7
Normal Means and Minimax Theory

In this chapter we will discuss the **many Normal means problem** which unifies some nonparametric problems and will be the basis for the methods in the next two chapters. The material in this chapter is more theoretical than in the rest of the book. If you are not interested in the theoretical details, I recommend reading sections 7.1, 7.2, and 7.3 and then skipping to the next chapter, referring back as needed. If you want more details on this topic, I recommend Johnstone (2003).

7.1 The Normal Means Model

Let $Z^n = (Z_1, \ldots, Z_n)$ where

$$Z_i = \theta_i + \sigma_n \, \epsilon_i, \quad i = 1, \ldots, n, \tag{7.1}$$

$\epsilon_1, \ldots, \epsilon_n$ are independent, Normal$(0, 1)$ random variables,

$$\theta^n = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$$

is a vector of unknown parameters and σ_n is assumed known. Typically $\sigma_n = \sigma/\sqrt{n}$ but we shall not assume this unless specifically noted. Sometimes we write Z^n and θ^n as Z and θ. The model may appear to be parametric but the number of parameters is increasing at the same rate as the number of

θ_1	θ_2	$\dots$	θ_i	$\dots$	θ_n
X_{11}	X_{21}	$\dots$	X_{i1}	$\dots$	X_{n1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1j}	X_{2j}	$\dots$	X_{ij}	$\dots$	X_{nj}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_{1n}	X_{2n}	$\dots$	X_{in}	$\vdots$	X_{nn}
Z_1	Z_2	$\dots$	Z_i	$\dots$	Z_n

FIGURE 7.1. The Normal means model. $X_{ij} = \theta_i + N(0,\sigma^2)$ and $Z_i = n^{-1}\sum_{j=1}^n X_{ij} = \theta_i + \sigma_n\epsilon_i$ where $\sigma_n = \sigma/\sqrt{n}$. Estimating the parameters $\theta_1,\dots,\theta_n$ from the n column means $Z_1,\dots,Z_n$ leads to the model (7.1) with $\sigma_n = \sigma/\sqrt{n}$.

data points. This model carries with it all the complexities and subtleties of a nonparametric problem. We will also consider an infinite-dimensional version of the model:

$$Z_i = \theta_i + \sigma_n\,\epsilon_i, \quad i = 1,2,\dots, \tag{7.2}$$

where now the unknown parameter is $\theta = (\theta_1,\theta_2,\dots)$.

Throughout this chapter we take σ_n^2 as known. In practice, we would need to estimate the variance using one of the methods discussed in Chapter 5. In this case, the exact results that follow may no longer hold but, under appropriate smoothness conditions, asymptotic versions of the results will hold.

7.3 Example. To provide some intuition for this model, suppose that we have data $X_{ij} = \theta_i + \sigma\delta_{ij}$ where $1 \le i,j \le n$ and the δ_{ij} are independent N(0,1) random variables. This is simply a one-way analysis of variance model; see Figure 7.1. Let $Z_i = n^{-1}\sum_{j=1}^n X_{ij}$. Then the model (7.1) holds with $\sigma_n = \sigma/\sqrt{n}$. We get the infinite version (7.2) by having infinitely many columns in Figure 7.1 (but still n rows). ■

Given an estimator $\widehat{\theta}^n = (\widehat{\theta}_1,\dots,\widehat{\theta}_n)$ we will use the squared error loss

$$L(\widehat{\theta}^n,\theta^n) = \sum_{i=1}^n (\widehat{\theta}_i - \theta_i)^2 = ||\widehat{\theta}^n - \theta^n||^2$$

with risk function

$$R(\widehat{\theta}^n,\theta^n) = \mathbb{E}_\theta\big(L(\widehat{\theta}^n,\theta^n)\big) = \sum_{i=1}^n \mathbb{E}_\theta(\widehat{\theta}_i - \theta_i)^2.$$

An obvious choice for an estimator of θ^n is $\widehat{\theta}^n = Z^n$. This estimator has impressive credentials: it is the maximum likelihood estimator, it is the min-

imum variance unbiased estimator and it is the Bayes estimator under a flat prior. Nonetheless, it is a poor estimator. Its risk is

$$R(Z^n, \theta^n) = \sum_{i=1}^{n} \mathbb{E}_\theta (Z_i - \theta_i)^2 = \sum_{i=1}^{n} \sigma_n^2 = n\sigma_n^2.$$

We shall see that there are estimators with substantially smaller risk.

Before we explain how we can improve on the MLE, let's first see how the normal means problem relates to nonparametric regression and density estimation. To do that, we need to review some theory about function spaces.

7.2 Function Spaces

Let $L_2(a, b)$ denote the set of functions $f : [a, b] \to \mathbb{R}$ such that $\int_a^b f^2(x)\, dx < \infty$. Unless otherwise indicated, assume that $a = 0$ and $b = 1$. The **inner product** between two functions f and g in $L_2(a, b)$ is $\int_a^b f(x)g(x)dx$ and the **norm** of f is $||f|| = \sqrt{\int_a^b f^2(x)\, dx}$. A sequence of functions $\phi_1, \phi_2, \ldots$ is called **orthonormal** if $||\phi_j|| = 1$ for all j (normalized) and $\int_a^b \phi_i(x)\phi_j(x)dx = 0$ for $i \neq j$ (orthogonal). The sequence is **complete** if the only function that is orthogonal to each ϕ_j is the zero function. A complete, orthonormal set of functions forms a **basis**, meaning that if $f \in L_2(a, b)$ then f can be expanded in the basis:

7.4 Theorem. *If $f \in L_2(a, b)$ then*[1]

$$f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x) \tag{7.5}$$

where

$$\theta_j = \int_a^b f(x)\phi_j(x)\, dx. \tag{7.6}$$

Furthermore,

$$\int_a^b f^2(x)dx = \sum_{j=1}^{\infty} \theta_j^2 \tag{7.7}$$

which is known as **Parseval's identity**.

[1]The equality sign in (7.5) means that $\int_a^b (f(x) - f_N(x))^2 dx \to 0$ as $N \to \infty$, where $f_N = \sum_{j=1}^{N} \theta_j \phi_j(x)$.

An example of an orthonormal basis for $L_2(0,1)$ is the **cosine basis**

$$\phi_0(x) = 1, \quad \phi_j(x) = \sqrt{2}\cos(2\pi jx), \qquad j = 1, 2, \ldots.$$

Another example is the **Legendre basis** defined on $(-1, 1)$:

$$P_0(x) = 1, \;\; P_1(x) = x, \;\; P_2(x) = \frac{1}{2}(3x^2 - 1), \;\; P_3(x) = \frac{1}{2}(5x^3 - 3x), \;\; \ldots$$

These polynomials are defined by the relation

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2 - 1)^n.$$

The Legendre polynomials are orthogonal but not orthonormal since

$$\int_{-1}^{1} P_n^2(x)dx = \frac{2}{2n+1}.$$

However, we can define modified Legendre polynomials $Q_n(x) = \sqrt{(2n+1)/2}$ $P_n(x)$ which then form an orthonormal basis for $L_2(-1, 1)$.

Next we introduce Sobolev spaces, which are sets of smooth functions. Let $D^j f$ denote the j^{th} weak derivative[2] of f.

7.8 Definition. *The* **Sobolev space of order** m, *is defined by*

$$W(m) = \Big\{ f \in L_2(0,1) : \; D^m f \in L_2(0,1) \Big\}.$$

The **Sobolev space of order** m **and radius** c, *is defined by*

$$W(m,c) = \Big\{ f : \; f \in W(m), \; ||D^m f||^2 \leq c^2 \Big\}.$$

The **periodic Sobolev class** *is*

$$\widetilde{W}(m,c) = \Big\{ f \in W(m,c) : \; D^j f(0) = D^j f(1), \; j = 0, \ldots, m-1 \Big\}.$$

An **ellipsoid** is a set of the form

$$\Theta = \left\{ \theta : \; \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq c^2 \right\} \tag{7.9}$$

where a_j is a sequence of numbers such that $a_j \to \infty$ as $j \to \infty$.

