JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 526, 697–712: Theory and Methods
https://doi.org/10.1080/01621459.2018.1429276

Taylor & Francis
Taylor & Francis Group

Excess Optimism: How Biased is the Apparent Error of an Estimator Tuned by SURE?

Ryan J. Tibshirani [a] and Saharon Rosset[b]

[a]Department of Statistics and Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA; [b]Department of Statistics, Tel Aviv University, Tel Aviv, Israel

ABSTRACT

Nearly all estimators in statistical prediction come with an associated tuning parameter, in one way or another. Common practice, given data, is to choose the tuning parameter value that minimizes a constructed estimate of the prediction error of the estimator; we focus on Stein's unbiased risk estimator, or SURE, which forms an unbiased estimate of the prediction error by augmenting the observed training error with an estimate of the degrees of freedom of the estimator. Parameter tuning via SURE minimization has been advocated by many authors, in a wide variety of problem settings, and in general, it is natural to ask: what is the prediction error of the SURE-tuned estimator? An obvious strategy would be simply use the apparent error estimate as reported by SURE, that is, the value of the SURE criterion at its minimum, to estimate the prediction error of the SURE-tuned estimator. But this is no longer unbiased; in fact, we would expect that the minimum of the SURE criterion is systematically biased downwards for the true prediction error. In this work, we define the excess optimism of the SURE-tuned estimator to be the amount of this downward bias in the SURE minimum. We argue that the following two properties motivate the study of excess optimism: (i) an unbiased estimate of excess optimism, added to the SURE criterion at its minimum, gives an unbiased estimate of the prediction error of the SURE-tuned estimator; (ii) excess optimism serves as an upper bound on the excess risk, that is, the difference between the risk of the SURE-tuned estimator and the oracle risk (where the oracle uses the best fixed tuning parameter choice). We study excess optimism in two common settings: shrinkage estimators and subset regression estimators. Our main results include a James–Stein-like property of the SURE-tuned shrinkage estimator, which is shown to dominate the MLE; and both upper and lower bounds on excess optimism for SURE-tuned subset regression. In the latter setting, when the collection of subsets is nested, our bounds are particularly tight, and reveal that in the case of no signal, the excess optimism is always in between 0 and 10 degrees of freedom, regardless of how many models are being selected from. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2017
Revised December 2017

KEYWORDS
Bootstrap; Model selection; Optimism; SURE

1. Introduction

Consider data $Y \in \mathbb{R}^n$, drawn from a generic model

$$Y \sim F, \quad \text{where } \mathbb{E}(Y) = \theta_0, \ \operatorname{cov}(Y) = \sigma^2 I. \tag{1}$$

The mean $\theta_0 \in \mathbb{R}^n$ is unknown, and the variance $\sigma^2 > 0$ is assumed to be known. Let $\hat{\theta} \in \mathbb{R}^n$ denote an estimator of the mean. Define the prediction error, also called test error or just error for short, of $\hat{\theta}$ by

$$\operatorname{Err}(\hat{\theta}) = \mathbb{E}\|Y^* - \hat{\theta}(Y)\|_2^2, \tag{2}$$

where $Y^* \sim F$ is independent of Y and the expectation is taken over all that is random (over both Y, Y^*). A remark about notation: we write $\hat{\theta}$ to denote an *estimator* (also called a rule, procedure, or algorithm), and $\hat{\theta}(Y)$ to denote an *estimate* (a particular realization given data Y). Hence, it is perfectly well-defined to write the error as $\operatorname{Err}(\hat{\theta})$; this is indeed a fixed (i.e., nonrandom) quantity, because $\hat{\theta}$ represents a rule, not a random variable. This will be helpful to keep in mind when our notation becomes a bit more complicated.

Estimating prediction error as in (2) is a classical problem in statistics. One convenient method that does not require the use of held-out data stems from the *optimism theorem*, which says that

$$\operatorname{Err}(\hat{\theta}) = \mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2 + 2\sigma^2 \operatorname{df}(\hat{\theta}), \tag{3}$$

where $\operatorname{df}(\hat{\theta})$, called the *degrees of freedom* of $\hat{\theta}$, is defined as

$$\operatorname{df}(\hat{\theta}) = \frac{1}{\sigma^2}\operatorname{tr}\big(\operatorname{cov}(\hat{\theta}(Y), Y)\big) = \frac{1}{\sigma^2}\sum_{i=1}^n \operatorname{cov}(\hat{\theta}_i(Y), Y_i). \tag{4}$$

Let us define the *optimism* of $\hat{\theta}$ as $\operatorname{Opt}(\hat{\theta}) = \mathbb{E}\|Y^* - \hat{\theta}(Y)\|_2^2 - \mathbb{E}\|Y - \hat{\theta}(Y)\|_2^2$, the difference in prediction and training errors. Then, we can rewrite (3) as

$$\operatorname{Opt}(\hat{\theta}) = 2\sigma^2 \operatorname{df}(\hat{\theta}), \tag{5}$$

which explains its name. A nice treatment of the optimism theorem can be found in Efron (2004), though the idea can be found much earlier, for example, Mallows (1973), Stein (1981), Efron

CONTACT Ryan J. Tibshirani ryantibs@cmu.edu Department of Statistics and Machine Learning Department, 229B Baker Hall, Carnegie Mellon University, Pittsburgh, PA 15213.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

(1986). In fact, Efron (2004) developed more general versions of the optimism theorem in (3), beyond the standard setup in (1), (2); we discuss extensions along these lines in the supplementary document.

The optimism theorem in (3) suggests an estimator for the error in (2), defined by

$$\widehat{\mathrm{Err}}(Y) = \|Y - \hat{\theta}(Y)\|_2^2 + 2\sigma^2 \widehat{\mathrm{df}}(Y), \tag{6}$$

where $\widehat{\mathrm{df}}$ is any unbiased estimator of the degrees of freedom of $\hat{\theta}$, as defined in (4), that is, it satisfies $\mathbb{E}[\widehat{\mathrm{df}}(Y)] = \mathrm{df}(\hat{\theta})$. Clearly, from (6) and (3), we see that

$$\mathbb{E}[\widehat{\mathrm{Err}}(Y)] = \mathrm{Err}(\hat{\theta}), \tag{7}$$

that is, $\widehat{\mathrm{Err}}$ is an unbiased estimator of the prediction error of $\hat{\theta}$. We will call the estimator $\widehat{\mathrm{Err}}$ in (6) *Stein's unbiased risk estimator*, or SURE, in honor of Stein (1981). This is somewhat of an abuse of notation, as $\widehat{\mathrm{Err}}$ is actually an estimate of prediction error, $\mathrm{Err}(\hat{\theta})$ in (2), and not risk,

$$\mathrm{Risk}(\hat{\theta}) = \mathbb{E}\|\theta_0 - \hat{\theta}(Y)\|_2^2. \tag{8}$$

However, the two are essentially equivalent notions, because $\mathrm{Err}(\hat{\theta}) = n\sigma^2 + \mathrm{Risk}(\hat{\theta})$. (As such, in what follows, we will occasionally focus on risk instead of prediction error, when it is convenient.)

We note that when $\hat{\theta}$ is a linear regression estimator (onto a fixed and full column rank design matrix), the degrees of freedom of $\hat{\theta}$ is simply p, the number of predictor variables in the regression, and SURE reduces to Mallows' C_p (Mallows 1973), or equivalently, AIC (Akaike 1973), since σ^2 is assumed to be known.

1.1. Stein's Formula

Stein (1981) studied a risk decomposition, as in (6), with the specific degrees of freedom estimator

$$\widehat{\mathrm{df}}(Y) = (\nabla \cdot \hat{\theta})(Y) = \sum_{i=1}^n \frac{\partial \hat{\theta}_i}{\partial Y_i}(Y), \tag{9}$$

called the divergence of the map $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$. Assuming a normal distribution $F = N(\theta_0, \sigma^2 I)$ for the data in (1) and regularity conditions on $\hat{\theta}$ (specifically, weak differentiability and an integrability condition on the components of the weak derivative), Stein showed that the divergence estimator in (9) is unbiased for $\mathrm{df}(\hat{\theta})$; to be explicit

$$\mathrm{df}(\hat{\theta}) = \mathbb{E}\left[\sum_{i=1}^n \frac{\partial \hat{\theta}_i}{\partial Y_i}(Y)\right]. \tag{10}$$

This elegant result has had a significant following in statistics (e.g., see the references below).

1.2. Parameter Tuning via SURE

Here and henceforth, we write $\hat{\theta}_s$ for the estimator of interest, where the subscript s highlights the dependence of this estimator on a tuning parameter, taking values in a set S. The term "tuning parameter" is used loosely, and we do not place any restrictions on S (e.g., this can be a continuous or a discrete collection of tuning parameter values). Abstractly, we can just think of $\{\hat{\theta}_s : s \in S\}$ as a family of estimators under consideration. We use $\widehat{\mathrm{Err}}_s$ to denote the prediction error estimator in (6) for $\hat{\theta}_s$, and $\widehat{\mathrm{df}}_s$ to denote an unbiased degrees of freedom estimator for $\hat{\theta}_s$.

One sensible strategy for choosing the tuning parameter s, associated with our estimator $\hat{\theta}_s$, is to select the value minimizing SURE in (6), denoted

$$\hat{s}(Y) = \operatorname*{argmin}_{s \in S} \widehat{\mathrm{Err}}_s(Y). \tag{11}$$

We can think of $\hat{s}$ as an estimator of some optimal tuning parameter value, namely, an estimator of

$$s_0 = \operatorname*{argmin}_{s \in S} \mathrm{Err}(\hat{\theta}_s), \tag{12}$$

the tuning parameter value that minimizes error. When $\hat{\theta}_s$ is the linear regression estimator onto a set of predictor variables indexed by the parameter s, the rule in (11) encompasses model selection via C_p minimization, which is a classical topic in statistics. In general, tuning parameter selection via SURE minimization has been widely advocated by authors across various problem settings, for example, Donoho and Johnstone (1995), Johnstone (1999), Zou, Hastie, and Tibshirani (2007), Zou and Yuan (2008), Tibshirani and Taylor (2011, 2012), Candes, Sing-Long, and Trzasko (2013), Ulfarsson and Solo (2013a, b), Chen, Lin, and Sen (2015), just to name a few.

1.3. What is the Error of the SURE-tuned Estimator?

Having decided to use $\hat{s}$ as a rule for choosing the tuning parameter, it is natural to ask: what is the error of the subsequent SURE-tuned estimator $\hat{\theta}_{\hat{s}}$? To be explicit, this estimator produces the estimate $\hat{\theta}_{\hat{s}(Y)}(Y)$ given data Y, where $\hat{s}(Y)$ is the tuning parameter value minimizing the SURE criterion, as in (11). Initially, it might seem reasonable to use the apparent error estimate given to us by SURE, that is, $\widehat{\mathrm{Err}}_{\hat{s}(Y)}(Y)$, to estimate the prediction error of $\hat{\theta}_{\hat{s}}$. To be explicit, this gives

$$\widehat{\mathrm{Err}}_{\hat{s}(Y)}(Y) = \|Y - \hat{\theta}_{\hat{s}(Y)}(Y)\|_2^2 + 2\sigma^2 \widehat{\mathrm{df}}_{\hat{s}(Y)}(Y)$$

at each given data realization Y. However, even though $\widehat{\mathrm{Err}}_s$ is unbiased for $\mathrm{Err}(\hat{\theta}_s)$ for each fixed $s \in S$, the estimator $\widehat{\mathrm{Err}}_{\hat{s}}$ is no longer generally unbiased for $\mathrm{Err}(\hat{\theta}_{\hat{s}})$, and commonly, it will be too optimistic, that is, we will commonly observe that

$$\mathbb{E}[\widehat{\mathrm{Err}}_{\hat{s}(Y)}(Y)] < \mathrm{Err}(\hat{\theta}_{\hat{s}}) = \mathbb{E}\|Y^* - \hat{\theta}_{\hat{s}(Y)}(Y)\|_2^2. \tag{13}$$

After all, for each data instance Y, the value $\hat{s}(Y)$ is specifically chosen to minimize $\widehat{\mathrm{Err}}_s(Y)$ over all $s \in S$, and thus we would expect $\widehat{\mathrm{Err}}_{\hat{s}}$ to be biased downward as an estimator of the error of $\hat{\theta}_{\hat{s}}$. Of course, the optimism of training error, as displayed in (3)–(5), is by now a central principle in statistics and (we believe) nearly all statisticians are aware of and account for this optimism in applied statistical modeling. But the optimism of the optimized SURE criterion itself, as suggested in (13), is more subtle and has received less attention.

1.4. Excess Optimism

In light of the above discussion, we define the *excess optimism* associated with $\hat{\theta}_{\hat{s}}$ by[1]

$$\mathrm{ExOpt}(\hat{\theta}_{\hat{s}}) = \mathrm{Err}(\hat{\theta}_{\hat{s}}) - \mathbb{E}[\widehat{\mathrm{Err}}_{\hat{s}(Y)}(Y)]. \tag{14}$$

We similarly define the *excess degrees of freedom* of $\hat{\theta}_{\hat{s}}$ by

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) = \mathrm{df}(\hat{\theta}_{\hat{s}}) - \mathbb{E}[\widehat{\mathrm{df}}_{\hat{s}(Y)}(Y)]. \tag{15}$$

The same motivation for excess optimism can be retold from the perspective of degrees of freedom: even though the degrees of freedom estimator $\widehat{\mathrm{df}}_s$ is unbiased for $\mathrm{df}(\hat{\theta}_s)$ for each $s \in S$, we should not expect $\widehat{\mathrm{df}}_{\hat{s}}$ to be unbiased for $\mathrm{df}(\hat{\theta}_{\hat{s}})$, and it will be commonly biased downward, that is, excess degrees of freedom in (15) will be commonly positive.

