28 A taste of nonparametrics, shrinkage and the James-Stein estimator

In this chapter, we will be concerned with regression problems, although similar reasonings extend to problems such as density estimation. Assume we observe data pairs $(u_1, y_1), \ldots, (u_N, y_N)$ whose relationship is determined by an unknown function f and noise ϵ, i.e.,

$$\mathsf{y}_n = f(u_n) + \epsilon_n, \qquad \epsilon_n \sim \mathtt{N}(0, \sigma^2). \tag{1}$$

To predict values y, we estimate the function f and express it via a basis or dictionary of functions ϕ_j:

$$\hat{f}(u) = \sum_{j=1}^{m} \phi_j(u)\hat{x}_j. \tag{2}$$

So, we actually estimate the weights $\hat{x}_j$. The functions ϕ_j can take many forms, for example:

- coordinates: if $\mathbf{u} \in \mathbb{R}^m$ is a vector, then $\phi_j(\mathbf{u}) = u^j$ may be the jth coordinate of $\mathbf{u}$.
- indicator functions: $\phi_j(u) = \mathbb{1}_{[a_j \le u \le b_j]}$ for some constants a_j, b_j. In this case, f will be piecewise constant.
- cosine basis: $\phi_0(u) = 1$ and $\phi_j(u) = \sqrt{2}\cos(2\pi j u)$ for $j > 1$.
- polynomials of increasing degree (e.g., Legendre polynomials)
- wavelets
- ...

This representation suggests that, depending on the ϕ_j, we can easily represent non-linear functions, and the resulting function is still linear in the $\hat{x}_j$.

Least-squares estimator. Our cost function will be least squares; this is equivalent to the log likelohood under the Gaussian noise model. The resulting least squares estimator (or MLE) $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_m)$ minimizes

$$\frac{1}{N}\sum_{n=1}^{N}(\hat{\mathsf{f}}(u_n) - \mathsf{y}_n)^2 = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=1}^{m}\phi_j(u_n)\hat{\mathsf{x}}_j - \mathsf{y}_n\right)^2. \tag{3}$$

We can rewrite this objective function in terms of the vector $\hat{\mathbf{x}}$ and the $n \times m$ matrix Φ, where the nth row of Φ is the vector $(\phi_1(u_n) \dots \phi_m(u_n))$:

$$\frac{1}{N}\sum_{n=1}^{N}(\hat{f}(u_n) - y_n)^2 = (\Phi\hat{\mathbf{x}} - \mathbf{y})^\top(\Phi\hat{\mathbf{x}} - \mathbf{y}). \tag{4}$$

This is a quadratic convex function, and we can minimize it by setting its derivative with respect to $\hat{\mathbf{x}}$ to zero. Doing so results in the equation

$$2\Phi^\top\Phi\hat{\mathbf{x}} - 2\Phi^\top\mathbf{y} = 0. \tag{5}$$

If $\Phi^\top\Phi$ is invertible (that means it has full rank m), then we can solve this as

$$\hat{\mathbf{x}}^{\mathrm{LS}} = \hat{\mathbf{x}}^{\mathrm{LS}}(\mathbf{y}, \Phi) = (\Phi^\top\Phi)^{-1}\Phi^\top\mathbf{y}. \tag{6}$$

Large m. We have not yet considered the importance of m. If the ϕ_j form a basis of a function space (we will see more of this in a bit), then we can express any function in the space if we just make m large enough. Hence, by the criterion (3), making m very large is beneficial: we can then express any function, easily interpolate between our observations y_n and make the objective function (3) arbitrarily close to zero.

What is the effect if m is large? Looking for example at the cosine basis, we see that the larger j becomes the more "wiggly" and non-smooth our function can become: it can fluctuate highly within a small neighborhood. Intuitively, for estimating such a function accurately, we will need more than one observation within any such small neighborhood. In fact, our derivation above shows this problem even more directly: we assumed we can invert the $m \times m$ matrix $\Phi^\top\Phi$. If $N < m$, then this matrix can at most have rank N and is not invertible. This means there are many different solutions $\hat{\mathbf{x}}$ that satisfy Equation (5). Or, in other words, there are many functions that exactly interpolate our few observations. Which one is the one to pick, if any at all?