[2] The weak derivative is defined in the appendix.

7.10 Definition. *If Θ is an ellipsoid and if $a_j^2 \sim (\pi j)^{2m}$ as $j \to \infty$, we call Θ a* **Sobolev ellipsoid** *or a* **Sobolev body** *and we denote it by $\Theta(m, c)$.*

Now we relate Sobolev spaces to Sobolev ellipsoids.

7.11 Theorem. *Let $\{\phi_j, j = 0, 1, \ldots\}$ be the* **Fourier basis***:*

$$\phi_1(x) = 1, \quad \phi_{2j}(x) = \frac{1}{\sqrt{2}} \cos(2j\pi x), \quad \phi_{2j+1}(x) = \frac{1}{\sqrt{2}} \sin(2j\pi x), \quad j = 1, 2, \ldots$$

Then,

$$\widetilde{W}(m, c) = \left\{ f : \ f = \sum_{j=1}^{\infty} \theta_j \phi_j, \ \sum_{j=1}^{\infty} a_j^2 \theta_j^2 \leq c^2 \right\} \tag{7.12}$$

where $a_j = (\pi j)^m$ for j even and $a_j = (\pi(j-1))^m$ for j odd.

Thus, a Sobolev space corresponds to a Sobolev ellipsoid with $a_j \sim (\pi j)^{2m}$. It is also possible to relate the class $W(m, c)$ to an ellipsoid although the details are more complicated; see Nussbaum (1985).

In Sobolev spaces, smooth functions have small coefficients θ_j when j is large, otherwise $\sum_j \theta_j^2 (\pi j)^{2m}$ will blow up. Thus, to smooth a function, we shrink the θ_js closer to zero. Hence:

> smoothing f corresponds to shrinking the θ_j's towards zero for large j.

A generalization of Sobolev spaces are **Besov spaces.** These include Sobolev spaces as a special case but they also include functions with less smoothness. We defer discussion of Besov spaces until Chapter 9.

7.3 Connection to Regression and Density Estimation

Consider the nonparametric regression model

$$Y_i = f(i/n) + \sigma \epsilon_i, \quad i = 1, \ldots, n \tag{7.13}$$

where $\epsilon_i \sim N(0, 1)$, σ is known and $f \in L_2(0, 1)$. Let $\phi_1, \phi_2, \ldots$ be an orthonormal basis and write $f(x) = \sum_{i=1}^{\infty} \theta_j \phi_j(x)$ where $\theta_j = \int f(x) \phi_j(x) dx$. First, approximate f by the finite series $f(x) \approx \sum_{i=1}^{n} \theta_j \phi_j(x)$. Now define

$$Z_j = \frac{1}{n} \sum_{i=1}^{n} Y_i \ \phi_j(i/n) \tag{7.14}$$

for $j = 1, \ldots, n$. The random variable Z_j has a Normal distribution since Z_j is a linear combination of Normals. The mean of Z_j is

$$\begin{aligned}\mathbb{E}(Z_j) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i)\phi_j(i/n) = \frac{1}{n}\sum_{i=1}^{n} f(i/n)\phi_j(i/n) \\ &\approx \int f(x)\,\phi_j(x)\,dx = \theta_j.\end{aligned}$$

The variance is

$$\begin{aligned}\mathbb{V}(Z_j) &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}(Y_i)\ \phi_j^2(i/n) = \frac{\sigma^2}{n}\frac{1}{n}\sum_{i=1}^{n}\phi_j^2(i/n) \\ &\approx \frac{\sigma^2}{n}\int \phi_j^2(x)dx = \frac{\sigma^2}{n} \equiv \sigma_n^2.\end{aligned}$$

A similar calculation shows that $\mathrm{Cov}(Z_j, Z_k) \approx 0$. We conclude that the Z_j are approximately independent and

$$Z_j \sim N(\theta_j, \sigma_n^2), \quad \sigma_n^2 = \frac{\sigma^2}{n}. \tag{7.15}$$

We have thus converted the problem of estimating f into the problem of estimating the means of n Normal random variables as in (7.1) with $\sigma_n^2 = \sigma^2/n$. Also, squared error loss for f corresponds to squared error loss for θ since, by Parseval's identity, if $\widehat{f}_n(x) = \sum_{j=1}^{\infty}\widehat{\theta}_j\phi_j(x)$,

$$||\widehat{f}_n - f||^2 = \int\left(\widehat{f}_n(x) - f(x)\right)^2 dx = \sum_{j=1}^{\infty}(\widehat{\theta}_j - \theta_j)^2 = ||\widehat{\theta} - \theta||^2 \tag{7.16}$$

where $||\theta|| = \sqrt{\sum_j \theta_j^2}$.

It turns out that other nonparametric problems, such as density estimation, can also be connected to the Normal means problem. In the case of density estimation, it is the square root of the density that appears in the white noise problem. In this sense, the many Normal means problem serves as a unifying framework for many nonparametric models. See Nussbaum (1996a), Claeskens and Hjort (2004) and the appendix for more details.

7.4 Stein's Unbiased Risk Estimator (SURE)

Let $\widehat{\theta}$ be an estimate of θ. It will be useful to have an estimate of the risk of $\widehat{\theta}$. In previous chapters we used cross-validation to estimate risk. In the present context there is a more elegant method for risk estimation due to Stein (1981) known as **Stein's unbiased risk estimator** or SURE.

7.17 Theorem (Stein). *Let $Z \sim N_n(\theta, V)$, let $\widehat{\theta} = \widehat{\theta}(Z)$ be an estimate of θ and let $g(Z_1, \ldots, Z_n) = \widehat{\theta} - Z$. Note that g maps $\mathbb{R}^n$ to $\mathbb{R}^n$. Define*

$$\widehat{R}(z) = \text{tr}(V) + 2\text{tr}\,(V\,D) + \sum_i g_i^2(z) \tag{7.18}$$

where tr *denotes the trace of a matrix, $g_i = \widehat{\theta}_i - Z_i$ and the (i,j) component of D is the partial derivative of the i^{th} component of $g(z_1, \ldots, z_n)$ with respect to z_j. If g is weakly differentiable[3] then*

$$\mathbb{E}_\theta(\widehat{R}(Z)) = R(\theta, \widehat{\theta}).$$

If we apply Theorem 7.17 to the model (7.1) we get the following.

The SURE Formula for the Normal Means Model

Let $\widehat{\theta}$ be a weakly differentiable estimator of θ in model (7.1). An unbiased estimate of the risk of $\widehat{\theta}$ is

$$\widehat{R}(z) = n\sigma_n^2 + 2\sigma_n^2 \sum_{i=1}^n D_i + \sum_{i=1}^n g_i^2 \tag{7.19}$$

where $g(Z_1, \ldots, Z_n) = \widehat{\theta}^n - Z^n$ and $D_i = \partial g(z_1, \ldots, z_n)/\partial z_i$.

