The Risk of Machine Learning

Alberto Abadie   Maximilian Kasy
MIT   Harvard University

August 4, 2017

Abstract

Many applied settings in empirical economics involve simultaneous estimation of a large number of parameters. In particular, applied economists are often interested in estimating the effects of many-valued treatments (like teacher effects or location effects), treatment effects for many groups, and prediction models with many regressors. In these settings, machine learning methods that combine regularized estimation and data-driven choices of regularization parameters are useful to avoid over-fitting. In this article, we analyze the performance of a class of machine learning estimators that includes ridge, lasso and pretest in contexts that require simultaneous estimation of many parameters. Our analysis aims to provide guidance to applied researchers on (i) the choice between regularized estimators in practice and (ii) data-driven selection of regularization parameters. To address (i), we characterize the risk (mean squared error) of regularized estimators and derive their relative performance as a function of simple features of the data generating process. To address (ii), we show that data-driven choices of regularization parameters, based on Stein's unbiased risk estimate or on cross-validation, yield estimators with risk uniformly close to the risk attained under the optimal (unfeasible) choice of regularization parameters. We use data from recent examples in the empirical economics literature to illustrate the practical applicability of our results.

Alberto Abadie, Department of Economics, Massachusetts Institute of Technology, abadie@mit.edu. Maximilian Kasy, Department of Economics, Harvard University, maximiliankasy@fas.harvard.edu. We thank Gary Chamberlain, Ellora Derenoncourt, Jiaying Gu, Jérémy, L'Hour, José Luis Montiel Olea, Jann Spiess, Stefan Wager and seminar participants at several institutions for helpful comments and discussions.

1 Introduction

Applied economists often confront problems that require estimation of a large number of parameters. Examples include (a) estimation of causal (or predictive) effects for a large number of treatments such as neighborhoods or cities, teachers, workers and firms, or judges; (b) estimation of the causal effect of a given treatment for a large number of subgroups; and (c) prediction problems with a large number of predictive covariates or transformations of covariates. The machine learning literature provides a host of estimation methods, such as ridge, lasso, and pretest, which are particularly well adapted to high-dimensional problems. In view of the variety of available methods, the applied researcher faces the question of which of these procedures to adopt in any given situation. This article provides guidance on this choice based on the study of the risk properties (mean squared error) of a class of regularization-based machine learning methods.

A practical concern that generally motivates the adoption of machine learning procedures is the potential for severe over-fitting in high-dimensional settings. To avoid over-fitting, most machine learning procedures for "supervised learning" (that is, regression and classification methods for prediction) involve two key features, (i) regularized estimation and (ii) data-driven choice of regularization parameters. These features are also central to more familiar non-parametric estimation methods in econometrics, such as kernel or series regression.

Main takeaways For the empirical practitioner, we would like to emphasize two main takeaway messages of this paper. First, there is no one method for regularization that is universally optimal. Different methods work well in different types of settings, and we provide guidance based on theoretical considerations, simulations, and empirical applications. Second, a choice of tuning parameters using cross-validation or Stein's unbiased risk estimate is guaranteed to work well in high-dimensional estimation and prediction settings, under mild conditions. For the econometric theorist, our main contribution are our results on uniform loss consistency for a general class of regularization procedures. In seeming contrast to some results in the literature, data-driven tuning performs uniformly well in high-dimensional settings. Further interesting findings include that (i) lasso is surprisingly robust in its performance across many settings, and (ii) flexible regularization methods such as nonparametric empirical Bayes dominate in very high-dimensional settings, but are dominated by more parsimonious regularizers in more moderate dimensions.

Setup In this article, we consider the canonical problem of estimating the unknown means, $\mu_1, \ldots, \mu_n$, of a potentially large set of observed random variables, $X_1, \ldots, X_n$. After some transformations, our setup covers applications (a)-(c) mentioned above and many others. For example, in the context of a randomized experiment with n subgroups, X_i is the difference in the sample averages of an outcome variable between treated and non-treated for subgroup i, and μ_i is the average treatment effect on the same outcome and subgroup. Moreover, as we discuss in Section 2.1, the many means problem analyzed in this article encompasses the problem of nonparametric estimation of a regression function.

We consider componentwise estimators of the form $\widehat{\mu}_i = m(X_i, \lambda)$, where λ is a non-negative regularization parameter. Typically, $m(x, 0) = x$, so that $\lambda = 0$ corresponds to the unregularized estimator $\widehat{\mu}_i = X_i$. Positive values of λ typically correspond to regularized estimators, which shrink towards zero, $|\widehat{\mu}_i| \leq |X_i|$. The value $\lambda = \infty$ typically implies maximal shrinkage: $\widehat{\mu}_i = 0$ for $i = 1, \ldots, n$. Shrinkage towards zero is a convenient normalization but it is not essential. Shifting X_i by a constant to $X_i - c$, for $i = 1, \ldots, n$, results in shrinkage towards c.

The risk function of regularized estimators Our article is structured according to the two mentioned features of machine learning procedures, regularization and data-driven choice of regularization parameters. We first focus on feature (i) and study the risk properties (mean squared error) of regularized estimators with fixed and with oracle-optimal regularization parameters. We show that for any given data generating process there is an (infeasible) risk-optimal regularized componentwise estimator. This estimator has the form of the posterior mean of μ_I given X_I and given the empirical distribution of $\mu_1, \ldots, \mu_n$, where I is a random variable with uniform distribution on the set of indices $\{1, 2, \ldots, n\}$.

The optimal regularized estimator is useful to characterize the risk properties of machine learning estimators. It turns out that, in our setting, the risk function of any regularized estimator can be expressed as a function of the distance between that regularized estimator and the optimal one.

Instead of conditioning on $\mu_1, \ldots, \mu_n$, one can consider the case where each (X_i, μ_i) is a realization of a random vector (X, μ) with distribution π and a notion of risk that is integrated over the distribution of μ in the population. For this alternative definition of risk, we derive results analogous to those of the previous paragraph.

We next turn to a family of parametric models for π. We consider models that allow for a probability mass at zero in the distribution of μ, corresponding to the notion of sparsity, while conditional on $\mu \neq 0$ the distribution of μ is normal around some grand mean. For these parametric models we derive analytic risk functions under oracle choices of risk minimizing values for λ, which allow for an intuitive discussion of the relative performance of alternative estimators. We focus our attention on three estimators that are widespread in the empirical machine learning literature: ridge, lasso, and pretest. When the point-mass of true zeros is small, ridge tends to perform better than lasso or pretest. When there is a sizable share of true zeros, the ranking of the estimators depends on the other characteristics of the distribution of μ: (a) if the non-zero parameters are smoothly distributed in a vicinity of zero, ridge still performs best; (b) if most of the distribution of non-zero parameters assigns large probability to a set well-separated from zero, pretest estimation tends to perform well; and (c) lasso tends to do comparatively well in intermediate cases that fall somewhere between (a) and (b), and overall is remarkably robust across the different specifications. This characterization of the relative performance of ridge, lasso, and pretest is consistent with the results that we obtain for the empirical applications discussed later in the article.

Data-driven choice of regularization parameters The second part the article turns to feature (ii) of machine learning estimators and studies the data-driven choice of regularization parameters. We consider choices of regularization parameters based on the minimization of a criterion function that estimates risk. Ideally, a machine learning estimator evaluated at a data-driven choice of the regularization parameter would have a risk function that is uniformly close to the risk function of the infeasible estimator using an oracle-optimal regularization parameter (which minimizes true risk). We show this type of uniform consistency can be achieved under fairly mild conditions whenever the dimension of the problem under consideration is large. This is in stark contrast to well-known results in Leeb and Pötscher (2006) for low-dimensional settings. We further provide fairly weak conditions under which machine learning estimators with data-driven choices of the regularization parameter, based on Stein's unbiased risk estimate (SURE) and on cross-validation (CV), attain uniform risk consistency. In addition to allowing data-driven selection of regularization parameters, uniformly consistent estimation of the risk of shrinkage estimators can be used to select among alternative shrinkage estimators on the basis of their estimated risk in specific empirical settings.

Applications We illustrate our results in the context of three applications taken from the empirical economics literature. The first application uses data from Chetty and Hendren (2015) to study the effects of locations on intergenerational earnings mobility of children. The second application uses data from the event-study analysis in Della Vigna and La Ferrara (2010) who investigate whether the stock prices of weapon-producing companies react to changes in the intensity of conflicts in countries under arms trade embargoes. The third application considers nonparametric estimation of a Mincer equation using data from the Current Population Survey (CPS), as in Belloni and Chernozhukov (2011). The presence of many neighborhoods in the first application, many weapon producing companies in the second one, and many series regression terms in the third one makes these estimation problems high-dimensional.

These examples showcase how simple features of the data generating process affect the relative performance of machine learning estimators. They also illustrate the way in which consistent estimation of the risk of shrinkage estimators can be used to choose regularization parameters and to select among different estimators in practice. For the estimation of location effects in Chetty and Hendren (2015) we find estimates that are not overly dispersed around their mean and no evidence of sparsity. In this setting, ridge outperforms lasso and pretest in terms of estimated mean squared error. In the setting of the event-study analysis in Della Vigna and La Ferrara (2010), our results suggest that a large fraction of values of parameters are closely concentrated around zero, while a smaller but non-negligible fraction of parameters are positive and substantially separated from zero. In this setting, pretest dominates. Similarly to the result for the setting in Della Vigna and La Ferrara (2010), the estimation of the parameters of a Mincer equation in Belloni and Chernozhukov (2011) suggests a sparse approximation to the distribution of parameters. Substantial shrinkage at the tails of the distribution is still helpful in this setting, so that lasso dominates.

Relation to the literature This paper builds on a substantial literature in statistical decision theory and machine learning which we review in Section 2.2 below. In order to provide accessible guidance to applied researchers our discussion includes some known results from the theoretical literature in addition to new findings. Variants of our Theorem 1, characterizing risk by the distance of an estimator to the optimal regularizer, can be found in Robbins (1951) and later references; see also Zhang (2003). Appendix 2 of Donoho and Johnstone (1994) provides analytic expressions for the componentwise risk of ridge, lasso, and pretest similar to those derived in section 3.2, but only for the case of unit variance. Our expressions for empirical Bayes risk of these estimators in the spike and normal setting appear to be new. More significantly, the same holds for our results regarding uniform loss consistency of data driven choices of λ using SURE or CV. There are some related discussions of SURE and CV in the literature, including Donoho and Johnstone (1995), Xie, Kou, and Brown (2012) and Chetverikov and Liao (2016). In contrast to much of the theoretical literature, our analysis is distinguished by focusing on risk functions rather than minimax risk over some class of data generating processes (DGPs) or on rates of convergence. Our more disaggregated perspective allows us to discuss the relative performance of alternative estimators as a function of the DGP, rather than recommending one estimator by virtue of its properties for some restricted class of DGPs.

In addition, one of the aims of this article is to provide guidance to empirical researchers. While the main themes of this article – shrinkage, empirical Bayes, and high-dimensional thresholding estimation – are areas or rapid expansion and increased interest in the methodological literatures in econometrics and statistics (see, e.g., Efron, 2010; Johnstone, 2015), these powerful techniques have not yet been widely adopted in the empirical practice in economics. This is despite the fact many empirical applications in economics involve the estimation of a large number of means or fixed effects (e.g., location effects), a setting the techniques studied in this article were designed for. It is, therefore, important to demonstrate the empirical potential and practical relevance of the methods that we analyze in this article. In order to do so we use detailed simulations and three actual empirical applications in economics.

Roadmap The rest of this article is structured as follows. Section 2 introduces our setup: the canonical problem of estimating a vector of means under quadratic loss. Section 2.1 discusses a series of examples from empirical economics that are covered by our setup. Section 2.2 discusses the setup of this article in the context of the machine learning literature and of the older literature on estimation of normal means. Section 3 provides characterizations of the risk function of regularized estimators in our setting. We derive a general characterization in Section 3.1. Sections 3.2 and 3.3 provide analytic formulas for risk under additional assumptions. In particular, in Section 3.3 we derive analytic formulas for risk in a spike-and-normal model . These characterizations allow for a comparison of the mean squared error of alternative procedures and yield recommendations for the choice of an estimator. Section 4 turns to data-driven choices of regularization parameters. We show uniform risk consistency results for Stein's unbiased risk estimate and for cross-validation. Section 5 discusses extensions and explains the apparent contradiction between our results and those in Leeb and Pötscher (2005). Section 6 reports simulation results. Section 7 discusses several empirical applications. Section 8 concludes. The appendix contains proofs and supplemental materials.

2 Setup

Throughout this paper, we consider the following setting. We observe a realization of an n-vector of real-valued random variables, $\boldsymbol{X} = (X_1, \ldots, X_n)'$, where the components of $\boldsymbol{X}$ are mutually independent with finite mean μ_i and finite variance σ_i^2, for $i = 1, \ldots, n$. Our goal is to estimate $\mu_1, \ldots, \mu_n$.

In many applications, the X_i arise as preliminary least squares estimates of the coefficients of interest, μ_i. Consider, for instance, a randomized controlled trial where randomization of treatment assignment is carried out separately for n non-overlapping subgroups. Within each subgroup, the difference in the sample averages between treated and control units, X_i, has mean equal to the average treatment effect for that group in the population, μ_i. Further examples are discussed in Section 2.1 below.

Componentwise estimators We restrict our attention to componentwise estimators of μ_i,

$$\widehat{\mu}_i = m(X_i, \lambda),$$

where $m : \mathbb{R} \times [0, \infty] \mapsto \mathbb{R}$ defines an estimator of μ_i as a function of X_i and a non-negative regularization parameter, λ. The parameter λ is common across the components i but might depend on the vector $\boldsymbol{X}$. We study data-driven choices $\widehat{\lambda}$ in Section 4 below, focusing in particular on Stein's unbiased risk estimate (SURE) and cross-validation (CV).

Popular estimators of this componentwise form are ridge, lasso, and pretest. They are defined as follows:

$$\begin{aligned} m_R(x, \lambda) &= \underset{m \in \mathbb{R}}{\operatorname{argmin}} \;\; (x - m)^2 + \lambda m^2 && \text{(ridge)} \\ &= \frac{1}{1 + \lambda} x, \end{aligned}$$

$$\begin{aligned} m_L(x, \lambda) &= \underset{m \in \mathbb{R}}{\operatorname{argmin}} \;\; (x - m)^2 + 2\lambda |m| && \text{(lasso)} \\ &= 1(x < -\lambda)(x + \lambda) + 1(x > \lambda)(x - \lambda), \end{aligned}$$

$$m_{PT}(x, \lambda) = \underset{m \in \mathbb{R}}{\operatorname{argmin}} \;\; (x - m)^2 + \lambda^2 1(m \neq 0) \qquad \text{(pretest)}$$

$$= 1(|x| > \lambda)x,$$

where $1(A)$ denotes the indicator function, which equals 1 if A holds and 0 otherwise. Figure 1 plots $m_R(x, \lambda)$, $m_L(x, \lambda)$ and $m_{PT}(x, \lambda)$ as functions of x. For reasons apparent in Figure 1, ridge, lasso, and pretest estimators are sometimes referred to as linear shrinkage, soft thresholding, and hard thresholding, respectively. As we discuss below, the problem of determining the optimal choice among these estimators in terms of minimizing mean squared error is equivalent to the problem of determining which of these estimators best approximates a certain optimal estimating function, m^*.

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ and $\widehat{\boldsymbol{\mu}} = (\widehat{\mu}_1, \ldots, \widehat{\mu}_n)'$, where for simplicity we leave the dependence of $\widehat{\mu}$ on λ implicit in our notation. Let $P_1, \ldots, P_n$ be the distributions of $X_1, \ldots, X_n$, and let $\boldsymbol{P} = (P_1, \ldots, P_n)$.

Loss and risk We evaluate estimates based on the squared error loss function, or compound loss,

$$L_n(\boldsymbol{X}, m(\cdot, \lambda), \boldsymbol{P}) = \frac{1}{n} \sum_{i=1}^{n} \big(m(X_i, \lambda) - \mu_i\big)^2,$$

where L_n depends on $\boldsymbol{P}$ via $\boldsymbol{\mu}$. We will use expected loss to rank estimators. There are different ways of taking this expectation, resulting in different risk functions, and the distinction between them is conceptually important.

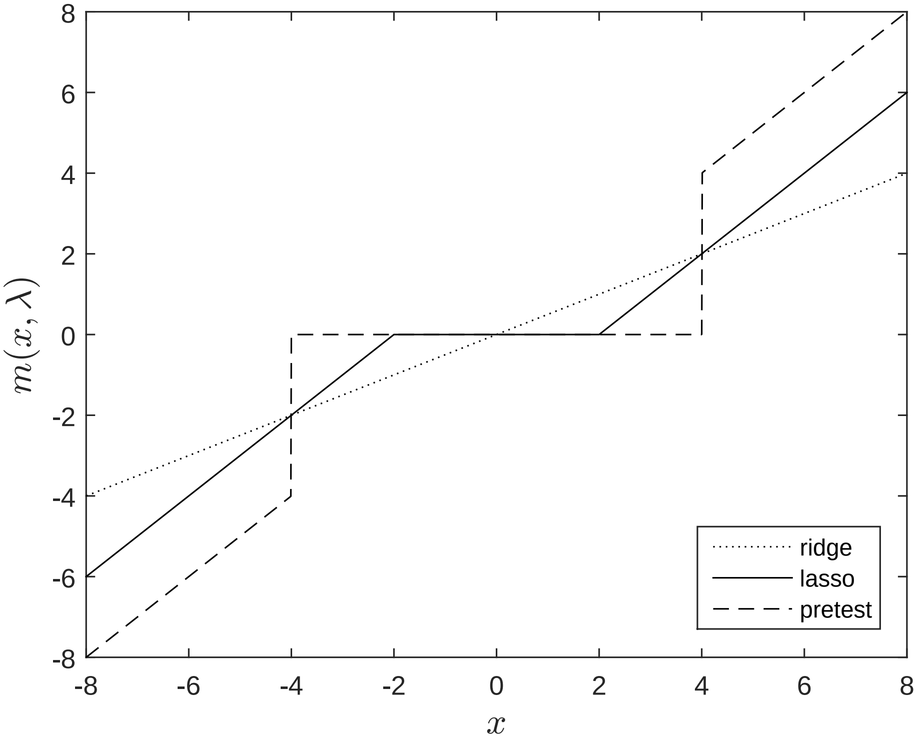
Componentwise risk fixes P_i and considers the expected squared error of $\widehat{\mu}_i$ as an estimator of μ_i,

$$R(m(\cdot, \lambda), P_i) = E[(m(X_i, \lambda) - \mu_i)^2 | P_i].$$

Compound risk averages componentwise risk over the empirical distribution of P_i across the components $i = i, \ldots, n$. Compound risk is given by the expectation of compound loss L_n given $\boldsymbol{P}$,

$$\begin{aligned} R_n(m(\cdot, \lambda), \boldsymbol{P}) &= E[L_n(\boldsymbol{X}, m(\cdot, \lambda), \boldsymbol{P}) | \boldsymbol{P}] \\ &= \frac{1}{n} \sum_{i=1}^{n} E[(m(X_i, \lambda) - \mu_i)^2 | P_i] \end{aligned}$$

Figure 1: Estimators



This graph plots $m_R(x, \lambda)$, $m_L(x, \lambda)$, and $m_{PT}(x, \lambda)$ as functions of x. The regularization parameters are $\lambda = 1$ for ridge, $\lambda = 2$ for lasso, and $\lambda = 4$ for pretest.

$$= \frac{1}{n} \sum_{i=1}^{n} R(m(\cdot, \lambda), P_i).$$

Finally, *integrated* (or *empirical Bayes*) *risk* considers $P_1, \ldots, P_n$ to be themselves draws from some population distribution, Π. This induces a joint distribution, π, for (X_i, μ_i). Throughout the article, we will often use a subscript π to denote characteristics of the joint distribution of (X_i, μ_i). Integrated risk refers to loss integrated over π or, equivalently, componentwise risk integrated over Π,

$$\begin{aligned}
\bar{R}(m(\cdot, \lambda), \pi) &= E_\pi[L_n(\boldsymbol{X}, m(\cdot, \lambda), \boldsymbol{P})] \\
&= E_\pi[(m(X_i, \lambda) - \mu_i)^2] \\
&= \int R(m(\cdot, \lambda), P_i) d\Pi(P_i). \qquad (1)
\end{aligned}$$

Notice the similarity between compound risk and integrated risk: they differ only by replacing an empirical (sample) distribution by a population distribution. For large n, the difference between the two vanishes, as we will explore in Section 4.

Regularization parameter Throughout, we will use $R_n(m(\cdot, \lambda), \boldsymbol{P})$ to denote the risk function of the estimator $m(\cdot, \lambda)$ with fixed (non-random) λ, and similarly for $\bar{R}(m(\cdot, \lambda), \pi)$. In contrast, $R_n(m(\cdot, \widehat{\lambda}_n), \boldsymbol{P})$ is the risk function taking into account the randomness of $\widehat{\lambda}_n$, where the latter is chosen in a data-dependent manner, and similarly for $\bar{R}(m(\cdot, \widehat{\lambda}_n), \pi)$.

For a given $\boldsymbol{P}$, we define the "oracle" selector of the regularization parameter as the value of λ that minimizes compound risk,

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

whenever the argmin exists. We use $\lambda^*_R(\boldsymbol{P})$, $\lambda^*_L(\boldsymbol{P})$ and $\lambda^*_{PT}(\boldsymbol{P})$ to denote the oracle selectors for ridge, lasso, and pretest, respectively. Analogously, for a given π, we define

$$\bar{\lambda}^*(\pi) = \underset{\lambda \in [0, \infty]}{\operatorname{argmin}} \; \bar{R}(m(\cdot, \lambda), \pi) \qquad (2)$$

whenever the argmin exists, with $\bar{\lambda}^*_R(\pi)$, $\bar{\lambda}^*_L(\pi)$, and $\bar{\lambda}^*_{PT}(\pi)$ for ridge, lasso, and pretest, respectively. In Section 3, we characterize compound and integrated risk for fixed λ and for the oracle-optimal λ. In Section 4 we show that data-driven choices $\widehat{\lambda}_n$ are, under certain conditions, as good as the oracle-optimal choice, in a sense to be made precise.