The problem of large m can arise even with the coordinate functions $\phi_j(\mathbf{u}) = u^j$. For example, in a real-world estimation problem, $\mathbf{u}$ may be a vector of thousands of genes, and we observe those genes only under a few hundred (N) conditions.

Intuitively, if there are many possible functions interpolating our observations, the "simplest" one seems the most plausible. Can we quantify this?

28.1 The bias-variance tradeoff

The problem with the mean-squared error (3) is that it is only a proxy. What we actually would like to minimize is the predictive risk. Define the point-wise risk at a given point u as

$$R(f(u), \hat{f}(u)) = \mathbb{E}[(f(u) - \hat{f}(u))^2] \tag{7}$$

$$= \underbrace{(\mathbb{E}[\hat{f}(u)] - f(u))^2}_{\mathrm{bias}_u^2} + \underbrace{\mathbb{E}[(\hat{f}(u) - \mathbb{E}[\hat{f}(u)])^2]}_{\mathrm{var}_u^2}. \tag{8}$$

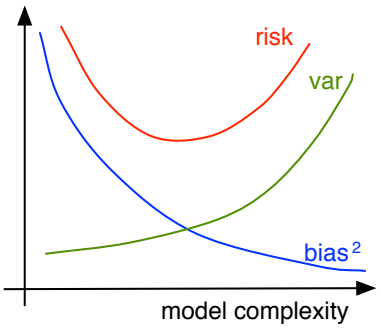

Figure 1: Bias-variance tradeoff (schematic)

For N new observations (u'_n, y'_n), we define the *predictive risk* as

$$\mathbb{E}[\tfrac{1}{N}\sum_{n=1}^{N}(y'_n - \hat{f}(u'_n))^2] = \mathbb{E}[\tfrac{1}{N}\sum_{n=1}^{N}(f(u'_n) + \epsilon'_n - \hat{f}(u'_n))^2] \tag{9}$$

$$= R(f, \hat{f}) + \mathbb{E}[\epsilon'_n] \tag{10}$$

$$= \text{bias}^2(\hat{f}) + \text{var}(\hat{f}) + \sigma^2, \tag{11}$$

where $R(f, \hat{f})$ is the average (or integrated) risk and σ^2 the variance of the noise. This means our actual objective, the predictive risk, has three components. An irreducible error σ^2 that is due to noise. The other terms, the bias and variance, are influenced by the model complexity m. As Figure 1 shows, they behave in opposite ways: if m is small, our model class may not be able to express the true underlying f, and no estimator $\hat{f}$ can be unbiased. As m increases, we can express a larger set of functions. If f is in our model class, then the least squares estimator is unbiased, and has minimum variance among all estimators. On the other hand, if we have a lot of flexibility in fitting $\hat{f}$, our estimate can very well adjust to noise and can therefore will have high variance. Indeed, we may "overfit" to the noise.

To minimize the predictive risk, we hence need to trade off bias and variance. We will admit some bias to reduce the variance and thereby the risk.

28.2 Penalized (regularized) least-squares

To trade off bias and variance, we add a penalty to the mean square error, and obtain the new objective

$$\min \quad \frac{1}{N}\sum_{n=1}^{N}(\hat{f}(u_n) - y_n)^2 + \lambda\Omega(\hat{f}). \tag{12}$$