PROOF OF THEOREM 7.17. We will prove the case where $V = \sigma^2 I$. If $X \sim N(\mu, \sigma^2)$ then $\mathbb{E}(g(X)(X-\mu)) = \sigma^2 \mathbb{E} g'(X)$. (This is known as Stein's Lemma and it can be proved using integration by parts. See Exercise 4.) Hence, $\sigma^2 \mathbb{E}_\theta D_i = \mathbb{E}_\theta g_i(Z_i - \theta)$ and

$$\begin{aligned}
\mathbb{E}_\theta(\widehat{R}(Z)) &= n\sigma^2 + 2\sigma^2 \sum_{i=1}^n \mathbb{E}_\theta D_i + \sum_{i=1}^n \mathbb{E}_\theta(\widehat{\theta}_i - Z_i)^2 \\
&= n\sigma^2 + 2\sum_{i=1}^n \mathbb{E}_\theta\left(g_i(Z_i - \theta_i)\right) + \sum_{i=1}^n \mathbb{E}_\theta(\widehat{\theta}_i - Z_i)^2 \\
&= \sum_{i=1}^n \mathbb{E}_\theta(Z_i - \theta_i)^2 + 2\sum_{i=1}^n \mathbb{E}_\theta\Big((\widehat{\theta}_i - Z_i)(Z_i - \theta_i)\Big) \\
&\quad + \sum_{i=1}^n \mathbb{E}_\theta(\widehat{\theta}_i - Z_i)^2 \\
&= \sum_{i=1}^n \mathbb{E}_\theta(\widehat{\theta}_i - Z_i + Z_i - \theta_i)^2 = \sum_{i=1}^n \mathbb{E}_\theta(\widehat{\theta}_i - \theta_i)^2 = R(\widehat{\theta}, \theta). \quad \blacksquare
\end{aligned}$$

[3]Weak differentiability is defined in the appendix.

7.20 Example. Let $V = \sigma^2 I$. Consider $\widehat{\theta} = Z$. Then $g(z) = (0, \ldots, 0)$ and $\widehat{R}(Z) = n\sigma^2$. In this case, $\widehat{R}$ is equal to the true risk. Now consider the linear estimator $\widehat{\theta} = bZ = (bZ_1, \ldots, bZ_n)$. Hence, $g(Z) = bZ - Z = (b-1)Z$ and $D_i = b - 1$. Therefore, $\widehat{R}(Z) = (2b-1)n\sigma^2 + (1-b)^2 \sum_{i=1}^n Z_i^2$. Next consider the **soft threshold estimator** defined by

$$\widehat{\theta}_i = \begin{cases} Z_i + \lambda & Z_i < -\lambda \\ 0 & -\lambda \leq Z_i \leq \lambda \\ Z_i - \lambda & Z_i > \lambda \end{cases} \tag{7.21}$$

where $\lambda > 0$ is a constant. We can write this estimator more succinctly as

$$\widehat{\theta}_i = \text{sign}(Z_i)(|Z_i| - \lambda)_+.$$

In Exercise 5 you will show that the SURE formula gives

$$\widehat{R}(Z) = \sum_{i=1}^n \Big(\sigma^2 - 2\sigma^2 I(|Z_i| \leq \lambda) + \min(Z_i^2, \lambda^2) \Big). \tag{7.22}$$

Finally, consider the **hard threshold estimator** defined by

$$\widehat{\theta}_i = \begin{cases} Z_i & |Z_i| > \lambda \\ 0 & |Z_i| \leq \lambda \end{cases} \tag{7.23}$$

where $\lambda > 0$ is a constant. It is tempting to use SURE but this is inappropriate because this estimator is not weakly differentiable. ∎

7.24 Example (Model selection)**.** For each $S \subset \{1, \ldots, n\}$ define

$$\widehat{\theta}_S = Z_i I(i \in S). \tag{7.25}$$

We can think of S as a submodel which says that $Z_i \sim N(\theta_i, \sigma_n^2)$ for $i \in S$ and $Z_i \sim N(0, \sigma_n^2)$ for $i \notin S$. Then $\widehat{\theta}_S$ is the estimator of θ assuming the model S. The true risk of $\widehat{\theta}_S$ is

$$R(\widehat{\theta}_S, \theta) = \sigma_n^2 |S| + \sum_{i \in S^c} \theta_i^2$$

where $|S|$ denotes the number of points in S. Replacing θ_i^2 in the risk with its unbiased estimator $Z_i^2 - \sigma_n^2$ yields the risk estimator

$$\widehat{R}_S = \sigma_n^2 |S| + \sum_{i \in S^c} (Z_i^2 - \sigma_n^2). \tag{7.26}$$

It is easy to check that this corresponds to the SURE formula. Now let $\mathcal{S}$ be some class of sets where each $S \in \mathcal{S}$ is a subset of $\{1, \ldots, n\}$. Choosing $S \in \mathcal{S}$ to minimize $\widehat{R}_S$ is an example of **model selection**. The special case where

$$\mathcal{S} = \Big\{ \emptyset, \{1\}, \{1, 2\}, \ldots, \{1, 2, \ldots, n\} \Big\}$$

is called **nested subset selection**. Taking $\mathcal{S}$ to be all subsets of $\{1, \ldots, n\}$ corresponds to **all possible subsets**. For any fixed model S, we expect that $\widehat{R}_S$ will be close to $R(\widehat{\theta}_S, \theta)$. However, this does not guarantee that $\widehat{R}_S$ is close to $R(\widehat{\theta}_S, \theta)$ uniformly over $\mathcal{S}$. See Exercise 10. ∎

7.5 Minimax Risk and Pinsker's Theorem

If Θ_n is a subset of $\mathbb{R}^n$, we define the **minimax risk** over Θ_n by

$$R_n \equiv R(\Theta_n) \equiv \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_n} R(\widehat{\theta}, \theta) \tag{7.27}$$

where the infimum is over all estimators. Two questions we will address are: (i) what is the value of the minimax risk $R(\Theta_n)$? and (ii) can we find an estimator that achieves this risk?

The following theorem[4] gives the exact, limiting form of the minimax risk for the L_2 ball

$$\Theta_n(c) = \left\{ (\theta_1, \ldots, \theta_n) : \sum_{i=1}^{n} \theta_i^2 \leq c^2 \right\}.$$

7.28 Theorem (Pinsker's theorem)**.** *Assume the model (7.1) with $\sigma_n^2 = \sigma^2/n$. For any $c > 0$,*

$$\liminf_{n \to \infty} \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_n(c)} R(\widehat{\theta}, \theta) = \frac{\sigma^2 c^2}{\sigma^2 + c^2}. \tag{7.29}$$

The right-hand side of (7.29) gives an exact expression for the (asymptotic) minimax risk. This expression is strictly smaller than σ^2 which is the risk for the maximum likelihood estimator. Later, we will introduce the James–Stein estimator which asymptotically achieves this risk. The proof of the theorem, which is in the appendix, is a bit technical and may be skipped without loss of continuity. Here is the basic idea behind the proof.

First, we note that the estimator with coordinates $\widehat{\theta}_j = c^2 Z_j/(\sigma^2 + c^2)$ has risk bounded above by $\sigma^2 c^2/(\sigma^2 + c^2)$. Hence,

$$R_n \leq \frac{\sigma^2 c^2}{\sigma^2 + c^2}. \tag{7.30}$$

If we could find a prior π on $\Theta_n(c)$ whose posterior mean $\widetilde{\theta}$ also has risk $\sigma^2 c^2/(\sigma^2 + c^2)$ then we could argue that, for any estimator $\widehat{\theta}$, we have

$$\frac{\sigma^2 c^2}{\sigma^2 + c^2} = \int R(\theta, \widetilde{\theta}) d\pi(\theta) \leq \int R(\theta, \widehat{\theta}) d\pi(\theta) \leq \sup_{\theta \in \Theta_n} R(\theta, \widehat{\theta}) = R_n. \tag{7.31}$$

[4] This is a finite-dimensional version of Pinsker's theorem. Theorem 7.32 is the usual version.

Combining (7.30) and (7.31) would yield $R_n = \sigma^2 c^2/(\sigma^2 + c^2)$. The proof is essentially an approximate version of this argument. One finds a prior over all of $\mathbb{R}^n$ whose risk is arbitrarily close to $\sigma^2 c^2/(\sigma^2 + c^2)$ and one then shows that the prior asymptotically concentrates on $\Theta_n(c)$.

