Stein's Unbiased Risk Estimate

Statistical Machine Learning, Spring 2015

Ryan Tibshirani (with Larry Wasserman)

1 Stein's lemma

- In a landmark paper, Stein (1981) derived a beautiful and simple lemma about the standard normal distribution. Indeed, Stein knew of this result much earlier and wrote about it in previous papers, but in Stein (1981), the author developed a multivariate extension of this lemma that led to a remarkable result on unbiased risk estimation. (And, an interesting note: the paper Stein (1981) itself was actually written in 1974, and rumor has it Stein wasn't planning on publishing it, until a colleague convinced him to do so in 1981...)
- We'll walk through Stein's univariate and multivariate lemmas on the normal distribution. Following this, we'll discuss how they apply to unbiased risk estimation. We note that the univariate lemma has a converse, and this has become extremely important in its own right, studied and further developed in probability theory for proving convergence to normality. Stein didn't write a lot of papers, but he was a pretty influential guy!

1.1 Univariate lemma

- First, the univariate result. Let $Z \sim N(0,1)$. Let $f : \mathbb{R} \to \mathbb{R}$ be absolutely continuous, with derivative f' (and assume that $\mathbb{E}|f'(Z)| < \infty$). Then

  $$\mathbb{E}[Zf(Z)] = \mathbb{E}[f'(Z)].$$

  Proof: the easiest way is to use integration by parts. In fact, this way, the proof is really just one line. Starting from the right-hand side above, with ϕ denoting the normal density,

  $$\int_{-\infty}^{\infty} f'(z)\phi(z)\,dz = f(z)\phi(z)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(z)\phi'(z)\,dz.$$

  The first term on the right vanishes, and the proof follows noting that $\phi'(z) = -z\phi(z)$

  Stein proves this result in an alternate way, that makes the upcoming multivariate proof more straightforward. Here is his argument:

  $$\begin{aligned}\int_{-\infty}^{\infty} f'(z)\phi(z)\,dz &= \int_0^{\infty} f'(z)\bigg\{\int_z^{\infty} t\phi(t)\,dt\bigg\}\,dz - \int_{-\infty}^{0} f'(z)\bigg\{\int_{-\infty}^{z} t\phi(t)\,dt\bigg\}\,dz \\ &= \int_0^{\infty} t\phi(t)\bigg\{\int_0^{t} f'(z)\,dz\bigg\}\,dt - \int_{-\infty}^{0} t\phi(t)\bigg\{\int_t^{0} f'(z)\,dz\bigg\}\,dt \\ &= \int_0^{\infty} t\phi(t)\big[f(t)-f(0)\big]\,dt - \int_{-\infty}^{0} t\phi(t)\big[f(0)-f(t)\big]\,dt \\ &= \int_{-\infty}^{\infty} t\phi(t)f(t)\,dt.\end{aligned}$$

  The first equality follows from $\phi'(t) = -t\phi(t)$, and the second is by Fubini's theorem

- We can extend this result to cover a normal variate with arbitrary mean and variance, $X \sim N(\mu, \sigma^2)$. In this case, we claim that
$$\frac{1}{\sigma^2}\mathbb{E}[(X-\mu)f(X)] = \mathbb{E}[f'(X)].$$
The proof follows by defining $Z = (X-\mu)/\sigma \sim N(0,1)$, and $\tilde{f} = f(\sigma z + \mu)$, and then applying the previous result to Z and $\tilde{f}$

- Before we move on to the multivariate case and unbiased risk estimation, we can already see how remarkable this last result is. Suppose that $X \sim N(\mu, 1)$, where μ is unknown, and we had a (potentially) complicated function f, delivering an estimate of μ. Suppose further that we wanted to estimate $\mathrm{Cov}(X, f(X)) = \mathbb{E}[(X-\mu)f(X)]$. To get an unbiased estimate of this covariance, from the definition, we'd have to either know μ, which recall is unknown, or we'd have to know $\mathbb{E}[f(X)]$, which again, will generically depend on the unknown μ (not to mention that it may be potentially intractable). On the other hand, Stein's lemma gives us the simple unbiased estimate: $f'(X)$! This is free from μ, and in many cases it is possible to calculate: just take the derivative of our estimator and evaluate it at the data

1.2 Multivariate lemma

- Now, let $X \sim N(\mu, \sigma^2 I)$, an n-dimensional normal variate, with mean $\mu \in \mathbb{R}^n$ and spherical covariance matrix $\sigma^2 I \in \mathbb{R}^{n\times n}$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that, for each $i = 1, \dots n$ and almost every $x_{-i} \in \mathbb{R}^{n-1}$, the function
$$f(\cdot, x_{-i}) : \mathbb{R} \to \mathbb{R}$$
is absolutely continuous. (Here we write $x = (x_i, x_{-i})$ to decompose a point $x \in \mathbb{R}^n$ in terms of its ith component x_i, and all other components x_{-i}. Hence $f(\cdot, x_{-i})$ refers to f as a function of its ith argument, with all other arguments fixed at x_{-i}.) Stein calls such a function f *almost differentiable*

- Note that an almost differentiable function f has partial derivatives almost everywhere; we will denote the collection of these by $\nabla f = (\partial f/\partial x_1, \dots \partial f/\partial x_n)$