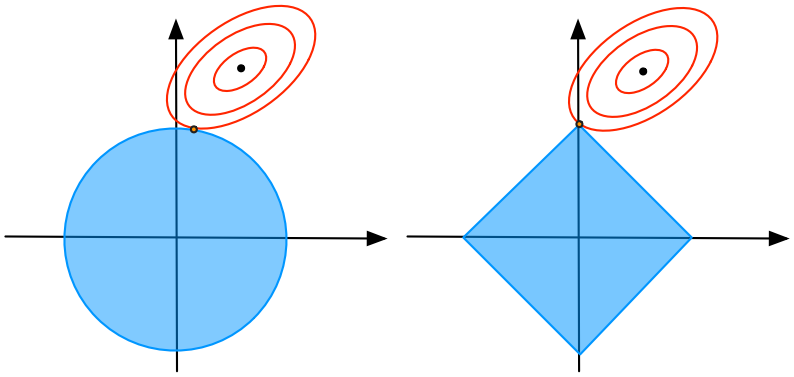

Figure 2: Geometry of penalized least squares (in $m = 2$ dimensions, the axes are $\hat{x}_1$ and $\hat{x}_2$). The red contours indicate sets with equal loss $\frac{1}{N}\sum_{n=1}^{N}(\hat{f}(u_n)-y_n)^2$. The unconstrained minimizer $\hat{\mathbf{x}}$ lies in the center of the ellipses. The blue regions indicate the feasible sets $\sum_j \hat{x}_j^2 \leq c$ (left) and $\sum_j |\hat{x}_j| \leq c$ (right). The penalized estimator is at the intersection of the blue region with the outmost contour line. In both cases, the coordinate values are shrunk. For Lasso (left), the intersection is at a vertex of the polytope (which is at the axis). This means some coordinates are shrunk to zero. (Figure adapted from [Hastie et al., 2009].)

The regularizer $\Omega(\hat{f})$ penalizes the smoothness of $\hat{f}$. Here, we will define the penalty on x directly:

$$\Omega_2(\hat{f}) = \sum\nolimits_j \hat{x}_j^2 = \|\hat{x}\|_2^2 \quad \text{or} \quad \Omega_1(\hat{f}) = \sum\nolimits_j |\hat{x}|_j = \|\hat{x}\|_1. \tag{13}$$

The left version is *ridge regression* and the right version the *Lasso*.

To understand the effect of these penalties, we view the function (12) as the Lagrangian of a constrained optimization problem:

$$\min \quad \frac{1}{N}\sum_{n=1}^{N}(\hat{f}(u_n) - y_n)^2 \tag{14}$$

$$\text{subject to } \|\hat{x}\|_2^2 \leq c \tag{15}$$

(and likewise for $\|\hat{x}\|_1 \leq c$). The penalized and constrained problems are equivalent in that for every λ, there is a corresponding c that leads to the same solution. Figure 2 illustrates the effect: the constraint says that $\hat{x}$ must lie in the ℓ_2 ball (ℓ_1 ball, respectively) around the origin. Hence, we will pick the vector $\hat{x}^{\text{PLS}}$ in the ball that minimizes the error. The point with minimum error is $\hat{x}^{\text{LS}}$. The elliptical countour lines around it indicate points with the same error value. The optimal constrained solution is the point on the innermost countour line within the ball. For the squared norm penalty, $\hat{x}^{\text{PLS}}$ is closer to the origin than $\hat{x}^{\text{LS}}$: the coordinates have shrunken. For Ω_1, the constrained solution $\hat{x}^{\text{PLS}}$ very often lies at a corner of the polyhedral ball:

some of the coordinates are shrunken to zero, the others are reduced too. Therefore, the resulting estimators are also called *shrinkage estimators*.