It should be noted that the two perspectives—excess optimism and excess degrees of freedom—are equivalent, as the optimism theorem in (3) (which holds for any estimator) applied to $\hat{\theta}_{\hat{s}}$ tells us that

$$\mathrm{Err}(\hat{\theta}_{\hat{s}}) = \mathbb{E}\|Y - \hat{\theta}_{\hat{s}(Y)}(Y)\|_2^2 + 2\sigma^2 \mathrm{df}(\hat{\theta}_{\hat{s}}).$$

Therefore, we have

$$\mathrm{ExOpt}(\hat{\theta}_{\hat{s}}) = 2\sigma^2 \mathrm{edf}(\hat{\theta}_{\hat{s}}),$$

analogous to the usual relationship between optimism and degrees of freedom.

It should also be noted that the focus on prediction error, rather than risk, is a decision based on ease of exposition, and that excess optimism can be equivalently expressed in terms of risk, that is,

$$\mathrm{ExOpt}(\hat{\theta}_{\hat{s}}) = \mathrm{Risk}(\hat{\theta}_{\hat{s}}) - \mathbb{E}[\widehat{\mathrm{Risk}}_{\hat{s}(Y)}(Y)], \tag{16}$$

where we define $\widehat{\mathrm{Risk}}_s = \widehat{\mathrm{Err}}_s - n\sigma^2$, an unbiased estimator of $\mathrm{Risk}(\hat{\theta}_s)$ in (8), for each $s \in S$.

Finally, a somewhat obvious but important point is the following: an unbiased estimator $\widehat{\mathrm{edf}}$ of excess degrees of freedom $\mathrm{edf}(\hat{\theta}_{\hat{s}})$ leads to an unbiased estimator of prediction error $\mathrm{Err}(\hat{\theta}_{\hat{s}})$, that is, $\widehat{\mathrm{Err}}_{\hat{s}} + 2\sigma^2\widehat{\mathrm{edf}}$, by construction of excess degrees of freedom in (15). Likewise, $\widehat{\mathrm{Risk}}_{\hat{s}} + 2\sigma^2\widehat{\mathrm{edf}}$ is an unbiased estimator of the risk $\mathrm{Risk}(\hat{\theta}_{\hat{s}})$.

1.5. Is Excess Optimism Always Nonnegative?

Intuitively, it seems that excess optimism should be always nonnegative, that is, for any "reasonable" class of estimators, the expectation of the SURE criterion at its minimum should be no larger than the actual error rate of the SURE-tuned estimator. However, we are not able to give a general proof of this claim. In each setting that we study in this work—shrinkage estimators, subset regression estimators, and soft-thresholding estimators—we prove that the excess degrees of freedom is nonnegative, albeit using different proof techniques. For "reasonable" classes of estimators, we have not seen evidence, either theoretical or empirical, that suggests excess degrees of freedom can be negative; but in the absence of a general result, of course, we cannot rule out the possibility that it is negative in some (likely pathological) situations.

[1] The excess optimism here is not only associated with $\hat{\theta}_{\hat{s}}$ itself, but also with the SURE family $\{\widehat{\mathrm{Err}}_s : s \in S\}$, used to define $\hat{s}$. This is meant to be implicit in our language and our notation.

1.6. Summary of Contributions

The goal of this work is to understand excess optimism, or equivalently, excess degrees of freedom, associated with estimators that are tuned by optimizing SURE. Below, we provide an outline of our results and contributions.

- In Section 2, we develop further motivation for the study of excess optimism, by showing that it upper bounds the excess risk, that is, the difference between the risk of the estimator in question and the oracle risk, in Theorem 2.1.
- In Section 3, we precisely characterize (and give an unbiased estimator for) the excess degrees of freedom of the SURE-tuned shrinkage estimator, both in a classical normal means problem setting and in a regression setting, in (24) and (32), respectively. This shows that the excess degrees of freedom in both of these settings is always nonnegative, and at most 2. Our analysis also reveals an interesting connection between SURE-tuned shrinkage estimation and James–Stein estimation.
- In Sections 4 and 5.4, we derive bounds on the excess degrees of freedom of the SURE-tuned subset regression estimator (or equivalently, the C_p-tuned subset regression estimator), using different approaches. Theorem 4.1 shows from first principles that, under reasonable conditions on the subset regression models being considered, the excess degrees of freedom of SURE-tuned subset regression is small compared to the oracle risk. Theorems 5.3 and 5.4 are derived using a more refined general result, from Mikkelsen and Hansen (2016), and present exact (though not always explicitly computable) expressions for excess degrees of freedom. Some implications for the excess degrees of freedom of the SURE-tuned subset regression estimator: we see that it is always nonnegative, and it is surprisingly small for nested subsets, for example, it is at most 10 for any nested collection of subsets (no matter the number of predictors) when $\theta_0 = 0$.
- In Section 5, we examine strategies for characterizing the excess degrees of freedom of generic estimators using Stein's formula, and extensions of Stein's formula for discontinuous mappings from Tibshirani (2015), Mikkelsen and Hansen (2016). We use the extension from Tibshirani (2015) in Section 5.3 to prove that excess degrees of freedom in SURE-tuned soft-thresholding is always nonnegative. We use that from Mikkelsen and Hansen (2016) in Section 5.4 to prove results on subset regression, already described.
- In Section 6, we study a simple bootstrap procedure for estimating excess degrees of freedom, which appears to work reasonably well in practice.
- In Section 7, we wrap up with a short discussion, and describe the implications of some of our technical results on the degrees of freedom of the best subset selection estimator. Extensions of our work, to heteroscedastic data,

and alternative loss functions (other than squared loss), are described in the supplement.

1.7. Related Work

There is a lot of work related to the topic of this article. In addition to the classical contributions of Mallows (1973), Stein (1981), Efron (1986, 2004), on optimism and degrees of freedom, that have already been discussed, it is worth mentioning Breiman (1992). In Section 2 of this work, the author warns precisely of the downward bias of SURE for estimating prediction error in regression models, when the former is evaluated at the model that minimizes SURE (or here, C_p). Breiman was thus keenly aware of excess optimism; he roughly calculated, for all subsets regression with p orthogonal variables, that the SURE-tuned subset regression estimator has an approximate excess optimism of $0.84p\sigma^2$, in the null case when $\theta_0 = 0$.

Several authors have addressed the problem of characterizing the risk of an estimator tuned by SURE (or a similar method) by uniformly controlling the deviations of SURE from its mean over all tuning parameter values $s \in S$, that is, by establishing that a quantity like $\sup_{s\in S} |\widehat{\mathrm{Risk}}_s(Y) - \mathrm{Risk}(\hat{\theta}_s)|$, in our notation, converges to zero in a suitable sense. Examples of this uniform control strategy are found in Li (1985, 1986, 1987), Kneip (1994), who study linear smoothers; Donoho and Johnstone (1995), who study wavelet smoothing; Cavalier et al. (2002), who study linear inverse problems in sequence space; and Xie, Kou, and Brown (2012), who study a family of shrinkage estimators in a heteroscedastic model. Notice that the idea of uniformly controlling the deviations of SURE away from its mean is quite different in spirit than our approach, in which we directly seek to understand the gap between $\mathbb{E}[\widehat{\mathrm{Risk}}_{\hat{s}(Y)}(Y)]$ and $\mathrm{Risk}(\hat{\theta}_{\hat{s}})$. It is not clear to us that uniform control of SURE deviations can be used in general to understand this gap precisely, that is, to understand excess optimism precisely.

Importantly, the strategy of uniform control can often be used to derive so-called oracle inequalities of the form

$$\mathrm{Risk}(\hat{\theta}_{\hat{s}}) \le (1 + o(1))\mathrm{Risk}(\hat{\theta}_{s_0}). \tag{17}$$

Such oracle inequalities were derived in Li (1985, 1986, 1987), Kneip (1994), Donoho and Johnstone (1995), Cavalier et al. (2002), Xie, Kou, and Brown (2012). In Section 2, we will return to the oracle inequality (17), and will show that (17) can be established in some cases via a bound on excess optimism.

When the data are normally distributed, that is, when $F = N(\theta_0, \sigma^2 I)$ in (1), one might think to use Stein's formula on the SURE-tuned estimator $\hat{\theta}_{\hat{s}}$ itself, to compute its proper degrees of freedom, and hence excess optimism. This idea is pursued in Section 5, where we also show that implicit differentiation can be applied to characterize the excess degrees of freedom, under some assumptions. These assumptions, however, are very strong. Stein's original work (Stein 1981) established the result in (10), when the estimator $\hat{\theta}$ is weakly differentiable, as a function of Y. But, even when $\hat{\theta}_s$ is itself continuous in Y for each $s \in S$, it is possible for the SURE-tuned estimator $\hat{\theta}_{\hat{s}}$ to be discontinuous in Y, and when the discontinuities become severe enough, weak differentiability fails and Stein's formula does not apply. Tibshirani (2015) and Mikkelsen and Hansen (2016) derive extensions of Stein's formula to deal with estimators having (specific types of) discontinuities. We leverage these extensions in Section 5.

A parallel problem is to study the excess optimism associated with parameter tuning by cross-validation, considered in Varma and Simon (2006), Tibshirani and Tibshirani (2009), Bernau, Augustin, and Boulesteix (2013), Krstajic et al. (2014), Tsamardinos, Rakhshani, and Lagani (2015). Since it is difficult to study cross-validation mathematically, these works do not develop formal characterizations or corrections and are mostly empirically driven.

Finally, it is worth mentioning that some of the motivation of Efron (2014) is similar to that in our article, though the focus is different: Efron focused on constructing proper estimates of standard error (and confidence intervals) for estimators that are defined with inherent parameter tuning (he used the term "model selection" rather than parameter tuning). Discontinuities play a major role in Efron (2014), as they do in ours (i.e., in our Section 5); Efron proposed to replace parameter-tuned estimators with bagged (bootstrap aggregated) versions, as the latter estimators are smoother and can lead to smaller standard errors (or shorter confidence intervals). More generally, post-selection inference, as studied in Berk et al. (2013), Lockhart et al. (2014), Lee et al. (2016), Tibshirani et al. (2016), Fithian, Sun, and Taylor (2014) and several other papers, is also related in spirit to our work, though our focus is on prediction error rather than inference. While post-selection prediction can also be studied from the conditional perspective that is often used in post-selection inference, this seems to be less common. A notable exception is Tian Harris (2016), who proposes a clever randomization scheme for estimating prediction error conditional on a model selection event, in regression.

2. An Upper Bound on the Oracle Gap

We derive a simple inequality that relates the error of the estimator $\hat{\theta}_{\hat{s}}$ to the error of what we may call the *oracle* estimator $\hat{\theta}_{s_0}$, where s_0 is the tuning parameter value that minimizes the (unavailable) true prediction error, as in (12). Observe that

$$\begin{aligned}\mathbb{E}[\widehat{\mathrm{Err}}_{\hat{s}(Y)}(Y)] &= \mathbb{E}\Big(\min_{s\in S} \widehat{\mathrm{Err}}_s(Y)\Big) \le \min_{s\in S} \mathbb{E}[\widehat{\mathrm{Err}}_s(Y)] \\ &= \min_{s\in S} \mathrm{Err}(\hat{\theta}_s) = \mathrm{Err}(\hat{\theta}_{s_0}).\end{aligned} \tag{18}$$

By adding $\mathrm{Err}(\hat{\theta}_{\hat{s}})$ to the left- and right-most expressions, and then rearranging, we have established the following result.

Theorem 2.1. For any family of estimators $\{\hat{\theta}_s : s \in S\}$, it holds that

$$\mathrm{Err}(\hat{\theta}_{\hat{s}}) \le \mathrm{Err}(\hat{\theta}_{s_0}) + \mathrm{ExOpt}(\hat{\theta}_{\hat{s}}). \tag{19}$$

Here, $\hat{s}$ is the tuning parameter rule defined by minimizing SURE, as in (11), s_0 is the oracle tuning parameter value minimizing prediction error, as in (12), and $\mathrm{ExOpt}(\hat{\theta}_{\hat{s}})$ is the excess optimism, as defined in (14).

Theorem 2.1 says that the excess optimism, which is a quantity that we can in principle calculate (or at least, estimate), serves as an upper bound for the gap between the prediction error of $\hat{\theta}_{\hat{s}}$ and the oracle error. This gives an interesting,

alternative motivation for excess optimism to that given in the introduction: excess optimism tells us how far the SURE-tuned estimator $\hat{\theta}_{\hat{s}}$ can be from the best member of the class $\{\hat{\theta}_s : s \in S\}$, in terms of prediction error. A few remarks are in order.

Remark 2.1 (Risk inequality). Recalling that excess optimism can be equivalently posed in terms of risk, as in (16), the bound in (19) can also be written in terms of risk, namely,

$$\text{Risk}(\hat{\theta}_{\hat{s}}) \le \text{Risk}(\hat{\theta}_{s_0}) + \text{ExOpt}(\hat{\theta}_{\hat{s}}), \tag{20}$$

which says the excess risk $\text{Risk}(\hat{\theta}_{\hat{s}}) - \text{Risk}(\hat{\theta}_{s_0})$ of the SURE-tuned estimator is upper bounded by its excess optimism, $\text{ExOpt}(\hat{\theta}_{\hat{s}})$. If we can show that this excess optimism is small compared to the oracle risk, in particular, if we can show that $\text{ExOpt}(\hat{\theta}_{\hat{s}}) = o(\text{Risk}(\hat{\theta}_{s_0}))$, then (20) implies the oracle inequality (17). We will revisit this idea in Sections 3 and 4.

Remark 2.2 (Beating the oracle). If $\text{ExOpt}(\hat{\theta}_{\hat{s}}) < 0$, then (19) implies $\hat{\theta}_{\hat{s}}$ outperforms the oracle, in terms of prediction error (or risk). Technically this is not impossible, as θ_{s_0} is the optimal fixed-parameter estimator, in the class $\{\theta_s : s \in S\}$, whereas $\hat{\theta}_{\hat{s}}$ is tuned in a data-dependent fashion. But it seems unlikely to us that excess optimism can be negative, recall Section 1.5.