Now let us see how minimax theory works for smooth functions.

7.32 Theorem (Pinsker's theorem for Sobolev ellipsoids)**.** *Let*

$$Z_j = \theta_j + \frac{\sigma}{\sqrt{n}}\epsilon_j, \quad j = 1, 2, \ldots \tag{7.33}$$

where $\epsilon_1, \epsilon_2, \ldots \sim N(0,1)$*. Assume that* $\theta \in \Theta(m, c)$*, a Sobolev ellipsoid (recall Definition 7.10). Let* R_n *denote the minimax risk over* $\Theta(m, c)$*. Then,*

$$\lim_{n\to\infty} n^{2m/(2m+1)} R_n = \left(\frac{\sigma}{\pi}\right)^{2m/(2m+1)} c^{2/(2m+1)} P_m \tag{7.34}$$

where

$$P_m = \left(\frac{m}{m+1}\right)^{2m/(2m+1)} (2m+1)^{1/(2m+1)} \tag{7.35}$$

is the **Pinsker constant.** *Hence, the minimax rate is* $n^{-2m/(2m+1)}$*, that is,*

$$0 < \lim_{n\to\infty} n^{2m/(2m+1)} R_n < \infty.$$

Here is a more general version of the theorem.

7.36 Theorem (Pinsker's theorem for ellipsoids)**.** *Let*

$$\Theta = \left\{ \theta : \sum_{j=1}^{\infty} a_j \theta_j^2 \le c^2 \right\}.$$

The set Θ *is called an* **ellipsoid***. Assume that* $a_j \to \infty$ *as* $j \to \infty$*. Let*

$$R_n = \inf_{\widehat{\theta}} \sup_{\theta\in\Theta} R(\widehat{\theta}, \theta)$$

denote the minimax risk and let

$$R_n^L = \inf_{\widehat{\theta}\in\mathcal{L}} \sup_{\theta\in\Theta} R(\widehat{\theta}, \theta)$$

denote the minimax linear risk where $\mathcal{L}$ *is the set of linear estimators of the form* $\widehat{\theta} = (w_1 Z_1, w_2 Z_2, \ldots)$*. Then:*

(1) linear estimators are asymptotically minimax: $R_n \sim R_n^L$ *as* $n \to \infty$*;*

(2) the minimax linear risk satisfies

$$R_n^L = \frac{\sigma^2}{n} \sum_i \left(1 - \frac{a_i}{\mu}\right)_+$$

where μ solves

$$\frac{\sigma^2}{n}\sum_i a_i(\mu - a_i)_+ = c^2.$$

(3) The linear minimax estimator is $\widehat{\theta}_i = w_i Z_i$ where $w_i = [1 - (a_i/\mu)]_+$.

(4) The linear minimax estimator is Bayes[5] for the prior with independent components such that $\theta_i \sim N(0, \tau_i^2)$, $\tau_i^2 = (\sigma^2/n)(\mu/a_i - 1)_+$.

7.6 Linear Shrinkage and the James–Stein Estimator

Let us now return to model (7.1) and see how we can improve on the MLE using linear estimators. A **linear estimator** is an estimator of the form $\widehat{\theta} = bZ = (bZ_1, \ldots, bZ_n)$ where $0 \leq b \leq 1$. Linear estimators are **shrinkage estimators** since they **shrink** Z towards the origin. We denote the set of linear shrinkage estimators by $\mathcal{L} = \{bZ : \ b \in [0,1]\}$.

The risk of a linear estimator is easy to compute. From the basic bias–variance breakdown we have

$$R(bZ, \theta) = (1-b)^2 ||\theta||_n^2 + nb^2\sigma_n^2 \tag{7.37}$$

where $||\theta||_n^2 = \sum_{i=1}^n \theta_i^2$. The risk is minimized by taking

$$b_* = \frac{||\theta||_n^2}{n\sigma_n^2 + ||\theta||_n^2}.$$

We call b_*Z the **ideal linear estimator**. The risk of this ideal linear estimator is

$$R(b_*Z, \theta) = \frac{n\sigma_n^2 ||\theta||_n^2}{n\sigma_n^2 + ||\theta||_n^2}. \tag{7.38}$$

Thus we have proved:

7.39 Theorem.

$$\inf_{\widehat{\theta} \in \mathcal{L}} R(\widehat{\theta}, \theta) = \frac{n\sigma_n^2 ||\theta||_n^2}{n\sigma_n^2 + ||\theta||_n^2}. \tag{7.40}$$

We can't use the estimator b_*Z because b_* depends on the unknown parameter θ. For this reason we call $R(b_*Z, \theta)$ the **linear oracular risk** since the risk could only be obtained by an "oracle" that knows $||\theta||_n^2$. We shall now show that the **James–Stein estimator** nearly achieves the risk of the ideal oracle.

[5] The Bayes estimator minimizes Bayes risk $\int R(\theta, \widehat{\theta}) d\pi(\theta)$ for a given prior π.

The James–Stein estimator of θ is defined by

$$\widehat{\theta}^{JS} = \left(1 - \frac{(n-2)\sigma_n^2}{\sum_{i=1}^n Z_i^2}\right) Z. \tag{7.41}$$

We'll see in Theorem 7.48 that this estimator is asymptotically optimal.

7.42 Theorem. *The risk of the James–Stein estimator satisfies the following bound:*

$$R(\widehat{\theta}^{JS}, \theta) \leq 2\sigma_n^2 + \frac{(n-2)\sigma_n^2 ||\theta||_n^2}{(n-2)\sigma_n^2 + ||\theta||_n^2} \leq 2\sigma_n^2 + \frac{n\sigma_n^2 ||\theta||_n^2}{n\sigma_n^2 + ||\theta||_n^2} \tag{7.43}$$

where $||\theta||_n^2 = \sum_{i=1}^n \theta_i^2$.

PROOF. Write $\widehat{\theta}^{JS} = Z + g(Z)$ where $g(z) = -(n-2)\sigma_n^2 z / \sum_i z_i^2$. Hence

$$D_i = \frac{\partial g_i}{\partial z_i} = -(n-2)\sigma_n^2 \left(\frac{1}{\sum_i z_i^2} - \frac{2z_i^2}{(\sum_i z_i^2)^2}\right)$$

and

$$\sum_{i=1}^n D_i = -\frac{(n-2)^2\sigma_n^2}{\sum_{i=1}^n z_i^2}.$$

Plugging this into the SURE formula (7.19) yields

$$\widehat{R}(Z) = n\sigma_n^2 - \frac{(n-2)^2\sigma_n^4}{\sum_i Z_i^2}.$$

Hence, the risk is

$$R(\widehat{\theta}^{JS}, \theta) = \mathbb{E}(\widehat{R}(Z)) = n\sigma_n^2 - (n-2)^2\sigma_n^4 \mathbb{E}\left(\frac{1}{\sum_i Z_i^2}\right). \tag{7.44}$$