2.1 Empirical examples

Our setup describes a variety of settings often encountered in empirical economics, where $X_1, \ldots, X_n$ are unbiased or close-to-unbiased but noisy least squares estimates of a set of parameters of interest, $\mu_1, \ldots, \mu_n$. As mentioned in the introduction, examples include (a) studies estimating causal or predictive effects for a large number of treatments such as neighborhoods, cities, teachers, workers, firms, or judges; (b) studies estimating the causal effect of a given treatment for a large number of subgroups; and (c) prediction problems with a large number of predictive covariates or transformations of covariates.

Large number of treatments Examples in the first category include Chetty and Hendren (2015), who estimate the effect of geographic locations on intergenerational mobility for a large number of locations. Chetty and Hendren use differences between the outcomes of siblings whose parents move during their childhood in order to identify these effects. The problem of estimating a large number of parameters also arises in the teacher value-added literature when the objects of interest are individual teachers' effects, see, for instance, Chetty, Friedman, and Rockoff (2014). In labor economics, estimation of firm and worker effects in studies of wage inequality has been considered in Abowd, Kramarz, and Margolis (1999). Another example within the first category is provided by Abrams, Bertrand, and Mullainathan (2012), who estimate differences in the effects of defendant's race on sentencing across individual judges.

Treatment for large number of subgroups Within the second category, which consists of estimating the effect of a treatment for many sub-populations, our setup can be applied to the estimation of heterogeneous causal effects of class size on student outcomes across many subgroups. For instance, project STAR (Krueger, 1999) involved experimental assignment of students to classes of different sizes in 79 schools. Causal effects for many subgroups are also of interest in medical contexts or for active labor market programs, where doctors / policy makers have to decide on treatment assignment based on individual characteristics. In some empirical settings, treatment impacts are individually estimated for each sample unit. This is often the case in empirical finance, where event studies are used to estimate reactions of stock market prices to newly available information. For example, Della Vigna and La Ferrara (2010) estimate the effects of changes in the intensity of armed conflicts in countries under arms trade embargoes on the stock market prices of arms-manufacturing companies.

Prediction with many regressors The third category is prediction with many regressors. This category fits in the setting of this article after orthogonalization of the regressors. Prediction with many regressors arises, in particular, in macroeconomic forecasting. Stock and Watson (2012), in an analysis complementing the present article, evaluate various procedures in terms of their forecast performance for a number of macroeconomic time series for the United States. Regression with many predictors also arises in series regression, where series terms are transformations of a set of predictors. Series regression and its asymptotic properties have been widely studied in econometrics (see for instance Newey, 1997). Wasserman (2006, Sections 7.2-7.3) provides an illuminating discussion of the equivalence between the normal means model studied in this article and nonparametric regression estimation. For that setting, $X_1, \dots, X_n$ and $\mu_1, \dots, \mu_n$ correspond to the estimated and true regression coefficients on an orthogonal basis of functions. Application of lasso and pretesting to series regression is discussed, for instance, in Belloni and Chernozhukov (2011). Appendix A.1 further discusses the relationship between the normal means model and prediction models.

In Section 7, we return to three of these applications, revisiting the estimation of location effects on intergenerational mobility, as in Chetty and Hendren (2015), the effect of changes in the intensity of conflicts in arms-embargo countries on the stock prices of arms manufacturers, as in Della Vigna and La Ferrara (2010), and nonparametric series estimation of a Mincer equation, as in Belloni and Chernozhukov (2011).

2.2 Statistical literature

Machine learning methods are becoming widespread in econometrics – see, for instance, Athey and Imbens (2015) and Kleinberg, Ludwig, Mullainathan, and Obermeyer (2015). A large number of estimation procedures are available to the applied researcher. Textbooks such as Hastie, Tibshirani, and Friedman (2009) or Murphy (2012) provide an introduction to machine learning. Lasso, which was first introduced by Tibshirani (1996), is becoming particularly popular in applied economics. Belloni and Chernozhukov (2011) provide a review of lasso including theoretical results and applications in economics.

Much of the research on machine learning focuses on algorithms and computational issues, while the formal statistical properties of machine learning estimators have received less attention. However, an older and superficially unrelated literature in mathematical statistics and statistical decision theory on the estimation of the normal means model has produced many deep results which turn out to be relevant for understanding the behavior of estimation procedures in non-parametric statistics and machine learning. A foundational article in this literature is James and Stein (1961), who study the case $X_i \sim N(\mu_i, 1)$. They show that the estimator $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}$ is inadmissible whenever $n \geq 3$. That is, there exists a (shrinkage) estimator that has mean squared error smaller than the mean squared error of $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}$ for all values of $\boldsymbol{\mu}$. Brown (1971) provides more general characterizations of admissibility and shows that this dependence on dimension is deeply connected to the recurrence or transience of Brownian motion. Stein et al. (1981) characterizes the risk function of arbitrary estimators, $\widehat{\boldsymbol{\mu}}$, and based on this characterization proposes an unbiased estimator of the mean squared error of a given estimator, labeled "Stein's unbiased risk estimator" or SURE. We return to SURE in Section 4.2 as a method to produce data-driven choices of regularization parameters. In section 4.3, we discuss cross-validation as an alternative method to obtain data-driven choices of regularization parameters in the context studied in this article.[1]

A general approach for the construction of regularized estimators, such as the one

[1] See, e.g., Arlot and Celisse (2010) for a survey on cross-validation methods for model selection.

proposed by James and Stein (1961), is provided by the empirical Bayes framework, first proposed in Robbins (1956) and Robbins (1964). A key insight of the empirical Bayes framework, and the closely related compound decision problem framework, is that trying to minimize squared error in higher dimensions involves a trade-off across components of the estimand. The data are informative about which estimators and regularization parameters perform well in terms of squared error and thus allow one to construct regularized estimators that dominate the unregularized $\widehat{\boldsymbol{\mu}} = \boldsymbol{X}$. This intuition is elaborated on in Stigler (1990). The empirical Bayes framework was developed further by Efron and Morris (1973) and Morris (1983), among others. Good reviews and introductions can be found in Zhang (2003) and Efron (2010).

In Section 4 we consider data-driven choices of regularization parameters and emphasize uniform validity of asymptotic approximations to the risk function of the resulting estimators. Lack of uniform validity of standard asymptotic characterizations of risk (as well as of test size) in the context of pretest and model-selection based estimators in low-dimensional settings has been emphasized by Leeb and Pötscher (2005).

The minimax risk properties and convergence rates of regularization procedures including the ones we consider (ridge, lasso, pretest, and tuning using SURE) have been investigated in the context of wavelet estimation by Donoho and Johnstone (1994) and Donoho and Johnstone (1995), among others; the textbook Johnstone (2015) provides an overview. The asymptotic properties of SURE and of CV in large dimensional settings have also been considered by Xie et al. (2012) and Chetverikov and Liao (2016). In our simulations and applications we consider nonparametric empirical Bayes (NPEB) as a point of comparison for more parsimonious regularization procedures. Jiang and Zhang (2009) show strong uniform risk consistency results for NPEB; Koenker and Mizera (2014) propose feasible procedures to implement NPEB in practice.

While in this article we study risk-optimal estimation of $\boldsymbol{\mu}$, a related literature has focused on the estimation of confidence sets for the same parameter. Wasserman (2006, Section 7.8) and Casella and Hwang (2012) surveys some results in this literature. Efron

(2010) studies hypotheses testing in high dimensional settings from an empirical Bayes perspective.

3 The risk function

We now turn to our first set of formal results, which pertain to the mean squared error of regularized estimators. Our goal is to guide the researcher's choice of estimator by describing the conditions under which each of the alternative machine learning estimators performs better than the others.

We first derive a general characterization of the mean squared error of regularized estimators. This characterization is based on the geometry of estimating functions m as depicted in Figure 1. It is a-priori not obvious which of these functions is best suited for estimation. We show that for any given data generating process there is an *optimal* function $m^*_{\boldsymbol{P}}$ that minimizes mean squared error. Moreover, we show that the mean squared error for an *arbitrary* m is equal, up to a constant, to the L^2 distance between m and $m^*_{\boldsymbol{P}}$. A function m thus yields a good estimator if it is able to approximate the shape of $m^*_{\boldsymbol{P}}$ well.

In Section 3.2, we provide analytic expressions for the componentwise risk of ridge, lasso, and pretest estimators, imposing the additional assumption of normality. Summing or integrating componentwise risk over some distribution for (μ_i, σ_i) delivers expressions for compound and integrated risk.

In Section 3.3, we turn to a specific parametric family of data generating processes where each μ_i is equal to zero with probability p, reflecting the notion of sparsity, and is otherwise drawn from a normal distribution with some mean μ_0 and variance σ_0^2. For this parametric family indexed by (p, μ_0, σ_0), we provide analytic risk functions and visual comparisons of the relative performance of alternative estimators. This allows us to identify key features of the data generating process which affect the relative performance of alternative estimators.

3.1 General characterization

Recall the setup introduced in Section 2, where we observe n jointly independent random variables $X_1, \ldots, X_n$, with means $\mu_1, \ldots, \mu_n$. We are interested in the mean squared error for the compound problem of estimating all $\mu_1, \ldots, \mu_n$ simultaneously. In this formulation of the problem, $\mu_1, \ldots, \mu_n$ are fixed unknown parameters.

Let I be a random variable with a uniform distribution over the set $\{1, 2, \ldots, n\}$ and consider the random component (X_I, μ_I) of $(\boldsymbol{X}, \boldsymbol{\mu})$. This construction induces a mixture distribution for (X_I, μ_I) (conditional on $\boldsymbol{P}$),

$$(X_I, \mu_I)|\boldsymbol{P} \sim \frac{1}{n}\sum_{i=1}^{n} P_i \delta_{\mu_i},$$

where $\delta_{\mu_1}, \ldots, \delta_{\mu_n}$ are Dirac measures at $\mu_1, \ldots, \mu_n$. Based on this mixture distribution, define the conditional expectation

$$m^*_{\boldsymbol{P}}(x) = E[\mu_I | X_I = x, \boldsymbol{P}]$$

and the average conditional variance

$$v^*_{\boldsymbol{P}} = E\big[\text{var}(\mu_I | X_I, \boldsymbol{P})|\boldsymbol{P}\big].$$

The next theorem characterizes the compound risk of an estimator in terms of the average squared discrepancy relative to $m^*_{\boldsymbol{P}}$, which implies that $m^*_{\boldsymbol{P}}$ is optimal (lowest mean squared error) for the compound problem.

Theorem 1 (Characterization of risk functions)
Under the assumptions of Section 2 and $\sup_{\lambda\in[0,\infty]} E[(m(X_I, \lambda))^2|\boldsymbol{P}] < \infty$*, the compound risk function* R_n *of* $\widehat{\mu}_i = m(X_i, \lambda)$ *can be written as*

$$R_n(m(\cdot, \lambda), \boldsymbol{P}) = v^*_{\boldsymbol{P}} + E\big[(m(X_I, \lambda) - m^*_{\boldsymbol{P}}(X_I))^2|\boldsymbol{P}\big],$$

which implies

$$\lambda^*(\boldsymbol{P}) = \underset{\lambda\in[0,\infty]}{\text{argmin}}\ E\big[(m(X_I, \lambda) - m^*_{\boldsymbol{P}}(X_I))^2|\boldsymbol{P}\big]$$

whenever $\lambda^*(\boldsymbol{P})$ *is well defined.*

The proof of this theorem and all further results can be found in the appendix.

The statement of this theorem implies that the risk of componentwise estimators is equal to an irreducible part $v^*_{\boldsymbol{P}}$, plus the L^2 distance of the estimating function $m(., \lambda)$ to the infeasible optimal estimating function $m^*_{\boldsymbol{P}}$. A given data generating process $\boldsymbol{P}$ maps into an optimal estimating function $m^*_{\boldsymbol{P}}$, and the relative performance of alternative estimators m depends on how well they approximate $m^*_{\boldsymbol{P}}$.

We can easily write $m^*_{\boldsymbol{P}}$ explicitly because the conditional expectation defining $m^*_{\boldsymbol{P}}$ is a weighted average of the values taken by μ_i. Suppose, for example, that $X_i \sim N(\mu_i, 1)$ for $i = 1 \ldots n$. Let ϕ be the standard normal probability density function. Then,

$$m^*_{\boldsymbol{P}}(x) = \frac{\displaystyle\sum_{i=1}^{n} \mu_i \, \phi(x - \mu_i)}{\displaystyle\sum_{i=1}^{n} \phi(x - \mu_i)}.$$

Theorem 1 conditions on the empirical distribution of $\mu_1, \ldots, \mu_n$, which corresponds to the notion of compound risk. Replacing this empirical distribution by the population distribution π, so that

$$(X_i, \mu_i) \sim \pi,$$

results analogous to those in Theorem 1 are obtained for the integrated risk and the integrated oracle selectors in equations (1) and (2). That is, let

$$\bar{m}^*_\pi(x) = E_\pi[\mu_i | X_i = x]$$

and

$$\bar{v}^*_\pi = E_\pi[\mathrm{var}_\pi(\mu_i | X_i)],$$

and assume $\sup_{\lambda \in [0,\infty]} E_\pi[(m(X_i, \lambda) - \mu_i)^2] < \infty$. Then

$$\bar{R}(m(\cdot, \lambda), \pi) = \bar{v}^*_\pi + E_\pi\left[(m(X_i, \lambda) - \bar{m}^*_\pi(X_i))^2\right]$$

and

$$\bar{\lambda}^*(\pi) = \underset{\lambda \in [0,\infty]}{\operatorname{argmin}} \; E_\pi\left[(m(X_i, \lambda) - \bar{m}^*_\pi(X_i))^2\right]. \tag{3}$$

The proof of these assertions is analogous to the proof of Theorem 1. $m^*_{\boldsymbol{P}}$ and $\bar{m}^*_\pi$ are optimal componentwise estimators or “shrinkage functions” in the sense that they minimize the compound and integrated risk, respectively.

3.2 Componentwise risk

The characterization of the risk of componentwise estimators in the previous section relies only on the existence of second moments. Explicit expressions for compound risk and integrated risk can be derived under additional structure. We shall now consider a setting in which the X_i are normally distributed,

$$X_i \sim N(\mu_i, \sigma_i^2).$$

This is a particularly relevant scenario in applied research, where the X_i are often unbiased estimators with a normal distribution in large samples (as in examples (a) to (c) in Sections 1 and 2.1). For concreteness, we will focus on the three widely used componentwise estimators introduced in Section 2, ridge, lasso, and pretest, whose estimating functions m were plotted in Figure 1. The following lemma provides explicit expressions for the componentwise risk of these estimators.

Lemma 1 (Componentwise risk)
Consider the setup of Section 2. Then, for $i = 1, \ldots, n$, the componentwise risk of ridge is:

$$R(m_R(\cdot, \lambda), P_i) = \left(\frac{1}{1+\lambda}\right)^2 \sigma_i^2 + \left(1 - \frac{1}{1+\lambda}\right)^2 \mu_i^2.$$

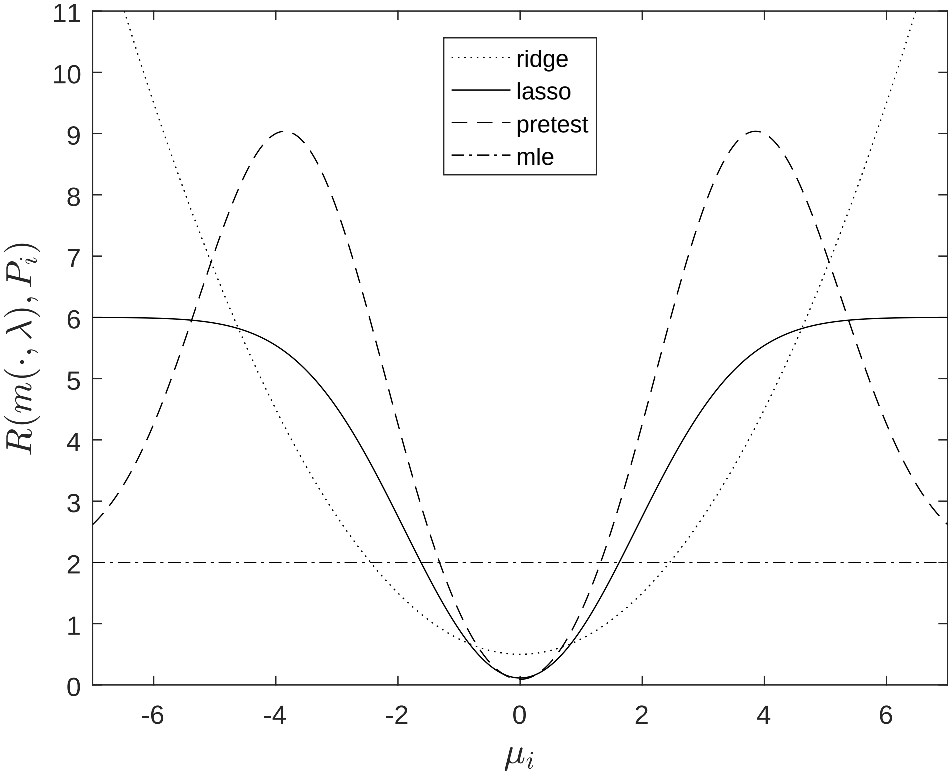
Assume in addition that X_i has a normal distribution. Then, the componentwise risk of lasso is

$$\begin{aligned} R(m_L(\cdot, \lambda), P_i) = & \left(1 + \Phi\Big(\frac{-\lambda - \mu_i}{\sigma_i}\Big) - \Phi\Big(\frac{\lambda - \mu_i}{\sigma_i}\Big)\right)(\sigma_i^2 + \lambda^2) \\ & + \left(\Big(\frac{-\lambda - \mu_i}{\sigma_i}\Big)\phi\Big(\frac{\lambda - \mu_i}{\sigma_i}\Big) + \Big(\frac{-\lambda + \mu_i}{\sigma_i}\Big)\phi\Big(\frac{-\lambda - \mu_i}{\sigma_i}\Big)\right)\sigma_i^2 \\ & + \left(\Phi\Big(\frac{\lambda - \mu_i}{\sigma_i}\Big) - \Phi\Big(\frac{-\lambda - \mu_i}{\sigma_i}\Big)\right)\mu_i^2. \end{aligned}$$

Under the same conditions, the componentwise risk of pretest is

$$R(m_{PT}(\cdot, \lambda), P_i) = \left(1 + \Phi\Big(\frac{-\lambda - \mu_i}{\sigma_i}\Big) - \Phi\Big(\frac{\lambda - \mu_i}{\sigma_i}\Big)\right)\sigma_i^2$$

Figure 2: Componentwise risk functions



This figure displays componentwise risk, $R(m(\cdot, \lambda))$, as a function of μ_i for componentwise estimators, where $\sigma_i^2 = 2$. "mle" refers to the maximum likelihood (unregularized) estimator, $\widehat{\mu}_i = X_i$, which has risk equal to $\sigma_i^2 = 2$. The regularization parameters are $\lambda = 1$ for ridge, $\lambda = 2$ for lasso, and $\lambda = 4$ for pretest, as in Figure 1.

$$+\left(\Big(\frac{\lambda-\mu_i}{\sigma_i}\Big)\phi\Big(\frac{\lambda-\mu_i}{\sigma_i}\Big)-\Big(\frac{-\lambda-\mu_i}{\sigma_i}\Big)\phi\Big(\frac{-\lambda-\mu_i}{\sigma_i}\Big)\right)\sigma_i^2$$
$$+\left(\Phi\Big(\frac{\lambda-\mu_i}{\sigma_i}\Big)-\Phi\Big(\frac{-\lambda-\mu_i}{\sigma_i}\Big)\right)\mu_i^2.$$

Figure 2 plots the componentwise risk functions in Lemma 1 as functions of μ_i (with $\lambda = 1$ for ridge, $\lambda = 2$ for lasso, and $\lambda = 4$ for pretest). It also plots the componentwise risk of the unregularized maximum likelihood estimator, $\widehat{\mu}_i = X_i$, which is equal to σ_i^2. As Figure 2 suggests, componentwise risk is large for ridge when $|\mu_i|$ is large. The same is true for lasso, except that risk remains bounded. For pretest, componentwise risk is large when $|\mu_i|$ is close to λ.

Notice that these functions are plotted for a *fixed* value of the regularization parameter. If λ is chosen *optimally* , then the componentwise risks of ridge, lasso, and pretest are no greater than the componentwise risk of the unregularized maximum likelihood estimator $\widehat{\mu}_i = X_i$, which is σ_i^2. The reason is that ridge, lasso, and pretest nest the unregularized estimator (as the case $\lambda = 0$).

3.3 Spike and normal data generating process

If we take the expressions for componentwise risk derived in Lemma 1 and average them over some population distribution of (μ_i, σ_i^2), we obtain the integrated, or empirical Bayes, risk. For parametric families of distributions of (μ_i, σ_i^2), this might be done analytically. We shall do so now, considering a family of distributions that is rich enough to cover common intuitions about data generating processes, but simple enough to allow for analytic expressions. Based on these expressions, we characterize scenarios that favor the relative performance of each of the estimators considered in this article.

We consider a family of distributions for (μ_i, σ_i) such that: (i) μ_i takes value zero with probability p and is otherwise distributed as a normal with mean value μ_0 and standard deviation σ_0, and (ii) $\sigma_i^2 = \sigma^2$. The following proposition derives the optimal estimating function $\bar{m}_\pi^*$, as well as integrated risk functions for this family of distributions.