Remark 2.3 (Beyond SURE). The argument in (18) and thus the validity of Theorem 2.1 only used the fact that $\hat{s}$ was defined by minimizing an unbiased estimator of prediction error, and SURE is not the only such estimator. For example, the result in Theorem 2.1 applies to the standard hold-out estimator of prediction error, when hold-out data $Y^* \sim F$ (independent of Y) is available. While the result does not exactly carry over to cross-validation (since the standard cross-validation estimator of prediction error is not unbiased in finite samples, at least not without additional corrections and assumptions), we can think of it as being true in some approximate sense.

3. Shrinkage Estimators

In this section, we focus on shrinkage estimators, and consider normal data, $Y \sim F = N(\theta_0, \sigma^2 I)$ in (1). Due to the simple form of the family of shrinkage estimators (and the normality assumption), we can compute an (exact) unbiased estimator of excess degrees of freedom, and excess optimism.

3.1. Shrinkage in Normal Means

First, we consider the simple family of shrinkage estimators

$$\hat{\theta}_s(Y) = \frac{Y}{1+s}, \quad \text{for } s \ge 0. \tag{21}$$

In this case, we can see that $\text{df}(\hat{\theta}_s) = n/(1+s)$ for each $s \ge 0$, and SURE in (6) is

$$\widehat{\text{Err}}_s(Y) = \|Y\|_2^2 \frac{s^2}{(1+s)^2} + 2\sigma^2 \frac{n}{1+s}. \tag{22}$$

The next lemma characterizes $\hat{s}$, the mapping defined by the minimizer of the above criterion. The proof is elementary; as with all proofs in this article, is given in the supplement.

Lemma 3.1. Define $g(x) = ax^2/(1+x)^2 + 2b/(1+x)$, where $a, b > 0$. Then, the minimizer of g over $x \ge 0$ is

$$x^* = \begin{cases} \frac{b}{a-b} & \text{if } a > b \\ \infty & \text{if } a \le b. \end{cases}$$

According to Lemma 3.1, the rule $\hat{s}$ defined by minimizing (22) is

$$\hat{s}(Y) = \begin{cases} \dfrac{n\sigma^2}{\|Y\|_2^2 - n\sigma^2} & \text{if } \|Y\|_2^2 > n\sigma^2 \\ \infty & \text{if } \|Y\|_2^2 \le n\sigma^2. \end{cases}$$

Plugging this in gives the SURE-tuned shrinkage estimate $\hat{\theta}_{\hat{s}(Y)}(Y) = Y/(1+\hat{s}(Y))$. Note that this is weakly differentiable as a function of Y, and so by Stein's formula (10), we can form an unbiased estimator of its degrees of freedom by computing its divergence. When $\hat{s}(Y) < \infty$, the divergence is

$$\begin{aligned} &\frac{n}{1+\hat{s}(Y)} - \sum_{i=1}^n \frac{Y_i}{(1+\hat{s}(Y))^2} \frac{\partial \hat{s}}{\partial Y_i}(Y) \\ &= \frac{n}{1+\hat{s}(Y)} + \sum_{i=1}^n \frac{Y_i}{(1+\hat{s}(Y))^2} \frac{n\sigma^2}{(\|Y\|_2^2 - n\sigma^2)^2} 2Y_i \\ &= \frac{n}{1+\hat{s}(Y)} + \frac{2\hat{s}(Y)}{1+\hat{s}(Y)}. \end{aligned} \tag{23}$$

When $\hat{s}(Y) = \infty$, the divergence is 0.

Hence, we can see directly that for the SURE-tuned shrinkage estimator $\hat{\theta}_{\hat{s}}$, we have the excess degrees of freedom bound

$$\text{edf}(\hat{\theta}_{\hat{s}}) = \mathbb{E}\left(\frac{2\hat{s}(Y)}{1+\hat{s}(Y)} ;\ \hat{s}(Y) < \infty \right) \le 2, \tag{24}$$

and so $\text{ExOpt}(\hat{\theta}_{\hat{s}}) \le 4\sigma^2$. A lot is known about shrinkage estimators in the current normal means problem that we are considering, dating back to the seminal work of James and Stein (1961); some excellent recent references are Chapter 1 of Efron (2010), and Chapter 2 of Johnstone (2015). It is easy to show that the oracle choice of tuning parameter in the current setting is $s_0 = n\sigma^2/\|\theta_0\|_2^2$, and so

$$\text{Risk}(\hat{\theta}_{s_0}) = \frac{n\sigma^2 \|\theta_0\|_2^2}{n\sigma^2 + \|\theta_0\|_2^2}. \tag{25}$$

By our excess optimism bound of $4\sigma^2$, and Theorem 2.1 (actually, (20), the risk version of the result in the theorem), the risk of the SURE-tuned shrinkage estimator $\hat{\theta}_{\hat{s}}$ satisfies

$$\text{Risk}(\hat{\theta}_{\hat{s}}) \le \frac{n\sigma^2 \|\theta_0\|_2^2}{n\sigma^2 + \|\theta_0\|_2^2} + 4\sigma^2. \tag{26}$$

Remark 3.1 (Oracle inequality for SURE-tuned shrinkage). For large $\|\theta_0\|_2^2$, the risk gap of $4\sigma^2$ for the SURE-tuned shrinkage estimator is negligible next to the oracle risk in (25). Specifically, if $\|\theta_0\|_2^2 \to \infty$ as $n \to \infty$ (with σ^2 held constant), then we see that (26) implies the oracle inequality (17) for the SURE-tuned shrinkage estimator.

3.2. Interlude: James-Stein Estimation

The SURE-tuned shrinkage estimator of the last subsection can be written as

$$\hat{\theta}_{\hat{s}(Y)}(Y) = \begin{cases} \dfrac{1}{1+\frac{n\sigma^2}{\|Y\|_2^2-n\sigma^2}} Y & \text{if } \|Y\|_2^2 > n\sigma^2 \\ 0 & \text{if } \|Y\|_2^2 \le n\sigma^2, \end{cases}$$

or more concisely, as

$$\hat{\theta}_{\hat{s}(Y)}(Y) = \left(1 - \frac{n\sigma^2}{\|Y\|_2^2}\right)_+ Y, \tag{27}$$

where we write $x_+ = \max\{x, 0\}$ for the positive part of x. Meanwhile, the positive part James–Stein estimator (James and Stein 1961; Baranchik 1964) is defined as

$$\hat{\theta}^{\text{JS+}}(Y) = \left(1 - \frac{(n-2)\sigma^2}{\|Y\|_2^2}\right)_+ Y, \tag{28}$$

so the two estimators (27) and (28) only differ by the appearance of n versus $n-2$ in the shrinkage factor. This connection—between SURE-tuned shrinkage estimation and positive part James–Stein estimation—seems to be not very well-known, and was a surprise to us; after writing an initial draft of this article, we found that this fact was mentioned in passing in Xie, Kou, and Brown (2012). We now give a few remarks.

Remark 3.2 (Dominating the MLE). It can be shown that the SURE-tuned shrinkage estimator in (27) dominates the MLE, that is, $\hat{\theta}^{\text{MLE}}(Y) = Y$, just like the positive part James–Stein estimator in (28). For this to be true of the former estimator, we require $n \ge 5$, while the latter estimator only requires $n \ge 3$.

Our proof of $\hat{\theta}_{\hat{s}}$ dominating $\hat{\theta}^{\text{MLE}}$ mimicks Stein's elegant proof for the James-Stein estimator, (Stein 1981). Consider SURE for $\hat{\theta}_{\hat{s}}$, which gives an unbiased estimator of the risk of $\hat{\theta}_{\hat{s}}$, provided we compute its divergence properly, as in (23). Write $\hat{R}$ for this unbiased risk estimator. If $\hat{s}(Y) < \infty$, that is, $\|Y\|_2^2 > n\sigma^2$, then

$$\begin{aligned} \hat{R}(Y) &= -n\sigma^2 + \frac{\hat{s}(Y)^2}{(1+\hat{s}(Y))^2}\|Y\|_2^2 \\ &\quad + 2\sigma^2\left(\frac{n}{1+\hat{s}(Y)} + \frac{2\hat{s}(Y)}{1+\hat{s}(Y)}\right) \\ &= -n\sigma^2 + \frac{(n\sigma^2)^2}{\|Y\|_2^2} + 2n\sigma^2\frac{\|Y\|_2^2 - n\sigma^2}{\|Y\|_2^2} + 4\sigma^2\frac{n\sigma^2}{\|Y\|_2^2} \\ &= n\sigma^2 - (n-4)\sigma^2\frac{n\sigma^2}{\|Y\|_2^2} < n\sigma^2. \end{aligned}$$

If $\hat{s}(Y) = \infty$, that is, $\|Y\|_2^2 \le n\sigma^2$, then we have $\hat{R}(Y) = -n\sigma^2 + \|Y\|_2^2 \le 0$. Taking an expectation, we thus see that $\text{Err}(\hat{\theta}_{\hat{s}}) = \mathbb{E}[\hat{R}(Y)] < n\sigma^2$, which establishes the result, as $n\sigma^2$ is the risk of the MLE.

Remark 3.3 (Risk of positive part James–Stein). A straightforward calculation, similar to that given above for $\hat{\theta}_{\hat{s}}$ (see also Theorem 5.3 of Donoho and Johnstone (1995)) shows that the risk of the positive part James–Stein estimator satisfies

$$\text{Risk}(\hat{\theta}^{\text{JS+}}) \le \frac{n\sigma^2\|\theta_0\|_2^2}{n\sigma^2 + \|\theta_0\|_2^2} + 2\sigma^2, \tag{29}$$

so it admits an even tighter gap to the oracle risk than does the SURE-tuned shrinkage estimator, recalling (26).

As for the risk of the positive part James–Stein estimator $\hat{\theta}^{\text{JS+}}$ versus that of the SURE-tuned shrinkage estimator $\hat{\theta}_{\hat{s}}$, neither one is always better than the other. When $\|\theta_0\|_2^2$ is small, the latter fares better since it shrinks more; when $\|\theta_0\|_2^2$ is small, the opposite is true. This can be confirmed via calculations with Stein's unbiased risk estimator (to bound the risks of $\hat{\theta}^{\text{JS+}}$, $\hat{\theta}_{\hat{s}}$, similar to the arguments in the previous remark).

3.3. Shrinkage in Regression

Now, we consider the family of regression shrinkage estimators

$$\hat{\theta}_s(Y) = \frac{P_X Y}{1+s}, \quad \text{for } s \ge 0, \tag{30}$$

where we write $P_X \in \mathbb{R}^{n\times n}$ for the projection matrix onto the column space of a predictor matrix $X \in \mathbb{R}^{n\times p}$, that is, $P_X = X(X^TX)^{-1}X^T$ if X has full column rank, and $P_X = X(X^TX)^+X^T$ otherwise (here and throughout, A^+ denotes the pseudoinverse of a matrix A).

Treating X as fixed (nonrandom), it is easy to check that SURE (6) for our regression shrinkage estimator is

$$\widehat{\text{Err}}_s(Y) = \|P_X Y\|_2^2\frac{s^2}{(1+s)^2} + 2\sigma^2\frac{r}{1+s}, \tag{31}$$

where $r = \text{rank}(X)$, the rank of X. This is directly analogous to (22) in the normal means setting, and Lemma 3.1 shows that the minimizer $\hat{s}$ of (31) is defined by

$$\hat{s}(Y) = \begin{cases} \dfrac{r\sigma^2}{\|P_X Y\|_2^2 - r\sigma^2} & \text{if } \|P_X Y\|_2^2 \ge r\sigma^2 \\ \infty & \text{if } \|P_X Y\|_2^2 < r\sigma^2. \end{cases}$$

The same arguments as in Section 3.1 then lead to the same excess degrees of freedom bound

$$\text{edf}(\hat{\theta}_{\hat{s}}) = \mathbb{E}\left(\frac{2\hat{s}(Y)}{1+\hat{s}(Y)};\ \hat{s}(Y) < \infty\right) \le 2, \tag{32}$$

thus $\text{ExOpt}(\hat{\theta}_{\hat{s}}) \le 4\sigma^2$. By direct calculation, the oracle tuning parameter is $s_0 = r\sigma^2/\|P_X\theta_0\|_2^2$, and now

$$\text{Risk}(\hat{\theta}_{s_0}) = \frac{r\sigma^2\|\theta_0\|_2^2 + \|P_X\theta_0\|_2^2(\|\theta_0\|_2^2 - \|P_X\theta_0\|_2^2)}{r\sigma^2 + \|P_X\theta_0\|_2^2}. \tag{33}$$

Combining our excess optimism bound of $4\sigma^2$ with Theorem 2.1 (i.e., combining it with (20)), we have

$$\text{Risk}(\hat{\theta}_{\hat{s}}) \le \frac{r\sigma^2\|\theta_0\|_2^2 + \|P_X\theta_0\|_2^2(\|\theta_0\|_2^2 - \|P_X\theta_0\|_2^2)}{r\sigma^2 + \|P_X\theta_0\|_2^2} + 4\sigma^2. \tag{34}$$

Remark 3.4 (Oracle inequality for SURE-tuned regression shrinkage). The risk gap of $4\sigma^2$, for the SURE-tuned regression shrinkage estimator, will be negligible next to the oracle risk (33) under various sufficient conditions. For example, if $\|\theta_0\|_2^2 \to \infty$ and $\|P_X\theta_0\|_2^2|\|\theta_0\|_2^2 - \|P_X\theta_0\|_2^2| = O(r)$ as $n, r \to \infty$ (and σ^2 is held constant), then it is not hard to check that (34) implies the oracle inequality (17) for the SURE-tuned regression shrinkage estimator.

3.4. Interlude: James–Stein and Ridge Regression

The SURE-tuned regression shrinkage estimator of the previous subsection can be expressed as

$$\hat{\theta}_{\hat{s}(Y)}(Y) = \left(1 - \frac{r\sigma^2}{\|P_X Y\|_2^2}\right)_+ P_X Y, \tag{35}$$

which resembles the positive part James–Stein regression estimator

$$\hat{\theta}^{\mathrm{JS+}}(Y) = \left(1 - \frac{(r-2)\sigma^2}{\|P_X Y\|_2^2}\right)_+ P_X Y. \tag{36}$$

As before, the SURE-tuned regression shrinkage estimator (35) dominates the MLE (i.e., the least squares regression estimator), $\hat{\theta}^{\mathrm{MLE}}(Y) = P_X Y$. The positive-part James–Stein estimator (36) also dominates the MLE, and neither the SURE-tuned regression shrinkage estimator nor the positive-part James–Stein regression estimator dominates the other.