- Stein's multivariate result: with such an X, and almost differentiable f (satisfying $\|f(X)\|_2 < \infty$), we have
$$\frac{1}{\sigma^2}\mathbb{E}[(X-\mu)f(X)] = \mathbb{E}[\nabla f(X)].$$
Proof: we will assume that $X \sim N(0, I)$, and then a standardization argument, as before, will give the result for an arbitrary mean and marginal variance. Fix some i, and X_{-i}; then the function $f(\cdot, X_{-i})$ is univariate and we can apply Stein's univariate lemma. Hence, using the independence of X_i and X_{-i},
$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\partial f}{\partial x_i}(z, X_{-i})\phi(z)\, dz \\
&\qquad = \int_0^{\infty} \frac{\partial f}{\partial x_i}(z, X_{-i})\bigg\{\int_z^{\infty} t\phi(t)\, dt\bigg\}\, dz - \int_{-\infty}^{0} \frac{\partial f}{\partial x_i}(z, X_{-i})\bigg\{\int_{-\infty}^{z} t\phi(t)\, dt\bigg\}\, dz \\
&\qquad = \int_0^{\infty} t\phi(t)\bigg\{\int_0^{t} \frac{\partial f}{\partial x_i}(z, X_{-i})\, dz\bigg\}\, dt - \int_{-\infty}^{0} t\phi(t)\bigg\{\int_t^{0} \frac{\partial f}{\partial x_i}(z, X_{-i})\, dz\bigg\}\, dt \\
&\qquad = \int_0^{\infty} t\phi(t)\big[f(t, X_{-i}) - f(0, X_{-i})\big]\, dt - \int_{-\infty}^{0} t\phi(t)\big[f(0, X_{-i}) - f(t, X_{-i})\big]\, dt \\
&\qquad = \int_{-\infty}^{\infty} t\phi(t) f(t, X_{-i})\, dt.
\end{aligned}$$

In other words, we have shown that

$$\mathbb{E}\bigg[\frac{\partial f}{\partial x_i}(X)\,\bigg|\,X_{-i}\bigg] = \mathbb{E}\big[X_i f(X_i, X_{-i})|X_{-i}\big].$$

Taking an expectation over X_{-i} gives the result

- A final remark about the case $X \sim N(\mu, \sigma^2 I)$, and a function $f : \mathbb{R}^n \to \mathbb{R}^n$; note now this function returns outputs in $\mathbb{R}^n$. Write $f = (f_1, \ldots f_n)$ for the coordinate functions. We will say that f is almost differentiable provided that each one of its coordinate functions is. Then, by the last result, for each $i = 1, \ldots n$,

  $$\frac{1}{\sigma^2}\mathbb{E}[(X-\mu)f_i(X)] = \mathbb{E}[\nabla f_i(X)].$$

  Taking the ith equality in the above, and then summing over all $i = 1, \ldots n$ gives

  $$\frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{Cov}\big(X_i, f_i(X)\big) = \frac{1}{\sigma^2}\sum_{i=1}^n \mathbb{E}[(X_i-\mu_i)f_i(X)] = \mathbb{E}\bigg[\sum_{i=1}^n \frac{\partial f_i}{\partial x_i}(X)\bigg].$$

  This should be starting to look familiar ...

2 Stein's unbiased risk estimate

- Given samples $y \sim N(\mu, \sigma^2 I)$, and a function $\hat{\mu} : \mathbb{R}^n \to \mathbb{R}^n$. You can think about $\hat{\mu}$ as a fitting procedure that, from y, provides an estimate $\hat{\mu}(y)$ of the underlying (unknown) mean μ. For simplicity in what follows, we will use $\hat{\mu}$ both to refer to this estimate, and to the function itself

- Stein's unbiased risk estimate starts by expanding

  $$\begin{aligned}\mathbb{E}\|\mu - \hat{\mu}\|_2^2 &= \mathbb{E}\|\mu - y + y - \hat{\mu}\|_2^2 \\ &= n\sigma^2 + \mathbb{E}\|y - \hat{\mu}\|_2^2 + 2\mathbb{E}(\mu - y)^T\big(y - \hat{\mu}\big) \\ &= -n\sigma^2 + \mathbb{E}\|y - \hat{\mu}\|_2^2 + 2\sum_{i=1}^n \mathrm{Cov}(y_i, \hat{\mu}_i)\end{aligned}$$

  We have already seen this decomposition, just expressed a little differently. It says that the risk $R = \mathbb{E}\|\mu - \hat{\mu}\|_2^2$ of $\hat{\mu}$ satisfies

  $$R = -n\sigma^2 + \mathbb{E}\|y - \hat{\mu}\|_2^2 + 2\sigma^2 \mathrm{df}(\hat{\mu}),$$

  where $\mathbb{E}\|y - \hat{\mu}\|_2^2$ is the expected training error of $\hat{\mu}$, and recall that its degrees of freedom is defined as

  $$\mathrm{df}(\hat{\mu}) = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{Cov}(y_i, \hat{\mu}_i)$$

- Stein's lemma, as we discussed in the last section, provides an explicit estimate of the degrees of freedom term $\mathrm{df}(\hat{\mu})$, and therefore the risk R. In particular, we know that if $\hat{\mu}$ is almost differentiable as a function of y, then

  $$\hat{R} = -n\sigma^2 + \|y - \hat{\mu}\|_2^2 + 2\sigma^2\sum_{i=1}^n \frac{\partial \hat{\mu}_i}{\partial y_i}(y)$$

  is an unbiased estimate for R, i.e., $\mathbb{E}(\hat{R}) = R$. The above estimate $\hat{R}$ is what we call *Stein's unbiased risk estimate*, or *SURE*