Proposition 1 (Spike and normal data generating process)

Assume π is such that (i) $\mu_1, \ldots, \mu_n$ are drawn independently from a distribution with probability mass p at zero, and normal with mean μ_0 and variance σ_0^2 elsewhere, and (ii) conditional on μ_i, X_i follows a normal distribution with mean μ_i and variance σ^2. Then, the optimal shrinkage function is

$$\bar{m}^*_\pi(x) = \frac{(1-p)\dfrac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\dfrac{x-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\dfrac{\mu_0\sigma^2 + x\sigma_0^2}{\sigma_0^2+\sigma^2}}{p\dfrac{1}{\sigma}\phi\left(\dfrac{x}{\sigma}\right) + (1-p)\dfrac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\dfrac{x-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)}.$$

The integrated risk of ridge is

$$\bar{R}(m_R(\cdot,\lambda),\pi) = \left(\frac{1}{1+\lambda}\right)^2\sigma^2 + (1-p)\left(\frac{\lambda}{1+\lambda}\right)^2(\mu_0^2+\sigma_0^2),$$

with

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

The integrated risk of lasso is given by

$$\bar{R}(m_L(\cdot,\lambda),\pi) = p\bar{R}_0(m_L(\cdot,\lambda),\pi) + (1-p)\bar{R}_1(m_L(\cdot,\lambda),\pi),$$

where

$$\bar{R}_0(m_L(\cdot,\lambda),\pi) = 2\Phi\Big(\frac{-\lambda}{\sigma}\Big)(\sigma^2+\lambda^2) - 2\Big(\frac{\lambda}{\sigma}\Big)\phi\Big(\frac{\lambda}{\sigma}\Big)\sigma^2,$$

and

$$\begin{aligned}\bar{R}_1(m_L(\cdot,\lambda),\pi) = {} & \left(1 + \Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right) - \Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\right)(\sigma^2+\lambda^2) \\ & + \left(\Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right) - \Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\right)(\mu_0^2+\sigma_0^2) \\ & - \frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)(\lambda+\mu_0)(\sigma_0^2+\sigma^2) \\ & - \frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)(\lambda-\mu_0)(\sigma_0^2+\sigma^2).\end{aligned}$$

Finally, the integrated risk of pretest is given by

$$\bar{R}(m_{PT}(\cdot,\lambda),\pi) = p\bar{R}_0(m_{PT}(\cdot,\lambda),\pi) + (1-p)\bar{R}_1(m_{PT}(\cdot,\lambda),\pi),$$

where

$$\bar{R}_0(m_{PT}(\cdot,\lambda),\pi) = 2\Phi\Big(\frac{-\lambda}{\sigma}\Big)\sigma^2 + 2\Big(\frac{\lambda}{\sigma}\Big)\phi\Big(\frac{\lambda}{\sigma}\Big)\sigma^2$$

and

$$\begin{aligned}\bar{R}_1(m_{PT}(\cdot,\lambda),\pi) = {} & \left(1 + \Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right) - \Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\right)\sigma^2 \\ & + \left(\Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right) - \Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\right)(\mu_0^2+\sigma_0^2) \\ & - \frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\big(\lambda(\sigma_0^2-\sigma^2)+\mu_0(\sigma_0^2+\sigma^2)\big) \\ & - \frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\big(\lambda(\sigma_0^2-\sigma^2)-\mu_0(\sigma_0^2+\sigma^2)\big).\end{aligned}$$

Notice that, even under substantial sparsity (that is, if p is large), the optimal shrinkage function, $\bar{m}^*_\pi$, never shrinks all the way to zero (unless, of course, $\mu_0 = \sigma_0 = 0$ or $p = 1$). This could in principle cast some doubts about the appropriateness of thresholding estimators, such as lasso or pretest, which induce sparsity in the estimated parameters. However, as we will see below, despite this stark difference between thresholding estimators and $\bar{m}^*_\pi$, lasso and, to a certain extent, pretest are able to approximate the integrated risk of $\bar{m}^*_\pi$ in the spike and normal model when the degree of sparsity in the parameters of interest is substantial.

Visual representations While it is difficult to directly interpret the risk formulas in Proposition 1, plotting these formulas as functions of the parameters governing the data generating process elucidates some crucial aspects of the risk of the corresponding estimators. Figure 3 does so, plotting the minimal integrated risk function of the different estimators. Each of the four subplots in Figure 3 is based on a fixed value of $p \in \{0, 0.25, 0.5, 0.75\}$, with μ_0 and σ_0^2 varying along the bottom axes. For each value of the triple (p, μ_0, σ_0), Figure 3 reports minimal integrated risk of each estimator (minimized over $\lambda \in [0, \infty]$). As a benchmark, Figure 3 reports the risk of the optimal shrinkage function, $\bar{m}^*_\pi$, simulated over 10 million repetitions. Figure 4 maps the regions of parameter values over which each of the three estimators, ridge, lasso, or pretest, performs best in terms of integrated risk.

Figures 3 and 4 provide some useful insights on the performance of shrinkage estimators. With no true zeros, ridge performs better than lasso or pretest. A clear advantage of ridge in this setting is that, in contrast to lasso or pretest, ridge allows shrinkage without shrinking some observations all the way to zero. As the share of true zeros increases, the relative performance of ridge deteriorates for pairs (μ_0, σ_0) away from the origin. Intuitively, linear shrinkage imposes a disadvantageous trade-off on ridge. Using ridge to heavily shrink towards the origin in order to fit potential true zeros produces large expected errors for observations with μ_i away from the origin. As a result, ridge performance suffers considerably unless much of the probability mass of the distribution of μ_i is tightly concentrated around zero. In the absence of true zeros, pretest performs particularly poorly unless the distribution of μ_i has much of its probability mass tightly concentrated around zero, in which case shrinking all the way to zero produces low risk. However, in the presence of true zeros, pretest performs well when much of the probability mass of the distribution of μ_i is located in a set that is well-separated from zero, which facilitates the detection of true zeros. Intermediate values of μ_0 coupled with moderate values of σ_0 produces settings where the conditional distributions $X_i|\mu_i = 0$ and $X_i|\mu_i \neq 0$ greatly overlap, inducing substantial risk for pretest estimation. The risk performance of lasso is particularly robust. It out-performs ridge and pretest for values of (μ_0, σ_0) at intermediate distances to the origin, and uniformly controls risk over the parameter space. This robustness of lasso may explain its popularity in empirical practice. Despite the fact that, unlike optimal shrinkage, thresholding estimators impose sparsity, lasso – and to a certain extent – pretest are able to approximate the integrated risk of the optimal shrinkage function over much of the parameter space.

All in all, the results in Figures 3 and 4 for the spike and normal case support the adoption of ridge in empirical applications where there are no reasons to presume the presence of many true zeros among the parameters of interest. In empirical settings where many true zeros may be expected, Figures 3 and 4 show that the choice among estimators

Figure 3: Risk for estimators in spike and normal setting

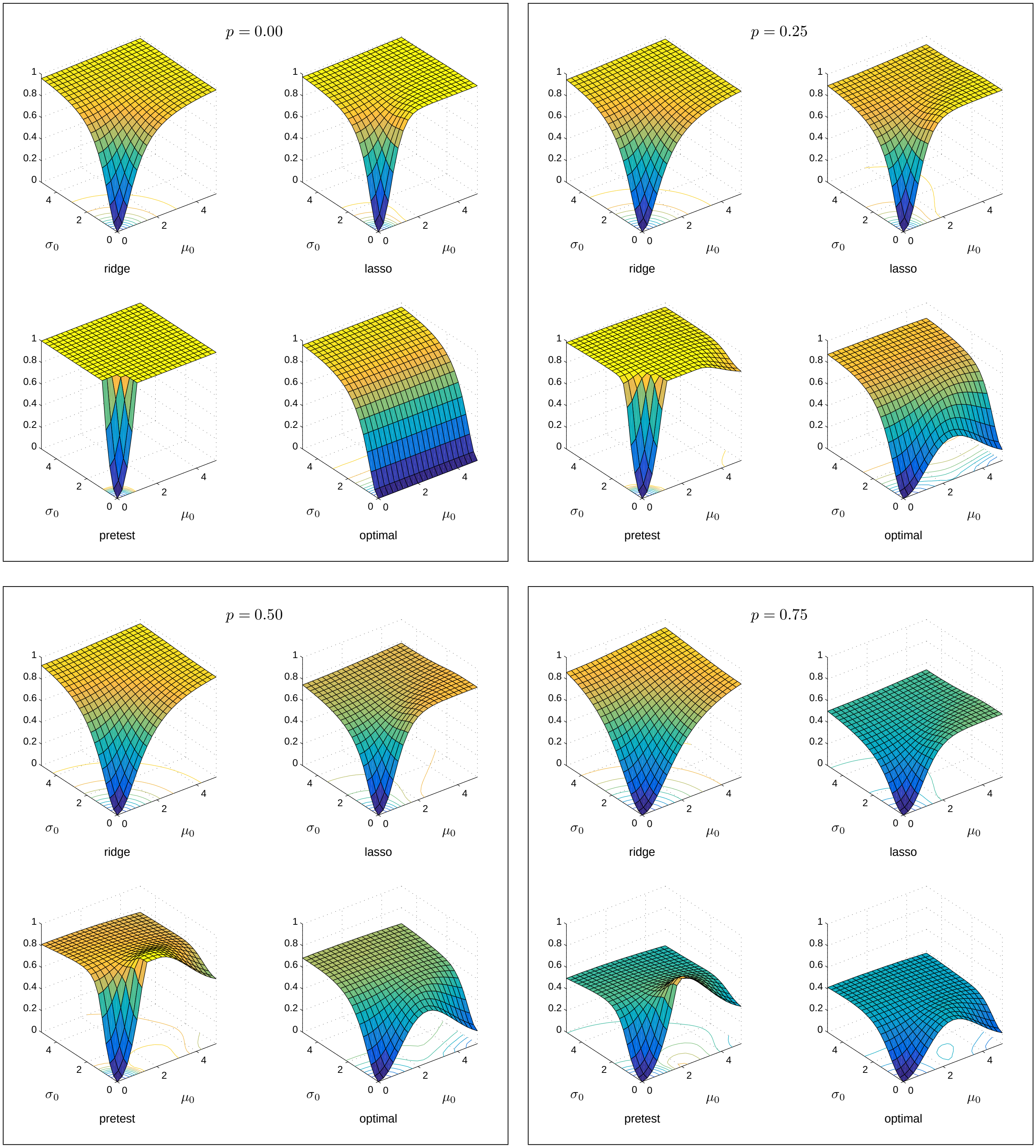
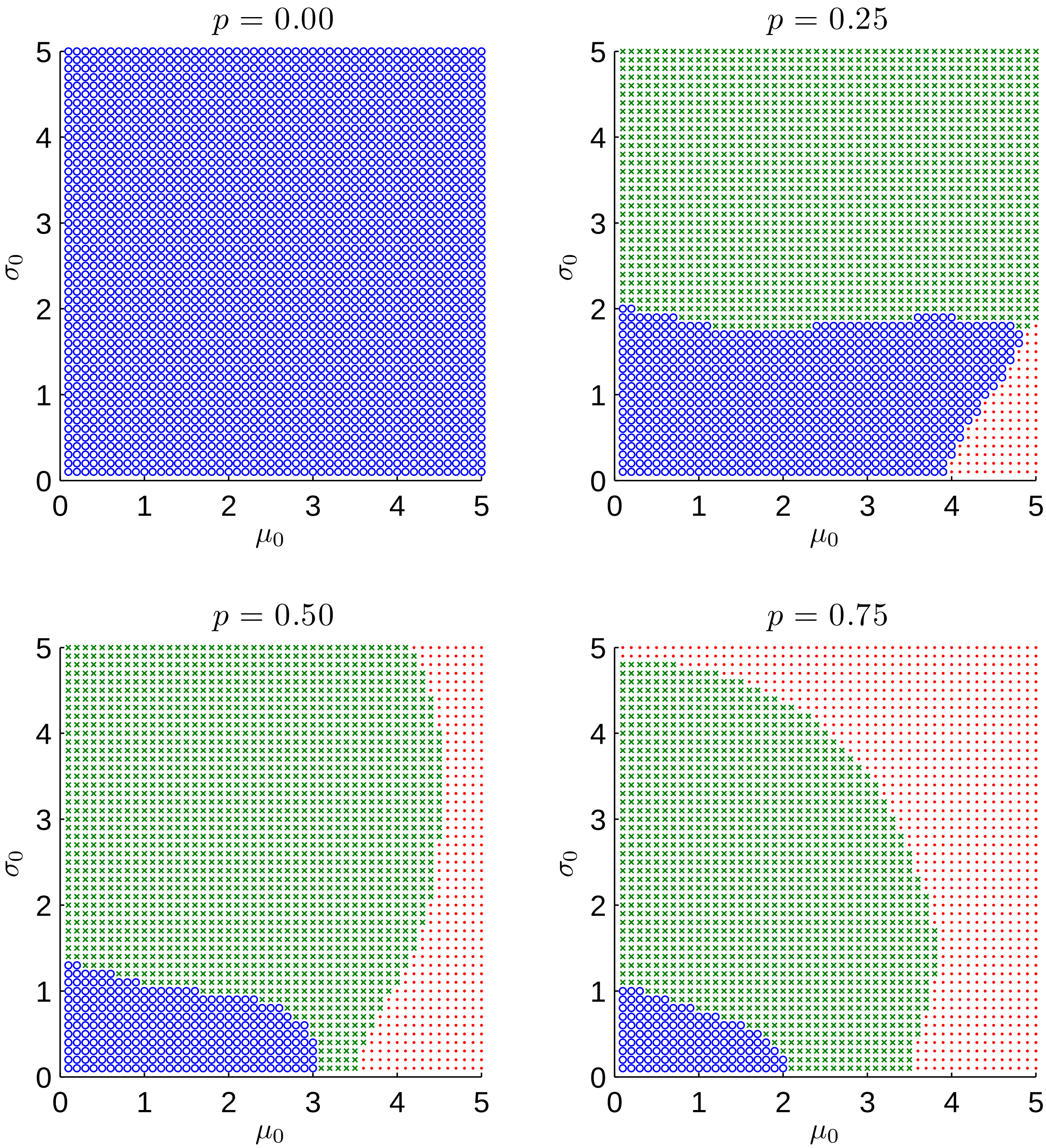


Figure 4: Best estimator in spike and normal setting



This figure compares integrated risk values attained by ridge, lasso, and pretest for different parameter values of the spike and normal specification in Section 3.3. Blue circles are placed at parameters values for which ridge minimizes integrated risk, green crosses at values for which lasso minimizes integrated risk, and red dots are parameters values for which pretest minimizes integrated risk.

in the spike and normal model depends on how well separated the distributions $X_i|\mu_i = 0$ and $X_i|\mu_i \neq 0$ are. Pretest is preferred in the well-separated case, while lasso is preferred in the non-separated case.

4 Data-driven choice of regularization parameters

In Section 3.3 we adopted a parametric model for the distribution of μ_i to study the risk properties of regularized estimators under an oracle choice of the regularization parameter, $\bar{\lambda}^*(\pi)$. In this section, we return to a nonparametric setting and show that it is possible to consistently estimate $\bar{\lambda}^*(\pi)$ from the data, $X_1, \ldots, X_n$, under some regularity conditions on π. We consider estimates $\widehat{\lambda}_n$ of $\bar{\lambda}^*(\pi)$ based on Stein's unbiased risk estimate and based on cross validation. The resulting estimators $m(X_i, \widehat{\lambda}_n)$ have risk functions which are *uniformly* close to those of the infeasible estimators $m(X_i, \bar{\lambda}^*(\pi))$.

The uniformity part of this statement is important and not obvious. Absent uniformity, asymptotic approximations might misleadingly suggest good behavior, while in fact the finite sample behavior of proposed estimators might be quite poor for plausible sets of data generating processes. This uniformity results in this section contrast markedly with other oracle approximations to risk, most notably approximations which assume that the true zeros, that is the components i for which $\mu_i = 0$, are known. Asymptotic approximations of this latter form are often invoked when justifying the use of lasso and pretest estimators. Such approximations are in general not uniformly valid, as emphasized by Leeb and Pötscher (2005) and others.

4.1 Uniform loss and risk consistency

For the remainder of the paper we adopt the following short-hand notation:

$$L_n(\lambda) = L_n(\boldsymbol{X}, m(\cdot, \lambda), \boldsymbol{P}) \qquad \text{(compound loss)}$$

$$R_n(\lambda) = R_n(m(\cdot, \lambda), \boldsymbol{P}) \qquad \text{(compound risk)}$$

$$\bar{R}_\pi(\lambda) = \bar{R}(m(\cdot, \lambda), \pi) \qquad \text{(empirical Bayes or integrated risk)}$$

We will now consider estimators $\widehat{\lambda}_n$ of $\bar{\lambda}^*(\pi)$ that are obtained by minimizing some empirical estimate of the risk function $\bar{R}_\pi$ (possibly up to a constant that depends only on π). The resulting $\widehat{\lambda}_n$ is then used to obtain regularized estimators of the form $\widehat{\mu}_i = m(X_i, \widehat{\lambda}_n)$. We will show that for large n the compound loss, the compound risk, and the integrated risk functions of the resulting estimators are uniformly close to the corresponding functions of the same estimators evaluated at oracle-optimal values of λ. As $n \to \infty$, the differences between L_n, R_n, and $\bar{R}_\pi$ vanish, so compound loss optimality, compound risk optimality, and integrated risk optimality become equivalent.

The following theorem establishes our key result for this section. Let $\mathcal{Q}$ be a set of probability distributions for (X_i, μ_i). Theorem 2 provides sufficient conditions for uniform loss consistency over $\pi \in \mathcal{Q}$, namely that (i) the supremum of the difference between the loss, $L_n(\lambda)$, and the empirical Bayes risk, $\bar{R}_\pi(\lambda)$, vanishes in probability uniformly over $\pi \in \mathcal{Q}$ and (ii) that $\widehat{\lambda}_n$ is chosen to minimize a uniformly consistent estimator, $r_n(\lambda)$, of the risk function, $\bar{R}_\pi(\lambda)$ (possibly up to a constant $\bar{v}_\pi$). Under these conditions, the difference between loss $L_n(\widehat{\lambda}_n)$ and the infeasible minimal loss $\inf_{\lambda \in [0,\infty]} L_n(\lambda)$ vanishes in probability uniformly over $\pi \in \mathcal{Q}$.

Theorem 2 (Uniform loss consistency)

Assume

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{\lambda \in [0,\infty]} \left| L_n(\lambda) - \bar{R}_\pi(\lambda) \right| > \epsilon \right) \to 0, \quad \forall \epsilon > 0. \tag{4}$$

Assume also that there are functions, $\bar{r}_\pi(\lambda)$, $\bar{v}_\pi$, and $r_n(\lambda)$ (of (π, λ), π, and $(\{X_i\}_{i=1}^n, \lambda)$, respectively) such that $\bar{R}_\pi(\lambda) = \bar{r}_\pi(\lambda) + \bar{v}_\pi$, and

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{\lambda \in [0,\infty]} \left| r_n(\lambda) - \bar{r}_\pi(\lambda) \right| > \epsilon \right) \to 0, \quad \forall \epsilon > 0. \tag{5}$$

Then,

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\left| L_n(\widehat{\lambda}_n) - \inf_{\lambda \in [0,\infty]} L_n(\lambda) \right| > \epsilon \right) \to 0, \quad \forall \epsilon > 0,$$

where $\widehat{\lambda}_n = \operatorname{argmin}_{\lambda \in [0,\infty]} r_n(\lambda)$.

The sufficient conditions given by this theorem, as stated in equations (4) and (5), are rather high-level. We shall now give more primitive conditions for these requirements to hold. In Sections 4.2 and 4.3 below, we propose suitable choices of $r_n(\lambda)$ based on Stein's unbiased risk estimator (SURE) and cross-validation (CV), and show that equation (5) holds for these choices of $r_n(\lambda)$.

The following Theorem 3 provides a set of conditions under which equation (4) holds, so the difference between compound loss and integrated risk vanishes uniformly. Aside from a bounded moment assumption, the conditions in Theorem 3 impose some restrictions on the estimating functions, $m(x, \lambda)$. Lemma 2 below shows that those conditions hold, in particular, for ridge, lasso, and pretest estimators.

Theorem 3 (Uniform L^2-convergence)
Suppose that

1. $m(x, \lambda)$ *is monotonic in* λ *for all* x *in* $\mathbb{R}$,

2. $m(x, 0) = x$ *and* $\lim_{\lambda\to\infty} m(x, \lambda) = 0$ *for all* x *in* $\mathbb{R}$,

3. $\sup_{\pi\in\mathcal{Q}} E_\pi[X^4] < \infty$.

4. *For any* $\epsilon > 0$ *there exists a set of regularization parameters* $0 = \lambda_0 < \ldots < \lambda_k = \infty$, *which may depend on* ϵ, *such that*

$$E_\pi[(|X-\mu| + |\mu|)|m(X, \lambda_j) - m(X, \lambda_{j-1})|] \le \epsilon$$

for all $j = 1, \ldots, k$ *and all* $\pi \in \mathcal{Q}$.

Then,

$$\sup_{\pi\in\mathcal{Q}} E_\pi\left[\sup_{\lambda\in[0,\infty]} \left(L_n(\lambda) - \bar{R}_\pi(\lambda)\right)^2\right] \to 0. \tag{6}$$

Notice that finiteness of $\sup_{\pi\in\mathcal{Q}} E_\pi[X^4]$ is equivalent to finiteness of $\sup_{\pi\in\mathcal{Q}} E_\pi[\mu^4]$ and $\sup_{\pi\in\mathcal{Q}} E_\pi[(X-\mu)^4]$ via Jensen's and Minkowski's inequalities.

Lemma 2
If $\sup_{\pi\in\mathcal{Q}} E_\pi[X^4] < \infty$, *then equation (6) holds for ridge and lasso. If, in addition,* X *is continuously distributed with a bounded density, then equation (6) holds for pretest.*

Theorem 2 provides sufficient conditions for uniform *loss* consistency. The following corollary shows that under the same conditions we obtain uniform *risk* consistency, that is, the integrated risk of the estimator based on the data-driven choice $\widehat{\lambda}_n$ becomes uniformly close to the risk of the oracle-optimal $\bar{\lambda}^*(\pi)$. For the statement of this corollary, recall that $\bar{R}(m(.,\widehat{\lambda}_n),\pi)$ is the integrated risk of the estimator $m(.,\widehat{\lambda}_n)$ using the stochastic (data-dependent) $\widehat{\lambda}_n$.