The estimator $\hat{x}^{\text{PLS}}$ for Ω_2 is derived like $\hat{x}^{\text{LS}}$ above. We write (12) in terms of matrices and vectors as

$$\frac{1}{N}\sum_{n=1}^{N}(\hat{f}(u_n) - y_n)^2 + \lambda\|\hat{\mathbf{x}}\|^2 = (\Phi\hat{\mathbf{x}} - \mathbf{y})^\top(\Phi\hat{\mathbf{x}} - \mathbf{y}) + \lambda\hat{\mathbf{x}}^\top\hat{\mathbf{x}} \tag{16}$$

$$= \hat{\mathbf{x}}^\top(\Phi^\top\Phi + \lambda\mathbf{I})\hat{\mathbf{x}} - 2\Phi\hat{\mathbf{x}} + \mathbf{y}^\top\mathbf{y}. \tag{17}$$

Setting the derivative to zero yields

$$\hat{\mathbf{x}}^{\text{PLS}} = (\Phi^\top\Phi + \lambda\mathbf{I})^{-1}\Phi^\top\mathbf{y}. \tag{18}$$

The matrix $(\Phi^\top\Phi + \lambda\mathbf{I})$ is invertible no matter how small $N \geq 1$ is. The effect of the penalty on $\hat{\mathbf{x}}^{\text{PLS}}$ is most easily seen if $\Phi = \mathbf{I}$ is the identity matrix; we will see this in Section 28.3.3.

Above, we claimed that shrinkage affects the variance of $\hat{f}(u)$. By our derivations, we know that for any u,

$$\hat{f}(u) = \sum_{j=1}^{m}\phi_j(u)\hat{x}_j = \sum_{j=1}^{m}\phi_j(u)\sum_{n=1}^{N}((\Phi^\top\Phi + \lambda\mathbf{I})^{-1}\Phi)_n y_n \tag{19}$$

$$= \sum_{n=1}^{N}\sum_{j=1}^{m}\phi_j(u)((\Phi^\top\Phi + \lambda\mathbf{I})^{-1}\Phi)_n y_n \tag{20}$$

$$= \sum_{n=1}^{N} s_n(u) y_n. \tag{21}$$

This is a linear function in $\mathbf{y}$, and therefore such $\hat{f}$ are also called linear smoothers. The variance of $\hat{f}(u)$ is

$$\text{var}(\hat{f}(u)) = \mathbb{E}[\sum_n s_n^2(u) y_n^2] = \sum_n s_n^2(u)\sigma^2. \tag{22}$$

The larger λ, the smaller is the the sum of $s_n^2(u)$.

28.2.1 Shrinkage and cross-validation error

We still need a good strategy to set the coefficient λ. The larger λ, the higher is the penalty, and the larger will be the shrinking effect. If $\lambda = 0$, we are back at the original least squares problem.

A typical strategy is to estimate the prediction risk via leave-one-out cross-validation: estimate the function from $N - 1$ points and compute the error on the left out point.

For linear smoothers, this looks as follows. By equation (21) we can write the vector of predictions $\hat{y}_n = \hat{f}(u_n)$ as

$$\widehat{\mathbf{y}} = \mathbf{S}\mathbf{y}, \tag{23}$$

where $\mathbf{S}$ is a matrix with entries $\mathbf{S}_{n,n'} = s_{n'}(u_n)$. We can estimate the leave-one-out error as

$$\widehat{R}_{\text{LOO}} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{f}(u_n))^2 = \frac{1}{N}\sum_{n=1}^{N}\frac{(y_n - \hat{f}(u_n))^2}{(1 - S_{nn})^2}. \tag{24}$$

The expectation $\mathbb{E}[\widehat{R}_{\text{LOO}}]$ is the predictive risk, so we have an unbiased estimator. A variant of this is the *generalized cross validation*, where we replace the diagonal terms S_{nn} by their average $\frac{\nu}{N} = \frac{1}{N}\sum_{n=1}^{N} S_{nn}$:

$$\text{GCV}(\hat{f}) = \frac{1}{N}\sum_{n=1}^{N}\frac{(y_n - \hat{f}(u_n))^2}{(1 - \frac{\nu}{N})^2} \cong \frac{1}{N}\sum_{n=1}^{N}(y_n - \hat{f}(u_n))^2\Big(1 + \frac{2\nu}{N}\Big). \tag{25}$$