- This can be an extremely useful tool. Aside from plainly estimating the risk of an estimator, we could also use it for model selection purposes: if our estimator depended on a tuning parameter $\lambda \in \Lambda$, denoted $\hat{\mu}_\lambda$, then we could choose this parameter to minimize SURE:

$$\hat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \ \|y - \hat{\mu}_\lambda\|_2^2 + 2\sigma^2 \sum_{i=1}^n \frac{\partial \hat{\mu}_{\lambda,i}}{\partial y_i}(y)$$

- Of course, in order for this to be useful, we need to figure out how to compute $\sum_{i=1}^n \partial \hat{\mu}_i / \partial y_i$ for the estimator $\hat{\mu}$ of interest (and, determine that $\hat{\mu}$ is almost differentiable so that Stein's lemma is applicable in the first place). This quantity is called the *divergence* of $\hat{\mu}$

- Furthermore, if we're minimizing SURE to choose a tuning parameter $\lambda \in \Lambda$, then we need some kind of concentration argument to show that the resulting parameter $\hat{\lambda}$ has good risk properties

- There is a considerable amount of classic literature that studies the minimization of a SURE-like risk estimate, for relatively simple procedures (such as linear smoothers) where divergence (or even exact degrees of freedom) is easily computable. Examples are: Li (1985), Li (1986), Li (1987), Johnstone (1986), Kneip (1994), Donoho & Johnstone (1995)

- Nowadays, with the more fancy estimators being of interest, it seems to be the trend to simply write papers about computing divergences. To give some examples: Efron et al. (2004), Zou et al. (2007), Tibshirani & Taylor (2011), Tibshirani & Taylor (2012) show how to compute divergences for lasso and generalized lasso estimators; Meyer & Woodroofe (2000) show how to compute divergences for convex-constrained regression estimators; Mukherjee et al. (2012) show how to compute divergences for reduced rank regression; Candes et al. (2012), Deledalle et al. (2012) show how to compute divergences for singular value thresholding ... An exception is a recent paper by Xie et al. (2012), who not only compute divergences, but also show a kind of uniform consistency property, for SURE in a hierarchical model

- In the last two sections here, we will walk through how to compute the divergence of the lasso estimator, and then we will show how concentration arguments can be used provide risk bounds for a special case: SURE-tuned thresholding estimates. In the next section, we show how Stein's unbiased risk estimate can be used to prove a very surprising but fundamental result about shrinkage and inadmissibility

3 Stein's paradox

- Let $X_1, \dots X_d$ be independent normal variates with unit variance, and means $\mu_1, \dots \mu_d$, respectively. Written differently, let $X \sim N(\mu, I)$, where $X = (X_1, \dots X_d)$. Let $\hat{\mu} = \hat{\mu}(X)$ be an estimator of μ. Recall that we say that another estimator $\tilde{\mu} = \tilde{\mu}(X)$ *strictly dominates* $\hat{\mu}$ (with respect to squared loss) if

$$\begin{aligned} \mathbb{E}|\mu - \tilde{\mu}\|_2^2 &\leq \mathbb{E}|\mu - \hat{\mu}\|_2^2 \quad \text{for all } \theta\text{, and} \\ \mathbb{E}|\mu - \tilde{\mu}\|_2^2 &< \mathbb{E}|\mu - \hat{\mu}\|_2^2 \quad \text{for some } \theta. \end{aligned}$$

In this case, we say that $\hat{\theta}$ is *inadmissible*. If no such other estimator $\tilde{\mu}$ exists, we say that $\hat{\theta}$ is *admissible*

- In our setup, the most natural estimator of μ is $\hat{\mu}_0(X) = X$. After all, this is the maximum likelihood estimator, and the unbiased estimator with the minimum variance. You may think that it would also be admissible. In fact, some folks proved this for the special case $d = 1$ in the 1950's. Stein (1956) showed that the same was true for $d = 2$. But in this paper, Stein shocked the statistics community when he showed that the identity estimator $\hat{\mu}_0$ was actually inadmissible when $d \geq 3$! This is known as *Stein's paradox*

- James & Stein (1961) provided an explicit estimator $\hat{\mu}^{\text{JS}}$ that strictly dominates $\hat{\mu}_0$, which is by now famous and known as the James-Stein shrinkage estimator. It is defined as

$$\hat{\mu}^{\text{JS}} = \left(1 - \frac{d-2}{\|X\|_2^2}\right) X.$$

  This can be viewed as taking the natural estimator $\hat{\mu}_0(X) = X$, and shrinking its components toward 0

- Note that, for each component i, the James-Stein estimator pools together the information in all of $X_1, \ldots X_d$ to form the estimate $\hat{\mu}_i^{\text{JS}}$ of the underlying mean μ_i. This is a surprising realization, because $X_1, \ldots X_d$ were independent. To give an example, suppose that we were estimating: the mean number of Justin Bieber records sold in the Bahamas, the mean profit Trader Joe's makes from its almond butter, and the mean number of deep learning papers appearing on arXiv each month. Why should our estimate for the number of Bieber albums sold have anything to do with how much Trader Joe's charges for its almond butter, or how many deep learning papers are written?