Corollary 1 (Uniform risk consistency)
Under the assumptions of Theorem 3,

$$\sup_{\pi\in\mathcal{Q}}\left|\bar{R}(m(.,\widehat{\lambda}_n),\pi)-\inf_{\lambda\in[0,\infty]}\bar{R}_\pi(\lambda)\right|\to 0. \tag{7}$$

In this section, we have shown that approximations to the risk function of machine learning estimators based on oracle-knowledge of λ are uniformly valid over $\pi\in\mathcal{Q}$ under mild assumptions. It is worth pointing out that such uniformity is not a trivial result. This is made clear by comparison to an alternative approximation, sometimes invoked to motivate the adoption of machine learning estimators, based on oracle-knowledge of true zeros among $\mu_1,\ldots,\mu_n$ (see, e.g., Fan and Li 2001). As shown in Appendix A.2, assuming oracle knowledge of zeros does not yield a uniformly valid approximation.

4.2 Stein's unbiased risk estimate

Theorem 2 provides sufficient conditions for uniform loss consistency using a general estimator r_n of risk. We shall now establish that our conditions apply to a particular estimator of r_n, known as Stein's unbiased risk estimate (SURE), which was first proposed by Stein et al. (1981). SURE leverages the assumption of normality to obtain an elegant expression of risk as an expected sum of squared residuals plus a penalization term.

SURE as originally proposed requires that m be piecewise differentiable as a function of x, which excludes discontinuous estimators such as the pretest estimator $m_{PT}(x,\lambda)$. We provide a generalization in Lemma 3 that allows for discontinuities. This lemma is stated in terms of integrated risk; with the appropriate modifications, the same result holds verbatim for compound risk.

Lemma 3 (SURE for piecewise differentiable estimators)

Suppose that $\mu \sim \vartheta$ *and*

$$X|\mu \sim N(\mu, 1).$$

Let $f_\pi = \vartheta * \phi$ *be the marginal density of* X*, where* ϕ *is the standard normal density. Consider an estimator* $m(X)$ *of* μ*, and suppose that* $m(x)$ *is differentiable everywhere in* $\mathbb{R}\backslash\{x_1, \ldots, x_J\}$*, but might be discontinuous at* $\{x_1, \ldots, x_J\}$*. Let* ∇m *be the derivative of* m *(defined arbitrarily at* $\{x_1, \ldots, x_J\}$*), and let* $\Delta m_j = \lim_{x \downarrow x_j} m(x) - \lim_{x \uparrow x_j} m(x)$ *for* $j \in \{1, \ldots, J\}$*. Assume that* $E_\pi[(m(X) - X)^2] < \infty$*,* $E_\pi[\nabla m(X)] < \infty$*, and* $(m(x) - x)\phi(x - \mu) \to 0$ *as* $|x| \to \infty$ ϑ*-a.s. Then,*

$$\bar{R}(m(.), \pi) = E_\pi[(m(X) - X)^2] + 2\left(E_\pi[\nabla m(X)] + \sum_{j=1}^{J} \Delta m_j f_\pi(x_j)\right) - 1.$$

The result of this lemma yields an objective function for the choice of λ of the general form we considered in Section 4.1, with $\bar{v}_\pi = -1$ and

$$\bar{r}_\pi(\lambda) = E_\pi[(m(X, \lambda) - X)^2] + 2\left(E_\pi[\nabla_x m(X, \lambda)] + \sum_{j=1}^{J} \Delta m_j(\lambda) f_\pi(x_j)\right), \tag{8}$$

where $\nabla_x m(x, \lambda)$ is the derivative of $m(x, \lambda)$ with respect to its first argument, and $\{x_1, \ldots, x_J\}$ may depend on λ. The expression in equation (8) can be estimated using its sample analog,

$$r_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(m(X_i, \lambda) - X_i)^2 + 2\left(\frac{1}{n}\sum_{i=1}^{n} \nabla_x m(X_i, \lambda) + \sum_{j=1}^{J} \Delta m_j(\lambda) \widehat{f}(x_j)\right), \tag{9}$$

where $\widehat{f}(x)$ is an estimator of $f_\pi(x)$. This expression can be thought of as a penalized least squares objective function. The following are explicit expressions for the penalty for the cases of ridge, lasso, and pretest.

ridge:	$\frac{2}{1+\lambda}$
lasso:	$\frac{2}{n}\sum_{i=1}^{n} 1(\lvert X_i\rvert > \lambda)$
prestest:	$\frac{2}{n}\sum_{i=1}^{n} 1(\lvert X_i\rvert > \lambda) + 2\lambda(\widehat{f}(-\lambda) + \widehat{f}(\lambda))$

The lasso penalty was previously derived in Donoho and Johnstone (1995). Our results allow to apply SURE estimation of risk to any machine learning estimator, as long as the conditions of Lemma 3 are satisfied.

To apply the uniform risk consistency in Theorem 2, we need to show that equation (5) holds. That is, we have to show that $r_n(\lambda)$ is uniformly consistent as an estimator of $\bar{r}_\pi(\lambda)$. The following lemma provides the desired result.

Lemma 4

Assume the conditions of Theorem 3. Then, equation (5) holds for $m(\cdot, \lambda)$ equal to $m_R(\cdot, \lambda)$, $m_L(\cdot, \lambda)$. If, in addition,

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{x \in \mathbb{R}} \left| |x| \widehat{f}(x) - |x| f_\pi(x) \right| > \epsilon \right) \to 0 \quad \forall \epsilon > 0,$$

then equation (5) holds for $m(\cdot, \lambda)$ equal to $m_{PT}(\cdot, \lambda)$.

Identification of $\bar{m}^*_\pi$ Under the conditions of Lemma 3 the optimal regularization parameter $\bar{\lambda}^*(\pi)$ is identified. In fact, under the same conditions, the stronger result holds that $\bar{m}^*_\pi$ as defined in Section 3.1 is identified as well (see, e.g., Brown, 1971; Efron, 2011). The next lemma states the identification result for $\bar{m}^*_\pi$.

Lemma 5

Under the conditions of Lemma 3, the optimal shrinkage function is given by

$$\bar{m}^*_\pi(x) = x + \nabla \log(f_\pi(x)). \tag{10}$$

Several nonparametric empirical Bayes estimators (NPEB) that target $\bar{m}^*_\pi(x)$ have been proposed (see Brown and Greenshtein, 2009; Jiang and Zhang, 2009, Efron, 2011, and Koenker and Mizera, 2014). In particular, Jiang and Zhang (2009) derive asymptotic optimality results for nonparametric estimation of $\bar{m}^*_\pi$ and provide an estimator based on the EM-algorithm. The estimator proposed in Koenker and Mizera (2014), which is based on convex optimization techniques, is particularly attractive, both in terms of computational properties and because it sidesteps the selection of a smoothing parameters (cf., e.g., Brown and Greenshtein, 2009). Both estimators, in Jiang and Zhang (2009) and Koenker and Mizera (2014), use a discrete distribution over a finite number of values to approximate the true distribution of μ. In sections 6 and 7, we will use the Koenker-Mizera estimator to visually compare the shape of this estimated $\bar{m}^*_\pi(x)$ to the shape of ridge, lasso and pretest estimating functions and to assess the performance of ridge, lasso and pretest relative to the performance of a nonparametric estimator of $\bar{m}^*_\pi$.

4.3 Cross-validation

A popular alternative to SURE is cross-validation, which chooses tuning parameters to optimize out-of-sample prediction. In this section, we investigate data-driven choices of the regularization parameter in a panel data setting, where multiple observations are available for each value of μ in the sample.

For $i = 1, \ldots, n$, consider i.i.d. draws, $(x_{1i}, \ldots, x_{ki}, \mu_i, \sigma_i)$, of a random variable $(x_1, \ldots, x_k, \mu, \sigma)$ with distribution $\pi \in \mathcal{Q}$. Assume that the components of $(x_1, \ldots, x_k)$ are i.i.d. conditional on (μ, σ^2) and that for each $j = 1, \ldots, k$,

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

Let

$$X_k = \frac{1}{k}\sum_{j=1}^{k} x_j \quad \text{and} \quad X_{ki} = \frac{1}{k}\sum_{j=1}^{k} x_{ji}.$$

For concreteness and to simplify notation, we will consider an estimator based on the first $k-1$ observations for each group $i = 1, \ldots, n$,

$$\widehat{\mu}_{k-1,i} = m(X_{k-1,i}, \lambda),$$

and will use observations x_{ki}, for $i = 1, \ldots n$, as a hold-out sample to choose λ. Similar results hold for alternative sample partitioning choices. The loss function and empirical Bayes risk function of this estimator are given by

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

and

$$\bar{R}_{\pi,k}(\lambda) = E_\pi[(m(X_{k-1},\lambda) - \mu)^2].$$

Consider the following cross-validation estimator

$$r_{n,k}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(m(X_{k-1,i},\lambda) - x_{ki}\right)^2.$$

Lemma 6

Assume Conditions 1 and 2 of Theorem 3 and $E_\pi[x_j^2] < \infty$, for $j = 1, \dots k$. Then,

$$E_\pi[r_{n,k}(\lambda)] = \bar{R}_{\pi,k}(\lambda) + E_\pi[\sigma^2].$$

That is, the cross validation yields an (up to a constant) unbiased estimator for the risk of the estimating function $m(X_{k-1},\lambda)$. The following theorem shows that this result can be strengthened to a uniform consistency result.

Theorem 4

Assume conditions 1 and 2 of Theorem 3 and $\sup_\pi E_\pi[x_j^4] < \infty$, for $j = 1, \dots k$. Let $\bar{v}_\pi = -E_\pi[\sigma^2]$,

$$\begin{aligned}\bar{r}_{\pi,k}(\lambda) &= E_\pi[r_{n,k}(\lambda)],\\ &= \bar{R}_{\pi,k}(\lambda) - \bar{v}_\pi,\end{aligned}$$

and $\widehat{\lambda}_n = \text{argmin}_{\lambda\in[0,\infty]} r_{n,k}(\lambda)$. Then, for ridge, lasso, and pretest,

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

and

$$\sup_{\pi\in\mathcal{Q}} P_\pi\left(\left|L_{n,k}\left(\widehat{\lambda}_n\right) - \inf_{\lambda\in[0,\infty]} L_{n,k}(\lambda)\right| > \epsilon\right) \to 0, \quad \forall\epsilon > 0.$$

Cross-validation has advantages as well as disadvantages relative to SURE. On the positive side, cross-validation does not rely on normal errors, while SURE does. Normality is less of an issue if k is large, so X_{ki} is approximately normal. On the negative side, however, cross-validation requires holding out part of the data from the second step estimation of $\boldsymbol{\mu}$, once the value of the regularization parameter has been chosen in a first step. This affects the essence of the cross-validation efficiency results, which apply to estimators of the form $m(X_{k-1,i}, \lambda)$, rather than to feasible estimators that use the entire sample in the second step, $m(X_{ki}, \lambda)$. Finally, cross-validation imposes greater data availability requirements, as it relies on availability of data on repeated realizations, $x_{1i}, \ldots, x_{ki}$, of a random variable centered at μ_i, for each sample unit $i = 1, \ldots, n$. This may hinder the practical applicability of cross-validation selection of regularization parameters in the context considered in this article.

5 Discussion and Extensions

5.1 Mixed estimators and estimators of the optimal shrinkage function

We have discussed criteria such as SURE and CV as means to select the regularization parameter, λ. In principle, these same criteria might also be used to choose among alternative estimators, such as ridge, lasso, and pretest, in specific empirical settings. Our uniform risk consistency results imply that such a mixed-estimator approach dominates each of the estimators which are being mixed, for n large enough. Going even further, one might aim to estimate the optimal shrinkage function, $\bar{m}^*_\pi$, using the result of Lemma 5, as in Jiang and Zhang (2009), Koenker and Mizera (2014)) and others. Under suitable consistency conditions, this approach will dominate all other componentwise estimators for large enough n (Jiang and Zhang, 2009). In practice, these results should be applied with some caution, as they are based on neglecting the variability in the choice of estimation procedure or in the estimation of $\bar{m}^*_\pi$. For small and moderate values of n, procedures with fewer degrees of freedom may perform better in practice. We return to this issue in section 6, where we compare the finite sample risk of the machine learning estimators considered in this article (ridge, lasso and pretest) to the finite sample risk of the NPEB estimator of Koenker and Mizera (2014).

5.2 Heteroskedasticity

While for simplicity many of our results are stated for the homoskedastic case, where $\mathrm{var}(X_i) = \sigma$ for all i, they easily generalize to heteroskedasticity.

The general characterization of compound risk in Theorem 1 does not use homoskedasticity, nor does the derivation of componentwise risk in Lemma 1. The analytical derivations of empirical Bayes risk for the spike and normal data generating process in Proposition 1, and the corresponding comparisons of risk in Figures 3 and 4 do rely on homoskedasticity. Similar formulas to those of Proposition 1 might be derived for other data generating processes with heteroskedasticity, but the rankings of estimators might change.

As for our proofs of uniform risk consistency, our general results (Theorem 2 and 3) do not require homoskedasticity, nor does the validity or consistency of crossvalidation, cf. Theorem 4. SURE, in the form we introduced in Lemma 3, does require homoskedasticity. However, the definition of SURE, and the corresponding consistency results, can be extended to the heteroskedastic case (see Xie et al., 2012).

5.3 Comparison with Leeb and Pötscher (2006)

Our results on the uniform consistency of estimators of risk such as SURE or CV appear to stand in contradiction to those of Leeb and Pötscher (2006). They consider the same setting as we do – estimation of normal means – and the same types of estimators, including ridge, lasso, and pretest. In this setting, Leeb and Pötscher (2006) show that no uniformly consistent estimator of risk exists for such estimators.

The apparent contradiction between our results and the results in Leeb and Pötscher (2006) is explained by the different nature of the asymptotic sequence adopted in this article to study the properties of machine learning estimators, relative to the asymptotic sequence adopted in Leeb and Pötscher (2006) for the same purpose. In this article, we consider the problem of estimating a large number of parameters, such as location effects for many locations or group-level treatment effects for many groups. This motivates the adoption of an asymptotic sequence along which the number of estimated parameters increases as

$n \to \infty$. In contrast, Leeb and Pötscher (2006) study the risk properties of regularized estimators embedded in a sequence along which the number of estimated parameters stays fixed as $n \to \infty$ and the estimation variance is of order $1/n$. We expect our approximation to work well when the dimension of the estimated parameter is large; the approximation of Leeb and Pötscher (2006) is likely to be more appropriate when the dimension of the estimated parameter is small while sample size is large.

In the simplest version of the setting in Leeb and Pötscher (2006) we observe a $(k \times 1)$ vector $\boldsymbol{X}_n$ with distribution $\boldsymbol{X}_n \sim N(\boldsymbol{\mu}_n, \boldsymbol{I}_k/n)$, where $\boldsymbol{I}_k$ is the identity matrix of dimension k. Let X_{ni} and μ_{ni} be the i-components of $\boldsymbol{X}_n$ and $\boldsymbol{\mu}_n$, respectively. Consider the componentwise estimator $m_n(X_{ni})$ of μ_{ni}. Leeb and Pötscher (2006) study consistent estimation of the normalized risk

$$\bar{R}_n^{LP} = nE\|\boldsymbol{m}_n(\boldsymbol{X}_n) - \boldsymbol{\mu}_n\|^2,$$

where $\boldsymbol{m}_n(\boldsymbol{X}_n)$ is a $(k \times 1)$ vector with i-th element equal to $m_n(X_{ni})$.

Adopting the re-parametrization, $\boldsymbol{Y}_n = \sqrt{n}\boldsymbol{X}_n$ and $\boldsymbol{h}_n = \sqrt{n}\boldsymbol{\mu}_n$, we obtain $\boldsymbol{Y}_n - \boldsymbol{h}_n \sim N(\boldsymbol{0}, \boldsymbol{I}_k)$. Notice that, for the maximum likelihood estimator, $\boldsymbol{m}_n(\boldsymbol{X}_n) - \boldsymbol{\mu}_n = (\boldsymbol{Y}_n - \boldsymbol{h}_n)/\sqrt{n}$ and $\bar{R}_n^{LP} = E\|\|m(\boldsymbol{Y}_n) - \boldsymbol{h}_n\|^2 = k$, so the risk of the maximum likelihood estimator does not depend on the sequence $\boldsymbol{h}_n$ and, therefore, can be consistently estimated. This is not the case for shrinkage estimators, however. Choosing $\boldsymbol{h}_n = \boldsymbol{h}$ for some fixed $\boldsymbol{h}$, the problem becomes invariant in n,

$$\boldsymbol{Y}_n \sim N(\boldsymbol{h}, \boldsymbol{I}_k).$$

In this setting, it is easy to show that the risk of machine learning estimators, such as ridge, lasso, and pretest depends on $\boldsymbol{h}$, and therefore it cannot be estimated consistently. For instance, consider the lasso estimator, $m_n(x) = m_L(x, \lambda_n)$, where $\sqrt{n}\lambda_n \to c$ with $0 < c < \infty$, as in Leeb and Pötscher (2006). Then, Lemma 1 implies that $\bar{R}_n^{LP}$ is constant in n and dependent on $\boldsymbol{h}$. As a result, $\bar{R}_n^{LP}$ cannot be estimated consistently.[2]

[2] This result holds more generally outside the normal error model. Let $\boldsymbol{m}_L(\boldsymbol{X}_n, \lambda)$ be the $(n \times 1)$ vector with i-th element equal to $m_L(X_i, \lambda)$. Consider the sequence of regularization parameters $\lambda_n = c/\sqrt{n}$, then $m_L(x, \lambda_n) = m_L(\sqrt{n}x, c)/\sqrt{n}$. This implies $\bar{R}_n^{LP} = E\|\|\boldsymbol{m}_L(\boldsymbol{Y}_n, c) - \boldsymbol{h}\|^2$, which is invariant in n.

Contrast the setting in Leeb and Pötscher (2006) to the one adopted in this article, where we consider a high dimensional setting, such that $\boldsymbol{X}$ and $\boldsymbol{\mu}$ have dimension equal to n. The pairs (X_i, μ_i) follow a distribution π which may vary with n. As n increases, π becomes identified and so does the average risk, $E_\pi[(m_n(X_i) - \mu_i)^2]$, of any componentwise estimator, $m_n(\cdot)$.

Whether the asymptotic approximation in Leeb and Pötscher (2006) or ours provides a better description of the performance of SURE, CV, or other estimators of risk in actual applications depends on the dimension of $\boldsymbol{\mu}$. If this dimension is large, as typical in the applications we consider in this article, we expect our uniform consistency result to apply: a "blessing of dimesionality". As demonstrated by Leeb and Pötscher, however, precise estimation of a fixed number of parameters does not ensure uniformly consistent estimation of risk.

6 Simulations

Designs To gauge the relative performance of the estimators considered in this article, we next report the results of a set of simulations that employ the spike and normal data generating process of Section 3.3. As in Proposition 1, we consider distributions π of (X, μ) such that μ is degenerate at zero with probability p and normal with mean μ_0 and variance σ_0^2 with probability $(1 - p)$. We consider all combinations of parameter values $p = 0.00, 0.25, 0.50, 0.75, 0.95$, $\mu_0 = 0, 2, 4$, $\sigma_0 = 2, 4, 6$, and sample sizes $n = 50, 200, 1000$.

Given a set of values $\mu_1, \ldots, \mu_n$, the values for $X_1, \ldots, X_n$ are generated as follows. To evaluate the performance of estimators based on SURE selectors and of the NPEB estimator of Koenker and Mizera (2014), we generate the data as

$$X_i = \mu_i + U_i, \tag{11}$$

where the U_i follow a standard normal distribution, independent of other components. To evaluate the performance of cross-validation estimators, we generate

$$x_{ji} = \mu_i + \sqrt{k} u_{ji}$$

for $j = 1, \ldots, k$, where the u_{ji} are draws from independent standard normal distributions. As a result, the averages

$$X_{ki} = \frac{1}{k}\sum_{j=1}^{k} x_{ji}$$

have the same distributions as the X_i in equation (11), which makes the comparison of between the cross-validation estimators and the SURE and NPEB estimators a meaninful one. For cross-validation estimators we consider $k = 4, 20$.

Estimators The SURE criterion function employed in the simulations is the one in equation (9) where, for the pretest estimator, the density of X is estimated with a normal kernel and the bandwidth implied by "Silverman's rule of thumb".[3] The cross-validation criterion function employed in the simulations is a leave-one-out version of the one considered in Section 4.3,

$$r_{n,k}(\lambda) = \sum_{j=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n}(m(X_{-ji}, \lambda) - x_{ji})^2\right), \tag{12}$$

where X_{-ji} is the average of $\{x_{1i}, \ldots, x_{ki}\}\backslash x_{ji}$. Notice that because of the result in Theorem 4 applies to each of the k terms on the right-hand-side of equation (12) it also applies to $r_{n,k}(\lambda)$ as defined on the left-hand-side of the same equation. The cross validation estimator employed in our simulations is $m(X_{ki}, \lambda)$, with λ evaluated at the minimizer of (12).

Results Tables 1, 2, and 3 report average compound risk across 1000 simulations for $n = 50$, $n = 200$ and $n = 1000$, respectively. Each row corresponds to a particular value of (p, μ_0, σ_0), and each column corresponds to a particular estimator/regularization criterion. The results are coded row-by-row on a continuous color scale which varies from dark blue (minimum row value) to light yellow (maximum row value).