Now $Z_i^2 = \sigma_n^2((\theta_i/\sigma_n) + \epsilon_i)^2$ and so $\sum_{i=1}^n Z_i^2 \sim \sigma_n^2 W$ where W is noncentral χ^2 with n degrees of freedom and noncentrality parameter $\delta = \sum_{i=1}^n (\theta_i^2/\sigma_n^2)$. Using a well-known result about noncentral χ^2 random variables, we can then write $W \sim \chi^2_{n+2K}$ where $K \sim \text{Poisson}(\delta/2)$. Recall that (for $n > 2$) $\mathbb{E}(1/\chi_n^2) = 1/(n-2)$. So,

$$\begin{aligned}
\mathbb{E}_\theta\left[\frac{1}{\sum_i Z_i^2}\right] &= \left(\frac{1}{\sigma_n^2}\right)\mathbb{E}\left[\frac{1}{\chi^2_{n+2K}}\right] = \left(\frac{1}{\sigma_n^2}\right)\mathbb{E}\left(E\left[\frac{1}{\chi^2_{n+2K}}\,\middle|\, K\right]\right) \\
&= \left(\frac{1}{\sigma_n^2}\right)\mathbb{E}\left[\frac{1}{n-2+2K}\right] \\
&\geq \left(\frac{1}{\sigma_n^2}\right)\frac{1}{(n-2)+\sigma_n^{-2}\sum_{i=1}^n \theta_i^2} \qquad \text{from Jensen's inequality} \\
&= \frac{1}{(n-2)\sigma_n^2 + \sum_{i=1}^n \theta_i^2}.
\end{aligned}$$

Substituting into (7.44) we get the first inequality. The second inequality follows from simple algebra. ■

7.45 Remark. The **modified James–Stein estimator** is defined by

$$\widehat{\theta} = \left(1 - \frac{n\sigma_n^2}{\sum_i Z_i^2}\right)_+ Z \tag{7.46}$$

where $(a)_+ = \max\{a, 0\}$. The change from $n-2$ to n leads to a simpler expression and for large n this has negligible effect. Taking the positive part of the shrinkage factor cannot increase the risk. In practice, the modified James–Stein estimator is often preferred.

The next result shows that the James–Stein estimator nearly achieves the risk of the linear oracle.

7.47 Theorem (James–Stein oracle inequality)**.** *Let $\mathcal{L} = \{bZ : b \in \mathbb{R}\}$ denote the class of linear estimators. For all $\theta \in \mathbb{R}^n$,*

$$\inf_{\widehat{\theta} \in \mathcal{L}} R(\widehat{\theta}, \theta) \leq R(\widehat{\theta}^{JS}, \theta) \leq 2\sigma_n^2 + \inf_{\widehat{\theta} \in \mathcal{L}} R(\widehat{\theta}, \theta).$$

PROOF. This follows from (7.38) and Theorem 7.42. ■

Here is another perspective on the James–Stein estimator. Let $\widehat{\theta} = bZ$. Stein's unbiased risk estimator is $\widehat{R}(Z) = n\sigma_n^2 + 2n\sigma_n^2(b-1) + (b-1)^2 \sum_{i=1}^n Z_i^2$ which is minimized at

$$\widehat{b} = 1 - \frac{n\sigma_n^2}{\sum_{i=1}^n Z_i^2}$$

yielding the estimator

$$\widehat{\theta} = \widehat{b}Z = \left(1 - \frac{n\sigma_n^2}{\sum_{i=1}^n Z_i^2}\right) Z$$

which is essentially the James–Stein estimator.

We can now show that the James–Stein estimator achieves the Pinsker bound (7.29) and so is asymptotically minimax.

7.48 Theorem. *Let $\sigma_n^2 = \sigma^2/n$. The James–Stein estimator is asymptotically minimax, that is,*

$$\lim_{n \to \infty} \sup_{\theta \in \Theta_n(c)} R(\widehat{\theta}^{JS}, \theta) = \frac{\sigma^2 c^2}{\sigma^2 + c^2}.$$

PROOF. Follows from Theorem 7.42 and 7.28. ■

7.49 Remark. The James–Stein estimator is **adaptive** in the sense that it achieves the minimax bound over $\Theta_n(c)$ without knowledge of the parameter c.

To summarize: the James–Stein estimator is essentially optimal over all linear estimators. Moreover, it is asymptotically optimal over all estimators, not just linear estimators. This also shows that the minimax risk and the linear minimax risk are asymptotically equivalent. This turns out to (sometimes) be a more general phenomenon, as we shall see.

7.7 Adaptive Estimation Over Sobolev Spaces

Theorem 7.32 gives an estimator that is minimax over $\Theta(m, c)$. However, the estimator is unsatisfactory because it requires that we know c and m.

Efromovich and Pinsker (1984) proved the remarkable result that there exists an estimator that is minimax over $\Theta(m, c)$ without requiring knowledge of m or c. The estimator is said to be **adaptively asymptotically minimax**. The idea is to divide the observations into blocks $B_1 = \{Z_1, \ldots, Z_{n_1}\}$, $B_2 = \{Z_{n_1+1}, \ldots, Z_{n_2}\}$, ... and then apply a suitable estimation procedure within blocks.

Here is particular block estimation scheme due to Cai et al. (2000). For any real number a let $[a]$ denote the integer part of a. Let $b = 1 + 1/\log n$ and let K_0 be an integer such that $[b^{K_0}] \geq 3$ and $[b^k] - [b^{k-1}] \geq 3$ for $k \geq K_0 + 1$. Let $B_0 = \{Z_i : \ 1 \leq i \leq [b^{K_0}]\}$ and let $B_k = \{Z_i : \ [b^{k-1}] < i \leq [b^k]\}$ for $k \geq K_0 + 1$. Let $\widehat{\theta}$ be the estimator obtained by applying the James–Stein estimator within each block B_k. The estimator is taken to be 0 for $i > [b^{K_1}]$ where $K_1 = [\log_b(n)] - 1$.

7.50 Theorem (Cai, Low and Zhao, 2000). *Let $\widehat{\theta}$ be the estimator above. Let $\Theta(m, c) = \{\theta : \ \sum_{i=1}^{\infty} a_i^2 \theta_i^2 \leq c^2\}$ where $a_1 = 1$ and $a_{2i} = a_{2i+1} = 1 + (2i\pi)^{2m}$. Let $R_n(m, c)$ denote the minimax risk over $\Theta(m, c)$. Then for all $m > 0$ and $c > 0$,*

$$\lim_{n \to \infty} \frac{\sup_{\theta \in \Theta(m,c)} R(\widehat{\theta}, \theta)}{R_n(m, c)} = 1.$$

7.8 Confidence Sets

In this section we discuss the construction of confidence sets for θ^n. It will now be convenient to write θ and Z instead of θ^n and Z^n.

Recall that $\mathcal{B}_n \subset \mathbb{R}^n$ is a $1-\alpha$ confidence set if

$$\inf_{\theta \in \mathbb{R}^n} \mathbb{P}_\theta(\theta \in \mathcal{B}_n) \geq 1 - \alpha. \tag{7.51}$$

We have written the probability distribution $\mathbb{P}_\theta$ with the subscript θ to emphasize that the distribution depends on θ. Here are some methods for constructing confidence sets.

Method I: The χ^2 Confidence Set. The simplest confidence set for θ is based on the fact that $||Z-\theta||^2/\sigma_n^2$ has a χ^2_n distribution. Let

$$\mathcal{B}_n = \left\{\theta \in \mathbb{R}^n : \ ||Z-\theta||^2 \leq \sigma_n^2 \, \chi^2_{n,\alpha}\right\} \tag{7.52}$$

where $\chi^2_{n,\alpha}$ is the upper α quantile of a χ^2 random variable with n degrees of freedom. It follows immediately that

$$\mathbb{P}_\theta(\theta \in \mathcal{B}_n) = 1 - \alpha, \quad \text{for all } \theta \in \mathbb{R}^n.$$

Hence, (7.51) is satisfied. The expected radius of this ball is $n\sigma_n^2$. We will see that we can improve on this.

Improving the χ^2 Ball by Pre-testing. Before discussing more complicated methods, here is a simple idea—based on ideas in Lepski (1999)—for improving the χ^2 ball. The methods that follow are generalizations of this method.