To obtain the last term, we used that $(1-a)^{-1} \cong 1 + 2a$. The right hand side of (25) is also called C_p statistic and was introduced by Colin Mallows. This statistic scales up the (empirical) mean squared error by a factor that depends on ν. When $\lambda = 0$, then $\nu = m$, and when $\lambda \to \infty$, then $\nu \to 0$. Hence, ν is an estimate of the effective degrees of freedom. The more degrees of freedom, the better the fit (the smaller the first part), but the larger the error is scaled up[1] (second part). We may pick λ to minimize this statistic. This statistic is reminiscent of the AIC criterion, where we penalize the size of the model, and the penalty term scales as $1/N$.

28.3 Regression with orthonormal basis functions

For the remainder of this chapter, we will assume that the ϕ_j are orthonormal basis functions for $L_2[0,1] = \{f : [0,1] \to \mathbb{R} \mid \int_0^1 f^2(u)du < \infty\}$ (e.g., the cosine bases). This means every function $f \in L_2[0,1]$ can be written as

$$f(u) = \sum_{j=1}^{\infty}\phi_j(u)\theta_j. \tag{26}$$

Just like for bases of a finite vector space, the coefficients θ_j can be found by a "projection" onto a basis function:

$$\theta_j = \int_0^1 f(u)\phi_j(u)du. \tag{27}$$

[1]Formally, $\nu = \sum_{i=1}^{N}\frac{\sigma_i^2}{\sigma_i^2+\lambda}$, where σ_i is the ith singular value of Φ.

It also holds that

$$\int_0^1 f^2(u) = \sum_{j=1}^{\infty} \theta_j^2. \tag{28}$$

We will use a truncation $f_m(u) = \sum_{j=1}^{m} \phi_j(u)\theta_j$. This truncation induces a bias

$$\int_0^1 (f(u) - f_m(u))^2 du = \sum_{j=m+1}^{\infty} \theta_j^2. \tag{29}$$

If f is smooth, this term is not too large.

Let us assume that our u_n are equi-spaced between 0 and 1, i.e., $u_n = \frac{n}{N}$. (The results here generalize to other, arbitrary u_n too, as described e.g. in [Wasserman, 2006].) Hence,

$$\mathsf{y}_n = f(\tfrac{n}{N}) + \epsilon_n, \quad \epsilon_n \sim \mathtt{N}(0, \sigma^2). \tag{30}$$

We set $m = N$, aiming to estimate the truncation $f_N(u) = \sum_{j=1}^{N} \phi_j(u) x_j$. The identity (27) gives a hint on how to estimate the coefficients x_j. We define the random variables

$$\mathsf{z}_j = \frac{1}{N} \sum_{n=1}^{N} \mathsf{y}_n \phi_j(\tfrac{n}{N}) = \frac{1}{N} \sum_{n=1}^{N} (f(\tfrac{n}{N}) + \epsilon_n)\phi_j(\tfrac{n}{N}). \tag{31}$$

Each z_j is a sum of Gaussian random variables and hence Gaussian. Its expectation is

$$\mathbb{E}[\mathsf{z}_j] = \frac{1}{N} \sum_{n=1}^{N} f(\tfrac{n}{N})\phi_j(\tfrac{n}{N}) \cong \int_0^1 f(u)\phi_j(u)du = x_j. \tag{32}$$

Moreover, one can show that $\mathrm{var}(\mathsf{z}_j) = \frac{\sigma^2}{N}$ and $\mathrm{cov}(\mathsf{z}_i, \mathsf{z}_j) \cong 0$ for all $i \neq j$. A final useful equivalence is that for $N \to \infty$, loss can be measured via the coefficients: $\int_0^1 (f(u) - \hat{f}(u))^2 du = \sum_j (x_j - \hat{x}_j)^2$.

In consequence, we have reduced our estimation problem to the following "normal means" problem (with $\tau^2 = \sigma^2/N$).