Several clear patterns emerge from the simulation results. First, even for a dimensionality as modest as $n = 50$, the patterns in Figure 3, which were obtained for oracle choices of regularization parameters, are reproduced in Tables 1 to 3 for the same estimators but using data-driven choices of regularization parameters. As in Figure 3, among ridge, lasso

[3] See Silverman (1986) equation (3.31).

and pretest, ridge dominates when there is little or no sparsity in the parameters of interest, pretest dominates when the distribution of non-zero parameters is substantially separated from zero, and lasso dominates in the intermediate cases. Second, while the results in Jiang and Zhang (2009) suggest good performance of nonparametric estimators of $\bar{m}^*_\pi$ for large n, the simulation results in Tables 1 and 2 indicate that the performance of NPEB may be substantially worse than the performance of the other machine learning estimators in the table, for moderate and small n. In particular, the performance of the NPEB estimator suffers in the settings with low or no sparsity, especially when the distribution of the non-zero values of $\mu_1, \ldots, \mu_n$ has considerable dispersion. This is explained by the fact that, in practice, the NPEB estimator approximates the distribution of μ using a discrete distribution supported on a small number of values. When most of the probability mass of the true distribution of μ is also concentrated around a small number of values (that is, when p is large or σ_0 is small), the approximation employed by the NPEB estimator is accurate and the performance of the NPEB estimator is good. This is not the case, however, when the true distribution of μ cannot be closely approximated with a small number of values (that is, when p is small and σ_0 is large). Lasso shows a remarkable degree of robustness to the value of (p, μ_0, σ_0), which makes it an attractive estimator in practice. For large n, as in Table 3, NPEB dominates except in settings with no sparsity and a large dispersion in μ ($p = 0$ and σ_0 large).

Table 1: Average Compound Loss Across 1000 Simulations with $N = 50$

			SURE			Cross-Validation ($k = 4$)			Cross-Validation ($k = 20$)			NPEB
p	μ_0	σ_0	ridge	lasso	pretest	ridge	lasso	pretest	ridge	lasso	pretest	
0.00	0	2	0.80	0.89	1.02	0.83	0.90	1.12	0.81	0.88	1.12	0.94
0.00	0	4	0.95	0.98	1.02	0.96	0.98	1.08	0.94	0.97	1.08	1.15
0.00	0	6	0.97	0.99	1.01	0.97	0.99	1.05	0.97	0.99	1.07	1.21
0.00	2	2	0.89	0.96	1.01	0.90	0.95	1.06	0.89	0.95	1.09	0.93
0.00	2	4	0.96	0.99	1.01	0.96	0.98	1.06	0.96	0.98	1.09	1.13
0.00	2	6	0.97	0.99	1.01	0.99	1.00	1.06	0.97	0.98	1.07	1.21
0.00	4	2	0.95	1.00	1.01	0.95	0.99	1.02	0.95	1.00	1.04	0.93
0.00	4	4	0.97	1.00	1.01	0.99	1.01	1.06	0.97	0.99	1.07	1.15
0.00	4	6	0.99	1.00	1.02	0.99	1.00	1.05	0.99	1.00	1.07	1.21
0.25	0	2	0.75	0.78	1.01	0.77	0.79	1.12	0.77	0.78	1.08	0.85
0.25	0	4	0.92	0.90	1.00	0.94	0.90	1.07	0.92	0.88	1.05	1.04
0.25	0	6	0.97	0.93	0.99	0.97	0.92	1.02	0.96	0.92	1.02	1.09
0.25	2	2	0.87	0.88	1.01	0.87	0.86	1.06	0.87	0.86	1.07	0.88
0.25	2	4	0.93	0.90	0.99	0.94	0.89	1.04	0.95	0.90	1.04	1.03
0.25	2	6	0.97	0.93	0.98	0.98	0.93	1.03	0.97	0.93	1.02	1.09
0.25	4	2	0.94	0.95	0.99	0.95	0.95	1.03	0.95	0.95	1.04	0.92
0.25	4	4	0.97	0.94	0.99	0.97	0.93	1.03	0.97	0.93	1.02	1.04
0.25	4	6	0.98	0.94	0.98	0.98	0.93	1.02	0.98	0.93	1.00	1.09
0.50	0	2	0.67	0.64	0.94	0.69	0.64	0.96	0.67	0.62	0.90	0.69
0.50	0	4	0.89	0.75	0.92	0.91	0.76	0.92	0.89	0.75	0.89	0.82
0.50	0	6	0.95	0.80	0.90	0.95	0.79	0.87	0.96	0.78	0.84	0.84
0.50	2	2	0.80	0.72	0.96	0.82	0.72	0.96	0.81	0.72	0.93	0.73
0.50	2	4	0.92	0.77	0.94	0.93	0.76	0.90	0.90	0.75	0.87	0.83
0.50	2	6	0.96	0.80	0.92	0.95	0.77	0.83	0.95	0.78	0.82	0.86
0.50	4	2	0.91	0.82	0.95	0.92	0.81	0.90	0.92	0.81	0.87	0.75
0.50	4	4	0.94	0.80	0.93	0.94	0.79	0.87	0.94	0.78	0.83	0.81
0.50	4	6	0.97	0.81	0.93	0.97	0.79	0.83	0.96	0.78	0.79	0.85
0.75	0	2	0.51	0.43	0.61	0.51	0.42	0.57	0.50	0.41	0.57	0.46
0.75	0	4	0.77	0.50	0.59	0.80	0.51	0.58	0.78	0.50	0.57	0.52
0.75	0	6	0.88	0.54	0.55	0.90	0.54	0.55	0.88	0.53	0.52	0.51
0.75	2	2	0.66	0.49	0.65	0.67	0.49	0.63	0.67	0.49	0.62	0.47
0.75	2	4	0.81	0.53	0.59	0.86	0.54	0.58	0.82	0.52	0.56	0.51
0.75	2	6	0.90	0.56	0.54	0.91	0.56	0.53	0.90	0.55	0.52	0.51
0.75	4	2	0.84	0.59	0.64	0.85	0.57	0.60	0.84	0.58	0.58	0.49
0.75	4	4	0.88	0.56	0.57	0.89	0.55	0.53	0.89	0.55	0.52	0.50
0.75	4	6	0.92	0.57	0.53	0.92	0.55	0.49	0.92	0.56	0.51	0.50
0.95	0	2	0.18	0.15	0.17	0.17	0.12	0.15	0.18	0.13	0.19	0.17
0.95	0	4	0.37	0.19	0.17	0.37	0.17	0.17	0.37	0.18	0.20	0.18
0.95	0	6	0.49	0.21	0.16	0.51	0.19	0.16	0.49	0.19	0.19	0.16
0.95	2	2	0.26	0.17	0.18	0.27	0.16	0.18	0.27	0.17	0.23	0.17
0.95	2	4	0.40	0.19	0.17	0.43	0.18	0.16	0.40	0.18	0.20	0.17
0.95	2	6	0.53	0.21	0.15	0.53	0.19	0.15	0.53	0.20	0.18	0.16
0.95	4	2	0.44	0.21	0.18	0.45	0.20	0.18	0.45	0.20	0.22	0.18
0.95	4	4	0.51	0.21	0.16	0.51	0.19	0.17	0.52	0.20	0.19	0.17
0.95	4	6	0.57	0.21	0.15	0.58	0.19	0.14	0.57	0.20	0.18	0.16

Table 2: Average Compound Loss Across 1000 Simulations with $N = 200$

			SURE			Cross-Validation ($k = 4$)			Cross-Validation ($k = 20$)			NPEB
p	μ_0	σ_0	ridge	lasso	pretest	ridge	lasso	pretest	ridge	lasso	pretest	
0.00	0	2	0.80	0.87	1.01	0.82	0.88	1.04	0.80	0.87	1.04	0.86
0.00	0	4	0.94	0.96	1.00	0.95	0.97	1.02	0.94	0.96	1.03	1.03
0.00	0	6	0.98	0.99	1.01	0.98	0.99	1.02	0.98	0.99	1.03	1.09
0.00	2	2	0.89	0.95	1.00	0.90	0.95	1.02	0.89	0.94	1.03	0.86
0.00	2	4	0.95	0.97	1.00	0.96	0.98	1.02	0.96	0.97	1.03	1.03
0.00	2	6	0.98	1.00	1.01	0.98	0.99	1.02	0.98	0.99	1.03	1.10
0.00	4	2	0.95	1.00	1.00	0.96	1.00	1.01	0.95	1.00	1.02	0.86
0.00	4	4	0.97	0.99	1.00	0.97	0.99	1.01	0.97	0.98	1.02	1.03
0.00	4	6	0.98	0.99	1.01	0.98	0.99	1.01	0.99	0.99	1.03	1.09
0.25	0	2	0.75	0.77	1.00	0.77	0.78	1.07	0.75	0.75	1.04	0.78
0.25	0	4	0.92	0.88	0.99	0.93	0.88	1.02	0.93	0.88	1.02	0.95
0.25	0	6	0.96	0.91	0.99	0.97	0.91	1.01	0.96	0.91	1.00	0.98
0.25	2	2	0.86	0.86	1.00	0.87	0.86	1.03	0.86	0.85	1.03	0.80
0.25	2	4	0.94	0.90	1.00	0.95	0.90	1.02	0.93	0.88	1.01	0.95
0.25	2	6	0.97	0.92	0.99	0.97	0.92	1.00	0.97	0.91	1.00	0.98
0.25	4	2	0.94	0.95	1.00	0.94	0.93	1.00	0.94	0.94	1.01	0.83
0.25	4	4	0.96	0.92	0.99	0.97	0.92	1.01	0.95	0.91	0.99	0.94
0.25	4	6	0.97	0.92	0.98	0.97	0.92	0.99	0.97	0.92	0.98	0.98
0.50	0	2	0.67	0.61	0.90	0.69	0.62	0.93	0.67	0.61	0.90	0.63
0.50	0	4	0.89	0.74	0.90	0.90	0.74	0.89	0.89	0.73	0.86	0.76
0.50	0	6	0.94	0.77	0.86	0.95	0.76	0.82	0.95	0.77	0.83	0.77
0.50	2	2	0.80	0.70	0.94	0.82	0.71	0.93	0.80	0.69	0.91	0.65
0.50	2	4	0.92	0.75	0.92	0.92	0.75	0.87	0.91	0.74	0.86	0.76
0.50	2	6	0.95	0.78	0.88	0.96	0.78	0.83	0.95	0.77	0.82	0.77
0.50	4	2	0.91	0.80	0.94	0.92	0.81	0.87	0.91	0.80	0.87	0.67
0.50	4	4	0.94	0.78	0.94	0.95	0.78	0.83	0.94	0.77	0.82	0.74
0.50	4	6	0.96	0.79	0.92	0.97	0.79	0.81	0.97	0.78	0.80	0.76
0.75	0	2	0.50	0.39	0.55	0.51	0.40	0.57	0.50	0.39	0.55	0.40
0.75	0	4	0.80	0.50	0.55	0.81	0.50	0.57	0.80	0.49	0.56	0.48
0.75	0	6	0.90	0.53	0.49	0.91	0.53	0.52	0.89	0.52	0.50	0.47
0.75	2	2	0.67	0.47	0.59	0.68	0.47	0.61	0.67	0.46	0.59	0.42
0.75	2	4	0.83	0.50	0.53	0.84	0.51	0.56	0.83	0.51	0.55	0.46
0.75	2	6	0.91	0.54	0.50	0.91	0.54	0.52	0.91	0.53	0.51	0.47
0.75	4	2	0.83	0.56	0.55	0.85	0.57	0.58	0.83	0.55	0.56	0.42
0.75	4	4	0.89	0.54	0.50	0.90	0.54	0.52	0.88	0.53	0.50	0.45
0.75	4	6	0.93	0.55	0.47	0.93	0.55	0.49	0.92	0.53	0.48	0.46
0.95	0	2	0.17	0.12	0.14	0.17	0.12	0.14	0.17	0.12	0.15	0.12
0.95	0	4	0.43	0.17	0.16	0.43	0.17	0.16	0.42	0.17	0.16	0.14
0.95	0	6	0.61	0.18	0.14	0.62	0.18	0.14	0.61	0.18	0.14	0.14
0.95	2	2	0.28	0.16	0.17	0.29	0.16	0.18	0.28	0.15	0.17	0.14
0.95	2	4	0.46	0.17	0.15	0.48	0.17	0.16	0.47	0.17	0.16	0.14
0.95	2	6	0.63	0.19	0.14	0.64	0.19	0.14	0.63	0.18	0.14	0.13
0.95	4	2	0.49	0.20	0.17	0.50	0.20	0.17	0.48	0.19	0.17	0.14
0.95	4	4	0.58	0.19	0.14	0.59	0.19	0.14	0.59	0.18	0.15	0.14
0.95	4	6	0.68	0.19	0.13	0.70	0.19	0.13	0.67	0.19	0.14	0.13

Table 3: Average Compound Loss Across 1000 Simulations with $N = 1000$

			SURE			Cross-Validation ($k = 4$)			Cross-Validation ($k = 20$)			NPEB
p	μ_0	σ_0	ridge	lasso	pretest	ridge	lasso	pretest	ridge	lasso	pretest	
0.00	0	2	0.80	0.87	1.01	0.81	0.87	1.01	0.80	0.86	1.01	0.82
0.00	0	4	0.94	0.96	1.00	0.95	0.97	1.01	0.94	0.96	1.00	0.97
0.00	0	6	0.97	0.98	1.00	0.98	0.98	1.00	0.97	0.98	1.01	1.02
0.00	2	2	0.89	0.94	1.00	0.90	0.95	1.00	0.89	0.94	1.01	0.82
0.00	2	4	0.95	0.97	1.00	0.96	0.97	1.00	0.95	0.97	1.01	0.98
0.00	2	6	0.97	0.98	1.00	0.98	0.99	1.00	0.97	0.98	1.01	1.02
0.00	4	2	0.95	1.00	1.00	0.96	1.00	1.00	0.95	0.99	1.00	0.82
0.00	4	4	0.97	0.99	1.00	0.97	0.99	1.01	0.97	0.99	1.01	0.97
0.00	4	6	0.98	0.99	1.00	0.98	0.99	1.00	0.98	0.99	1.01	1.02
0.25	0	2	0.75	0.76	1.00	0.76	0.77	1.02	0.75	0.75	1.01	0.74
0.25	0	4	0.92	0.88	0.99	0.93	0.88	1.00	0.92	0.87	1.00	0.89
0.25	0	6	0.97	0.91	0.99	0.97	0.91	0.99	0.96	0.91	0.99	0.92
0.25	2	2	0.86	0.85	1.00	0.87	0.86	1.01	0.86	0.84	1.01	0.76
0.25	2	4	0.94	0.89	1.00	0.94	0.89	1.00	0.94	0.89	1.00	0.89
0.25	2	6	0.97	0.91	0.99	0.97	0.91	0.99	0.97	0.91	0.99	0.92
0.25	4	2	0.94	0.94	0.99	0.94	0.94	0.99	0.94	0.93	0.99	0.79
0.25	4	4	0.96	0.92	0.99	0.96	0.91	0.99	0.96	0.91	0.99	0.88
0.25	4	6	0.98	0.92	0.99	0.98	0.92	0.98	0.97	0.92	0.98	0.91
0.50	0	2	0.67	0.60	0.87	0.68	0.61	0.90	0.67	0.60	0.87	0.60
0.50	0	4	0.89	0.73	0.85	0.90	0.73	0.86	0.89	0.72	0.85	0.71
0.50	0	6	0.95	0.77	0.81	0.95	0.77	0.82	0.95	0.76	0.81	0.72
0.50	2	2	0.80	0.70	0.90	0.81	0.71	0.90	0.80	0.69	0.89	0.62
0.50	2	4	0.91	0.74	0.85	0.92	0.75	0.85	0.91	0.74	0.84	0.70
0.50	2	6	0.95	0.77	0.80	0.96	0.78	0.81	0.95	0.77	0.80	0.71
0.50	4	2	0.91	0.80	0.87	0.92	0.80	0.84	0.91	0.80	0.84	0.63
0.50	4	4	0.94	0.77	0.88	0.95	0.78	0.81	0.94	0.77	0.80	0.68
0.50	4	6	0.96	0.78	0.87	0.97	0.78	0.79	0.96	0.78	0.78	0.70
0.75	0	2	0.50	0.38	0.54	0.51	0.40	0.56	0.50	0.38	0.54	0.38
0.75	0	4	0.80	0.49	0.53	0.81	0.50	0.55	0.80	0.48	0.53	0.44
0.75	0	6	0.90	0.52	0.49	0.91	0.53	0.51	0.90	0.52	0.49	0.43
0.75	2	2	0.67	0.46	0.57	0.68	0.47	0.59	0.67	0.46	0.58	0.40
0.75	2	4	0.83	0.50	0.52	0.85	0.51	0.55	0.83	0.50	0.53	0.44
0.75	2	6	0.91	0.53	0.48	0.92	0.53	0.50	0.91	0.52	0.48	0.43
0.75	4	2	0.83	0.55	0.53	0.85	0.56	0.55	0.83	0.55	0.54	0.39
0.75	4	4	0.89	0.53	0.49	0.90	0.54	0.51	0.89	0.52	0.49	0.41
0.75	4	6	0.93	0.54	0.46	0.94	0.54	0.48	0.93	0.53	0.47	0.42
0.95	0	2	0.17	0.11	0.14	0.17	0.12	0.14	0.17	0.11	0.14	0.11
0.95	0	4	0.44	0.16	0.15	0.45	0.16	0.16	0.44	0.16	0.15	0.13
0.95	0	6	0.63	0.18	0.13	0.65	0.18	0.14	0.64	0.17	0.14	0.12
0.95	2	2	0.28	0.15	0.16	0.29	0.15	0.18	0.29	0.14	0.17	0.12
0.95	2	4	0.49	0.16	0.14	0.50	0.17	0.16	0.50	0.16	0.15	0.12
0.95	2	6	0.66	0.18	0.13	0.67	0.18	0.14	0.66	0.18	0.13	0.12
0.95	4	2	0.50	0.19	0.16	0.51	0.19	0.17	0.50	0.19	0.16	0.12
0.95	4	4	0.61	0.18	0.14	0.62	0.18	0.14	0.61	0.18	0.14	0.12
0.95	4	6	0.72	0.18	0.13	0.73	0.19	0.13	0.71	0.18	0.13	0.12

7 Applications

In this section, we apply our results to three data sets from the empirical economics literature. The first application, based on Chetty and Hendren (2015), estimates the effect of living in a given commuting zone during childhood on intergenerational income mobility. The second application, based on Della Vigna and La Ferrara (2010), estimates changes in the stock prices of arms manufacturers following changes in the intensity of conflicts in countries under arms trade embargoes. The third application uses data from the 2000 census of the US, previously employed in Angrist, Chernozhukov, and Fernández-Val (2006) and Belloni and Chernozhukov (2011), to estimate a nonparametric Mincer regression equation of log wages on education and potential experience.

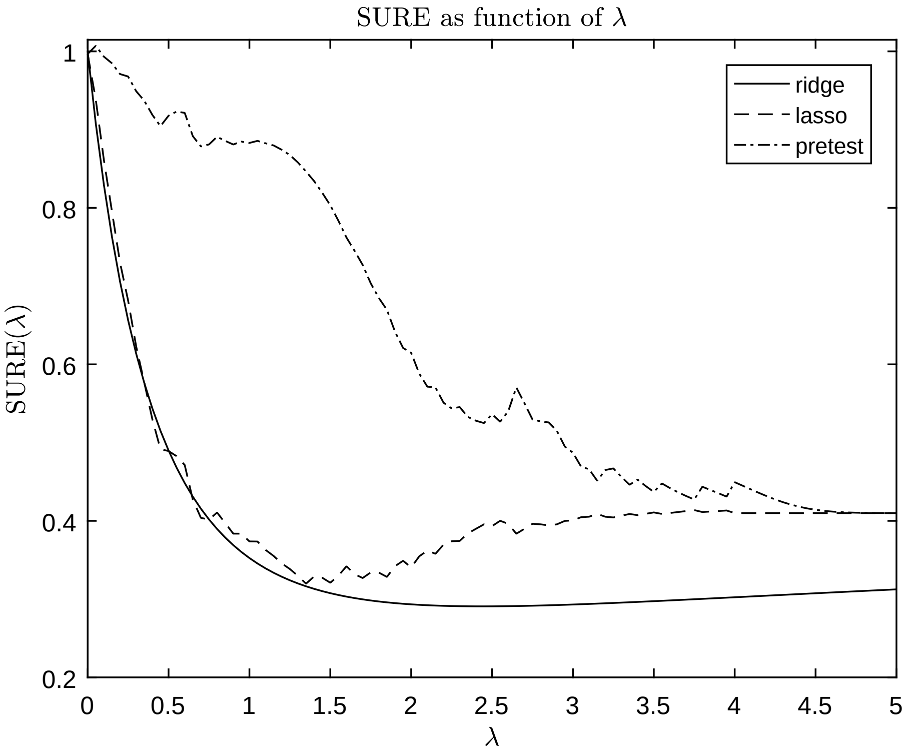
For all applications we normalize the observed X_i by their estimated standard error. Note that this normalization (i) defines the implied loss function, which is quadratic error loss for estimation of the normalized latent parameter μ_i, and (ii) defines the class of estimators considered, which are componentwise shrinkage estimators based on the normalized X_i.

7.1 Neighborhood Effects: Chetty and Hendren (2015)

Chetty and Hendren (2015) use information on income at age 26 for individuals who moved between commuting zones during childhood to estimate the effects of location on income. Identification comes from comparing differently aged children of the same parents, who are exposed to different locations for different durations in their youth. In the context of this application, X_i is the (studentized) estimate of the effect of spending an additional year of childhood in commuting zone i, conditional on parental income rank, on child income rank relative to the national household income distribution at age 26.[4] In this setting, the point zero has no special role; it is just defined, by normalization, to equal the average of commuting zone effects. We therefore have no reason to expect sparsity, nor the presence

[4] The data employed in this section were obtained from `http://www.equality-of-opportunity.org/images/nbhds_online_data_table3.xlsx`. We focus on the estimates for children with parents at the 25th percentile of the national income distribution among parents with children in the same birth cohort.