We point out a connection to penalized regression. For any fixed tuning parameter value $s \geq 0$, we can express the estimate in (30) as $\hat{\theta}_s(Y) = X\hat{\beta}_s(Y)$, where $\hat{\beta}_s(Y)$ solves the convex (though not necessarily strictly convex) penalized regression problem,

$$\hat{\beta}_s(Y) \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + s\|X\beta\|_2^2. \tag{37}$$

Hence, an alternative interpretation for the estimator $\hat{\theta}_{\hat{s}}$ in (35) (whose close cousin is the positive part James–Stein regression estimator $\hat{\theta}^{\mathrm{JS+}}$ in (36)) is that we are using SURE to select the tuning parameter over the family of penalized regression estimators in (37), for $s \geq 0$. This has the precise risk guarantee in (34) (and $\hat{\theta}^{\mathrm{JS+}}$ enjoys an even stronger guarantee, with $2\sigma^2$ in place of $4\sigma^2$).

Compared to (37), a more familiar penalized regression problem to most statisticians is perhaps the ridge regression problem (Hoerl and Kennard 1970),

$$\hat{\beta}_s^{\mathrm{ridge}}(Y) = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2 + s\|\beta\|_2^2. \tag{38}$$

Several differences between (37) and (38) can be enumerated; one interesting difference is that the solution in the former problem shrinks uniformly across all dimensions $1, \ldots, p$, whereas that in the latter problem shrinks less in directions of high variance and more in directions of low variance, defined with respect to the predictor variables (i.e., shrinks less in the top eigendirections of $X^T X$).

It is generally accepted that neither regression shrinkage estimator, in (37) and (38), is better than the other.[2] But, we have seen that SURE-tuning in the first problem (37) provides us with an estimator $\hat{\theta}_{\hat{s}} = X\hat{\beta}_{\hat{s}}$ that has a definitive risk guarantee (34) and provably dominates the MLE. The story for ridge regression is less clear; to quote Efron and Hastie (2016), Chapter 7.3: "*There is no [analogous] guarantee for ridge regression, and no foolproof way to choose the ridge parameter.*" Of course, if we could bound the excess degrees of freedom for SURE-tuned ridge regression, then this could lead (depending on the size of the bound) to a useful risk guarantee, providing some rigorous backing to SURE tuning for ridge regression. However, characterizing excess degrees of freedom for ridge regression is far from straightforward, for two reasons: (i) the SURE-optimal tuning parameter map $\hat{s}$ is not available in a closed form for ridge regression, and (ii) the SURE-tuned ridge estimator $\hat{\theta}_{\hat{s}}^{\mathrm{ridge}}$ is not necessarily continuous with respect to the data Y, thus (supposing the discontinuities are severe enough to violate weak differentiability) Stein's formula cannot be used to compute an unbiased estimator of its degrees of freedom. (Specifically, it is unclear whether the SURE-optimal ridge parameter map $\hat{s}$ is itself continuous with respect to Y, as it is defined by the minimizer of a possibly multimodal SURE criterion; see Figure 1.)

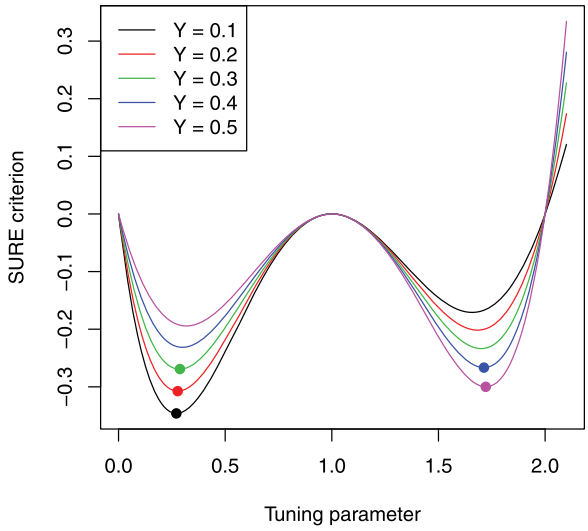

Figure 1. An illustration of a discontinuous mapping $\hat{s}$. Each curve represents the SURE criterion $G(Y, \cdot)$, as a function of the tuning parameter s, at nearby values of the (one-dimensional) data realization Y. As Y varies, $G(Y, \cdot)$ changes smoothly, but its minimizer $\hat{s}(Y)$ jumps discontinuously, from about 0.75 at $Y = 0.3$ (green curve) to 1.75 at $Y = 0.4$ (blue curve).

The second reason above, that is (possibly severe) discontinuities in $\hat{\theta}_{\hat{s}}^{\mathrm{ridge}}$, is what truly complicates the analysis. Even when $\hat{s}$ cannot be expressed in closed form, implicit differentiation can be used to compute the divergence of $\hat{\theta}_{\hat{s}}^{\mathrm{ridge}}$, as we explain in Section 5.1; but this divergence will not generally be enough to characterize the degrees of freedom (and thus excess degrees of freedom) of $\hat{\theta}_{\hat{s}}^{\mathrm{ridge}}$ in the presence of discontinuities. Extensions of Stein's divergence formula from Tibshirani (2015) and Mikkelsen and Hansen (2016) can be used to characterize degrees of freedom for estimators having certain types of discontinuities, which we review in Section 5.2. Generally speaking, these extensions involve sophisticated calculations. In the supplement, we revisit the ridge regression problem, and compute the divergence of the SURE-tuned ridge estimator via implicit differentiation, but we leave proper treatment of its discontinuities to future work.

4. Subset Regression Estimators

Here, we study subset regression estimators, and again consider normal data, $Y \sim F = N(\theta_0, \sigma^2 I)$ in (1). Our family of estimators is defined by regression onto subsets of the columns of a

[2] It is worth pointing out that the former problem (37) does not give a well-defined, that is, unique solution for the coefficients when $\mathrm{rank}(X) < p$, and the latter problem (38) does, when $s > 0$.

predictor matrix $X \in \mathbb{R}^{n\times p}$, that is,

$$\hat{\theta}_s(Y) = P_{X_s} Y \quad \text{for } s \in S, \tag{39}$$

where each $s = \{j_1, \ldots, j_{p_s}\}$ is an arbitrary subset of $\{1, \ldots, p\}$ of size p_s, $X_s \in \mathbb{R}^{n\times p_s}$ denotes the columns of X indexed by elements of s, P_{X_s} denotes the projection matrix onto the column space of X_s, and S denotes a collection of subsets of $\{1, \ldots, p\}$. We will abbreviate $P_s = P_{X_s}$, and we will assume, without any real loss of generality, that for each $s \in S$, the matrix X_s has full column rank (otherwise, simply replace each instance of p_s below with $r_s = \text{rank}(X_s)$).

SURE in (6) is now the familiar C_p criterion

$$\widehat{\text{Err}}_s(Y) = \|Y - P_s Y\|_2^2 + 2\sigma^2 p_s. \tag{40}$$

As S is discrete, it is not generally possible to express the minimizer $\hat{s}(Y)$ of the above criterion in closed form, and so, unlike the previous section, not generally possible to analytically characterize the excess degrees of freedom of the SURE-tuned subset regression estimator $\hat{\theta}_{\hat{s}}$. In what follows, we derive an upper bound on the excess degrees of freedom, using elementary arguments (note that our approach is roughly in line with the general strategy of uniform deviations control, see the bound used in (42)). Later in Section 5.4, we give a lower bound and a more sophisticated upper bound, by leveraging a powerful tool from Mikkelsen and Hansen (2016).

4.1. Upper Bounds for Excess Degrees of Freedom in Subset Regression

Note that we can write the excess degrees of freedom as

$$\begin{aligned}\text{edf}(\hat{\theta}_{\hat{s}}) &= \frac{1}{\sigma^2}\mathbb{E}\big[(P_{\hat{s}(Y)}(Y))^T(Y-\theta_0)\big] - \mathbb{E}(p_{\hat{s}(Y)}) \\ &= \frac{1}{\sigma^2}\mathbb{E}\|P_{\hat{s}(Y)}Z\|_2^2 - \mathbb{E}(p_{\hat{s}(Y)}),\end{aligned} \tag{41}$$

where $Z = Y - \theta_0 \sim N(0, \sigma^2 I)$. Furthermore, defining $W_s = \|P_s Z\|_2^2/\sigma^2 \sim \chi^2_{p_s}$ for $s \in S$, we have

$$\text{edf}(\hat{\theta}_{\hat{s}}) = \mathbb{E}(W_{\hat{s}(Y)} - p_{\hat{s}(Y)}) \le \mathbb{E}\Big[\max_{s\in S}(W_s - p_s)\Big]. \tag{42}$$

The next lemma provides a useful upper bound for the right-hand side above.

Lemma 4.1. Let $W_s \sim \chi^2_{p_s}$, $s \in S$. This collection need not be independent. Then for any $0 \le \delta < 1$,

$$\mathbb{E}\Big[\max_{s\in S}(W_s - p_s)\Big] \le \frac{2}{1-\delta}\log\sum_{s\in S}(\delta e^{1-\delta})^{-p_s/2}. \tag{43}$$

The proof of the above lemma relies only on the moment generating function of the chi-squared distribution, and so our assumption of normality for the data Y could be weakened. For example, a similar result to that in Lemma 4.1 can be derived when W_s, $s \in S$ each have subexponential tails (generalizing the chi-squared assumption). For simplicity, we do not pursue this.

Combining (42) and (43) gives an upper bound on the excess degrees of freedom of $\hat{\theta}_{\hat{s}}$,

$$\text{edf}(\hat{\theta}_{\hat{s}}) \le \frac{2}{1-\delta}\log\sum_{s\in S}(\delta e^{1-\delta})^{-p_s/2}. \tag{44}$$

To make this more explicit, we denote by $|S|$ the size of S, and $p_{\max} = \max_{s\in S} p_s$, and consider a simple upper bound for the right-hand side in (44),

$$\text{edf}(\hat{\theta}_{\hat{s}}) \le \frac{2}{1-\delta}\log|S| + p_{\max}\bigg(\frac{\log(1/\delta)}{1-\delta} - 1\bigg). \tag{45}$$

This simplification should be fairly tight, that is, the right-hand side in (45) should be close to that in (44), when $|S|$ and $\max_{s\in S} p_s - \min_{s\in S} p_s$ are both not very large. Now, any choice of $0 \le \delta < 1$ can be used to give a valid bound in (45). As an example, taking $\delta = 9/10$ gives

$$\text{edf}(\hat{\theta}_{\hat{s}}) \le 20\log|S| + 0.054 p_{\max}.$$

By (20), the risk reformulation of the result in Theorem 2.1, we get the finite-sample risk bound

$$\text{Risk}(\hat{\theta}_{\hat{s}}) \le \|(I - P_{s_0})\theta_0\|_2^2 + \sigma^2(p_{s_0} + 0.108 p_{\max}) + 40\sigma^2\log|S|,$$

where we have explicitly written the oracle risk as $\text{Risk}(\hat{\theta}_{s_0}) = \|(I - P_{s_0})\theta_0\|_2^2 + \sigma^2 p_{s_0}$.

4.2. Oracle Inequality for SURE-tuned Subset Regression

The optimal choice of δ, that is, the choice giving the tightest bound in (45) (and so, the tightest risk bound), will depend on $|S|$ and $p_{\max}$. The analytic form of such a value of δ is not clear from the form of the bound in (45). But, we can adopt an asymptotic perpsective: if $\log|S|$ is small compared to the oracle risk $\text{Risk}(\hat{\theta}_{s_0})$, and $p_{\max}$ is not too large compared to the oracle risk, then (45) implies $\text{edf}(\hat{\theta}_{\hat{s}}) = o(\text{Risk}(\hat{\theta}_{s_0}))$. We state this formally next.

Theorem 4.1. Assume that $Y \sim N(\theta_0, \sigma^2 I)$, and that there is a sequence $a_n > 0$, $n = 1, 2, 3, \ldots$ with $a_n \to 0$ as $n \to \infty$, such that the risk of the oracle subset regression estimator $\hat{\theta}_{s_0}$ satisfies

$$\frac{1}{a_n}\frac{\log|S|}{\text{Risk}(\hat{\theta}_{s_0})} \to 0 \quad \text{and} \quad a_n\frac{p_{\max}}{\text{Risk}(\hat{\theta}_{s_0})} \to 0 \quad \text{as } n \to \infty. \tag{46}$$

Then, there is a sequence $0 \le \delta_n < 1, n = 1, 2, 3, \ldots$ with $\delta_n \to 1$ as $n \to \infty$, such that

$$\bigg[\frac{2}{1-\delta_n}\log|S| + p_{\max}\bigg(\frac{\log(1/\delta_n)}{1-\delta_n} - 1\bigg)\bigg]/\text{Risk}(\hat{\theta}_{s_0}) \to 0 \quad \text{as } n \to \infty.$$

Plugging this into the bound in (45) shows that $\text{edf}(\hat{\theta}_{\hat{s}})/\text{Risk}(\hat{\theta}_{s_0}) \to 0$, so $\text{ExOpt}(\hat{\theta}_{\hat{s}})/\text{Risk}(\hat{\theta}_{s_0}) \to 0$ as well, establishing the oracle inequality (17) for the SURE-tuned subset regression estimator.