28.3.1 The Normal means problem

Assume we observe m independent Gaussian random variables $\mathsf{z}_j \sim \mathtt{N}(x_j, \tau^2)$, and we aim to estimate the unknown x_j. Our compound loss function is $C(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{j=1}^{m} (x_j - \hat{x}_j)^2$, and the risk, as before, $\mathbb{E}[C(\mathbf{x}, \hat{\mathbf{x}})]$.

Since our problem actually consists of m independent estimation problems, an obvious estimator is $\hat{x}_j^{\mathrm{LS}} = z_j$. This is the maximum likelihood estimator, and the minimum variance unbiased estimator – this sounds like a good choice. Its risk is

$$R(\mathbf{x}, \hat{\mathbf{x}}) = \sum_j \mathbb{E}[(x_j - \hat{x}_j)^2] = m\tau^2. \tag{33}$$

In 1961, James and Stein published a paper [James and Stein, 1961] showing that this common least squares estimator is not the "optimal" estimator: the following estimator has lower risk than $\hat{\mathbf{x}}^{\mathrm{LS}}$ for *any* $\mathbf{x}$ (if $m \geq 3$)!

$$\hat{x}_j^{JS} = \left(1 - \frac{m-2}{\sum_{j=1}^m z_j^2}\right) z_j \tag{34}$$

This was a remarkable result, questioning ideas in classical statistics. (Stigler calls it "one of the most provocative results in mathematical statistics of the past 35 years" [Stigler, 1990]). The estimator itself looks very surprising too: even though the m estimation problems may be completely unrelated, information of all of them is pooled together for estimating a single x_j. Moreover, $\hat{x}^{JS}$ is a shrinkage estimator, it shrinks the estimate towards zero. (It is also possible to "shrink" z_j to other values.)

This estimator can be explained in at least two ways:

1. by a Bayesian argument: Assume that $x_j \sim \mathtt{N}(0, \tau_2^2)$ and $z_j \sim \mathtt{N}(x_j, 1)$. Then the posterior mean is $\mathbb{E}[x_j | z_j = z_j] = (1 - \frac{1}{1+\tau_2^2}) z_j$. We still need to determine τ_2. An *empirical Bayes* approach estimates this hyperparameter from the data. Under the Bayesian model, the z_j are iid distributed as $\mathtt{N}(0, \tau_2^2 + 1)$. The James-Stein estimator results from estimating $\frac{1}{1+\tau^2}$ by $\frac{m-2}{\sum_{j=1}^m z_j^2}$. Details of this viewpoint are described e.g. in [Keener, 2010].

2. via Stein's unbiased risk estimate (SURE). We will talk about this next, and will see that $\hat{\mathbf{x}}^{JS}$ approximately minimizes SURE. (Even more, it is asymptotically minimax optimal.)

More discussion and other interesting viewpoints of the James-Stein estimator may be found in the papers by Efron and Morris [1977] and Stigler [1990].

28.3.2 Stein's unbiased risk estimate (SURE)

The following theorem can be used to compute the risk of the James-Stein (and other) estimators. The quantity $\widehat{R}(\mathbf{z})$ is Stein's unbiased risk estimate.

Theorem 1. *Let $z_j \sim \mathtt{N}(x_j, \tau^2)$, let $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{z})$ be an estimator of $\mathbf{x}$ and $h(\mathbf{z}) = \hat{\mathbf{x}} - \mathbf{z}$. Assume h is (weakly) differentiable, and define*

$$\widehat{R}(\mathbf{z}) = m\tau^2 + \sum_{j=1}^m h_j(\mathbf{z})^2 + 2\tau^2 \sum_{j=1}^m \frac{\partial h_j(\mathbf{z})}{\partial z_j}. \tag{35}$$

Then $\mathbb{E}[\widehat{R}(\mathbf{z})] = \mathbb{E}[\|\hat{\mathbf{x}} - \mathbf{x}\|^2] = R(\hat{\mathbf{x}}, \mathbf{x})$.