Figure 5: Neighborhood Effects: SURE Estimates

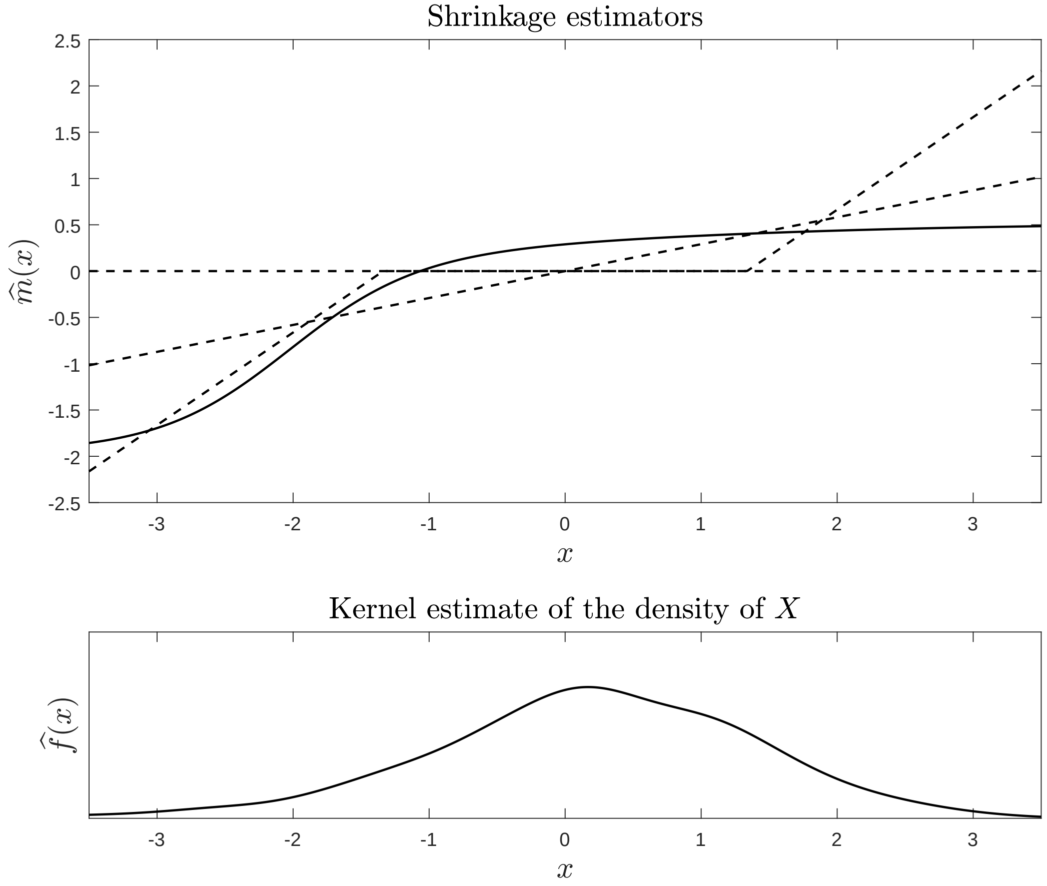


of a set of effects well separated from zero. Our discussion in Section 3 would thus lead us to expect that ridge will perform well, and this is indeed what we find.

Figure 5 reports SURE estimates of risk for ridge, lasso, and pretest estimators, as functions of λ. Among the three estimators, minimal estimated risk is equal to 0.29, and it is attained by ridge for $\widehat{\lambda}_{R,n} = 2.44$. Minimal estimated risk for lasso and pretest are 0.31 and 0.41, respectively. The relative performance of the three shrinkage estimators reflects the characteristics of the example and, in particular, the very limited evidence of sparsity in the data.

The first panel of Figure 6 shows the Koenker-Mizera NPEB estimator (solid line) along with the ridge, lasso, and pretest estimators (dashed lines) evaluated at SURE-minimizing values of the regularization parameters. The identity of the estimators can be easily recognized from their shape. The ridge estimator is linear, with positive slope equal to estimated risk, 0.29. Lasso has the familiar piecewise linear shape, with kinks

Figure 6: Neighborhood Effects: Shrinkage Estimators



The first panel shows the Koenker-Mizera NPEB estimator (solid line) along with the ridge, lasso, and pretest estimators (dashed lines) evaluated at SURE-minimizing values of the regularization parameters. The ridge estimator is linear, with positive slope equal to estimated risk, 0.29. Lasso is piecewise linear, with kinks at the positive and negative versions of the SURE-minimizing value of the regularization parameter, $\widehat{\lambda}_{L,n} = 1.34$. Pretest is flat at zero, because SURE is minimized for values of λ higher than the maximum absolute value of $X_1, \ldots, X_n$. The second panel shows a kernel estimate of the distribution of X.

at the positive and negative versions of the SURE-minimizing value of the regularization parameter, $\widehat{\lambda}_{L,n} = 1.34$. Pretest is flat at zero, because SURE is minimized for values of λ higher than the maximum absolute value of $X_1, \ldots, X_n$. The second panel shows a kernel estimate of the distribution of X.[5] Among ridge, lasso, and pretest, ridge best approximates the optimal shrinkage estimator over most of the estimated distribution of X. Lasso comes a close second, as evidenced in the minimal SURE values for the three estimators, and pretest is way off. Despite substantial shrinkage, these estimates suggest considerable heterogeneity in the effects of childhood neighborhood on earnings. In addition, as expected given the nature of this application, we do not find evidence of sparsity in the location effects estimates.

7.2 Detecting Illegal Arms Trade: Della Vigna and La Ferrara (2010)

Della Vigna and La Ferrara (2010) use changes in stocks prices of arms manufacturing companies at the time of large changes in the intensity of conflicts in countries under arms-trade embargoes to detect illegal arms trade. In this section, we apply the estimators in Section 4 to data from the Della Vigna and La Ferrara study.[6]

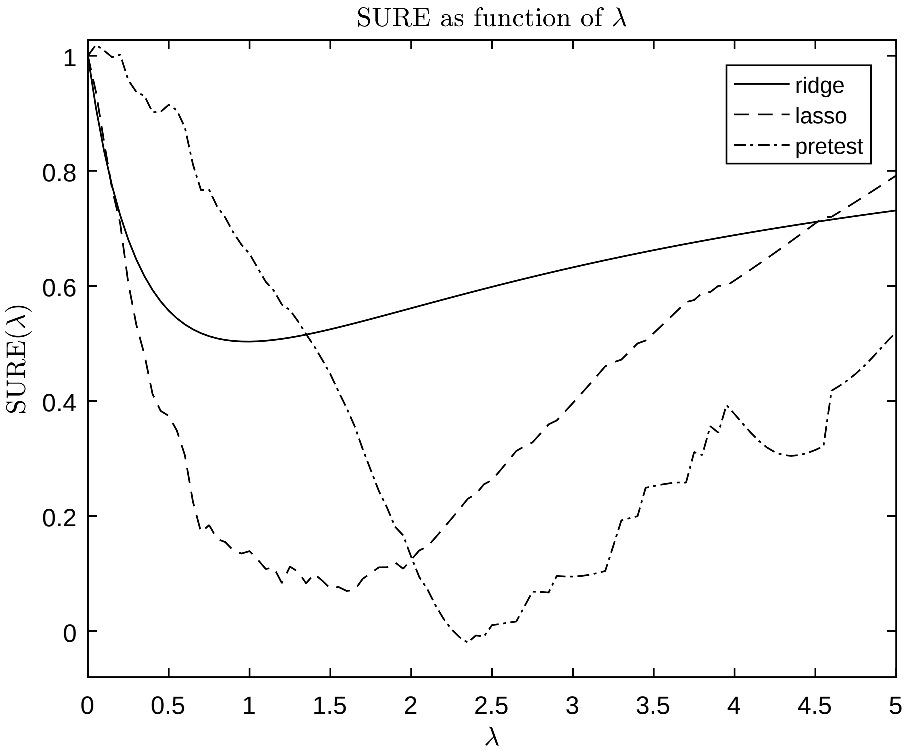
In contrast to the location effects example in Section 7.1, in this application there are reasons to expect a certain amount of sparsity, if changes in the intensity of the conflicts in arms-embargo areas do not affect the stock prices of arms manufacturers that comply with the embargoes.[7] Economic theory would suggest this to be the case if there are fixed costs for violating the embargo. In this case, our discussion of Section 3 would lead us to

[5] To produce a smooth depiction of densities, for the panels reporting densities in this section we use the normal reference rule to choose the bandwidth. See, e.g., Silverman (1986) equation (3.28).

[6] Della Vigna and La Ferrara (2010) divide their sample of arms manufacturers in two groups, depending on whether the company is head-quartered in a country with a high or low level of corruption. They also divide the events of changes in the intensity of the conflicts in embargo areas in two groups, depending on whether the intensity of the conflict increased or decreased at the time of the event. For concreteness, we use the 214 event study estimates for events of increase in the intensity of conflicts in arms embargo areas and for companies in high-corruption countries. The data for this application is available at `http://eml.berkeley.edu/~sdellavi/wp/AEJDataPostingZip.zip`.

[7] In the words of Della Vigna and La Ferrara (2010): "If a company is not trading or trading legally, an event increasing the hostilities should not affect its stock price or should affect it adversely, since it delays the removal of the embargo and hence the re-establishment of legal sales. Conversely, if a company is trading illegally, the event should increase its stock price, since it increases the demand for illegal weapons."

Figure 7: Arms Event Study: SURE Estimates



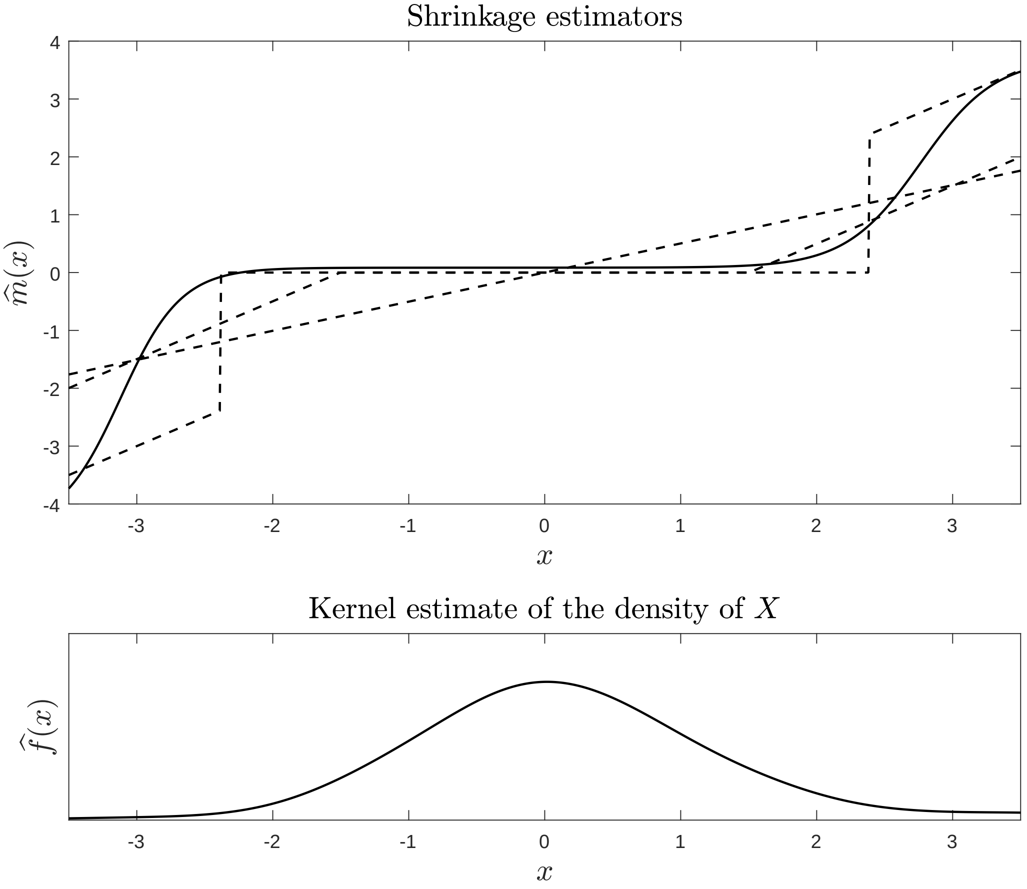
expect that pretest might be optimal, which is again what we find.

Figure 7 shows SURE estimates for ridge, lasso, and pretest. Pretest has the lowest estimated risk, for $\widehat{\lambda}_{PT,n} = 2.39$,[8] followed by lasso, for $\widehat{\lambda}_{L,n} = 1.50$.

Figure 8 depicts the different shrinkage estimators and shows that lasso and especially pretest closely approximate the NPEB estimator over a large part of the distribution of X. The NPEB estimate suggests a substantial amount of sparsity in the distribution of μ. There is, however, a subset of the support of X around $x = 3$ where the estimate of the optimal shrinkage function implies only a small amount of shrinkage. Given the shapes of the optimal shrinkage function estimate and of the estimate of the distribution of X, it is not surprising that the minimal values of SURE in Figure 7 for lasso and pretest are considerably lower than for ridge.

[8] Notice that the pretest's SURE estimate attains a negative minimum value. This could be a matter of estimation variability, of inappropriate choice of bandwidth for the estimation of the density of X in small samples, or it could reflect misspecification of the model (in particular, Gaussianity of X given μ).

Figure 8: Arms Event Study: Shrinkage Estimators



The first panel shows the Koenker-Mizera NPEB estimator (solid line) along with the ridge, lasso, and pretest estimators (dashed lines) evaluated at SURE-minimizing values of the regularization parameters. The ridge estimator is linear, with positive slope equal to estimated risk, 0.50. Lasso is piecewise linear, with kinks at the positive and negative versions of the SURE-minimizing value of the regularization parameter, $\widehat{\lambda}_{L,n} = 1.50$. Pretest is discontinuous at $\widehat{\lambda}_{PT,n} = 2.39$ and $-\widehat{\lambda}_{PT,n} = -2.39$.

7.3 Nonparametric Mincer equation: Belloni and Chernozhukov (2011)

In our third application, we use data from the 2000 US Census in order to estimate a nonparametric regression of log wages on years of education and potential experience, similar to the example considered in Belloni and Chernozhukov (2011).[9] We construct a set of 66 regressors by taking a saturated basis of linear splines in education, fully interacted with the terms of a 6-th order polynomial in potential experience. We orthogonalize these regressors and take the coefficients X_i of an OLS regression of log wages on these orthogonalized regressors as our point of departure. We exclude three coefficients of very large magnitude,[10] which results in $n = 63$. In this application, economics provides less intuition as to what distribution of coefficients to expect. Based on functional analysis considerations, Belloni and Chernozhukov (2011) argue that for plausible families of functions containing the true conditional expectation function, sparse approximations of the coefficients of series regression as induced by the lasso penalty, have low mean squared error.

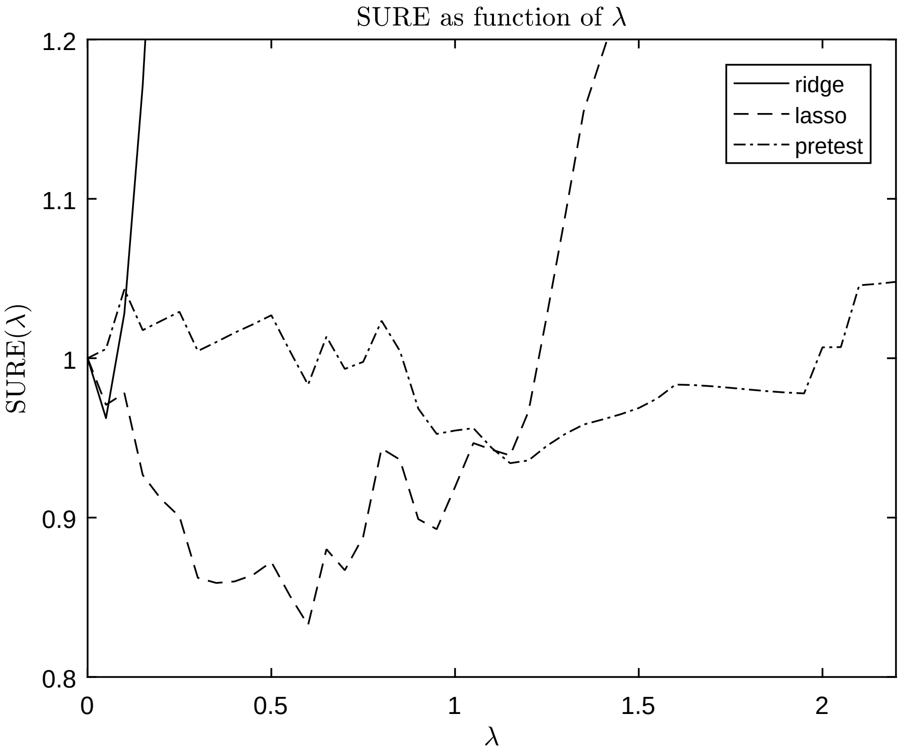
Figure 9 reports SURE estimates of risk for ridge, lasso and pretest. In this application, estimated risk for lasso is substantially smaller than for ridge or pretest.

The top panel of Figure 10 reports the three regularized estimators, ridge, lasso, and pretest, evaluated at the data-driven choice of regularization parameter, along with the Koenker-Mizera NPEB estimator. In order to visualize the differences between the estimates close to the origin, where most of the coefficients are, we report the value of the estimates for $x \in [-10, 10]$. The bottom panel of Figure 10 reports an estimate of the density of X. Locally, the shape of the NPEB estimate looks similar to a step function. This behavior is explained by the fact that the NPEB estimator is based on an approximation to the distribution of μ that is supported on a finite number of values. However, over the whole range of x in the Figure 10, the NPEB estimate is fairly linear. In view of this close-to-linear behavior of NPEB in the $[10, 10]$ interval, the very poor risk performance

[9] The data for this application are available at `http://economics.mit.edu/files/384`.

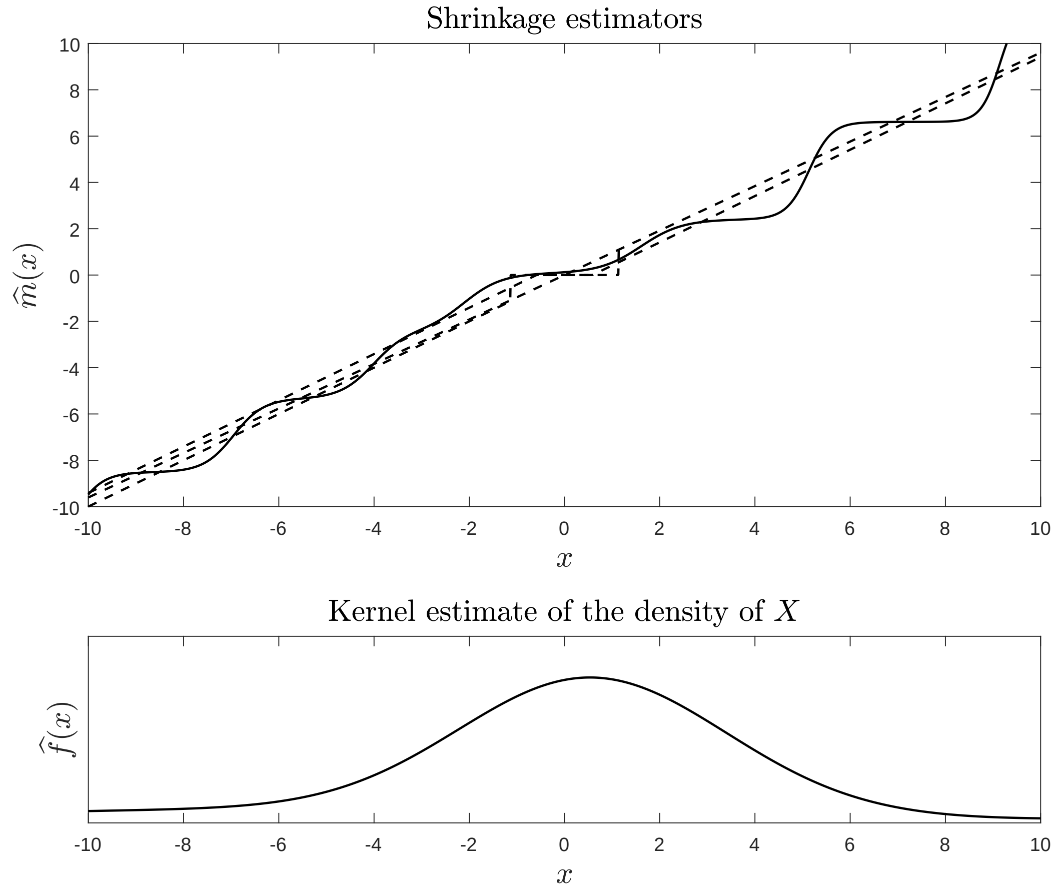
[10] The three excluded coefficients have values, 2938.04 (the intercept), 98.19, and -77.35. The largest absolute value among the included coefficients is -21.06. Most of the included coefficients are small in absolute value. About 40 percent of them have absolute values smaller than one, and about 60 percent of them have absolute value smaller than two.

Figure 9: Nonparametric Mincer Equation: SURE Estimates



of ridge relative to lasso and pretest, as evidenced in Figure 9, may appear surprising. This is explained by the fact that in this application, some of the values in $X_1, \ldots, X_n$ fall exceedingly far from the origin. Linearly shrinking those values towards zero induces severe loss. As a result, ridge attains minimal risk for a close-to-zero value of the regularization parameter, $\widehat{\lambda}_{R,n} = 0.04$, resulting in negligible shrinkage. Among ridge, lasso, and pretest, minimal estimated risk is attained by lasso for $\widehat{\lambda}_{L,n} = 0.59$, which shrinks about 24 percent of the regression coefficients all the way to zero. Pretest induces higher sparsity ($\widehat{\lambda}_{PT,n} = 1.14$, shrinking about 49 percent of the coefficients all the way to zero) but does not improve over lasso in terms of risk.

Figure 10: Nonparametric Mincer Equation: Shrinkage Estimators



The first panel shows the Koenker-Mizera NPEB estimator (solid line) along with the ridge, lasso, and pretest estimators (dashed lines) evaluated at SURE-minimizing values of the regularization parameters. The ridge estimator is linear, with positive slope equal to estimated risk, 0.996. Lasso is piecewise linear, with kinks at the positive and negative versions of the SURE-minimizing value of the regularization parameter, $\lambda = 0.59$. Pretest is discontinuous at $\widehat{\lambda}_{PT,n} = 1.14$ and $-\widehat{\lambda}_{PT,n} = -1.14$. The second panel shows a kernel estimate of the distribution of X.