- Counterintuitive as it may seem, at its heart the James-Stein estimator is a shrinkage device to reduce variance at the expense of introducing a little bias. What is remarkable, though, is that this tradeoff is so elegantly navigated that in the end the James-Stein estimator strictly dominates the identity estimator

- There is an empirical Bayes interpretation of the James-Stein estimator, where we place a prior $\mu \sim N(0, \tau^2 I)$ on the underlying mean, and estimate τ from the observed data X. Some people say that this perspective is misleading, since the prior encodes some similarity in the mean components (they share the same marginal variance) but the original paradox holds in a frequentist setting where the means are fixed and completely unrelated

- Variants of the usual James-Stein estimator: we do not need to shrink towards zero; we can actually shrink towards any fixed $\mu_0 \in \mathbb{R}^d$, as in $\mu_0 + (1 - (d-2)/\|X - \mu_0\|_2^2) \cdot (X - \mu_0)$, and this would still strictly dominate the identity estimator. Moreover, we can shrink towards the sample mean $\bar{X} = \frac{1}{d}\sum_{i=1}^d X_i$, and still strictly dominate the identity estimator, but here we would replace $d-2$ by $d-3$, and would hence require $d \geq 4$. Finally, we can just take the positive part of the shrinkage factor $(1 - (d-2)/\|X\|^2)_+$, i.e., truncate it at zero if it were to go negative, and this positive-part James-Stein estimator actually strictly dominates the James-Stein estimator (but is still inadmissible itself!)

- We will not go into many more details about the James-Stein estimator and Stein's paradox, but it is a fascinating topic and there is much supporting and related literature. A classic, nontechnical reference is Efron & Morris (1977); another friendly reference is Samworth (2012)

- But, we will prove that the James-Stein estimator strictly dominates the natural estimator. How? With SURE! First note that the risk of the identity estimator is

$$\mathbb{E}\|\hat{\mu}_0 - \mu\|_2^2 = \mathbb{E}\|X - \mu\|_2^2 = d.$$

  Now let's form Stein's unbiased risk estimate for $\hat{\mu}^{\text{JS}}$:

$$\hat{R} = -d + \left\| \left(1 - \frac{d-2}{\|X\|_2^2}\right) X - X \right\|_2^2 + 2 \cdot \sum_{i=1}^d \frac{\partial \hat{\mu}_i^{\text{JS}}}{\partial X_i}.$$

The middle term, i.e., the expected training error, is

$$\left\| \frac{d-2}{\|X\|_2^2} X \right\|_2^2 = \frac{(d-2)^2}{\|X\|_2^2}.$$

For the last term, i.e., the divergence term, we compute

$$\frac{\partial \hat{\mu}_i^{\mathrm{JS}}}{\partial X_i} = 1 - \frac{d-2}{\|X\|_2^2} + \frac{d-2}{\|X\|_2^4} 2X_i \cdot X_i,$$

and therefore

$$\sum_{i=1}^d \frac{\partial \hat{\mu}_i^{\mathrm{JS}}}{\partial X_i} = d - \frac{d(d-2)}{\|X\|_2^2} + \frac{2(d-2)}{\|X\|_2^2} = d - \frac{(d-2)^2}{\|X\|_2^2}.$$

Adding these together, we get

$$\hat{R} = -d + \frac{(d-2)^2}{\|X\|_2^2} + 2d - 2\frac{(d-2)^2}{\|X\|_2^2} = d - \frac{(d-2)^2}{\|X\|_2^2},$$

and finally, the risk of the James-Stein estimate is

$$\mathbb{E}(\hat{R}) = d - (d-2)^2 \cdot \mathbb{E}\left(\frac{1}{\|X\|_2^2} \right) < d$$

4 Lasso degrees of freedom

- In this section we show how to use Stein's lemma to compute the divergence, i.e., the degrees of freedom, of the lasso estimator. Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, an outcome vector and predictor matrix, the lasso problem is
$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1,$$
where $\lambda \geq 0$ is a tuning parameter. We will assume that X has columns in general position, which recall, implies that the lasso solution $\hat{\beta}$ is unique

- We will assume that $y \sim N(\mu, \sigma^2 I)$, and consider the lasso fit $\hat{\mu} = X\hat{\beta}$ as an estimate of the underlying mean μ. We will verify that $X\hat{\beta}$ is almost differentiable as a function of y, and then use Stein's lemma to compute its divergence. This yields an unbiased estimate for the degrees of freedom of the lasso fit, and hence its risk. Some mathematical details are skipped in what follows; see Tibshirani & Taylor (2012)

4.1 The lasso dual to the rescue

- The first thing we must do is to check that the lasso fit $X\hat{\beta}$, is almost differentiable. This is actually nonobvious and nontrivial