In the risk estimate, we recognize the irreducible variance, an empirical squared loss and a term that measures how much $(\hat{\mathbf{x}} - \mathbf{z})$ varies as $\mathbf{z}$ varies.

Proof. The proof uses Stein's Lemma, which states that $\sum_j \mathbb{E}[\frac{\partial h_j(\mathbf{z})}{\partial z_j}] = \mathbb{E}[\sum_j h_j(\mathbf{z})(z_j - x_j)]$. With this, we obtain

$$\mathbb{E}[\widehat{R}(\mathbf{z})] = m\sigma^2 + \sum_{j=1}^{m} \mathbb{E}[h_j(\mathbf{z})^2] + 2\tau^2 \sum_{j=1}^{m} \mathbb{E}[h_j(\mathbf{z})(z_j - x_j)] \tag{36}$$

$$= \sum_j \mathbb{E}[(z_j - x_j)^2 + \sum_j \mathbb{E}[(\hat{x}_j - z_j)^2] + 2\tau^2 \mathbb{E}[(\hat{x}_j - z_j)(z_j - x_j)] \tag{37}$$

$$= \sum_j \mathbb{E}[(\hat{x}_j - z_j + z_j - x_j)^2] \tag{38}$$

$$= R(\hat{\mathbf{x}}, \mathbf{x}). \tag{39}$$

□

For the James-Stein estimator, we get $h(\mathbf{z}) = -\frac{(m-2)\tau^2 \mathbf{z}}{\sum_j z_j^2}$, and

$$\widehat{R}(\mathbf{z}) = m\tau^2 - \frac{(m-2)\tau^2}{\sum_j z_j^2}. \tag{40}$$

Since, for $m \geq 3$, we always subtract a positive term, it clearly holds that $R(\mathbf{x}, \hat{\mathbf{x}}^{\text{JS}}) = \mathbb{E}[\widehat{R}(\mathbf{z})] < m\tau^2 = R(\mathbf{x}, \hat{\mathbf{x}}^{\text{LS}})$. The difference in risk between the average and the James-Stein estimator is largest when $\mathbf{x} = 0$.

Let us draw some more connections between the normal means problem and our discussion of linear smoothers. Both $\hat{\mathbf{x}}^{\text{JS}}$ and $\hat{\mathbf{x}}^{\text{LS}} = \mathbf{z}$ are linear smoothers of the form $\hat{\mathbf{x}}^b = b\mathbf{z}$ for some scalar $0 \leq b \leq 1$. For any such estimator, $h(\mathbf{z}) = (1-b)\mathbf{z}$ and $\partial h_j(\mathbf{z})/\partial z_j = b - 1$. Using SURE, we can estimate the risk of any such estimator as

$$\widehat{R}(\mathbf{z}) = m\tau^2 + \sum_j (b-1)^2 z_j^2 + 2\tau^2 \sum_j (b-1)\tau^2. \tag{41}$$

The b that minimizes $\widehat{R}(\mathbf{z})$ is $b^* = (1 - \frac{m\tau^2}{\sum_j z_j^2})$. This looks very close to the James-Stein estimator.

28.3.3 Penalized least squares for normal means

We can use SURE to estimate the risk of a range of other shrinkage estimators. Those include shrinkage estimators that can be obtained from penalized least-squares formulations. In the normal means case, Equation (12) becomes

$$\min_{\hat{\mathbf{x}}} \sum_{j=1}^{m} (z_j - \hat{x}_j)^2 + \lambda \Omega(\hat{\mathbf{x}}). \tag{42}$$

If $\lambda = 0$, then the $\hat{\mathbf{x}}$ minimizing (42) is $\hat{\mathbf{x}}^{\mathrm{LS}} = \mathbf{z}$. If $\Omega(\hat{\mathbf{x}}) = \sum_j x_j^2$, then the minimizer is given by

$$\hat{x}_j = \frac{1}{1+\lambda} z_j = \left(1 - \frac{1}{1+\lambda}\right) z_j. \tag{43}$$

This is a linear smoother whose risk estimate is given above in (41), with $b = (1 - \frac{1}{1+\lambda})$. SURE helps set λ.