8 Conclusion

The interest in adopting machine learning methods in economics is growing rapidly. Two common features of machine learning algorithms are regularization and data-driven choice of regularization parameters. We study the properties of such procedures. We consider, in particular, the problem of estimating many means μ_i based on observations X_i. This problem arises often in economic applications. In such applications, the "observations" X_i are usually equal to preliminary least squares coefficient estimates, like fixed effects.

Our goal is to provide guidance for applied researchers on the use of machine learning estimators. Which estimation method should one choose in a given application? And how should one choose regularization parameters? To the extent that researchers care about the squared error of their estimates, procedures are preferable if they have lower mean squared errors than the competitors.

Based on our results, ridge appears to dominate the alternatives considered when the true effects μ_i are smoothly distributed, and there is no point mass of true zeros. This is likely to be the case in applications where the objects of interests are the effects of many treatments, such as locations or teachers, and applications that estimate effects for many subgroups. Pretest appears to dominate if there are true zeros and non-zero effects are well separated from zero. This happens in economic applications when there are fixed costs for agents who engage in non-zero behavior. Lasso finally dominates for intermediate cases and appears to do well for series regression, in particular.

Regarding the choice of regularization parameters, we prove a series of results which show that data-driven choices are almost optimal (in a uniform sense) for large-dimensional problems. This is the case, in particular, for choices of regularization parameters that minimize Stein's Unbiased Risk Estimate (SURE), when observations are normally distributed, and for Cross Validation (CV), when repeated observations for a given effect are available. Although not explicitly analyzed in this article, equation (3) suggests a new empirical selector of regularization parameters based on the minimization of the sample mean square discrepancy between $m(X_i, \lambda)$ and NPEB estimates of $\bar{m}^*_\pi(X_i)$.

There are, of course, some limitations to our analysis. First, we focus on a restricted class of estimators, those which can be written in the componentwise shrinkage form $\widehat{\mu}_i = m(X_i, \widehat{\lambda})$. This covers many estimators of interest for economists, most notably ridge, lasso, and pretest estimation. Many other estimators in the machine learning literature, such as random forests or neural nets, do not have this tractable form. The analysis of the risk properties of such estimators constitutes an interesting avenue of future research.

Finally, we focus on mean square error. This loss function is analytically quite convenient and amenable to tractable results. Other loss functions might be of practical interest, however, and might be studied using numerical methods. In this context, it is also worth emphasizing again that we were focusing on point estimation, where all coefficients μ_i are simultaneously of interest. This is relevant for many practical applications such as those discussed above. In other cases, however, one might instead be interested in the estimates $\widehat{\mu}_i$ solely as input for a lower-dimensional decision problem, or in (frequentist) testing of hypotheses on the coefficients μ_i. Our analysis of mean squared error does not directly speak to such questions.

Appendix

A.1 Relating prediction problems to the normal means model setup

We have introduced our setup in the canonical form of the problem of estimating many means. Machine learning methods are often discussed in terms of the problem of minimizing out-of-sample prediction error. The two problems are closely related. Consider the linear prediction model

$$Y = \boldsymbol{W}'\boldsymbol{\beta} + \epsilon,$$

where Y is a scalar random variable, $\boldsymbol{W}$ is an $(n \times 1)$ vector of covariates (features), and $\epsilon|\boldsymbol{W} \sim N(0, \sigma^2)$.[11] The machine learning literature is often concerned with the problem of predicting the value of Y of a draw of $(Y, \boldsymbol{W})$ using

$$\widehat{Y} = \boldsymbol{W}'\widehat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\beta}}$ is an estimator of $\boldsymbol{\beta}$ based on N ($N \geq n$) previous independent draws, $(Y_1, \boldsymbol{W}_1), \ldots, (Y_N, \boldsymbol{W}_N)$, from the distribution of $(Y, \boldsymbol{W})$, so $\widehat{\beta}$ is independent of $(Y, \boldsymbol{W})$. We evaluate out-of-sample predictions based on the squared prediction error,

$$\tilde{L} = (\widehat{Y} - Y)^2 = \left(\boldsymbol{W}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)^2 + \epsilon^2 + 2\left(\boldsymbol{W}'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)\epsilon.$$

Suppose that the features $\boldsymbol{W}$ for prediction are drawn from the empirical distribution of $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N$,[12] and that Y is drawn from the conditional population distribution of Y given $\boldsymbol{W}$. The expected squared prediction error, $\tilde{R} = E[\tilde{L}]$, is then equal to

$$\tilde{R} = \operatorname{tr}\left(\boldsymbol{\Omega} \cdot E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})']\right) + E[\epsilon^2],$$

where

$$\boldsymbol{\Omega} = \frac{1}{N}\sum_{j=1}^{N} \boldsymbol{W}_j \boldsymbol{W}_j'.$$

In the special case where the components of $\boldsymbol{W}$ are orthonormal in the sample, $\boldsymbol{\Omega} = \boldsymbol{I}_n$, this immediately yields

$$\tilde{R} = \sum_{i=1}^{n} E[(\widehat{\beta}_i - \beta_i)^2] + E[\epsilon^2],$$

where $\widehat{\beta}_i$ and β_i are the i-th components of $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$, respectively. In this special case, we thus get that the risk function for out of sample prediction and the mean squared error for coefficient estimation are the same, up to a constant.

More generally, assume that $\boldsymbol{\Omega}$ has full rank, define $\boldsymbol{V} = \boldsymbol{\Omega}^{-1/2}\boldsymbol{W}$, $\boldsymbol{\mu} = \boldsymbol{\Omega}^{1/2}\boldsymbol{\beta}$, and let $\boldsymbol{X}$ be the coefficients of an ordinary least squares regression of $Y_1, \ldots, Y_N$ on $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_N$. This change of coordinates yields, conditional on $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_N$,

$$\boldsymbol{X} \sim N\left(\boldsymbol{\mu}, \frac{\sigma^2}{N}\boldsymbol{I}_n\right),$$

so that the assumptions of our setup regarding $\boldsymbol{X}$ and $\boldsymbol{\mu}$ hold. Regularized estimators $\widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ can be formed by componentwise shrinkage of $\boldsymbol{X}$. For any estimator $\widehat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}$ we can furthermore write the corresponding risk for out of sample prediction as

$$\tilde{R} = E[(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})'(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})] + E[\epsilon^2].$$

[11] Linearity of the conditional expectation and normality are assumed here for ease of exposition; both could in principle be dropped in an asymptotic version of the following argument.

[12] This assumption is again made for convenience, to sidestep asymptotic approximations

To summarize: After orthogonalizing the regressors for a linear regression problem, the assumptions of the many means setup apply to the vector of ordinary least squares coefficients. The risk function for out of sample prediction is furthermore the same as the risk function of the many means problem, if we assume the features for prediction are drawn from the empirical distribution of observed features.

A.2 Assuming oracle knowledge of zeros is not uniformly valid

Consider the pretest estimator, $m_{PT}(X_i, \widehat{\lambda}_n)$. An alternative approximation to the risk of the pretest estimator is given by the risk of the infeasible estimator based on oracle-knowledge of true zeros,

$$m_{PT}^{0,\mu}(X_i) = 1(\mu_i \neq 0)X_i.$$

As we show now, this approximation is not uniformly valid, which illustrates that uniformity is not a trivial requirement. Consider the following family $\mathcal{Q}$ of data generating processes,

$$\begin{aligned} X|\mu &\sim N(\mu, 1), \\ P(\mu = 0) &= p, \\ P(\mu = \mu_0) &= 1 - p. \end{aligned}$$

It is easy to check that

$$\bar{R}(m_{PT}^{0,\mu}(\cdot), \pi) = 1 - p,$$

for all $\pi \in \mathcal{Q}$. By Proposition 1, for $\pi \in \mathcal{Q}$, the integrated risk of the pretest estimator is

$$\begin{aligned} \bar{R}(m_{PT}(\cdot, \lambda), \pi) = {} & 2\Big(\Phi(-\lambda) + \lambda\phi(\lambda)\Big)p \\ & + \Big(1 + \Phi(-\lambda - \mu_0) - \Phi(\lambda - \mu_0) + (\Phi(\lambda - \mu_0) - \Phi(-\lambda - \mu_0))\mu_0^2 \\ & - \phi(\lambda - \mu_0)\big(-\lambda + \mu_0\big) - \phi(-\lambda - \mu_0)\big(-\lambda - \mu_0\big)\Big)(1 - p). \end{aligned}$$

We have shown above that data-driven choices of λ are uniformly risk consistent, so their integrated risk is asymptotically equal to $\min_{\lambda \in [0,\infty]} R(m_{PT}(\cdot, \lambda), \pi)$. It follows that the risk of $m_{PT}^{0,\mu}(\cdot)$ provides a uniformaly valid approximation to the risk of $m_{PT}(\cdot, \widehat{\lambda})$ if and only if

$$\min_{\lambda \in [0,\infty]} \bar{R}(m_{PT}(\cdot, \lambda), \pi) = 1 - p, \quad \forall \pi \in \mathcal{Q}. \tag{A.1}$$

It is easy to show that equation (A.1) is violated. Consider, for example, $(p, \mu_0) = (1/2, \sqrt{2})$. Then, the minimum value of $\bar{R}(m_{PT}(, \lambda), \pi)$ is equal to one (achieved at $\lambda = 0$ and $\lambda = \infty$). Therefore,

$$\min_{\lambda \in [0,\infty]} \bar{R}(m_{PT}(, \lambda), \pi) = 1 > 0.5 = \bar{R}(m_{PT}^{0,\mu}(\cdot), \pi).$$

Moreover, equation (A.1) is also violated in the opposite direction. Notice that

$$\lim_{\lambda \to \infty} \bar{R}(m_{PT}(\cdot, \lambda), \pi) = (1 - p)\mu_0^2.$$

As a result, if $|\mu_0| < 1$ we obtain

$$\min_{\lambda \in [0,\infty]} \bar{R}(m_{PT}(, \lambda), \pi) < 1 - p = \bar{R}(m_{PT}^{0,\mu}(\cdot), \pi),$$

which violates equation (A.1).

A.3 Proofs

Proof of Theorem 1:

$$\begin{aligned}
R_n(m(.,\lambda),\boldsymbol{P}) &= \frac{1}{n}\sum_{i=1}^{n} E[(m(X_i,\lambda)-\mu_i)^2|P_i] \\
&= E\big[(m(X_I,\lambda)-\mu_I)^2|\boldsymbol{P}\big] \\
&= E\big[E[(m^*_{\boldsymbol{P}}(X_I)-\mu_I)^2|X_I,\boldsymbol{P}]|\boldsymbol{P}\big] + E\big[(m(X_I,\lambda)-m^*_{\boldsymbol{P}}(X_I))^2|\boldsymbol{P}\big] \\
&= v^*_{\boldsymbol{P}} + E\big[(m(X_I,\lambda)-m^*_{\boldsymbol{P}}(X_I))^2|\boldsymbol{P}\big].
\end{aligned}$$

The second equality in this proof is termed *the fundamental theorem of compound decisions* in Jiang and Zhang (2009), who credit Robbins (1951). Finiteness of $\mu_1,\dots,\mu_n$, and $\sup_{\lambda\in[0,\infty]} E[(m(X_I,\lambda))^2|\boldsymbol{P}]$ implies that all relevant expectations are finite. □

Proof of Lemma 1: Notice that

$$m_R(x,\lambda)-\mu_i = \left(\frac{1}{1+\lambda}\right)(x-\mu_i) - \left(\frac{\lambda}{1+\lambda}\right)\mu_i.$$

The result for ridge equals the second moment of this expression. For pretest, notice that

$$m_{PT}(x,\lambda)-\mu_i = 1(|x|>\lambda)(x-\mu_i) - 1(|x|\le\lambda)\mu_i.$$

Therefore,

$$R(m_{PT}(\cdot,\lambda),P_i) = E\big[(X_i-\mu_i)^2 1(|X_i|>\lambda)\big] + \mu_i^2 \Pr\big(|X_i|\le\lambda\big). \tag{A.2}$$

Using the fact that $\phi'(v) = -v\phi(v)$ and integrating by parts, we obtain

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

Now,

$$\begin{aligned}
E\big[(X_i-\mu_i)^2 1(|X_i|>\lambda)\big] &= \sigma_i^2 E\left[\left(\frac{X_i-\mu_i}{\sigma_i}\right)^2 1(|X_i|>\lambda)\right] \\
&= \left(1+\Phi\Big(\frac{-\lambda-\mu_i}{\sigma_i}\Big)-\Phi\Big(\frac{\lambda-\mu_i}{\sigma_i}\Big)\right)\sigma_i^2 \\
&\quad + \left(\Big(\frac{\lambda-\mu_i}{\sigma_i}\Big)\phi\Big(\frac{\lambda-\mu_i}{\sigma_i}\Big)-\Big(\frac{-\lambda-\mu_i}{\sigma_i}\Big)\phi\Big(\frac{-\lambda-\mu_i}{\sigma_i}\Big)\right)\sigma_i^2.
\end{aligned} \tag{A.3}$$

The result for the pretest estimator now follows easily from equations (A.2) and (A.3). For lasso, notice that

$$\begin{aligned}
m_L(x,\lambda)-\mu_i &= 1(x<-\lambda)(x+\lambda-\mu_i) + 1(x>\lambda)(x-\lambda-\mu_i) - 1(|x|\le\lambda)\mu_i \\
&= 1(|x|>\lambda)(x-\mu_i) + (1(x<-\lambda)-1(x>\lambda))\lambda - 1(|x|\le\lambda)\mu_i.
\end{aligned}$$

Therefore,

$$\begin{aligned}
R(m_L(\cdot,\lambda),P_i) &= E\big[(X_i-\mu_i)^2 1(|X_i|>\lambda)\big] + \lambda^2 E[1(|X_i|>\lambda)] + \mu_i^2 E[1(|X_i|\le\lambda)] \\
&\quad + 2\lambda\Big(E\big[(X_i-\mu_i)1(X_i<-\lambda)\big] - E\big[(X_i-\mu_i)1(X_i>\lambda)\big]\Big)
\end{aligned}$$

$$
\begin{aligned}
&= R(m_{PT}(\cdot,\lambda),P_i) + \lambda^2 E[1(|X_i|>\lambda)] \\
&+ 2\lambda\Big(E\big[(X_i-\mu_i)1(X_i<-\lambda)\big] - E\big[(X_i-\mu_i)1(X_i>\lambda)\big]\Big).
\end{aligned} \tag{A.4}
$$

Notice that

$$\int_a^b v\phi(v)dv = \phi(a)-\phi(b).$$

As a result,

$$E\big[(X_i-\mu_i)1(X_i<-\lambda)\big] - E\big[(X_i-\mu_i)1(X_i>\lambda)\big] = -\sigma_i\left(\phi\Big(\frac{-\lambda-\mu_i}{\sigma_i}\Big)+\phi\Big(\frac{\lambda-\mu_i}{\sigma_i}\Big)\right). \tag{A.5}$$

Now, the result for lasso follows from equations (A.4) and (A.5). □

Proof of Proposition 1: The results for ridge are trivial. For lasso, first notice that the integrated risk at zero is:

$$R_0(m_L(\cdot,\lambda),\pi) = 2\Phi\Big(\frac{-\lambda}{\sigma}\Big)(\sigma^2+\lambda^2) - 2\Big(\frac{\lambda}{\sigma}\Big)\phi\Big(\frac{\lambda}{\sigma}\Big)\sigma^2.$$

Next, notice that

$$
\begin{aligned}
\int \Phi\Big(\frac{-\lambda-\mu}{\sigma}\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu_0-\mu}{\sigma_0}\Big)d\mu &= \Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right), \\
\int \Phi\Big(\frac{\lambda-\mu}{\sigma}\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu_0-\mu}{\sigma_0}\Big)d\mu &= \Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right), \\
\int \Big(\frac{-\lambda-\mu}{\sigma}\Big)\phi\Big(\frac{\lambda-\mu}{\sigma}\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu_0-\mu}{\sigma_0}\Big)d\mu &= -\left(\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\Big(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\Big)\right)\left(\lambda+\frac{\mu_0\sigma^2+\lambda\sigma_0^2}{\sigma_0^2+\sigma^2}\right) \\
\int \Big(\frac{-\lambda+\mu}{\sigma}\Big)\phi\Big(\frac{-\lambda-\mu}{\sigma}\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu_0-\mu}{\sigma_0}\Big)d\mu &= -\left(\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\Big(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\Big)\right)\left(\lambda-\frac{\mu_0\sigma^2-\lambda\sigma_0^2}{\sigma_0^2+\sigma^2}\right).
\end{aligned}
$$

The integrals involving μ^2 are more involved. Let v be a Standard normal variable independent of μ. Notice that,

$$
\begin{aligned}
\int \mu^2\Phi\Big(\frac{\lambda-\mu}{\sigma}\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)d\mu &= \int \mu^2\Big(\int I_{[v\le(\lambda-\mu)/\sigma]}\phi(v)dv\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)d\mu \\
&= \int\Big(\int \mu^2 I_{[\mu\le\lambda-\sigma v]}\frac{1}{\sigma_0}\phi\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)d\mu\Big)\phi(v)dv.
\end{aligned}
$$

Using the change of variable $u=(\mu-\mu_0)/\sigma_0$, we obtain,

$$
\begin{aligned}
\int \mu^2 I_{[\mu\le\lambda-\sigma v]}\frac{1}{\sigma_0}\phi\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)d\mu &= \int(\mu_0+\sigma_0 u)^2 I_{[u\le(\lambda-\mu_0-\sigma v)/\sigma_0]}\phi(u)du \\
&= \Phi\Big(\frac{\lambda-\mu_0-\sigma v}{\sigma_0}\Big)\mu_0^2 - 2\phi\Big(\frac{\lambda-\mu_0-\sigma v}{\sigma_0}\Big)\sigma_0\mu_0 \\
&+ \left(\Phi\Big(\frac{\lambda-\mu_0-\sigma v}{\sigma_0}\Big) - \Big(\frac{\lambda-\mu_0-\sigma v}{\sigma_0}\Big)\phi\Big(\frac{\lambda-\mu_0-\sigma v}{\sigma_0}\Big)\right)\sigma_0^2 \\
&= \Phi\Big(\frac{\lambda-\mu_0-\sigma v}{\sigma_0}\Big)(\mu_0^2+\sigma_0^2) - \phi\Big(\frac{\lambda-\mu_0-\sigma v}{\sigma_0}\Big)\sigma_0(\lambda+\mu_0-\sigma v).
\end{aligned}
$$

Therefore,

$$\int \mu^2\Phi\Big(\frac{\lambda-\mu}{\sigma}\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)d\mu = \Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)(\mu_0^2+\sigma_0^2)$$

$$
\begin{aligned}
&- \frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)(\lambda+\mu_0)\sigma_0^2 \\
&+ \sigma_0^2\sigma^2\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\left(\frac{\lambda-\mu_0}{\sigma_0^2+\sigma^2}\right).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\int \mu^2\Phi\Big(\frac{-\lambda-\mu}{\sigma}\Big)\frac{1}{\sigma_0}\phi\Big(\frac{\mu-\mu_0}{\sigma_0}\Big)d\mu = \Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)&(\mu_0^2+\sigma_0^2) \\
- \frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)&(-\lambda+\mu_0)\sigma_0^2 \\
+ \sigma_0^2\sigma^2\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)&\left(\frac{-\lambda-\mu_0}{\sigma_0^2+\sigma^2}\right).
\end{aligned}
$$

The integrated risk conditional on $\mu \neq 0$ is

$$
\begin{aligned}
R_1(m_L(\cdot,\lambda),\pi) = &\left(1+\Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)-\Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\right)(\sigma^2+\lambda^2) \\
&+\left(\Phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)-\Phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)\right)(\mu_0^2+\sigma_0^2) \\
&-\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)(\lambda+\mu_0)(\sigma_0^2+\sigma^2) \\
&-\frac{1}{\sqrt{\sigma_0^2+\sigma^2}}\phi\left(\frac{-\lambda-\mu_0}{\sqrt{\sigma_0^2+\sigma^2}}\right)(\lambda-\mu_0)(\sigma_0^2+\sigma^2).
\end{aligned}
$$

The results for pretest follow from similar calculations. □

The next lemma is used in the proof of Theorem 2.

Lemma A.1
For any two real-valued functions, f and g,

$$
\Big|\inf f - \inf g\Big| \leq \sup|f-g|.
$$

Proof: The result of the lemma follows directly from

$$
\inf f \geq \inf g - \sup|f-g|,
$$

and

$$
\inf g \geq \inf f - \sup|f-g|.
$$

□

Proof of Theorem 2: Because $\bar{v}_\pi$ does not depend on λ, we obtain

$$
\begin{aligned}
\Big(L_n(\lambda)-L_n(\widehat{\lambda}_n)\Big)-\Big(r_n(\lambda)-r_n(\widehat{\lambda}_n)\Big) = &\Big(L_n(\lambda)-\bar{R}_\pi(\lambda)\Big)-\Big(L_n(\widehat{\lambda}_n)-\bar{R}_\pi(\widehat{\lambda}_n)\Big) \\
&+\Big(\bar{r}_\pi(\lambda)-r_n(\lambda)\Big)-\Big(\bar{r}_\pi(\widehat{\lambda}_n)-r_n(\widehat{\lambda}_n)\Big).
\end{aligned}
$$

Applying Lemma A.1 we obtain

$$\Big|\Big(\inf_{\lambda\in[0,\infty]} L_n(\lambda) - L_n(\widehat{\lambda}_n)\Big) - \Big(\inf_{\lambda\in[0,\infty]} r_n(\lambda) - r_n(\widehat{\lambda}_n)\Big)\Big| \le 2\sup_{\lambda\in[0,\infty]}\Big|L_n(\lambda) - \bar{R}_\pi(\lambda)\Big| \\ + 2\sup_{\lambda\in[0,\infty]}\Big|\bar{r}_\pi(\lambda) - r_n(\lambda)\Big|.$$

Given that $\widehat{\lambda}_n$ is the value of λ at which $r_n(\lambda)$ attains its minimum, the result of the theorem follows. □

The following preliminary lemma will be used in the proof of Theorem 3.