- E.g., is the lasso fit even continuous in y? Recall that we write $A = \mathrm{supp}(\hat{\beta})$ for the active set of the lasso solution, and correspondingly, the fit is $X\hat{\beta} = X_A\hat{\beta}_A$ (where X_A indexes the columns of X in A, and $\hat{\beta}_A$ the components of $\hat{\beta}$ in A). As we change y, the active set A can change, and so you might think that the lasso fit can jump discontinuously at this point. Certainly this would happen with best subset selection: if we change y, and the active set A changes, then the best subset selection fit—which is just given by least squares regression on X_A—would change discontinuously. With the lasso, we can express the fit as
$$X\hat{\beta} = X_A(X_A^T X_A)^{-1}(X_A^T y - \lambda s_A),$$

where $s_A = \mathrm{sign}(\hat{\beta}_A)$ are the signs of active lasso coefficients. This is a sum of two terms; the first term $X_A(X_A^T X_A)^{-1} X_A^T y$ is exactly the least squares fit of y on X_A, and the second term $\lambda X_A(X_A^T X_A)^{-1} s_A$ is a shrinkage term, a result of using an ℓ_1 penalty. Now, if this shrinkage term happened to be defined "just right", then as we change y and change the active set A, perhaps the coefficients of variables to leave the active set would gracefully and continosly drop to zero, and coefficients of variables to enter the active set would continuously move form zero. This would make whole the lasso fit $X\hat{\beta} = X_A \hat{\beta}_A$ continuous

- Fortuitously, this is indeed the case, and so the lasso fit is continuous at every y. It might seem a daunting task to prove this, but the lasso dual provides a very simple proof. The dual perspective even proves that the lasso fit is Lipschitz continuous, which makes it almost differentiable according to Stein's notion, taking care of the regularity conditions we need for Stein's lemma

- Recall that the lasso dual is

$$\min_{u \in \mathbb{R}^n} \ \|y - u\|_2^2 \ \text{ subject to } \ \|X^T u\|_\infty \le \lambda,$$

and the relationship between the dual $\hat{u}$ and primal $\hat{\beta}$ solutions is

$$X\hat{\beta} = y - \hat{u}.$$

Note that the dual solution $\hat{u}$ is nothing more than the projection of y onto the convex polyhedron $C = \{u : \|X^T u\|_\infty \le \lambda\}$, which we will denote by $\hat{u} = P_C(y)$. Therefore, the lasso fit is simply

$$X\hat{\beta} = (I - P_C)(y),$$

the residual from projecting y onto C. A picture to show this (just look at the left panel for now) is given in Figure 1

- The projection onto any convex set is a nonexpansive mapping. In other words, $\|P_C(y) - P_C(y')\|_2 \le \|y - y'\|_2$ for any y, y'. This should be visually clear from the picture. Actually, the same is true with the residual map: $I - P_C$ is also nonexpansive, and hence the lasso fit is a Lipschitz continuous function (with Lipschitz constant 1). This implies that it meets our requirement of almost differentiability (a result sometimes called Rademacher's theorem)

4.2 More from the polyhedral perspective

- Viewing the lasso fit as the residual from projection onto a convex polyhedron, as developed in above, is actually an even more fruitful perspective. Write this polyhedron as

$$C = (X^T)^{-1}\{v : \|v\|_\infty \le \lambda\},$$

where $(X^T)^{-1}$ denotes the preimage operator under the linear map X^T. The set $\{v : \|v\|_\infty \le \lambda\}$ is a hypercube in a $\mathbb{R}^p$. Each face of this cube is determined by a subset $A \subseteq \{1, \ldots p\}$ of dimensions (that achieve the maximum value $|\lambda|$) and signs $s_A \in \{-1, 1\}^{|A|}$ (that tell which side of the cube the face will lie on, for each dimension). Faces of C are just faces of $\{v : \|v\|_\infty \le \lambda\}$ run through the inverse image transformation, and so this means that each face of C is also indexed by a set $A \subseteq \{1, \ldots p\}$ and signs $s_A \in \{-1, 1\}^{|A|}$. The picture above attempts to convey this relationship with the colored black face in each of the panels

- Now imagine projecting y onto C; it will land on some face. We have just argued that this face corresponds to a set A and signs s_A. One can show that this set A is exactly the active set of the lasso solution at y, and s_A are exactly the active signs

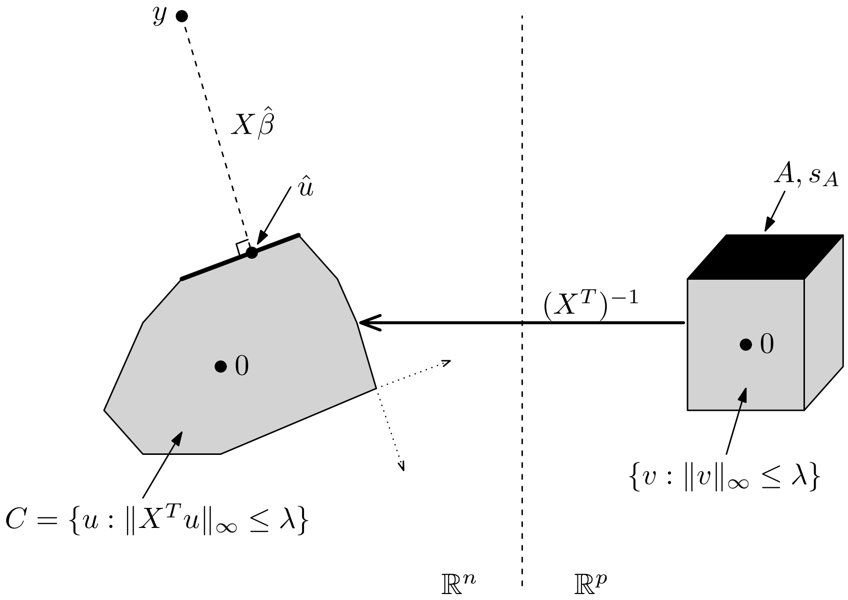

Figure 1: *An illustration of the primal-dual relationship for the lasso problem.*