Note that the χ^2 ball $\mathcal{B}_n$ has a fixed radius $s_n = \sigma_n\sqrt{n}$. When applied to function estimation, $\sigma_n = O(1/\sqrt{n})$ so that $s_n = O(1)$ and hence the radius of the ball does not even converge to zero as $n \to \infty$. The following construction makes the radius smaller. The idea is to test the hypothesis that $\theta = \theta_0$. If we accept the null hypothesis, we use a smaller ball centered at θ_0. Here are the details.

First, test the hypothesis that $\theta = (0, \ldots, 0)$ using $\sum_i Z_i^2$ as a test statistic. Specifically, reject the null when

$$T_n = \sum_i Z_i^2 > c_n^2$$

and c_n is defined by

$$\mathbb{P}\left(\chi^2_n > \frac{c_n^2}{\sigma_n^2}\right) = \frac{\alpha}{2}.$$

By construction, the test has type one error rate $\alpha/2$. If Z denotes a N(0,1) random variable then

$$\frac{\alpha}{2} = \mathbb{P}\left(\chi^2_n > \frac{c_n^2}{\sigma_n^2}\right) = \mathbb{P}\left(\frac{\chi^2_n - n}{\sqrt{2n}} > \frac{\frac{c_n^2}{\sigma_n^2} - n}{\sqrt{2n}}\right) \approx \mathbb{P}\left(Z > \frac{\frac{c_n^2}{\sigma_n^2} - n}{\sqrt{2n}}\right)$$

implying that

$$c_n^2 \approx \sigma_n^2(n + \sqrt{2n} z_{\alpha/2}).$$

Now we compute the power of this test when $||\theta|| > \Delta_n$ where

$$\Delta_n = \sqrt{2\sqrt{2} z_{\alpha/2}}\, n^{1/4}\sigma_n.$$

Write $Z_i = \theta_i + \sigma_n \epsilon_i$ where $\epsilon_i \sim N(0,1)$. Then,

$$\begin{aligned}
\mathbb{P}_\theta(T_n > c_n^2) &= \mathbb{P}_\theta\left(\sum_i Z_i^2 > c_n^2\right) = \mathbb{P}_\theta\left(\sum_i (\theta_i + \sigma_n\epsilon_i)^2 > c_n^2\right) \\
&= \mathbb{P}_\theta\left(||\theta||^2 + 2\sigma_n \sum_i \theta_i\epsilon_i + \sigma_n^2 \sum_i \epsilon_i^2 > c_n^2\right).
\end{aligned}$$

Now, $||\theta||^2 + 2\sigma_n \sum_i \theta_i\epsilon_i + \sigma_n^2 \sum_i \epsilon_i^2$ has mean $||\theta||^2 + n\sigma_n^2$ and variance $4\sigma_n^2||\theta||^2 + 2n\sigma_n^4$. Hence, with Z denoting a N(0,1) random variable,

$$\begin{aligned}
\mathbb{P}_\theta(T_n > c_n^2) &\approx \mathbb{P}\Big(||\theta||^2 + n\sigma_n^2 + \sqrt{4\sigma_n^2||\theta||^2 + 2n\sigma_n^4} Z > c_n^2\Big) \\
&\approx \mathbb{P}\Big(||\theta||^2 n\sigma_n^2 + \sqrt{4\sigma_n^2||\theta||^2 + 2n\sigma_n^4} Z > \sigma_n^2(n + \sqrt{2n} z_{\alpha/2})\Big) \\
&= \mathbb{P}\left(Z > \frac{\left(\sqrt{2} z_{\alpha/2} - \frac{||\theta||^2}{\sqrt{n}\sigma_n^2}\right)}{2 + \frac{4||\theta||^2}{n\sigma_n^2}}\right) \geq \mathbb{P}\left(Z > \frac{\left(\sqrt{2} z_{\alpha/2} - \frac{||\theta||^2}{\sqrt{n}\sigma_n^2}\right)}{2}\right) \\
&\geq 1 - \frac{\alpha}{2}
\end{aligned}$$

since $||\theta|| > \Delta_n$ implies that

$$\frac{\left(\sqrt{2} z_{\alpha/2} - \frac{||\theta||^2}{\sqrt{n}\sigma_n^2}\right)}{2} \geq -z_{\alpha/2}.$$

In summary, the test has type-one error $\alpha/2$ and type-two error no more than $\alpha/2$ for all $||\theta|| > \Delta_n$.

Next we define the confidence procedure as follows. Let $\phi = 0$ if the test accepts and $\phi = 1$ if the test rejects. Define

$$R_n = \begin{cases} \mathcal{B}_n & \text{if } \phi = 1 \\ \Big\{\theta:\ ||\theta|| \leq \Delta_n\Big\} & \text{if } \phi = 0. \end{cases}$$

Thus, R_n is a random radius confidence ball. The radius is the same as the χ^2 ball when $\phi = 1$ but when $\phi = 0$, the radius is Δ_n which is much smaller. Let us now verify that the ball has the right coverage.

The noncoverage of this ball when $\theta = (0, \ldots, 0)$ is

$$\begin{aligned} \mathbb{P}_0(\theta \notin R) &= \mathbb{P}_0(\theta \notin R, \phi = 0) + \mathbb{P}_0(\theta \notin R, \phi = 1) \\ &\leq 0 + \mathbb{P}_0(\phi = 1) = \frac{\alpha}{2}. \end{aligned}$$

The noncoverage of this ball when $\theta \neq (0, \ldots, 0)$ and $||\theta|| \leq \Delta_n$ is

$$\begin{aligned} \mathbb{P}_\theta(\theta \notin R) &= \mathbb{P}_\theta(\theta \notin R, \phi = 0) + \mathbb{P}_\theta(\theta \notin R, \phi = 1) \\ &\leq 0 + \mathbb{P}_\theta(\theta \notin B) = \frac{\alpha}{2}. \end{aligned}$$

The noncoverage of this ball when $\theta \neq (0, \ldots, 0)$ and $||\theta|| > \Delta_n$ is

$$\begin{aligned} \mathbb{P}_\theta(\theta \notin R) &= \mathbb{P}_\theta(\theta \notin R, \phi = 0) + \mathbb{P}_\theta(\theta \notin R, \phi = 1) \\ &\leq \mathbb{P}_\theta(\phi = 0) + \mathbb{P}_\theta(\theta \notin B) \leq \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha. \end{aligned}$$

In summary, by testing whether θ is close to $(0, \ldots, 0)$ and using a smaller ball centered at $(0, \ldots, 0)$ when the test accepts, we get a ball with proper coverage and whose radius is sometimes smaller than the χ^2 ball. The message is that:

> a random radius confidence ball can have an expected radius that is smaller than a fixed confidence ball at some points in the parameter space.

The next section generalizes this idea.

Method II: The Baraud Confidence Set. Here we discuss the method due to Baraud (2004) which builds on Lepski (1999), as discussed above. We begin with a class $\mathcal{S}$ of linear subspaces of $\mathbb{R}^n$. Let Π_S denote the projector onto S. Thus, for any vector $Z \in \mathbb{R}^n$, $\Pi_S Z$ is the vector in S closest to Z.

For each subspace S, we construct a ball $\mathcal{B}_S$ of radius ρ_S centered at an estimator in S, namely,

$$\mathcal{B}_S = \Big\{\theta : \ ||\theta - \Pi_S Z|| \leq \rho_S \Big\}. \tag{7.53}$$

For each $S \in \mathcal{S}$, we test whether θ is close to S using $||Z - \Pi_S Z||$ as a test statistic. We then take the smallest confidence ball $\mathcal{B}_S$ among all unrejected

subspaces S. The key to making this work is this: the radius ρ_S is chosen so that

$$\max_\theta \mathbb{P}_\theta(S \text{ is not rejected and } \theta \notin \mathcal{B}_S) \leq \alpha_S \tag{7.54}$$

where $\sum_{S\in\mathcal{S}} \alpha_S \leq \alpha$. The resulting confidence ball has coverage at least $1-\alpha$ since

$$\begin{aligned} \max_\theta \mathbb{P}_\theta(\theta \notin \mathcal{B}) &\leq \sum_S \max_\theta \mathbb{P}_\theta(S \text{ is not rejected and } \theta \notin \mathcal{B}_S) \\ &= \sum_S \alpha_S \leq \alpha. \end{aligned}$$

We will see that the n-dimensional maximization over $\theta \in \mathbb{R}^n$ can be reduced to a one-dimensional maximization since the probabilities only depend on θ through the quantity $z = ||\theta - \Pi_S \theta||$.