The assumptions (46) may look abstract, but are not strong and are satisfied under fairly simple conditions. For example, if we assume that $\|(I - P_{s_0})\theta_0\|_2^2 = 0$ (which means there

is no bias), and as $n \to \infty$ (with σ^2 constant) it holds that $(\log|S|)/p_{s_0} \to 0$ and $p_{\max}/p_{s_0} = O(1)$ (which means the number $|S|$ of candidate models is much smaller than 2^{p_0}, and we are not searching over much larger models than the oracle), then it is easy to check (46) is satisfied, say, with $a_n = \sqrt{(\log|S|)/p_{s_0}}$. Assumptions (46) can accommodate more general settings, for example, in which there is bias, or in which $p_{\max}/p_{s_0}$ diverges, as long as these quantities scale at appropriate rates.

Theorem 4.1 establishes the classical oracle inequality (17) for the SURE-tuned subset regression estimator, which is nothing more than the C_p-tuned (or AIC-tuned, as σ^2 is assumed to be known) subset regression estimator. This of course is not really a new result; cf. classical theory on model selection in regression, as in Corollary 2.1 of Li (1987). This author established a result similar to (17) for the C_p-tuned subset regression estimator, chosen over a family of nested regression models, and showed asymptotic equivalence of the attained loss to the oracle loss (rather than the attained and oracle risks), in probability.

We remark that a similar analysis to that above, where we upper bound the excess degrees of freedom and risk, should be possible for a general discrete family of linear smoothers, beyond linear regression estimators. This would cover, for example, s-nearest neighbor regression estimators across various choices $s = 1, 2, 3, \ldots, |S|$. The linear smoother setting is studied by Li (1987), and would make for another demonstration of our excess optimism theory, but we do not pursue it.

5. Characterizing Excess Degrees of Freedom with (Extensions of) Stein's Formula

In this section, we keep the normal assumption, $Y \sim F = N(\theta_0, \sigma^2 I)$ in (1), and we move beyond individual families of estimators, by studying the use of Stein's formula (and extensions thereof) for calculating excess degrees of freedom, in an effort to understand this quantity in some generality.

5.1. Stein's Formula, for Smooth Estimators

We consider the case in which $S \subseteq \mathbb{R}$ is an open interval, so $\hat{\theta}_s$ is defined over a continuously-valued (rather than a discrete) tuning parameter $s \in S$. We make the following assumption.

Assumption 5.1. The map $\hat{s} : \mathbb{R}^n \to S$ is differentiable.

It is worth noting that Assumption 5.1 seems strong. In particular, it is not implied by the SURE criterion in (6) being smooth in (Y, s) jointly, that is, by the map $G : \mathbb{R}^n \times S \to \mathbb{R}$, defined by

$$G(Y, s) = \|Y - \hat{\theta}_s(Y)\|_2^2 + 2\sigma^2 \widehat{\mathrm{df}}_s(Y), \tag{47}$$

being smooth. When $G(Y, \cdot)$ is multimodal over $s \in S$, its minimizer $\hat{s}(Y)$ can jump discontinuously as Y varies, even if G itself varies smoothly. Figure 1 provides an illustration of this phenomenon. Notably, the SURE criterion for the family of shrinkage estimators we considered in Section 3.1 (as well as Section 3.3) was unimodal, and Assumption 5.1 held in this setting; however, we see no reason for this to be true in general. Thus, we will use Assumption 5.1 to develop a characterization of excess degrees of freedom, shedding light on the nature of this quantity, but should keep in mind that our assumptions may represent a somewhat restricted setting.

It is now helpful to define a "parent" mapping $\widehat{\Theta} : \mathbb{R}^n \times S \to \mathbb{R}^n$ by $\hat{\theta}_s = \widehat{\Theta}(\cdot, s)$ for each $s \in S$, and $h : \mathbb{R}^n \to \mathbb{R}^n \times S$ by $h(Y) = (Y, \hat{s}(Y))$. In this notation, the SURE-tuned estimator is given by the composition $\hat{\theta}_{\hat{s}} = \widehat{\Theta} \circ h$. The following is our assumption on $\widehat{\Theta}$.

Assumption 5.2. The function $\widehat{\Theta} : \mathbb{R}^n \times S \to \mathbb{R}^n$ is differentiable and satisfies the integrability condition $\mathbb{E}[\sup_{s \in S} \sum_{i=1}^n |\partial \widehat{\Theta}_i(Y, s)/\partial Y_i|] < \infty$.

Differentiability of both $\widehat{\Theta}$ and h implies differentiability of their composition $\hat{\theta}_{\hat{s}} = \widehat{\Theta} \circ h$; in addition, we know from the integrability condition in Assumption 5.2 that $\mathbb{E}[\sum_{i=1}^n |\partial \hat{\theta}_{\hat{s},i}(Y)/\partial Y_i|] < \infty$, so Stein's formula is applicable to $\hat{\theta}_{\hat{s}}$.

Finally, we consider the following assumption on the SURE criterion G in (47).

Assumption 5.3. The map $G : \mathbb{R}^n \times S \to \mathbb{R}$ is twice differentiable, and for each point $Y \in \mathbb{R}^n$, the minimizer $\hat{s}(Y)$ of $G(Y, \cdot)$ is the unique value satisfying

$$\frac{\partial G}{\partial s}(Y, \hat{s}(Y)) = 0, \tag{48}$$

$$\frac{\partial^2 G}{\partial s^2}(Y, \hat{s}(Y)) > 0. \tag{49}$$

As in our comment following Assumption 5.1, we must point out that Assumption 5.3 seems quite strong, and as far as we can tell, in a generic problem setting there seems to be nothing preventing $G(Y, \cdot)$ from being multimodal, which would violate Assumption 5.3. Still, we will use it to develop insight on the nature of excess degrees of freedom.

The following is the main result of this subsection.

Theorem 5.1. Under $Y \sim N(\theta_0, \sigma^2 I)$ and Assumptions 5.1 and 5.2, the excess degrees of freedom of the SURE-tuned estimator $\hat{\theta}_{\hat{s}}$ is given by

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) = \mathbb{E}\left(\sum_{i=1}^n \frac{\partial \widehat{\Theta}_i}{\partial s}(Y, \hat{s}(Y)) \frac{\partial \hat{s}}{\partial Y_i}(Y) \right). \tag{50}$$

Additionally, under Assumption 5.3, it is given by

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) = -\mathbb{E}\left[\left(\frac{\partial^2 G}{\partial s^2}(Y, \hat{s}(Y)) \right)^{-1} \times \sum_{i=1}^n \left(\frac{\partial \widehat{\Theta}_i}{\partial s}(Y, \hat{s}(Y)) \frac{\partial^2 G}{\partial Y_i \partial s}(Y, \hat{s}(Y)) \right) \right]. \tag{51}$$

The advantage of (51) over (50) is that the former is in general easier to compute. Computing $\partial \widehat{\Theta}_i/\partial s$, $i = 1, \ldots, n$ in (50) is often easy, at least when the estimator $\hat{\theta}_s$ (for fixed s) is available in closed form. But computing $\partial \hat{s}/\partial Y_i$, $i = 1, \ldots, n$ in (50) is typically much harder; even for simple problems, the SURE-optimal tuning parameter $\hat{s}$ often cannot be written in the closed form.

A straightforward calculation shows that, for the classes of shrinkage estimators in Sections 3.1 and 3.3, both (50) and (51) match the excess degrees of freedom results derived in

these sections. In principle, whenever its assumptions hold, Theorem 5.1 gives explicitly computable unbiased estimators for excess degrees of freedom, that is, the quantities inside the expectations in (50) and (51). It is unclear to us (as we have already discussed) to what extent these assumptions hold in general, but we can still use the results, particularly (51) to derive some helpful intuition on excess degrees of freedom. Roughly speaking:

- if (on average) $(\partial^2 G/\partial s^2)(Y, \hat{s}(Y))$ is large, that is, $G(Y, \cdot)$ is sharply curved around its minimum, that is, SURE sharply identifies the optimal tuning parameter value $\hat{s}(Y)$ given Y, then this drives the excess degrees of freedom to be smaller;
- if (on average) $|(\partial^2 G/\partial Y_i \partial s)(Y, \hat{s}(Y))|$ is large, that is, $|(\partial G/\partial s)(Y, \hat{s}(Y))|$ varies quickly with Y_i, that is, the function whose root in (48) determines $\hat{s}(Y)$ changes quickly with Y_i, then this drives the excess degrees of freedom to be larger;
- the pair of terms in the summand in (51) tend to have opposite signs (their specific signs are a reflection of the tuning parameterization associated with $s \in S$), which cancels out the -1 in front, and makes the excess degrees of freedom positive.

5.2. Extensions of Stein's Formula, for Nonsmooth Estimators

When an estimator has severe enough discontinuities, it will not be weakly differentiable, and then Stein's formula (10) cannot be directly applied. This is especially relevant to the topic of our article, as the SURE-tuned estimator $\hat{\theta}_{\hat{s}}$ can itself be discontinuous in Y even if each member of the family $\{\hat{\theta}_s : s \in S\}$ is continuous in Y (due to discontinuities in the SURE-optimal tuning parameter map $\hat{s}$). Note, this will always be the case for a discrete tuning parameter set S; it can also be the case for a continuous tuning parameter set S, recall Figure 1.

Fortunately, extensions of Stein's formula have been recently developed, to account for discontinuities of certain types. Tibshirani (2015) established an extension for estimators that are piecewise smooth. To define this notion of piecewise smoothness precisely, we must introduce some notation. Given an estimator $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$, we write $\hat{\theta}_i(\cdot, Y_{-i}) : \mathbb{R} \to \mathbb{R}$ for the ith component function $\hat{\theta}_i$ of $\hat{\theta}$ acting on the ith coordinate of the input alone, with all other $n-1$ coordinates fixed at Y_{-i}. We also write $\mathcal{D}(\hat{\theta}_i(\cdot, Y_{-i}))$ to denote the set of dicontinuities of the map $\hat{\theta}_i(\cdot, Y_{-i})$. In this notation, the estimator $\hat{\theta}$ is said to be *p-almost differentiable* if, for each $i = 1, \ldots, n$ and (Lebesgue) almost every $Y_{-i} \in \mathbb{R}^{n-1}$, the map $\hat{\theta}_i(\cdot, Y_{-i}) : \mathbb{R} \to \mathbb{R}$ is absolutely continuous on each of the open intervals $(-\infty, \delta_1), (\delta_2, \delta_3), \ldots, (\delta_m, \infty)$, where $\delta_1 < \delta_2 < \cdots < \delta_m$ are the sorted elements of $\mathcal{D}(\hat{\theta}_i(\cdot, Y_{-i}))$, assumed to be a finite set. For p-almost differentiable $\hat{\theta}$, Tibshirani (2015) proved that

$$\mathrm{df}(\hat{\theta}) = \mathbb{E}\bigg[\sum_{i=1}^n \frac{\partial \hat{\theta}_i}{\partial Y_i}(Y)\bigg] + \frac{1}{\sigma}\mathbb{E}\bigg[\sum_{i=1}^n \sum_{\delta \in \mathcal{D}(\hat{\theta}_i(\cdot, Y_{-i}))} \phi\bigg(\frac{\delta - \theta_{0,i}}{\sigma}\bigg)\big[\hat{\theta}_i(\delta, Y_{-i})_+ - \hat{\theta}_i(\delta, Y_{-i})_-\big]\bigg], \tag{52}$$

under some regularity conditions that ensure the second term on the right-hand side is well-defined. Above, we denote one-sided limits from above and from below by $\hat{\theta}_i(\delta, Y_{-i})_+ = \lim_{t \downarrow \delta} \hat{\theta}_i(t, Y_{-i})$ and $\hat{\theta}_i(\delta, Y_{-i})_- = \lim_{t \uparrow \delta} \hat{\theta}_i(t, Y_{-i})$, respectively, for the map $\hat{\theta}_i(\cdot, Y_{-i})$, $i = 1, \ldots, n$, and we denote by ϕ the univariate standard normal density.

A difficulty with (52) is that it is often hard to compute or characterize the extra term on the right-hand side. Mikkelsen and Hansen (2016) derived an alternate extension of Stein's formula for piecewise Lipschitz estimators. While this setting is more restricted than that in Tibshirani (2015), the resulting characterization is more "global" (instead of being based on discontinuities along the coordinate axes), and thus it can be more tractable in some cases. Formally, Mikkelsen and Hansen (2016) considered an estimator $\hat{\theta} : \mathbb{R}^n \to \mathbb{R}^n$ with associated regular open sets $U_j \subseteq \mathbb{R}^n$, $j = 1, \ldots, J$ whose closures cover $\mathbb{R}^n$ (i.e., $\cup_{j=1}^J \bar{U}_j = \mathbb{R}^n$), such that each map $\hat{\theta}^j := \hat{\theta}|_{U_j}$ (the restriction of $\hat{\theta}$ to U_j) is locally Lipschitz continuous. The authors proved that, for such an estimator $\hat{\theta}$,

$$\mathrm{df}(\hat{\theta}) = \mathbb{E}\bigg[\sum_{i=1}^n \frac{\partial \hat{\theta}_i}{\partial Y_i}(Y)\bigg] + \frac{1}{2}\sum_{j \neq k} \int_{\bar{U}_j \cap \bar{U}_k} \Big\langle \hat{\theta}^k(y) - \hat{\theta}^j(y), \eta_j(y) \Big\rangle \phi_{\theta_0, \sigma^2 I}(y)\, d\mathcal{H}^{n-1}(y), \tag{53}$$

again under some further regularity conditions that ensure the second term on the right-hand side is well-defined. Above, $\eta_j(y)$ denotes the outer unit normal vector to ∂U_j (the boundary of U_j) at a point y, $j = 1, \ldots, J$, $\phi_{\theta_0, \sigma^2 I}$ is the density of a normal variate with mean θ_0 and covariance $\sigma^2 I$, and $\mathcal{H}^{n-1}$ denotes the $(n-1)$-dimensional Hausdorff measure.

Our interest in (52), (53) is in applying these extensions to $\hat{\theta} = \hat{\theta}_{\hat{s}}$, the SURE-tuned estimator defined from a family $\{\hat{\theta}_s : s \in S\}$. A general formula for excess degrees of freedom, following from (52) or (53), would be possible, but also complicated in terms of the required regularity conditions. Here is a high-level discussion, to reiterate motivation for (52), (53) and outline their applications. We discuss the discrete and continuous tuning parameter settings separately.