- Looking at the picture: we can that see that as we wiggle y around, it will project to the same face. From the correspondence between faces and active set and signs of lasso solutions, this means that A, s_A do not change as we perturb y, i.e., they are locally constant
- But this isn't true for all points y, e.g., if y lies on one of the rays emanating from the lower right corner of the polyhedron in the picture, then we can see that small perturbations of y do actually change the face that it projects to, which invariably changes the active set and signs of the lasso solution. However, this is somewhat of an exceptional case, in that such points can be shown to form a set of Lebesgue measure zero, and therefore we can assure ourselves that the active set and signs A, s_A are locally constant almost everywhere in y
- Finally, let's re-examine the lasso fit

  $$X\hat{\beta} = X_A(X_A^T X_A)^{-1}(X_A^T y - \lambda s_A).$$

  We have reasoned that for almost every y, the pair A, s_A is locally constant, and therefore the divergence of the lasso fit is simply given by

  $$\sum_{i=1}^n \frac{\partial (X\hat{\beta})_i}{\partial y_i} = \mathrm{tr}\Big(X_A(X_A^T X_A)^{-1} X_A^T\Big) = |A|,$$

  almost everywhere. Stein's lemma tells us that this means that lasso degrees of freedom is

  $$\mathrm{df}(X\hat{\beta}) = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{Cov}\big(y_i, (X\hat{\beta})_i\big) = \mathbb{E}|A|.$$

  Accordingly, an unbiased estimate for its risk is

  $$\hat{R} = -n\sigma^2 + \|y - X\hat{\beta}\|_2^2 + 2\sigma^2|A|$$

4.3 Wait, what? That seems wrong

- To repeat the conclusion of the last subsection, in words, the degrees of freedom of the lasso fit is the number of selected variables, in expectation. This seems like it can't be true, as the lasso uses the data to choose which variables to put in the model, and, least squares regression on a fixed subset of variables already has degrees of freedom equal to the number of variables. So how can the degrees of freedom of the lasso fit be equal to the (average) number of selected variables, and not more?
- The key realization is that the lasso shrinks the coefficients of the selected variables towards zero, instead of perfoming a full least squares fit. We have seen this already in the expression for the lasso fitted values. Remarkably, the "surplus" from adaptively building the model is exactly accounted for by the "deficit" from shrinking the coefficients, so that altogether (in expectation), the degrees of freedom is simply the number of variables in the model
- Mathematically, it was the shrinkage that was responsible for the continuity, and ultimately, the almost differentiability of the lasso fit, which is what we needed in order to apply Stein's lemma
- What about variable selection procedures that do not employ shrinkage, such as best subset selection? Intuitively, it seems like their degrees of freedom should be much larger than the (expected) number of selected variables. But studying this rigorously is a bit difficult because Stein's lemma does not generically apply here; these estimators are typically discontinuous in y. For a modest (and not completely useable) extension of Stein's lemma, in an attempt to study such estimators, see Tibshirani (2014)

5 Risk of SURE-tuned thresholding

5.1 Uniform deviation bound for SURE

- Let's consider a special case of the lasso problem, when $X = I$:

$$\min_{\beta} \frac{1}{2}\|y-\beta\|_2^2 + \lambda\|\beta\|_1$$

In this case, as we've seen before, the lasso fit is explicitly given by soft-thresholding at the level λ, i.e.,

$$\hat{\beta}_{\lambda,i} = S_\lambda(y_i) = \begin{cases} y_i - \lambda & \text{if } y_i > \lambda \\ 0 & \text{if } y_i \in [-\lambda, \lambda] \\ y_i + \lambda & \text{if } y_i < \lambda \end{cases}, \quad i = 1, \ldots n.$$

- We follow Donoho & Johnstone (1995). Assuming that $\sigma^2 = 1$ for simplicity, note that SURE becomes

$$\begin{aligned} \hat{R}(\lambda) &= -n + \|y - \hat{\beta}_\lambda\|_2^2 + 2|A_\lambda| \\ &= n + \|y - \hat{\beta}_\lambda\|_2^2 - 2(n - |A_\lambda|) \\ &= n + \sum_{i=1}^n \Big(\min\{y_i^2, \lambda^2\} - 2 \cdot 1\{y_i^2 \le \lambda^2\} \Big) \end{aligned}$$

It will be useful to instead consider a different scaling,

$$\hat{U}(\lambda) = \frac{\hat{R}(\lambda)}{n} = \frac{1}{n}\sum_{i=1}^n \Big(1 + \min\{y_i^2, \lambda^2\} - 2 \cdot 1\{y_i^2 \le \lambda^2\} \Big)$$

- Now with $r(\lambda) = R(\lambda)/n = \mathbb{E}\|\mu - \hat{\beta}_\lambda\|_2^2/n$ denoting the scaled true risk, inspect

$$\begin{aligned}\hat{U}(\lambda) - r(\lambda) &= \frac{1}{n}\sum_{i=1}^n \Big(1 + \min\{y_i^2, \lambda^2\} - 2 \cdot 1\{y_i^2 \le \lambda^2\}\Big) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}(\mu_i - \hat{\beta}_{\lambda,i})^2 \\ &= \frac{1}{n}\sum_{i=1}^n W_i(\lambda)\end{aligned}$$

  where the summands $W_i(\lambda)$, $i = 1, \ldots n$ are independent and have mean zero. This comes from the fact that the estimation problem here competely decomposes over $i = 1, \ldots n$, and the unbiased property of SURE