Lemma A.2
For any finite set of regularization parameters, $0 = \lambda_0 < \ldots < \lambda_k = \infty$, let

$$u_j = \sup_{\lambda\in[\lambda_{j-1},\lambda_j]} L(\lambda)$$
$$l_j = \inf_{\lambda\in[\lambda_{j-1},\lambda_j]} L(\lambda),$$

where $L(\lambda) = (\mu - m(X,\lambda))^2$. Suppose that for any $\epsilon > 0$ there is a finite set of regularization parameters, $0 = \lambda_0 < \ldots < \lambda_k = \infty$ (where k may depend on ϵ), such that

$$\sup_{\pi\in\mathcal{Q}} \max_{1\le j\le k} E_\pi[u_j - l_j] \le \epsilon \tag{A.6}$$

and

$$\sup_{\pi\in\mathcal{Q}} \max_{1\le j\le k} \max\{\mathrm{var}_\pi(l_j), \mathrm{var}_\pi(u_j)\} < \infty. \tag{A.7}$$

Then, equation (6) holds.

Proof: We will use E_n to indicate averages over $(\mu_1, X_1), \ldots, (\mu_n, X_n)$. Let $\lambda \in [\lambda_{j-1}, \lambda_j]$. By construction

$$E_n[L(\lambda)] - E_\pi[L(\lambda)] \le E_n[u_j] - E_\pi[l_j] \le E_n[u_j] - E_\pi[u_j] + E_\pi[u_j - l_j]$$
$$E_n[L(\lambda)] - E_\pi[L(\lambda)] \ge E_n[l_j] - E_\pi[u_j] \ge E_n[l_j] - E_\pi[l_j] - E_\pi[u_j - l_j]$$

and thus

$$\begin{aligned}\sup_{\lambda\in[0,\infty]} &(E_n[L(\lambda)] - E_\pi[L(\lambda)])^2 \\ &\le \max_{1\le j\le k} \max\{(E_n[u_j] - E_\pi[u_j])^2, (E_n[l_j] - E_\pi[l_j])^2\} + \Big(\max_{1\le j\le k} E_\pi[u_j - l_j]\Big)^2 \\ &+ 2\max_{1\le j\le k} \max\{|E_n[u_j] - E_\pi[u_j]|, |E_n[l_j] - E_\pi[l_j]|\} \max_{1\le j\le k} E_\pi[u_j - l_j] \\ &\le \sum_{j=1}^{k} \Big((E_n[u_j] - E_\pi[u_j])^2 + (E_n[l_j] - E_\pi[l_j])^2\Big) + \epsilon^2 \\ &+ 2\epsilon \sum_{j=1}^{k} \Big(|E_n[u_j] - E_\pi[u_j]| + |E_n[l_j] - E_\pi[l_j]|\Big).\end{aligned}$$

Therefore,

$$E_\pi\Big[\sup_{\lambda\in[0,\infty]} (E_n[L(\lambda)] - E_\pi[L(\lambda)])^2\Big]$$

$$\leq \sum_{j=1}^{k} \Big(E_\pi[(E_n[u_j] - E_\pi[u_j])^2] + E_\pi[E_n[l_j] - E_\pi[l_j])^2] \Big) + \epsilon^2$$
$$+ 2\epsilon \sum_{j=1}^{k} E_\pi[|E_n[u_j] - E_\pi[u_j]| + |E_n[l_j] - E_\pi[l_j]|]$$
$$\leq \sum_{j=1}^{k} \Big(\mathrm{var}_\pi(u_j)/n + \mathrm{var}_\pi(l_j)/n \Big) + \epsilon^2$$
$$+ 2\epsilon \sum_{j=1}^{k} \Big(\sqrt{\mathrm{var}_\pi(u_j)/n} + \sqrt{\mathrm{var}_\pi(l_j)/n} \Big).$$

Now, the result of the lemma follows from the assumption of uniformly bounded variances. □

Proof of Theorem 3: We will show that the conditions of the theorem imply equations (A.6) and (A.7) and, therefore, the uniform convergence result in equation (6). Using conditions 1 and 2, along with the convexity of 4th powers, we immediately get bounded variances. Because the maximum of a convex function is achieved at the boundary,

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

Notice also that

$$\mathrm{var}_\pi(l_j) \leq E_\pi[l_j^2] \leq E_\pi[u_j^2].$$

Now, condition 3 implies equation (A.7) in Lemma A.2.

It remains to find a set of regularization parameters such that $E_\pi[u_j - l_j] < \epsilon$ for all j. Using again the monotonicity of $m(X, \lambda)$ in λ and convexity of the square function, we have that the supremum defining u_j is achieved at the boundary,

$$u_j = \max\{L(\lambda_{j-1}), L(\lambda_j)\},$$

while

$$l_j = \min\{L(\lambda_{j-1}), L(\lambda_j)\}$$

if $\mu \notin [m(X, \lambda_{j-1}), m(X, \lambda_j)]$ and $l_j = 0$, otherwise. In the former case,

$$u_j - l_j = |L(\lambda_j) - L(\lambda_{j-1})|,$$

and in the latter case, $u_j - l_j = \max\{L(\lambda_{j-1}), L(\lambda_j)\}$. Consider first the case of $\mu \notin [m(X, \lambda_{j-1}), m(X, \lambda_j)]$. Using the formula $a^2 - b^2 = (a+b)(a-b)$ and the shorthand $m_j = m(X, \lambda_j)$, we obtain

$$\begin{aligned} u_j - l_j &= \big|(m_j - \mu)^2 - (m_{j-1} - \mu)^2\big| \\ &= \big|\big((m_j - \mu) + (m_{j-1} - \mu)\big)\big(m_j - m_{j-1}\big)\big| \\ &\leq (|m_j - \mu| + |m_{j-1} - \mu|)|m_j - m_{j-1}|. \end{aligned}$$

To check that the same bound applies to the case $\mu \in [m(X, \lambda_{j-1}), m(X, \lambda_j)]$, notice that

$$\max\{|m_j - \mu|, |m_{j-1} - \mu|\} \leq |m_j - \mu| + |m_{j-1} - \mu|$$

and because $\mu \in [m(X, \lambda_{j-1}), m(X, \lambda_j)]$,

$$\max\{|m_j - \mu|, |m_{j-1} - \mu|\} \leq |m_j - m_{j-1}|.$$

Monotonicity, boundary conditions, and the convexity of absolute values allow one to bound further,

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

Now, condition 4 in Theorem 3 implies equation (A.6) in Lemma A.2 and, therefore, the result of the theorem. □

Proof of Lemma 2: Conditions 1 and 2 of Theorem 3 are easily verified to hold for ridge, lasso, and the pretest estimator. Let us thus discuss condition 4.

Let $\Delta m_j = m(X, \lambda_j) - m(X, \lambda_{j-1})$, and $\Delta\lambda_j = \lambda_j - \lambda_{j-1}$. For ridge, Δm_j is given by

$$\Delta m_j = \left(\frac{1}{1+\lambda_j} - \frac{1}{1+\lambda_{j-1}}\right) X$$

so that the requirement follows from finite variances if we choose a finite set of regularization parameters such that

$$\left|\frac{1}{1+\lambda_j} - \frac{1}{1+\lambda_{j-1}}\right| \sup_{\pi\in\mathcal{Q}} E\big[(|X-\mu|+|\mu|)|X|\big] < \epsilon$$

for all $j = 1, \ldots, k$, which is possible by the uniformly bounded moments condition.

For lasso, notice that $|\Delta m_k| = (|X| - \lambda_{k-1})\, 1(|X| > \lambda_{k-1}) \leq |X|\, 1(|X| > \lambda_{k-1})$, and $|\Delta m_j| \leq \Delta\lambda_j$ for $j = 1, \ldots, k-1$. We will first verify that for any $\epsilon > 0$ there is a finite λ_{k-1} such that condition 4 of the lemma holds for $j = k$. Notice that for any pair of non-negative random variables (ξ, ζ) such that $E[\xi\,\zeta] < \infty$ and for any positive constant, c, we have that

$$E[\xi\zeta] \geq E[\xi\zeta\, 1(\zeta > c)] \geq cE[\xi 1(\zeta > c)]$$

and, therefore,

$$E[\xi 1(\zeta > c)] \leq \frac{E[\xi\zeta]}{c}.$$

As a consequence of this inequality, and because $\sup_{\pi\in\mathcal{Q}} E_\pi[(|X-\mu|+|\mu|)|X|^2] < \infty$ (implied by condition 3), then for any $\epsilon > 0$ there exists a finite positive constant, λ_{k-1} such that condition 4 of the lemma holds for $j = k$. Given that λ_{k-1} is finite, $\sup_{\pi\in\mathcal{Q}} E_\pi[|X-\mu|+|\mu|] < \infty$ and $|\Delta m_j| \leq \Delta\lambda_j$ imply condition 4 for $j = 1, \ldots, k-1$.

For pretest,

$$|\Delta m_j| = |X|\, 1(|X| \in (\lambda_{j-1}, \lambda_j]),$$

so that we require that for any $\epsilon > 0$ we can find a finite number of regularization parameters, $0 = \lambda_0 < \lambda_1 < \ldots < \lambda_{k-1} < \lambda_k = \infty$, such that

$$E_\pi[(|X-\mu|+|\mu|)|X|\, 1(|X| \in (\lambda_{j-1}, \lambda_j])] < \epsilon,$$

for $j = 1, \ldots, k$. Applying the Cauchy-Schwarz inequality and uniform boundedness of fourth moments, this condition is satisfied if we can choose uniformly bounded $P_\pi(|X| \in (\lambda_{j-1}, \lambda_j])$, which is possible under the assumption that X is continuously distributed with a (version of the) density that is uniformly bounded. □

Proof of Corollary 1: From Theorem 2 and Lemma A.1, it follows immediately that

$$\sup_{\pi\in\mathcal{Q}} P_\pi\left(\left|L_n(\widehat{\lambda}_n) - \inf_{\lambda\in[0,\infty]} \bar{R}_\pi(\lambda)\right| > \epsilon\right) \to 0.$$

By definition,

$$\bar{R}(m(.,\widehat{\lambda}_n), \pi) = E_\pi[L_n(\widehat{\lambda}_n)].$$

Equation (7) thus follows if we can strengthen uniform convergence in probability to uniform L^1 convergence. To do so, we need to show uniform integrability of $L_n(\widehat{\lambda}_n)$, as per Theorem 2.20 in van der Vaart (1998).

Monotonicity, convexity of loss, and boundary conditions imply

$$L_n(\widehat{\lambda}_n) \leq \frac{1}{n} \sum_{i=1}^{n} \Big(\mu_i^2 + (X_i - \mu_i)^2 \Big).$$

Uniform integrability along arbitrary sequences π_n, and thus L^1 convergence, follows from the assumed bounds on moments. □

Proof of Lemma 3: Recall the definition $\bar{R}(m(.), \pi) = E_\pi[(m(X) - \mu)^2]$. Expanding the square yields

$$\begin{aligned} E_\pi[(m(X) - \mu)^2] &= E_\pi[(m(X) - X + X - \mu)^2] \\ &= E_\pi[(X - \mu)^2] + E_\pi[(m(X) - X)^2] + 2E_\pi[(X - \mu)(m(X) - X)]. \end{aligned}$$

By the form of the standard normal density,

$$\nabla_x \phi(x - \mu) = -(x - \mu)\phi(x - \mu).$$

Partial integration over the intervals $]x_j, x_{j+1}[$ (where we let $x_0 = -\infty$ and $x_{J+1} = \infty$) yields

$$\begin{aligned} E_\pi[(X - \mu)(m(X) - X)] &= \int_{\mathbb{R}} \int_{\mathbb{R}} (x - \mu)\,(m(x) - x)\,\phi(x - \mu)\,dx\,d\pi(\mu) \\ &= -\sum_{j=0}^{J} \int_{\mathbb{R}} \int_{x_j}^{x_{j+1}} (m(x) - x)\,\nabla_x \phi(x - \mu)\,dx\,d\pi(\mu) \\ &= \sum_{j=0}^{J} \int_{\mathbb{R}} \Bigg[\int_{x_j}^{x_{j+1}} (\nabla m(x) - 1)\,\phi(x - \mu)\,dx \\ &+ \lim_{x \downarrow x_j} (m(x) - x)\phi(x - \mu) - \lim_{x \uparrow x_{j+1}} (m(x) - x)\phi(x - \mu) \Bigg] d\pi(\mu) \\ &= E_\pi[\nabla m(X)] - 1 + \sum_{j=1}^{J} \Delta m_j f(x_j). \end{aligned}$$

□

Proof of Lemma 4: Uniform convergence of the first term follows by the exact same arguments we used to show uniform convergence of $L_n(\lambda)$ to $\bar{R}_\pi(\lambda)$ in Theorem 3. We thus focus on the second term, and discuss its convergence on a case-by-case basis for our leading examples.

For ridge, this second term is equal to the constant

$$2\,\nabla_x m_R(x, \lambda) = \frac{2}{1 + \lambda},$$

and uniform convergence holds trivially.

For lasso, the second term is equal to

$$2\,E_n[\nabla_x m_L(X, \lambda)] = 2\,P_n(|X| > \lambda).$$

To prove uniform convergence of this term we slightly modify the proof of the Glivenko-Cantelly Theorem (e.g., van der Vaart (1998), Theorem 19.1). Let F_n be the cumulative distribution function of $X_1, \ldots, X_n$, and let F_π be its population counterpart. It is enough to prove uniform convergence of $F_n(\lambda)$,

$$\sup_{\pi \in \mathcal{Q}} P_\pi \left(\sup_{\lambda \in [0, \infty]} |F_n(\lambda) - F_\pi(\lambda)| > \epsilon \right) \to 0 \quad \forall \epsilon > 0.$$

Using Chebyshev's inequality and $\sup_{\pi \in \mathcal{Q}} \operatorname{var}_\pi(1(X \leq \lambda)) \leq 1/4$ for every $\lambda \in [0, \infty]$, we obtain

$$\sup_{\pi \in \mathcal{Q}} |F_n(\lambda) - F_\pi(\lambda)| \xrightarrow{p} 0,$$

for every $\lambda \in [0, \infty]$. Next, we will establish that for any $\epsilon > 0$, it is possible to find a finite set of regularization parameters $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_k = \infty$ such that

$$\sup_{\pi \in \mathcal{Q}} \max_{1 \leq j \leq k} \{F_\pi(\lambda_j) - F_\pi(\lambda_{j-1})\} < \epsilon.$$

This assertion follows from the fact that $f_\pi(x)$ is uniformly bounded by $\phi(0)$. The rest of the proof proceeds as in the proof of Theorem 19.1 in van der Vaart (1998).

Let us finally turn to pre-testing. The objective function for pre-testing is equal to the one for lasso, plus additional terms for the jumps at $\pm\lambda$; the penalty term equals

$$2P_n(|X| > \lambda) + 2\lambda(\widehat{f}(-\lambda) + \widehat{f}(\lambda)).$$

Uniform convergence of the SURE criterion for pre-testing thus holds if (i) the conditions for lasso are satisfied, and (ii) we have a uniformly consistent estimator of $|x|\widehat{f}(x)$. □

Proof of Lemma 6: First, notice that the assumptions of the lemma plus convexity of the square function make $E_\pi[r_{n,k}(\lambda)]$ finite. Now, i.i.d.-ness of $(x_{1i}, \ldots, x_{ki}, \mu_i, \sigma_i)$ and mutual independence of $(x_1, \ldots, x_k)$ conditional on (μ, σ^2) imply,

$$\begin{aligned}
E_\pi[r_{n,k}(\lambda)] &= E_\pi\left[\left(m(X_{k-1}, \lambda) - x_k\right)^2\right] \\
&= E_\pi\left[\left(m(X_{k-1}, \lambda) - \mu\right)^2\right] + E_\pi\left[\left(x_k - \mu\right)^2\right] \\
&= \bar{R}_{\pi,k}(\lambda) + E_\pi[\sigma^2].
\end{aligned}$$

□

Proof of Theorem 4: We can decompose

$$\begin{aligned}
r_{n,k}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\Big[\left(m(X_{k-1,i}, \lambda) - \mu_i\right)^2 + (x_{ki} - \mu_i)^2 + 2\left(m(X_{k-1,i}, \lambda) - \mu_i\right)(x_{ki} - \mu_i)\Big] \\
&= L_{n,k}(\lambda) + \frac{1}{n}\sum_{i=1}^{n}(x_{ki} - \mu_i)^2 - \frac{2}{n}\sum_{i=1}^{n}\left(m(X_{k-1,i}, \lambda) - \mu_i\right)(x_{ki} - \mu_i). \qquad \text{(A.8)}
\end{aligned}$$

Theorem 3 and Lemma 2 imply that the first term on the last line of equation (A.8) converges uniformly in quadratic mean to $\bar{R}_{\pi,k}(\lambda)$. The second term does not depend on λ. Uniform convergence in quadratic mean of this term to $-\bar{v}_\pi = E_\pi[\sigma_i^2]$ follows immediately from the assumption that $\sup_{\pi \in \mathcal{Q}} E_\pi[x_k^4] < \infty$. To prove uniform convergence to zero in quadratic mean of the third term, notice that,

$$\begin{aligned}
E_\pi\left[\left(\frac{1}{n}\sum_{i=1}^{n}(m(X_{k-1,i}, \lambda) - \mu_i)(x_{ki} - \mu_i)\right)^2\right] &= \frac{1}{n^2}\sum_{i=1}^{n}E_\pi\left[(m(X_{k-1,i}, \lambda) - \mu_i)^2(x_{ki} - \mu_i)^2\right] \\
&\leq \frac{1}{n}\left(E_\pi\left[(m(X_{k-1}, \lambda) - \mu)^4\right]E_\pi\left[(x_k - \mu)^4\right]\right)^{1/2} \\
&\leq \frac{1}{n}\left(E_\pi\left[(X_{k-1} - \mu)^4 + \mu^4\right]E_\pi\left[(x_k - \mu)^4\right]\right)^{1/2}.
\end{aligned}$$

The condition $\sup_{\pi \in \mathcal{Q}} E[x_j^4] < \infty$ for $j = 1, \ldots k$ guarantees that the two expectations on the last line of the last equation are uniformly bounded in $\pi \in \mathcal{Q}$, which yields the first result of the theorem.

The second result follows from Theorem 2. □

References

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Abrams, D., M. Bertrand, and S. Mullainathan (2012). Do judges vary in their treatment of race? *Journal of Legal Studies 41*(2), 347–383.

Angrist, J., V. Chernozhukov, and I. Fernández-Val (2006). Quantile regression under misspecification, with an application to the U.S. wage structure. *Econometrica 74*(2), 539–563.

Arlot, S. and A. Celisse (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys 4*, 40–79.

Athey, S. and G. W. Imbens (2015). 2015 NBER Summer institute methods lectures. `http://www.nber.org/econometrics_minicourse_2015/`.

Belloni, A. and V. Chernozhukov (2011). High dimensional sparse econometric models: An introduction. In P. Alquier, E. Gautier, and G. Stoltz (Eds.), *Inverse Problems and High-Dimensional Estimation: : Stats in the Château Summer School, August 31 - September 4, 2009*, Volume 203 of *Lecture Notes in Statistics*, Chapter 3, pp. 121–156. Berlin: Springer.

Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics*, 855–903.

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

Casella, G. and J. T. G. Hwang (2012). Shrinkage confidence procedures. *Statistical Science 27*(1), 51–60.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014). Measuring the impacts of teachers II: Teacher value-added and student outcomes in adulthood. *American Economic Review 104*(9), 2633–2679.

Chetty, R. and N. Hendren (2015). The impacts of neighborhoods on intergenerational mobility: Childhood exposure effects and county-level estimates. Working Paper.

Chetverikov, D. and Z. Liao (2016). On cross-validated lasso. *arXiv preprint arXiv:1605.02214*.

Della Vigna, S. and E. La Ferrara (2010). Detecting illegal arms trade. *American Economic Journal: Economic Policy 2*(4), 26–57.

Donoho, D. L. and I. M. Johnstone (1995). Adapting to unkown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*(432), 1200–1224.

Donoho, D. L. and J. M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *biometrika 81*(3), 425–455.

Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction.* Institute of mathematical statistics monographs. Cambridge: Cambridge University Press.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association 106*(496), 1602–1614.

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

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

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2 ed.). Springer series in statistics Springer, Berlin.

James, W. and C. Stein (1961). Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pp. 361–379.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical Bayes estimation of normal means. *Annals of Statistics 37*(4), 1647–1684.

Johnstone, I. M. (2015). *Gaussian estimation: Sequence and wavelet models.*

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015). Prediction policy problems. *American Economic Review: Papers and Proceedings 105*(5), 491–495.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association 109*(506), 674–685.

Krueger, A. B. (1999). Experimental estimates of education production functions. *Quarterly Journal of Economics 114*(2), 497–532.

Leeb, H. and B. M. Pötscher (2005). Model selection and inference: Facts and fiction. *Econometric Theory 21*(1), 21–59.

Leeb, H. and B. M. Pötscher (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.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* Cambridge: The MIT Press.

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics 79*(1), 147–168.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 131–149. Berkeley: University of California Press.

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

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *Annals of Mathematical Statistics 35*, 1–20.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall/CRC.

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

Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science 5*(1), 147–155.

Stock, J. H. and M. W. Watson (2012). Generalized shrinkage methods for forecasting using many predictors. *Journal of Business & Economic Statistics 30*(4), 481–493.

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

van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.

Wasserman, L. (2006). *All of Nonparametric Statistics*. New York: Springer.

Xie, X., S. Kou, and L. D. Brown (2012). SURE estimates for a heteroscedastic hierarchical model. *Journal of the American Statistical Association 107*(500), 1465–1479.

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