- When the tuning parameter s takes discrete values (i.e., S is a discrete set), extensions such as (52), (53) are needed to characterize excess degrees freedom, because the estimator $\hat{\theta}_{\hat{s}}$ is generically discontinuous and Stein's original formula cannot be used. In the discrete setting, the first term on the right-hand side of both (52), (53) (when $\hat{\theta} = \hat{\theta}_{\hat{s}}$) is $\mathbb{E}[\widehat{\mathrm{df}}_{\hat{s}(Y)}(Y)]$, in the notation of (15), and thus the second term on the right-hand side of either (52), (53) (when $\hat{\theta} = \hat{\theta}_{\hat{s}}$) gives precisely the excess degrees of freedom.
- When s takes continuous values (i.e., S is a connected subset of Euclidean space), extensions as in (52), (53) are not strictly speaking always needed, though it seems likely to us that they will be needed in many cases, because the SURE-tuned estimator $\hat{\theta}_{\hat{s}}$ can inherit discontinuities from the SURE-optimal parameter map $\hat{s}$ (recall Figure 1). In the continous tuning parameter case, both the first and second

terms on the right-hand sides of (52), (53) (when $\hat{\theta} = \hat{\theta}_{\hat{s}}$) can contribute to excess degrees of freedom; that is, excess degrees of freedom is given by the second term plus any terms left over from applying the chain-rule for differentiation in the first term.

Over the next two subsections, we demonstrate the usefulness of the extensions in (52) and (53) by applying them in two specific settings.

5.3. Soft-Thresholding Estimators

Consider the family of soft-thresholding estimators with component functions

$$\hat{\theta}_{s,i}(Y) = \operatorname{sign}(Y_i)(|Y_i| - s)_+, \quad i = 1, \dots, n, \quad \text{for } s \geq 0. \tag{54}$$

In this setting, SURE in (6) is

$$\widehat{\operatorname{Err}}_s(Y) = \sum_{i=1}^{n} \min\{Y_i^2, s^2\} + 2\sigma^2 |\{i : |Y_i| \geq s\}|. \tag{55}$$

Soft-thresholding estimators, like the shrinkage estimators of Section 3.1, have been studied extensively in the statistical literature; some key references that study risk properties of soft-thresholding estimators are Donoho and Johnstone (1994, 1995, 1998), and Chapters 8 and 9 of Johnstone (2015) give a thorough summary.

The extension of Stein's formula from Tibshirani (2015), as given in (52), can be used to prove that the excess degrees of freedom of the SURE-tuned soft-thresholding estimator is nonnegative. The key realization is as follows: if a component function $\hat{\theta}_{\hat{s},i}$ of the SURE-tuned soft-thresholding estimator jumps discontinuously as we move Y along the ith coordinate axes, then the sign of this jump must match the direction in which Y_i is moving, thus the latter term on the right-hand side of (52) is always nonnegative.

Theorem 5.2. The SURE-tuned soft-thresholding estimator $\hat{\theta}_{\hat{s}}$ is p-almost differentiable. Moreover, for each $i = 1, \dots, n$, each $Y_{-i} \in \mathbb{R}^{n-1}$, and each discontinuity point δ of $\hat{\theta}_{\hat{s}(\cdot, Y_{-i}),i}(\cdot, Y_{-i}) : \mathbb{R} \to \mathbb{R}$, it holds that

$$\big[\hat{\theta}_{\hat{s}(\delta, Y_{-i}),i}(\delta, Y_{-i})\big]_+ - \big[\hat{\theta}_{\hat{s}(\delta, Y_{-i}),i}(\delta, Y_{-i})\big]_- \geq 0. \tag{56}$$

Therefore, when $Y \sim N(\theta_0, \sigma^2 I)$, we have from (52) that $\operatorname{edf}(\hat{\theta}_{\hat{s}}) \geq 0$ and so $\operatorname{df}(\hat{\theta}_{\hat{s}}) \geq \mathbb{E}\big|\big\{i : |Y_i| \geq \hat{s}(Y)\big\}\big|$.

The proof of Theorem 5.2 examines the discontinuities in the SURE-tuned soft-thresholding estimator; in particular, it shows that for each $i = 1, \dots, n$ and $Y_{-i} \in \mathbb{R}^n$, the map $\hat{\theta}_{\hat{s}(\cdot, Y_{-i}),i}(\cdot, Y_{-i})$ has at most two discontinuity points, and at a discontinuity point δ, the magnitude of the jump is itself bounded by δ. This can be used to show that $\operatorname{edf}(\hat{\theta}_{\hat{s}}) \leq \sqrt{2/(\pi e)}n \approx 0.484n$ in the null case, $\theta_0 = 0$. We note that this upper bound is likely very loose (e.g., see Figure 3, where the excess degrees of freedom is seen empirically to scale as $\log n$). A tighter upper bound should be possible with more refined calculations, but we do not pursue this here.

5.4. Subset Regression Estimators, Revisited

We return to the setting of Section 4, that is, we consider the family of subset regression estimators in (39), which we can abbreviate by $\hat{\theta}_s(Y) = P_s Y$, $s \in S$, using the notation of the latter section. In Section 4.1, recall, we derived upper bounds on the excess degrees of freedom of the SURE-tuned subset regression estimator $\operatorname{edf}(\hat{\theta}_{\hat{s}})$. Here, we apply the extension of Stein's formula from Mikkelsen and Hansen (2016), as stated in (53), to represent excess degrees of freedom for SURE-tuned subset regression in an alternative and (in principle) exact form. The calculation of the second-term on the right-hand side in (53) for the SURE-tuned subset regression estimator, which yields the result (58) in the next theorem, can already be found in Mikkelsen and Hansen (2016) (in their study of best subset selection). A complete proof is given in the supplement nonetheless.

Theorem 5.3 (Mikkelsen and Hansen 2016). The SURE-tuned subset regression estimator $\hat{\theta}_{\hat{s}}$ is piecewise Lipschitz (in fact, piecewise linear) over regular open sets U_s, $s \in S$, whose closures cover $\mathbb{R}^n$. For $s, t \in S$, the outer unit normal vector $\eta_s(y)$ to ∂U_s at a point $y \in \bar{U}_s \cap \bar{U}_t$ is given by

$$\eta_s(y) = \frac{(P_t - P_s)y}{\|(P_t - P_s)y\|_2}. \tag{57}$$

Therefore, when $Y \sim N(\theta_0, \sigma^2 I)$, we have from (53) that

$$\operatorname{edf}(\hat{\theta}_{\hat{s}}) = \frac{1}{2} \sum_{s \neq t} \int_{\bar{U}_s \cap \bar{U}_t} \|(P_t - P_s)y\|_2 \, \phi_{\theta_0, \sigma^2 I}(y) \, d\mathcal{H}^{n-1}(y). \tag{58}$$

An important implication is $\operatorname{edf}(\hat{\theta}_{\hat{s}}) \geq 0$, which implies that $\operatorname{df}(\hat{\theta}_{\hat{s}}) \geq \mathbb{E}(p_{\hat{s}(Y)})$.

While the integral (58) is hard to evaluate in general, it is somewhat more tractable in the case of nested regression models. In the present setting each $s \in S$, recall, is identified with a subset of $\{1, \dots, p\}$. We say the collection S is *nested* if for each pair $s, t \in S$, we have either $s \subseteq t$ or $t \subseteq s$. The next result shows that for a nested collection of regression models, the integral expression (58) for excess degrees of freedom simplifies considerably, and can be upper bounded in terms of surface areas of balls under an appropriate Gaussian probability measure.

Before stating the result, it helps to introduce some notation. For a matrix A, we write $A_{j:k}$ as shorthand for $A_{\{j, j+1, \dots, k\}}$, that is, the submatrix given by extracting columns j through k. Likewise, for a vector a, we write $a_{j:k}$ as shorthand for $(a_j, a_{j+1}, \dots, a_k)$. When s is identified with a nonempty subset $\{1, \dots, j\}$, we write P_s, U_s, η_s as P_j, U_j, η_j respectively, and use $P_j^\perp$ for the orthogonal projector to P_j. Finally, we refer to the *Gaussian surface measure* Γ_d, defined over (Borel) sets $A \subseteq \mathbb{R}^d$ as

$$\Gamma_d(A) = \liminf_{\delta \to 0} \frac{\mathbb{P}(Z \in A_\delta \setminus A)}{\delta},$$

where $Z \sim N(0, I)$ denotes a d-dimensional standard normal variate, and $A_\delta = A + B_d(0, \delta)$ is the Minkowski sum of A and the d-dimensional ball $B_d(0, \delta)$ centered at the origin with radius δ. For a set A with smooth boundary ∂A, an equivalent definition is $\Gamma_d(A) = \int_{\partial A} \phi_{0,I}(x) \, d\mathcal{H}^{d-1}(x)$, where $\phi_{0,I}$ is the

density of Z, and $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure. Helpful references on Gaussian surface area include Ball (1993), Nazarov (2003), Klivans, O'Donnell, and Servedio (2008). We now state our main result of this subsection.

Theorem 5.4. Assume that $Y \sim N(\theta_0, \sigma^2 I)$, and that all models in the collection S are nested. Then, the excess degrees of freedom of the SURE-tuned subset regression estimator $\hat{\theta}_{\hat{s}}$ is

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) = \sqrt{2}\sigma \sum_{s \subseteq t} \sqrt{p_t - p_s} \int_{\tilde{U}_s \cap \tilde{U}_t} \phi_{\theta_0, \sigma^2 I}(y)\, d\mathcal{H}^{n-1}(y). \tag{59}$$

Now, without a loss of generality (otherwise, the only real adjustment is notational), let us identify each s with a subset $\{1, \ldots, j\}$. Then, the excess degrees of freedom is upper bounded by

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) \leq \sum_{d=1}^{p} \sqrt{2d}(d+1) \max_{j=1,\ldots,d} \Gamma_d\Big(B_d\big(\mu_{(j+1):(j+d)}, \sqrt{2d}\big)\Big), \tag{60}$$

where $\mu = V^T \theta_0/\sigma$, and $V \in \mathbb{R}^{n \times p}$ is an orthogonal matrix with columns $v_j = P_{j-1}^{\perp} X_j / \|P_{j-1}^{\perp} X_j\|_2$, $j = 1, \ldots, p$ (where we let $P_0 = 0$ for notational convenience). Also, recall that $\Gamma_d(B_d(u, r))$ denotes the d-dimensional Gaussian surface area of a ball $B_d(u, r)$ centered at u with radius r. When $\theta_0 = 0$, the result in (60) can be sharpened and simplified, giving

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) \leq \sum_{d=1}^{p} \sqrt{2d}\left(1 + \frac{1}{d}\right)\Gamma_d\big(B_d(0, \sqrt{2d})\big) < 10. \tag{61}$$

Though it is established in a restricted setting, $\theta_0 = 0$, the result in (61) is quite interesting, as it shows that the excess degrees of freedom of the SURE-tuned subset regression is bounded by the constant 10, and therefore its excess optimism is bounded by the constant $20\sigma^2$, regardless of the number of predictors p in the regression problem.

The derivation of (61) from (60) relies on two key facts: (i) the null case, $\theta_0 = 0$, admits a kind of symmetry that allows us to apply a classic result in combinatorics (the gas stations problem) to compute the exact probability of a collection of chi-squared inequalities, which leads to a reduction in the factor of $d+1$ in each summand of (60) to a factor of $1 + 1/d$ in each summand of (61); and (ii) the balls in the null case, in the summands of (61), are centered at the origin, so their Gaussian surface areas can be explicitly computed as in Ball (1993), Klivans, O'Donnell, and Servedio (2008).

Neither fact is true in the nonnull case, $\theta_0 \neq 0$, making it more difficult to derive a sharp upper bound on excess degrees of freedom. We finish with a couple remarks on the nonnull setting; more serious investigation of explicitly bounding and/or improving (60) is left to future work.

Remark 5.1 (Nonnull case: two models). When our collection is composed of just two nested models that are separated by a single variable, that is, $S = \{\{1, \ldots, p-1\}, \{1, \ldots, p\}\}$, straightforward inspection of the proof of Theorem 5.3 reveals that (60) becomes $\mathrm{edf}(\hat{\theta}_{\hat{s}}) = \sqrt{2}\Gamma_1(B_1(v_2^T \theta_0/\sigma, \sqrt{2}))$ (i.e., note the equality), where $v_2 = P_{p-1}^{\perp} X_p / \|P_{p-1}^{\perp} X_p\|_2$. The Gaussian surface measure is trivial to compute here (under an arbitrary mean θ_0) because it reduces to two evaluations of the Gaussian density, and thus we see that

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) = \sqrt{2}\phi\left(\sqrt{2} - v_2^T \theta_0/\sigma\right) + \sqrt{2}\phi\left(\sqrt{2} + v_2^T \theta_0/\sigma\right),$$

where ϕ is the standard (univariate) normal density. When $\theta_0 = 0$, the excess degrees of freedom is $2\sqrt{2}\phi(\sqrt{2}) \approx 0.415$. For general θ_0, it is upper bounded by $\max_{u \in \mathbb{R}} \sqrt{2}\phi(\sqrt{2} - u) + \sqrt{2}\phi(\sqrt{2} + u) \approx 0.575$.