The confidence set has coverage $1-\alpha$ even if θ is not close to one of the subspaces in $\mathcal{S}$. However, if it is close to one of the subspaces in $\mathcal{S}$, then the confidence ball will be smaller than the χ^2 ball.

For example, suppose we expand a function $f(x) = \sum_j \theta_j \phi_j(x)$ in a basis, as in Section 7.3. Then, the θ_is correspond to the coefficients of f in this basis. If the function is smooth, then we expect that θ_i will be small for large i. Hence, θ might be well approximated by a vector of the form $(\theta_1, \ldots, \theta_m, 0, \ldots, 0)$. This suggests that we could test whether θ is close to the subspace S_m of the vectors of the form $(\theta_1, \ldots, \theta_m, 0, \ldots, 0)$, for $m = 0, \ldots, n$. In this case we would take the class of subspaces to be $\mathcal{S} = \{S_0, \ldots, S_n\}$.

Before we proceed with the details, we need some notation. If $X_j \sim N(\mu_j, 1)$, $j = 1, \ldots, k$ are IID, then $T = \sum_{j=1}^k X_j^2$ has a noncentral χ^2 distribution with noncentrality parameter $d = \sum_j \mu_j^2$ and k degrees of freedom and we write $T \sim \chi^2_{d,k}$. Let $G_{d,k}$ denote the CDF of this random variable and let $q_{d,k}(\alpha) = G^{-1}_{d,k}(1-\alpha)$ denote the upper α quantile. By convention, we define $q_{d,k}(\alpha) = -\infty$ for $\alpha \geq 1$.

Let $\mathcal{S}$ be a finite collection of linear subspaces of $\mathbb{R}^n$. We assume that $\mathbb{R}^n \in \mathcal{S}$. Let $d(S)$ be the dimension of $S \in \mathcal{S}$ and let $e(S) = n - d(S)$. Fix $\alpha \in (0,1)$ and $\gamma \in (0,1)$ where $\gamma < 1-\alpha$. Let

$$\mathcal{A} = \left\{ S : \ \frac{||Z - \Pi_S Z||^2}{\sigma_n^2} \leq c(S) \right\} \tag{7.55}$$

where

$$c(S) = q_{0,e(S)}(\gamma). \tag{7.56}$$

Think of $||Z-\Pi_S Z||^2$ as a test statistic for the hypothesis that $\theta \in S$. Then $\mathcal{A}$ is the set of nonrejected subspaces. Note that $\mathcal{A}$ always includes the subspace $S = \mathbb{R}^n$ since, when $S = \mathbb{R}^n$, $\Pi_S Z = Z$ and $||Z - \Pi_S Z||^2 = 0$.

Let $(\alpha_S : \ S \in \mathcal{S})$ be a set of numbers such that $\sum_{S\in\mathcal{S}} \alpha_S \leq \alpha$. Now define the ρ_S as follows:

$$\rho_S^2 = \sigma_n^2 \times \begin{cases} \inf_{z>0}\left\{G_{z,n}(q_{0,n}(\gamma)) \leq \alpha_S\right\} & \text{if } d(S) = 0 \\ \sup_{z>0}\left\{z + q_{0,d(S)}\left(\frac{\alpha_S}{G_{z,e(S)}(c(S))}\right)\right\} & \text{if } 0 < d(S) < n \\ \rho_S^2 = \sigma_n^2 q_{0,n}(\alpha_S) & \text{if } d(S) = n. \end{cases} \tag{7.57}$$

Define

$$\widehat{S} = \text{argmin}_{S\in\mathcal{A}}\ \rho_S,$$

$\widehat{\theta} = \Pi_{\widehat{S}} Z$, and $\widehat{\rho} = \rho_{\widehat{S}}$. Finally, define

$$\mathcal{B}_n = \left\{\theta \in \mathbb{R}^n : \ ||\theta - \widehat{\theta}||^2 \leq \widehat{\rho}^2\right\}. \tag{7.58}$$

7.59 Theorem (Baraud 2004). *The set $\mathcal{B}_n$ defined in (7.58) is a valid confidence set:*

$$\inf_{\theta\in\mathbb{R}^n} \mathbb{P}_\theta(\theta \in \mathcal{B}_n) \geq 1 - \alpha. \tag{7.60}$$

PROOF. Let $\mathcal{B}_S = \{\theta : \ ||\theta - \Pi_S Z||^2 \leq \rho_S^2\}$. Then,

$$\begin{aligned} \mathbb{P}_\theta(\theta \notin \mathcal{B}_n) &\leq \mathbb{P}_\theta(\theta \notin \mathcal{B}_S \text{ for some } S \in \mathcal{A}) \\ &\leq \sum_S \mathbb{P}_\theta(||\theta - \Pi_S Z|| > \rho_S, \ \widehat{S} \in \mathcal{A}) \\ &= \sum_S \mathbb{P}_\theta\big(||\theta - \Pi_S Z|| > \rho_S, \ ||Z - \Pi_S Z||^2 \leq c(S)\sigma_n^2\big). \end{aligned}$$

Since $\sum_S \alpha_S \leq \alpha$, it suffices to show that $a(S) \leq \alpha_S$ for all $S \in \mathcal{S}$, where

$$a(S) \equiv \mathbb{P}_\theta\left(||\theta - \Pi_S Z|| > \rho_S, \ ||Z - \Pi_S Z||^2 \leq \sigma_n^2\, c(S)\right). \tag{7.61}$$

When $d(S) = 0$, $\Pi_S Z = (0, \ldots, 0)$. If $||\theta|| \leq \rho_S$ then $a(0) = 0$ which is less than α_S. If $||\theta|| > \rho_S$, then

$$\begin{aligned} a(S) &= \mathbb{P}_\theta\left(\sum_{i=1}^n Z_i^2 \leq \sigma_n^2 q_{0,n}(\gamma)\right) \\ &= G_{||\theta||^2/\sigma_n^2,n}(q_{0,n}(\gamma)) \leq G_{\rho_0^2/\sigma_n^2,n}(q_{0,n}(\gamma)) \\ &\leq \alpha_S \end{aligned}$$

since $G_{d,n}(u)$ is decreasing in z for all u and from the definition of ρ_0^2.

Now consider the case where $0 < d(S) < n$. Let

$$A = \frac{||\theta - \Pi_S Z||^2}{\sigma_n^2} = z + \sum_{j=1}^{m} \epsilon_j^2, \qquad B = \frac{||\widehat{\theta} - Z||^2}{\sigma_n^2}$$

where $z = ||\theta - \Pi_S \theta||^2/\sigma_n^2$. Then A and B are independent, $A \sim z + \chi^2_{0,d(S)}$, and $B \sim \chi^2_{z,e(S)}$. Hence,

$$\begin{aligned} a(S) &= \mathbb{P}_\theta\left(A > \frac{\rho_m^2}{\sigma_n^2},\ B < c(S)\right) \\ &= \mathbb{P}_\theta\left(z + \chi^2_{d(S)} > \frac{\rho_S^2}{\sigma_n^2},\ \chi^2_{z,e(S)} < c(S)\right) \qquad (7.62) \\ &= \left(1 - G_{0,d(S)}\left(\frac{\rho_S^2}{\sigma_n^2} - z\right)\right) \times G_{z,e(S)}\Big(c(S)\Big). \qquad (7.63) \end{aligned}$$

From the definition of ρ_S^2,

$$\frac{\rho_S^2}{\sigma_n^2} - z \geq q_{0,d(S)}\left(\frac{\alpha_S}{G_{z,e(S)}(c(S))} \wedge 1\right)$$

and hence,

$$\begin{aligned} 1 - G_{0,d(S)}\left(\frac{\rho_S^2}{\sigma_n^2} - z\right) &\leq 1 - G_{0,d(S)}\left(q_{0,d(S)}\left(\frac{\alpha_S}{G_{z,e(S)}(c(S))}\right)\right) \\ &= \frac{\alpha_S}{G_{z,e(S)}(c(S))}. \qquad (7.64) \end{aligned}$$

It then follows (7.63) and (7.64) that $a(S) \leq \alpha_S$.