- We will prove below that, uniformly over all $\mu \in \mathbb{R}^n$,

$$\max_{\lambda \in [0,\lambda_0]} \big|\hat{U}(\lambda) - r(\lambda)\big| = O_{\mathbb{P}}\Big(\frac{\log^{3/2} n}{n^{1/2}}\Big),$$

  where $\lambda_0 = \sqrt{2 \log n}$. Therefore if SURE happens to pick $\hat{\lambda} \in [0, \lambda_0]$, then we can be pretty confident (sure! no pun intended) of the risk $r(\hat{\lambda})$ at the selected threshold level

- Interestingly, the SURE rule actually does pretty poorly in the very sparse regime, when the selected value $\hat{\lambda}$ should be close to λ_0. Donoho & Johnstone (1995) hence propose a hybrid rule, which essentially tests for this very sparse regime: it uses a global threshold $\lambda_0 = \sqrt{2 \log n}$ if this passes, else it lets SURE select the threshold $\hat{\lambda}$. The resulting procedure, called *SureShrink*, is used to threshold wavelet coefficients, and this has a provable minimax optimal error rate, with a completely data adaptive threshold

5.2 Proof of uniform deviation bound

- Now we prove the uniform deviation result. Observe the simple bounds

$$\big|1 + \min\{y_i^2, \lambda^2\} - 2 \cdot 1\{y_i^2 \le \lambda^2\}\big| \le 1 + \lambda^2$$

  and

$$\begin{aligned}\mathbb{E}(\mu_i - \hat{\beta}_{\lambda,i})^2 &= \mathbb{E}\big(\mu_i - S_\lambda(y_i)\big)^2 \\ &= \mathbb{E}\big(\mu_i - y_i + y_i - S_\lambda(y_i)\big)^2 \\ &\le 2\mathbb{E}(\mu_i - y_i)^2 + 2\mathbb{E}\big(y_i - S_\lambda(y_i)\big)^2 \\ &= 2 + 2\mathbb{E}\Big(y_i^2 \cdot 1\{y_i^2 \le \lambda^2\} + \lambda^2 \cdot 1\{y_i^2 > \lambda^2\}\Big) \\ &\le 2(1 + \lambda^2)\end{aligned}$$

  Therefore

$$|W_i(\lambda)| = \big|1 + \min\{y_i^2, \lambda^2\} - 2 \cdot 1\{y_i^2 \le \lambda^2\} - \mathbb{E}(\mu_i - \hat{\beta}_{\lambda,i})^2\big| \le 3(1 + \lambda^2),$$

  holding for each $i = 1, \ldots n$

- A sum of mean zero, bounded random variables ... so let's apply Hoeffding's inequality to the quantity $\hat{U}(\lambda) - r(\lambda) = \sum_{i=1}^n W_i(\lambda)/n$, giving

$$\mathbb{P}\Big(\big|\hat{U}(\lambda) - r(\lambda)\big| > \frac{\epsilon}{\sqrt{n}}\Big) \le 2\exp\Big(-\frac{2\epsilon^2}{9(1+\lambda^2)^2}\Big)$$

  where $\epsilon > 1$ is arbitrary, for now

- For distinct $\lambda < \lambda'$, we let $N(\lambda, \lambda') = \#\{i : \lambda^2 < y_i^2 \le \lambda'^2\}$, and bound

$$\begin{aligned}\left|\hat{U}(\lambda) - \hat{U}(\lambda')\right| &= \left| \frac{1}{n} \sum_{i=1}^{n} \Big(2 \cdot 1\{\lambda^2 < y_i^2 \le \lambda'^2\} + \min\{y_i^2, \lambda^2\} - \min\{y_i^2, \lambda'^2\} \Big) \right| \\ &\le \frac{2}{n} N(\lambda, \lambda') + \lambda'^2 - \lambda^2.\end{aligned}$$

  Furthermore, a tedious but not particularly interesting calculation involving integrals shows that $|\frac{d}{d\lambda} r(\lambda)| \le 5\lambda$. Therefore, restricting our attention to $\lambda, \lambda' \in [0, \lambda_0]$, with $\delta = \lambda' - \lambda > 0$, we have

$$\begin{aligned}\left|\hat{U}(\lambda) - r(\lambda)\right| &\le \frac{2}{n} N(\lambda, \lambda') + (\lambda' - \lambda)(\lambda' + \lambda) + 5\lambda_0(\lambda' - \lambda) \\ &\le \frac{2}{n} N(\lambda, \lambda') + 7\delta\lambda_0\end{aligned}$$

- Set $\lambda_j = j\delta$, across $j = 1, 2, 3, \dots$ such that $\lambda_j \in [0, \lambda_0]$. Then

$$A = \Big\{ \max_{\lambda \in [0, \lambda_0]} \left|\hat{U}(\lambda) - r(\lambda)\right| \ge \frac{3\epsilon}{\sqrt{n}} \Big\} \subseteq D \cup E,$$

  where

$$D = \Big\{ \max_{j} \left|\hat{U}(\lambda) - r(\lambda)\right| \ge \frac{\epsilon}{\sqrt{n}} \Big\}$$

  and

$$E = \Big\{ \max_{j} \max_{|\lambda - \lambda_j| \le \delta} \left|\hat{U}(\lambda) - r(\lambda)\right| \ge \frac{2\epsilon}{\sqrt{n}} \Big\}$$

