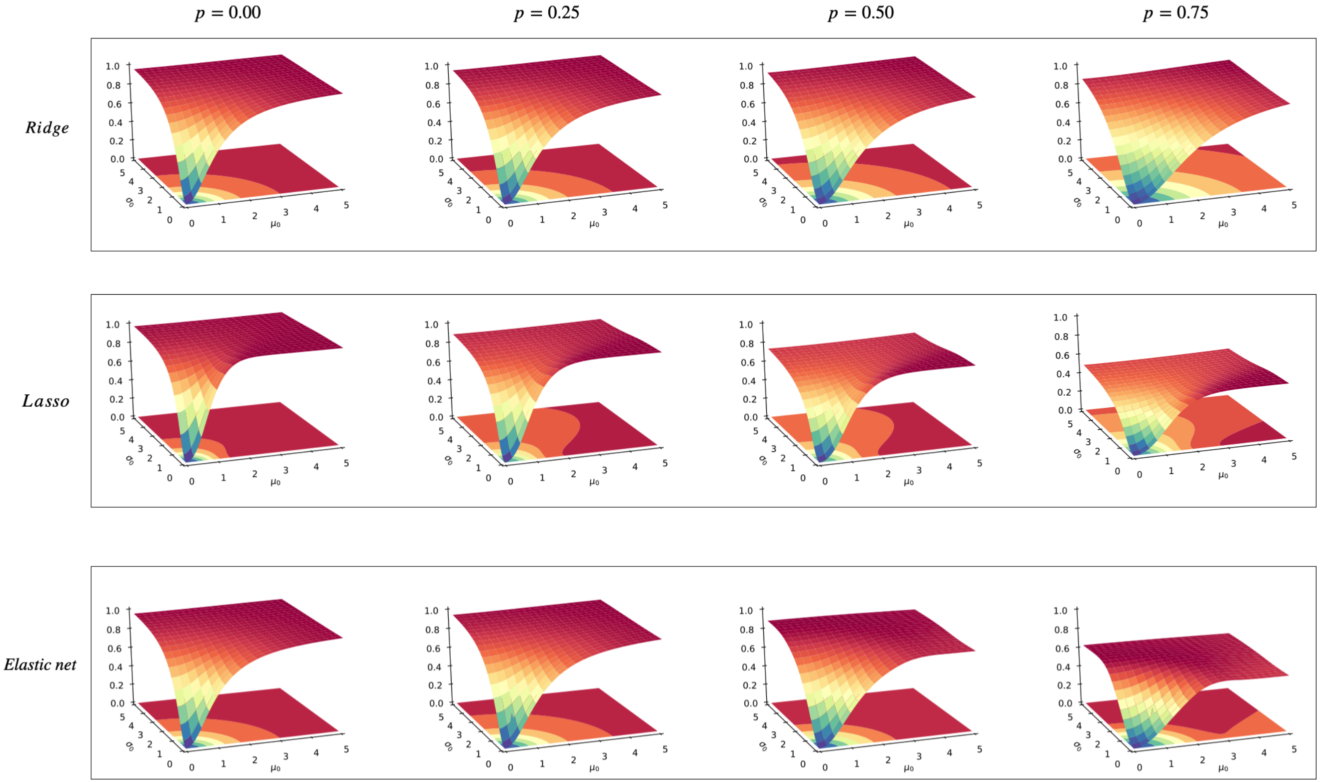

FIGURE A.1. Integrated risks for ridge, lasso, and elastic net for the spike and normal distribution over a (μ_0, σ_0) grid. Contours of the integrated risk surface are shown in the (μ_0, σ_0) plane. A color version of this figure can be found in the online Appendix.