Remark 5.2 (Nonnull case: general bounds). For an arbitrary collection S of nested models and arbitrary mean θ_0, a very loose upper bound on the right-hand side in (60) is $\sqrt{2p}p(p+1)$, which follows as the Gaussian surface measure of any ball is at most 1, as shown in Klivans, O'Donnell, and Servedio (2008). Under restrictions on θ_0, tighter bounds on the Gaussian surface measures of the appropriate balls should be possible. Furthermore, the multiplicative factor of $d+1$ in each summand of (60) is also likely larger than it needs to be; we note that an alternate excess degrees of freedom bound to that in (60) (following from similar arguments) is

$$\begin{aligned}\mathrm{edf}(\hat{\theta}_{\hat{s}}) \leq \sqrt{2} \sum_{j<k} \sqrt{k-j}\, \mathbb{P}\big(W_j(\|\mu_{1:j}\|_2^2) > 2(j-1)\big) \\ \times P\big(W_{p-k}(\|\mu_{(k+1):p}\|_2^2) < 2(p-k)\big) \cdot \\ \times \Gamma_{k-j}\Big(B_{k-j}\big(\mu_{(j+1):k}, \sqrt{2(k-j)}\big)\Big),\end{aligned} \tag{62}$$

where $W_d(\lambda)$ denotes a chi-squared random variable, with d degrees of freedom and noncentrality parameter λ. Sharp bounds on the noncentral chi-squared tails could deliver a useful upper bound on the right-hand side in (62); we do not expect the final bound reduce to a constant (independent of p) as it did in (61) in the null case, but it could certainly improve on the results in Section 4.1, that is, the bound in (45), which is on the order of $p_{\max}$ (the largest subset size in S).

6. Estimating Excess Degrees of Freedom with the Bootstrap

We discuss bootstrap methods for estimating excess degrees of freedom. As we have thus far, we assume normality, $Y \sim F = N(\theta_0, \sigma^2 I)$ in (1), but in what follows this assumption is used mostly for convenience,and can be relaxed (we can of course replace the normal distribution in the parameteric bootstrap with any known data distribution, or in general, use the residual bootstrap). The main ideas in this section are fairly simple, and follow naturally from standard ideas for estimating optimism using the bootstrap, for example, Breiman (1992), Ye (1998), Efron (2004).

6.1. Parametric Bootstrap Procedure

First, we descibe a parametric bootstrap procedure. We draw

$$Y^{*,b} \sim N(\hat{\theta}_{\hat{s}(Y)}(Y), \sigma^2 I), \quad b = 1, \ldots, B, \tag{63}$$

where B is some large number of bootstrap repetitions, for example, $B = 1000$. Our bootstrap estimate for the excess degrees of

freedom $\mathrm{edf}(\hat{\theta}_{\hat{s}})$ is then

$$\widehat{\mathrm{edf}}(Y) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\theta}_{\hat{s}(Y^{*,b}),i}(Y^{*,b})(Y_i^{*,b} - \bar{Y}_i^*) - \frac{1}{B}\sum_{b=1}^{B}\widehat{\mathrm{df}}_{\hat{s}(Y^{*,b})}(Y^{*,b}), \tag{64}$$

where we write $\bar{Y}_i^* = (1/B)\sum_{b=1}^{B} Y_i^{*,b}$ for $i = 1, \ldots, n$, and $\widehat{\mathrm{df}}_s$ is our estimator for the degrees of freedom of $\hat{\theta}_s$, unbiased for each $s \in S$. Note that in (64), for each bootstrap draw $b = 1, \ldots B$, we compute the SURE-optimal tuning parameter value $\hat{s}(Y^{*,b})$ for the given bootstrap data $Y^{*,b}$, and we compare the sum of empirical covariances (first term) to the plug-in degrees of freedom estimate (second term). We can express the definition of excess degrees of freedom in (15) as

$$\mathrm{edf}(\hat{\theta}_{\hat{s}}) = \mathbb{E}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\theta}_{\hat{s}(Y),i}(Y)(Y_i - \theta_{0,i})\right) - \mathbb{E}[\widehat{\mathrm{df}}_{\hat{s}(Y)}(Y)], \tag{65}$$

making it clear that (64) estimates (65). Fortuitously, the validity of the bootstrap approximation (64), as noted by Efron (2004), does not depend on the smoothness of $\hat{\theta}_{\hat{s}}$ as a function of Y. This makes it appropriate for estimating excess degrees of freedom, even when $\hat{\theta}_{\hat{s}}$ is discontinuous (e.g., due to discontinuities in the SURE-optimal parameter mapping $\hat{s}$), which can be difficult to handle analytically (recall Sections 5.2–5.4).

It should be noted, however, that typical applications of the bootstrap for estimating optimism, as reviewed in Efron (2004), consider low-dimensional problems, and it is not clear that (64) will be appropriate for high-dimensional problems. Indeed, we shall see in the examples in Section 6.2 that the bootstrap estimate for the degrees of freedom $\mathrm{df}(\hat{\theta}_{\hat{s}})$,

$$\widehat{\mathrm{df}}(Y) = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{\sigma^2}\sum_{i=1}^{n}\hat{\theta}_{\hat{s}(Y^{*,b}),i}(Y^{*,b})(Y_i^{*,b} - \bar{Y}_i^*), \tag{66}$$

can be poor in the high-dimensional settings being considered, which is not unexpected. But (perhaps) unexpectedly, in these same settings we will also see that the *difference* between (66) and the baseline estimate $(1/B)\sum_{b=1}^{B}\widehat{\mathrm{df}}_{\hat{s}(Y^{*,b})}(Y^{*,b})$, that is, the bootstrap excess degrees of freedom estimate, $\widehat{\mathrm{edf}}(Y)$ in (64), can still be reasonably accurate.

Alternatives to the parametric bootstrap procedure are discussed in the supplement.

6.2. Simulated Examples

We empirically evaluate the excess degrees of freedom of the SURE-tuned shrinkage estimator and the SURE-tuned soft-thresholding estimator, across different configurations for the data generating distribution, and evaluate the performance of the parametric bootstrap estimator for excess degrees of freedom. Specifically, our simulation setup can be described as follows.

- We consider 10 sample sizes n, log-spaced in between 10 and 5000.
- We consider 3 settings for the mean parameter θ_0: the null setting, where we set $\theta_0 = 0$; the weak sparsity setting, where $\theta_{0,i} = 4i^{-1/2}$ for $i = 1, \ldots, n$; and the strong sparsity setting, where $\theta_{0,i} = 4$ for $i = 1, \ldots, \lfloor \log n \rfloor$ and $\theta_{0,i} = 0$ for $i = \lfloor \log n \rfloor + 1, \ldots, n$.
- For each sample size n and mean θ_0, we draw observations Y from the normal data model in (1) with $\sigma^2 = 1$, for a total of 5000 repetitions.
- For each Y, we compute the SURE-tuned estimate over the shrinkage family in (21), and the SURE-tuned estimate over the soft-thresholding family in (54).
- For each SURE-tuned estimator $\hat{\theta}_{\hat{s}}$, we record various estimates of degrees of freedom, excess degrees of freedom, and prediction error (details given below).

The simulation results are displayed in Figures 2 and 3; for brevity, we only report on the null and weak sparsity settings for the shrinkage family, and the null and strong sparsity settings for the soft-thresholding family. All degrees of freedom, excess degrees of freedom, and prediction error estimates (except the Monte Carlo estimates) were averaged over the 5000 repetitions; the plots all display the averages along with ± 1 standard error bars.

Figure 2 shows the results for the shrinkage family, with the first row covering the null setting, and the second row the weak sparsity setting. The left column shows the excess degrees of freedom of the SURE-tuned shrinkage estimator, for growing n. Four types of estimates of excess degrees of freedom are considered: Monte Carlo, computed from the 5000 repetitions (drawn in black); the unbiased estimate from Stein's formula, that is, $2\hat{s}(Y)/(1 + \hat{s}(Y))$ (in red); the bootstrap estimate (64) (in green); and the observed (scaled) excess optimism, that is, $(\|Y^* - \hat{\theta}_{\hat{s}(Y)}(Y)\|_2^2 - \widehat{\mathrm{Err}}_{\hat{s}(Y)}(Y))/(2\sigma^2)$, where Y^* is an independent copy of Y (in gray). The middle column shows similar estimates, but for degrees of freedom; here, the naive estimate is $\widehat{\mathrm{df}}_{\hat{s}(Y)}(Y) = n/(1 + \hat{s}(Y))$; the unbiased estimate is $n/(1 + \hat{s}(Y)) + 2\hat{s}(Y)/(1 + \hat{s}(Y))$; the naive bootstrap estimate is the second term in (64); and the bootstrap estimate is the first term in (64), that is, as given in (66). Finally, the right column shows the analogous quantities, but for estimating prediction error. The error metric is normalized by the sample size n for visualization purposes.

We can see that the unbiased estimate of excess degrees of freedom is quite accurate (i.e., close to the Monte Carlo gold standard) throughout. The bootstrap estimate is also accurate in the null setting, but somewhat less accurate in the weak sparsity setting, particularly for large n. However, comparing it to the observed (scaled) excess optimism—which relies on test data and thus may not be available in practice—the bootstrap estimate still appears reasonable accurate, and more stable. While all estimates of degrees of freedom are quite accurate in the null setting, we can see that the two bootstrap degrees of freedom estimates are far too small in the weak sparsity setting. This can be attributed to the high-dimensionality of the problem (estimating n means from n observations). Fortuitously, we can see that the *difference* between the bootstrap and naive bootstrap degrees of freedom estimates, that is, the bootstrap excess degrees of freedom estimate, is still relatively accurate even when the original two are so highly inaccurate. Lastly, the error plots show that

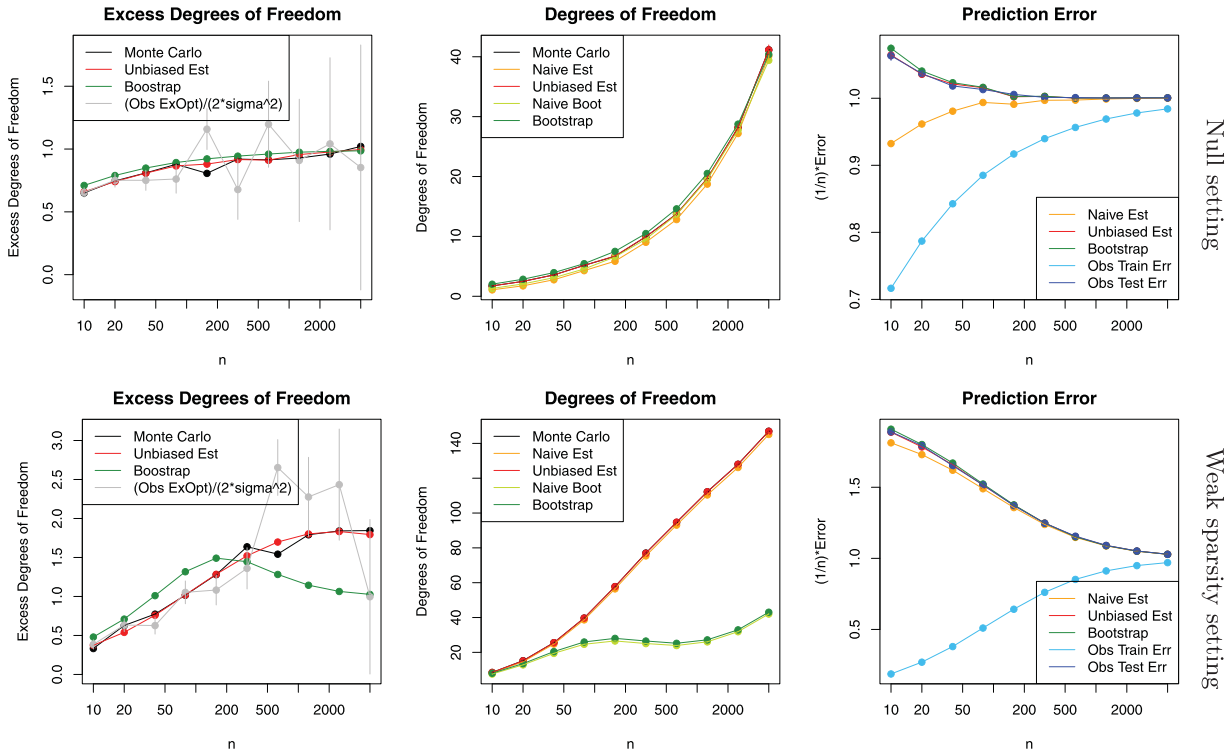

Figure 2. Simulation results for SURE-tuned shrinkage.

the correction for excess optimism is more significant (i.e., the gap between the naive error estimate and observed test error is larger) in the null setting than in the weak sparsity setting.

Figure 3 shows the results for the soft-thresholding family. The layout of plots is the same as that for the shrinkage family (note that the unbiased estimates of excess degrees of freedom and of degrees of freedom are not available for soft-thresholding). The summary of results is also similar: we can see that the bootstrap excess degrees of freedom estimate is fairly accurate in general, and less accurate in the nonnull case with larger n. One noteworthy difference between Figures 2 and 3: for the soft-thresholding family, we can see that the excess degrees of freedom estimates appear to be growing with n, rather than remaining upper bounded by 2, as they are for the

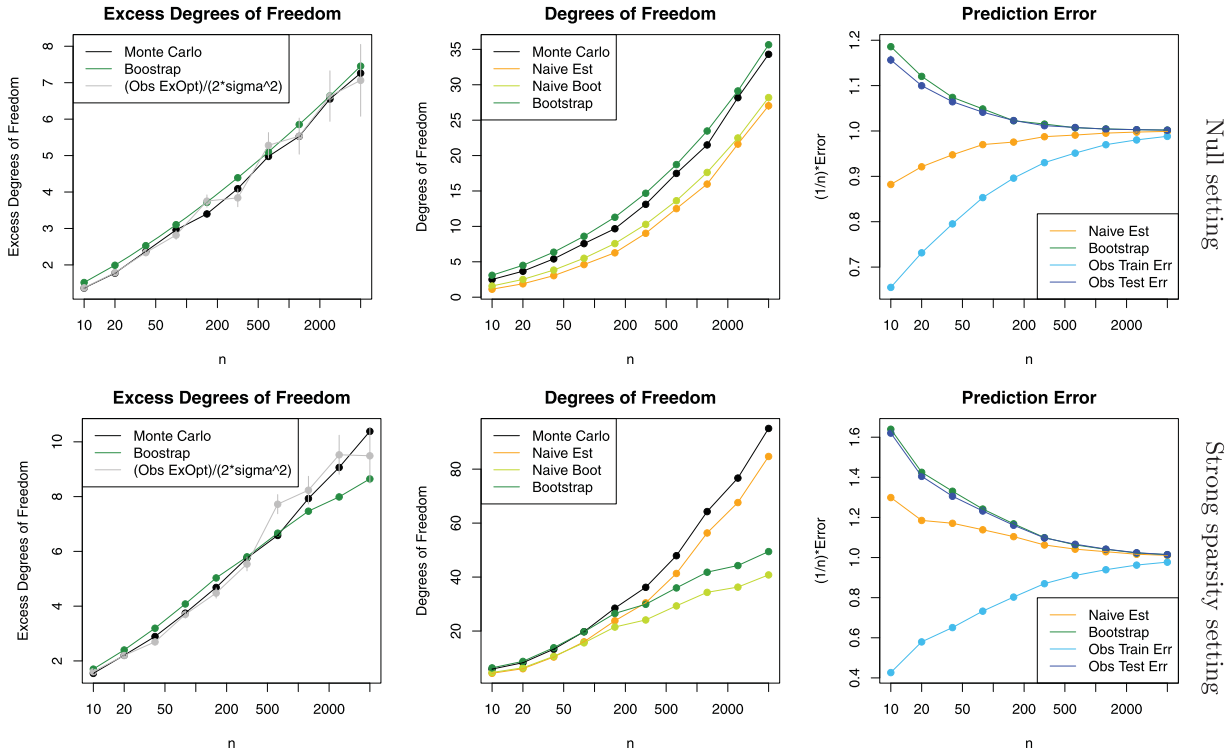

Figure 3. Simulation results for SURE-tuned soft-thresholding.

shrinkage family (recall also that this is clearly implied by the characterization in (24)). However, the growth rate is slow: the linear trend in the leftmost plots in Figure 3 suggests that the excess degrees of freedom scales as $\log n$ (noting that the x-axis is on a log scale).