If $\Omega(\hat{\mathbf{x}}) = \sum_j |x_j|$, then we obtain

$$\hat{x}_j = \begin{cases} z_j + \lambda & \text{if } z_j < -\lambda \\ 0 & \text{if } -\lambda \le z_j \le \lambda \\ z_j - \lambda & \text{if } z_j > \lambda. \end{cases} \tag{44}$$

Here, h is weakly differentiable so the estimator can still be analyzed using SURE (see e.g. [Wasserman, 2006]).

Finally, if $\Omega(\hat{\mathbf{x}}) = |\{j \mid \hat{x}_j \neq 0\}|$ is a penalty on the number of $\hat{x}_j$ that are not zero, then we obtain a hard threshold estimator:

$$\hat{x}_j = \begin{cases} z_j & \text{if } |z_j| > \lambda \\ 0 & \text{otherwise.} \end{cases} \tag{45}$$

Here, SURE does not apply any more, but it is still a shrinkage estimator.

28.3.4 Optimality

What would be the best possible estimator? Let us first look at linear estimators of the form $\hat{\mathbf{x}} = b\mathbf{z}$. The risk of such an estimator is

$$R(b\mathbf{z}, \mathbf{x}) = (1-b)^2 \sum_j x_j^2 + m b^2 \tau^2. \tag{46}$$

Minimizing this with respect to b, we find $b^* = \frac{\sum_j x_j^2}{m\tau^2 + \sum_j x_j^2}$, and $R(b^*\mathbf{z}, \mathbf{x}) = \frac{\tau^2 \sum_j x_j^2}{\tau^2 + \sum_i x_j^2}$. Of course, we cannot achieve this risk since we do not know $\mathbf{x}$. Hence, this risk is also called the *oracle risk*. However, one can show that the James-Stein estimator comes very close to this.

Pinsker's theorem says something even stronger: asymptotically, linear shrinkage estimators are optimal. Formally, it gives the following minimax result. Let $\mathcal{X}_m(c) = \{\mathbf{x} \mid \sum_{j=1}^m x_j^2 \le c\}$. The theorem says that the optimal possible risk (in a minimax sense) looks like the above oracle risk:

$$\lim_{m\to\infty} \inf_{\hat{\mathbf{x}}} \sup_{\mathbf{x} \in \mathcal{X}_m(c)} R(\hat{\mathbf{x}}, \mathbf{x}) = \frac{c^2 \tau^2}{\tau^2 + c^2}. \tag{47}$$

This lower bound holds for any estimator, not only linear ones.

Asymptotically, the James-Stein estimator achieves this (minimax) optimal risk. It does so adaptively without knowing $\sum_j x_j^2$.

28.4 Further Reading

The books by Hastie et al. [2009] and Wasserman [2006] offer more material on non-parametric and high-dimensional regression and shrinkage. The original papers on the James-Stein estimator are [James and Stein, 1961, Stein, 1956]. Efron and Morris [1977] give an introduction to the James-Stein estimator, and Stigler [1990] discusses it in greater detail.

References

B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236(5): 119–127, 1977.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Theory*. Springer, 2 edition, 2009.

W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Stat. Probab.*, 1961.

R.W. Keener. *Theoretical Statistics – Topics for a Core course*. Springer, 2010.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proc. Third Berkeley Symp. Math. Stat. Probab.*, 1956.

S. Stigler. The 1988 Neyman Memorial Lecture: A Galtonian perspective on shrinkage estimators. *Statistical Science*, 1990.

L. Wasserman. *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer, 2006.