Table A.1. MSE from CV for different spike and normal settings

			$N = 50$			$N = 250$			$N = 1,000$		
p	μ_0	σ_0	L	R	EN	L	R	EN	L	R	EN
0.0	0	1	0.521	**0.511**	0.519	**0.505**	0.512	0.511	0.502	0.508	**0.5**
0.0	2	1	0.528	0.518	**0.501**	0.51	0.51	**0.498**	**0.499**	0.502	0.502
0.0	4	1	**0.526**	0.533	0.535	**0.505**	0.516	0.509	0.507	**0.502**	0.505
0.0	0	3	0.961	0.928	**0.921**	0.927	0.904	**0.895**	0.904	0.9	**0.898**
0.0	2	3	0.98	0.989	**0.929**	0.922	0.927	**0.92**	0.913	0.906	**0.895**
0.0	4	3	**0.953**	1.014	0.954	**0.906**	0.909	0.912	**0.9**	0.9	0.904
0.0	0	5	1.039	**1.034**	1.066	**0.972**	0.996	0.978	0.97	0.968	**0.962**
0.0	2	5	**1.003**	1.021	1.01	0.985	**0.953**	0.955	0.965	**0.961**	0.963
0.0	4	5	1.021	**1.02**	1.052	0.964	**0.963**	0.976	0.969	0.966	**0.965**
0.25	0	1	**0.463**	0.48	0.467	**0.435**	0.436	0.437	0.427	0.427	**0.426**
0.25	2	1	0.653	**0.609**	0.636	0.606	**0.596**	0.599	0.6	**0.597**	0.603
0.25	4	1	0.837	0.813	**0.811**	0.808	0.797	**0.79**	0.795	**0.791**	0.794
0.25	0	3	0.927	**0.916**	0.938	0.887	**0.884**	0.896	**0.871**	0.871	0.878
0.25	2	3	**0.92**	0.931	0.962	**0.892**	0.897	0.9	**0.883**	0.888	0.896
0.25	4	3	**0.971**	1.018	0.988	0.909	0.91	**0.906**	0.906	**0.904**	0.906
0.25	0	5	**0.947**	0.976	1.023	**0.96**	0.966	0.966	0.955	0.95	**0.946**
0.25	2	5	1.003	**0.983**	1.0	**0.949**	0.973	0.955	0.962	**0.951**	0.953
0.25	4	5	1.018	1.035	**1.009**	0.97	0.966	**0.952**	**0.958**	0.958	0.963
0.5	0	1	0.4	**0.36**	0.369	0.343	0.342	**0.334**	0.339	**0.33**	0.336
0.5	2	1	0.624	**0.613**	0.632	**0.607**	0.608	0.625	**0.594**	0.609	0.601
0.5	4	1	**0.869**	0.917	0.892	0.836	**0.826**	0.827	**0.821**	0.827	0.825
0.5	0	3	0.886	**0.843**	0.871	0.839	0.836	**0.814**	0.822	**0.814**	0.816
0.5	2	3	**0.876**	0.915	0.941	0.858	0.847	**0.839**	**0.844**	0.849	0.85
0.5	4	3	**0.98**	0.986	1.012	0.909	0.903	**0.9**	0.896	**0.889**	0.89
0.5	0	5	**0.992**	1.039	1.07	0.965	0.948	**0.917**	**0.927**	0.931	0.935
0.5	2	5	0.966	1.017	**0.964**	**0.933**	0.936	0.938	**0.928**	0.934	0.942
0.5	4	5	1.016	**1.007**	1.043	**0.95**	0.956	0.956	0.951	0.946	**0.945**
0.75	0	1	**0.197**	0.235	0.24	0.202	0.203	**0.2**	0.203	0.201	**0.2**
0.75	2	1	0.549	**0.521**	0.537	0.504	**0.5**	0.509	0.502	**0.501**	0.507
0.75	4	1	0.831	0.849	**0.824**	**0.764**	0.764	0.788	0.767	0.767	**0.763**
0.75	0	3	**0.814**	0.817	0.862	**0.703**	0.718	0.711	0.698	**0.693**	0.696
0.75	2	3	0.859	**0.833**	0.944	0.773	0.784	**0.767**	0.758	**0.755**	0.756
0.75	4	3	**0.898**	0.911	0.926	**0.849**	0.857	0.865	0.843	0.85	**0.836**
0.75	0	5	1.05	1.033	**0.967**	0.876	0.888	**0.864**	**0.857**	0.868	0.876
0.75	2	5	**0.977**	1.035	1.03	0.922	0.895	**0.881**	0.893	0.876	**0.869**
0.75	4	5	**0.917**	1.103	1.079	**0.916**	0.918	0.926	0.906	0.904	**0.894**

A.13. Cubic Net (Figure A.2)

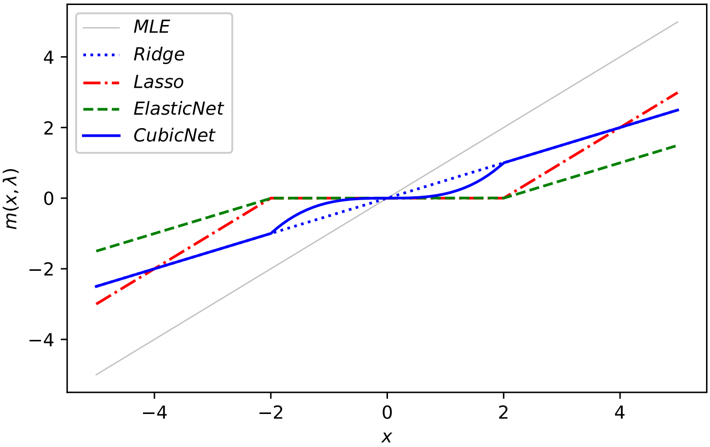

Figure A.2. Componentwise estimator functions for the MLE, ridge, lasso, elastic net, and cubic net. For the ridge, lasso, elastic net, and cubic net, the regularization parameters are set to $\lambda_1 = 1$ and $\lambda_2 = 2$. A color version of this figure can be found in the online Appendix.

REFERENCES

Abadie, A. & A. Kasy (2019) Choosing among regularized estimators in empirical economics: The risk of machine learning. *The Review of Economic and Statistics* 101(5), 743–762.

Chernozhukov, V., C. Hansen & Y. Liao (2017) A lava attack on the recovery of sums of dense and sparse signals. *Annals of Statistics* 45(1), 39–76.

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

James, W. & C. Stein (1961) Estimation with quadratic loss. In ed. J. Neyman, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379. University of California.

Jia, J. & B. Yu (2010) On model selection consistency of the elastic net when $p \gg n$. *Statistica Sinica* 20, 595–611.

Leeb, H. & B. M. Potscher (2006) Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* 22(1), 69–97.

Morris, C. N. (1983) Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association* 78(381), 47–55.

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

Stein, C. M. (1956) Inadmissibility of the usual estimator for the mean of a multivariate distribution. In ed. J. Neyman, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pp. 197–206. University of California Press.

van der Vaart, A. (1998) *Asymptotic Statistics* Cambridge University Press.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press

Zou, H. (2006) The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* 101(476), 1418–1429.

Zou, H. & T. Hastie (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, 301–320.

https://doi.org/10.1017/S0266466621000347 Published online by Cambridge University Press