7. Discussion

We have proposed and studied a concept called excess optimism, in (14), which captures the added optimism of a SURE-tuned estimator, beyond what is prescribed by SURE itself. By construction, an unbiased estimator of excess optimism leads to an unbiased estimator of the prediction error of the rule tuned by SURE. Further motivation for the study of excess optimism comes from its close connection to oracle estimation, as given in Theorem 2.1, where we showed that the excess optimism upper bounds the excess risk, that is, the difference between the risk of the SURE-tuned estimator and the risk of the oracle estimator. Hence, if the excess optimism is shown to be sufficiently small next to the oracle risk, then this establishes the oracle inequality (17) for the SURE-tuned estimator.

Interestingly, excess optimism can be exactly characterized for a family of shrinkage estimators, as studied in Section 3, where we showed that the excess optimism (and hence the excess risk) of a class of shrinkage estimators—in both simple normal means and regression settings—is at most $4\sigma^2$. For a family of subset regression estimators, such a precise characterization is not possible, but we showed in Section 4 that upper bounds on the excess optimism can be formed that imply the oracle inequality (17) for the SURE-tuned (here, C_p-tuned) subset regression estimator.

Characterizating excess optimism—equivalently excess degrees of freedom, in (15), which is just a constant multiple of the former quantity—is a difficult task in general, due to discontinuities that can exist in the SURE-tuned estimator. Severe enough discontinuities will imply the SURE-tuned estimator is not weakly differentiable and disallow the use of Stein's formula for estimating excess degrees of freedom. Section 5 discussed recently developed extensions of Stein's formula that handle certain types of discontinuities. As an example application, we proved that one of these extensions can be used to bound the excess optimism of the SURE-tuned subset regression estimator, over a family of nested subsets, by $20\sigma^2$, in the null case when $\theta_0 = 0$. Finally, in Section 6, we showed that estimation of excess degrees of freedom using the bootstrap is conceptually straightforward, and appears to works reasonably well (but, it tends to underestimate excess degrees of freedom in high-dimensional settings with nontrivial signal present in θ_0).

There are a number of interesting directions for extensions of our work. Two such directions, on heteroscedastic data, and alternative loss functions (other than squared loss), are described in the supplement. We finish here by noting an implication of some of our technical results in Sections 4.1 and 5.4 on the degrees of freedom of the (Lagrangian version of the) best subset selection estimator, defined by

$$\hat{\beta}_\lambda^{\text{subset}}(Y) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_0, \tag{67}$$

where recall, the ℓ_0 norm is defined by $\|\beta\|_0 = \sum_{j=1}^p 1\{\beta_j \neq 0\}$. Here $\lambda \geq 0$ is a tuning parameter. The best subset selection estimator in (67) can be seen as minimizing a SURE-like criterion, cf. the SURE criterion in (40), where we define the collection S to contain all subsets of $\{1, \ldots, p\}$, and we replace the multiplier $2\sigma^2$ in (40) with a generic parameter, $\lambda \geq 0$, used to weight the complexity penalty. Combining Lemma 4.1 (for the upper bound) and Theorem 5.3 (for the lower bound) provides the following result for best subset selection.

Theorem 7.1. Assume that $Y \sim N(\theta_0, \sigma^2 I)$. For any fixed value of $\lambda \geq 0$, the degrees of freedom of the best subset selection estimator in (67) satisfies

$$\mathbb{E}\|\hat{\beta}_\lambda^{\text{subset}}(Y)\|_0 \leq \text{df}(X\hat{\beta}_\lambda^{\text{subset}}) \leq \mathbb{E}\|\hat{\beta}_\lambda^{\text{subset}}(Y)\|_0 + 2.29p. \tag{68}$$

In the language of Tibshirani (2015), the result in (68) proves the search degrees of freedom of best subset selection—the difference between $\text{df}(X\hat{\beta}_\lambda^{\text{subset}})$ and $\mathbb{E}\|X\hat{\beta}_\lambda^{\text{subset}}(Y)\|_0$— is nonnegative, and at most $2.29p$. Nonnegativity of search degrees of freedom here was conjectured by Tibshirani (2015) but not established in full generality (i.e., for general X); to be fair, Mikkelsen and Hansen (2016) should be credited with establishing this nonegativity, since, recall, the lower bound in (68) comes from Theorem 5.3, a result of these authors. The upper bound in (68), as far as we can tell, is new. Though it may seem loose, it implies that the degrees of freedom of the Lagrangian form of best subset selection is at most $3.29p$—in comparison, Janson, Fithian, and Hastie (2015) proved that best subset selection in constrained form (for a specific configuration of the mean particular θ_0) has degrees of freedom approaching ∞ as $\sigma \to 0$. This could be a reason to prefer the Lagrangian formulation (67) over its constrained counterpart.

Supplementary Materials

The online supplementary materials contain additional proofs and models.

Acknowledgments

We thank the (anonymous) Associate Editor and Referees who provided very helpful feedback that led to improvements in the article.

ORCID

Ryan J. Tibshirani http://orcid.org/0000-0002-2158-8304

References

Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," *Second International Symposium on Information Theory*, pp. 267–281. [698]

Ball, K. (1993), "The Reverse Isoperimetric Problem for Gaussian Measure," *Discrete & Computational Geometry*, 10, 411–420. [708]

Baranchik, A. (1964), "Multiple Regression and Estimation of the Mean of a Multivariate Normal Distribution," Technical Report, Stanford University. [702]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *Annals of Statistics*, 41, 802–837. [700]

Bernau, C., Augustin, T., and Boulesteix, A.-L. (2013), "Correcting the Optimal Resampling-Based Error Rate by Estimating the Error Rate of Wrapper Algorithms," *Biometrics*, 69, 693–702. [700]

Breiman, L. (1992), "The Little Bootstrap and Other Methods for Dimensionality Selection in Regression: X-Fixed Prediction Error," *Journal of the American Statistical Society*, 87, 738–754. [700,708]

Candes, E. J., Sing-Long, C. M., and Trzasko, J. D. (2013), "Unbiased Risk Estimates for Singular Value Thresholding and Spectral Estimators," *IEEE Transactions on Signal Processing*, 61, 4643–4657. [698]

Cavalier, L., Golubev, Y., Picard, D., and Tsybakov, A. (2002), "Oracle Inequalities for Inverse Problems," *Annals of Statistics*, 30, 843–874. [700]

Chen, X., Lin, Q., and Sen, B. (2015), "On Degrees of Freedom of Projection Estimators with Applications to Multivariate Shape Restricted Regression," arXiv: 1509.01877. [698]

Donoho, D. L., and Johnstone, I. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455. [708]

——— (1995), "Adapting to Unknown Smoothness via Wavelet Shrinkage," *Journal of the American Statistical Association*, 90, 1200–1224. [698,700,702,708]

——— (1998), "Minimax Estimation via Wavelet Shrinkage," *Annals of Statistics*, 26, 879–921. [707]

Efron, B. (1986), "How Biased is the Apparent Error Rate of a Prediction Rule?" *Journal of the American Statistical Association*, 81, 461–470. [698,700]

——— (2004), "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," *Journal of the American Statistical Association*, 99, 619–632. [697,700,708,709]

——— (2010), *Large-scale Simultaneous Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, New York: Cambridge University Press. [701]

——— (2014), "Estimation and Accuracy after Model Selection," *Journal of the American Statistical Association*, 109, 991–1007. [700]

Efron, B., and Hastie, T. (2016), *Computer Age Statistical Inference: Algorithms, Inference, and Data Science*, New York: Cambridge University Press. [703]

Fithian, W., Sun, D., and Taylor, J. (2014), "Optimal Inference after Model Selection," arXiv: 1410.2597. [700]

Hoerl, A., and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67. [703]

James, W., and Stein, C. (1961), "Estimation with Quadratic Loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 361–379. [701,702]

Janson, L., Fithian, W., and Hastie, T. (2015), "Effective Degrees of Freedom: A Flawed Metaphor," *Biometrika*, 102, 479–485. [711]

Johnstone, I. M. (1999), "Wavelet Shrinkage for Correlated Data and Inverse Problems: Adaptivity Results," *Statistica Sinica*, 9, 51–83. [698]

——— (2015), *Gaussian Estimation: Sequence and Wavelet Models*, New York: Cambridge University Press, draft version. [701,707]

Klivans, A., O'Donnell, R., and Servedio, R. (2008), "Learning Geometric Concepts via Gaussian Surface Area," *Foundations of Computer Science*, 49, 541–550. [708]

Kneip, A. (1994), "Ordered Linear Smoothers," *Annals of Statistics*, 22, 835–866. [700]

Krstajic, D., Buturovic, L., Leahy, D., and Thomas, S. (2014), "Cross-Validation pitfalls when Selecting and Assessing Regression and Classification Models," *Journal of Cheminformatics*, 6. [700]

Lee, J., Sun, D., Sun, Y., and Taylor, J. (2016), "Exact Post-Selection Inference, with Application to the Lasso," *Annals of Statistics*, 44, 907–927. [700]

Li, K.-C. (1985), "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross-Validation," *Annals of Statistics*, 14, 1352–1377. [700]

——— (1986), "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing," *Annals of Statistics*, 14, 1101–1112. [700]

——— (1987), "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *Annals of Statistics*, 15, 958–975. [700,705]

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014), "A Significance Test for the Lasso," *Annals of Statistics*, 42, 413–468. [700]

Mallows, C. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675. [697,698,700]

Mikkelsen, F. R., and Hansen, N. R. (2016), "Degrees of Freedom for Piecewise Lipschitz Estimators," arXiv: 1601.03524. [700,703,704,706,707,711]

Nazarov, F. (2003), "On the Maximal Perimeter of a Convex set in $\mathbb{R}^n$ with Respect to Gaussian Measure," *Geometric Aspects of Functional Analysis*, 1806, 169–187. [708]

Stein, C. (1981), "Estimation of the Mean of a Multivariate Normal Distribution," *Annals of Statistics*, 9, 1135–1151. [697,698,700,702]

Tian Harris, X. (2016), "Prediction Error after Model Selection," arXiv: 1610.06107. [700]

Tibshirani, R. J. (2015), "Degrees of Freedom and Model Search," *Statistica Sinica*, 25, 1265–1296. [700,703,706,707,711]

Tibshirani, R. J., and Taylor, J. (2011), "The Solution Path of the Generalized Lasso," *Annals of Statistics*, 39, 1335–1371. [698]

——— (2012), "Degrees of Freedom in Lasso Problems," *Annals of Statistics*, 40, 1198–1232. [698]

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016), "Exact Post-Selection Inference for Sequential Regression Procedures," *Journal of the American Statistical Association*, 111, 600–620. [700]

Tibshirani, R. J., and Tibshirani, R. (2009), "A Bias Correction for the Minimum Error Rate in Cross-Validation," *Annals of Applied Statistics*, 3, 822–829. [700]

Tsamardinos, I., Rakhshani, A., and Lagani, V. (2015), "Performance-Estimation Properties of Cross-Validation-Based Protocols with Simultaneous Hyper-Parameter Optimization," *International Journal on Artificial Intelligence Tools*, 24. [700]

Ulfarsson, M. O., and Solo, V. (2013a), "Tuning Parameter Selection for Nonnegative Matrix Factorization," *IEEE International Conference on Acoustics, Speech and Signal Processing*. [698]

——— (2013b), "Tuning Parameter Selection for Underdetermined Reduced-Rank Regression," *IEEE Signal Processing Letters*, 20, 881–884. [698]

Varma, S., and Simon, R. (2006), "Bias in Error Estimation When Using Cross-Validation for Model Selection," *BMC Bioinformatics*, 7. [700]

Xie, X., Kou, S., and Brown, L. (2012), "SURE Estimates for a Heteroscedastic Hierarchical Model," *Journal of the American Statistical Association*, 107, 1465–1479. [700,702]

Ye, J. (1998), "On Measuring and Correcting the Effects of Data Mining and Model Selection," *Journal of the American Statistical Society*, 93, 120–131. [708]

Zou, H., Hastie, T., and Tibshirani, R. (2007), "On the 'Degrees of Freedom' of the Lasso," *Annals of Statistics*, 35, 2173–2192. [698]

Zou, H., and Yuan, M. (2008), "Regularized Simultaneous Model Selection in Multiple Quantiles Regression," *Computational Statistics and Data Analysis*, 52, 5296–5304. [698]