- Choosing δ so that $\delta\lambda_0 = o(1/\sqrt{n})$, we see that, for large enough n,

$$E \subseteq E' = \Big\{ \max_{j} \frac{2}{n} N(\lambda_j, \lambda_j + \delta) \ge \frac{\epsilon}{\sqrt{n}} \Big\}.$$

  Bound $\mathbb{E}N(\lambda_j, \lambda_j + \delta) = O(\delta n)$, so that

$$E' \subseteq E'' = \Big\{ \max_{j} \frac{1}{n} \Big(N(\lambda_j, \lambda_j + \delta) - \mathbb{E}N(\lambda_j, \lambda_j + \delta) \Big) \ge \frac{\epsilon}{3\sqrt{n}} \Big\}$$

  where we again used the fact that $\delta = o(1/\sqrt{n})$

- Another application of Hoeffding's inequality gives us

$$\mathbb{P}\Big(\frac{1}{n} \Big| N(\lambda_j, \lambda_j + \delta) - \mathbb{E}N(\lambda_j, \lambda_j + \delta) \Big| \ge \frac{\epsilon}{3\sqrt{n}} \Big) \le 2\exp(-2\epsilon^2/9)$$

- Finally, using the cardinality of the set $\{j : t_j = j\delta \in [0, \lambda_0]\}$ and our two Hoeffding bounds,

$$\begin{aligned}P(A) &\le P(D) + P(E'') \\ &\le \frac{2\lambda_0}{\delta} \bigg(\exp\Big(-\frac{2\epsilon^2}{9(1 + \lambda_0^2)^2} \Big) + \exp(-2\epsilon^2/9) \bigg)\end{aligned}$$

  Choose $\epsilon = s\sqrt{9/2 \cdot \log n}(1 + \lambda_0^2)$. Then

$$P(A) \le \frac{4\lambda_0}{\delta} n^{-s^2}$$

- Rephrased, this says that

$$\mathbb{P}\bigg(\max_{\lambda \in [0, \lambda_0]} \frac{\sqrt{2n}}{\log n(1 + \log^2 n)} \left|\hat{U}(\lambda) - r(\lambda)\right| \ge s \bigg) \le \frac{4\lambda_0}{\delta} n^{-s^2}$$

  proving the desired result, taking, e.g., $\delta = n^{-1/4}$

References

Candes, E. J., Sing-Long, C. & Trzasko, J. (2012), Unbiased risk estimates for singular value thresholding and spectral estimators. arXiv: 1210.4139.

Deledalle, C.-A., Vaiter, S., Peyre, G., Fadili, J. & Dossal, C. (2012), Risk estimation for matrix recovery with spectral regularization. arXiv: 1205.1482.

Donoho, D. & Johnstone, I. (1995), 'Adapting to unknown smoothness via wavelet shrinkage', *Journal of the American Statistical Association* **90**(432), 1200–1224.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**(2), 407–499.

Efron, B. & Morris, C. (1977), 'Stein's paradox in statistics', *Scientific American* **236**, 119–127.

James, W. & Stein, C. (1961), 'Estimation with quadratic loss', *Proceeds of the Fourth Berkeley Symposium* **1**, 361–380.

Johnstone, I. (1986), On imadmissibility of some unbiased estimates of loss. Technical Report, Stanford University.

Kneip, A. (1994), 'Ordered linear smoothers', *Annals of Statistics* **22**(2), 835–866.

Li, K.-C. (1985), 'From Stein's unbiased risk estimates to the method of generalized cross-validation', *Annals of Statistics* **14**(4), 1352–1377.

Li, K.-C. (1986), 'Asymptotic optimality of c_l and generalized cross-validation in ridge regression with application to spline smoothing', *Annals of Statistics* **14**(3), 1101–1112.

Li, K.-C. (1987), 'Asymptotic optimailty for c_p, c_l, cross-validation and generalized cross-validation: discrete index set', *Annals of Statistics* **15**(3), 958–975.

Meyer, M. & Woodroofe, M. (2000), 'On the degrees of freedom in shape-restricted regression', *Annals of Statistics* **28**(4), 1083–1104.

Mukherjee, A., Wang, N. & Zhu, J. (2012), Degrees of freedom of the reduced rank regression. arXiv: 1210.2464.

Samworth, R. (2012), 'Stein's paradox', *Eureka* **62**, 38–41.

Stein, C. (1956), 'Inadmissibility of the usual estimator of the mean of a multivariate normal distribution', *Proceedings of the Third Berkeley Symposium* **1**, 197–206.

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *Annals of Statistics* **9**(6), 1135–1151.

Tibshirani, R. J. (2014), Degrees of freedom and model search. arXiv: 1402.1920.

Tibshirani, R. J. & Taylor, J. (2011), 'The solution path of the generalized lasso', *Annals of Statistics* **39**(3), 1335–1371.

Tibshirani, R. J. & Taylor, J. (2012), 'Degrees of freedom in lasso problems', *Annals of Statistics* **40**(2), 1198–1232.

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

Zou, H., Hastie, T. & Tibshirani, R. (2007), 'On the "degrees of freedom" of the lasso', *Annals of Statistics* **35**(5), 2173–2192.