For the case $d(S) = n$, $\Pi_S Z = Z$, and $||\theta - \Pi_S Z||^2 = \sigma_n^2 \sum_{i=1}^n \epsilon_i^2 \stackrel{d}{=} \sigma_n^2 \chi_n^2$ and so

$$a(S) = \mathbb{P}_\theta\big(\sigma_n^2 \chi_n^2 > q_{0,n}(\alpha_S)\sigma_n^2\big) = \alpha_S$$

by the definition of $q_{0,n}$. ∎

When σ_n is unknown we estimate the variance using one of the methods discussed in Chapter 5 and generally the coverage is only asymptotically correct. To see the effect of having uncertainty about σ_n, consider the idealized case where σ_n is known to lie with certainty in the interval $I = [\sqrt{1-\eta_n}\tau_n, \tau_n]$. (In practice, we would construct a confidence interval for σ and adjust the

level α of the confidence ball appropriately.) In this case, the radii ρ_S are now defined by:

$$\rho_S^2 = \begin{cases} \inf_{z>0}\left\{\sup_{\sigma_n \in I} G_{z/\sigma_n^2,n}(q_{0,n}(\gamma)\tau_n^2/\sigma_n^2) \leq \alpha_S\right\} & \text{if } d(S)=0 \\ \sup_{z>0,\sigma_n\in I}\left\{z\sigma_n^2 + \sigma_n^2 q_{0,d(S)}(h_S(z,\sigma_n))\right\} & \text{if } 0<d(S)<n \\ q_{0,n}(\alpha_S)\tau_n^2 & \text{if } d(S)=n \end{cases} \tag{7.65}$$

where

$$h_S(z,\sigma) = \frac{\alpha_S}{G_{z,e(S)}\left(G_{z,e(S)}(q_{0,e(S)}(\gamma)\tau_n^2/\sigma^2)\right)} \tag{7.66}$$

and $\mathcal{A}$ is now defined by

$$\mathcal{A} = \left\{S \in \mathcal{S} :\ ||Z - \Pi_S Z||^2 \leq q_{0,e(S)}(\gamma)\tau_n^2\right\}. \tag{7.67}$$

Beran–Dümbgen–Stein Pivotal method. Now we discuss a different approach due to Stein (1981) and developed further by Li (1989), Beran and Dümbgen (1998), and Genovese and Wasserman (2005). The method is simpler than the Baraud–Lepski approach but it uses asymptotic approximations. This method is considered in more detail in the next chapter but here is the basic idea.

Consider nested subsets $\mathcal{S} = \{S_0, S_1, \ldots, S_n\}$ where

$$S_j = \left\{\theta = (\theta_1, \ldots, \theta_j, 0, \ldots, 0) :\ (\theta_1, \ldots, \theta_j) \in \mathbb{R}^j\right\}.$$

Let $\widehat{\theta}_m = (Z_1, \ldots, Z_m, 0, \ldots, 0)$ denote the estimator under model S_m. The loss function is

$$L_m = ||\widehat{\theta}_m - \theta||^2.$$

Define the **pivot**

$$V_m = \sqrt{n}(L_m - \widehat{R}_m) \tag{7.68}$$

where $\widehat{R}_m = m\sigma_n^2 + \sum_{j=m+1}^n (Z_j^2 - \sigma_n^2)$ is sure. Let $\widehat{m}$ minimize $\widehat{R}_m$ over m. Beran and Dümbgen (1998) show that $V_{\widehat{m}}/\widehat{\tau} \rightsquigarrow N(0,1)$ where

$$\tau_m^2 = \mathbb{V}(V_m) = 2n\sigma_n^2\left(n\sigma_n^2 + 2\sum_{j=m+1}^n \theta_j^2\right)$$

and

$$\widehat{\tau}^2 = 2n\sigma_n^2\left(n\sigma_n^2 + 2\sum_{j=\widehat{m}+1}^n (Z_j^2 - \sigma_n^2)\right).$$

Let

$$r_n^2 = \widehat{R}_m + \frac{\widehat{\tau} z_\alpha}{\sqrt{n}}$$

and define

$$\mathcal{B}_n = \left\{ \theta \in \mathbb{R}^n : \ ||\theta_m - \widehat{\theta}||^2 \leq r_n^2 \right\}.$$

Then,

$$\begin{aligned} \mathbb{P}_\theta(\theta \in \mathcal{B}_n) &= \mathbb{P}_\theta(||\theta - \widehat{\theta}||^2 \leq r_n^2) = \mathbb{P}_\theta(L_m \leq r_n^2) \\ &= \mathbb{P}_\theta\left(L_m \leq \widehat{R}_m + \frac{\widehat{\tau} z_\alpha}{\sqrt{n}} \right) = \mathbb{P}_\theta\left(\frac{V_{\widehat{m}}}{\widehat{\tau}} \leq z_\alpha \right) \\ &\to 1 - \alpha. \end{aligned}$$

A practical problem with this method is that r_n^2 can be negative. This is due to the presence of the term $\sum_{j=m+1}^n (Z_j^2 - \sigma_n^2)$ in $\widehat{R}$ and τ. We deal with this by replacing such terms with $\max\{\sum_{j=m+1}^n (Z_j^2 - \sigma_n^2), 0\}$. This can lead to over-coverage but at least leads to well-defined radii.

7.69 Example. Consider nested subsets $\mathcal{S} = \{S_0, S_1, \ldots, S_n\}$ where $S_0 = \{(0, \ldots, 0)\}$ and

$$S_j = \left\{ \theta = (\theta_1, \ldots, \theta_j, 0, \ldots, 0) : \ (\theta_1, \ldots, \theta_j) \in \mathbb{R}^j \right\}.$$

We take $\alpha = 0.05$, $n = 100$, $\sigma_n = 1/\sqrt{n}$, and $\alpha_S = \alpha/(n+1)$ for all S so that $\sum \alpha_S = \alpha$ as required. Figure 7.2 shows ρ_S versus the dimension of S for $\gamma = 0.05, 0.15, 0.50, 0.90$. The dotted line is the radius of the χ^2 ball. One can show that

$$\frac{\rho_0}{\rho_n} = O\left(n^{-1/4} \right) \tag{7.70}$$

which shows that shrinking towards lower-dimensional models leads to smaller confidence sets. There is an interesting tradeoff. Setting γ large makes ρ_0 small leading to a potentially smaller confidence ball. However, making γ large increases the set $\mathcal{A}$ which diminishes the chances of choosing a small ρ. We simulated under the model $\theta = (10, 10, 10, 10, 10, 0, \ldots, 0)$. See Table 7.1 for a summary. In this example, the pivotal method seems to perform the best. ■

7.9 Optimality of Confidence Sets

How small can we make the confidence set while still maintaining correct coverage? In this section we will see that if $\mathcal{B}_n$ is a confidence ball with radius

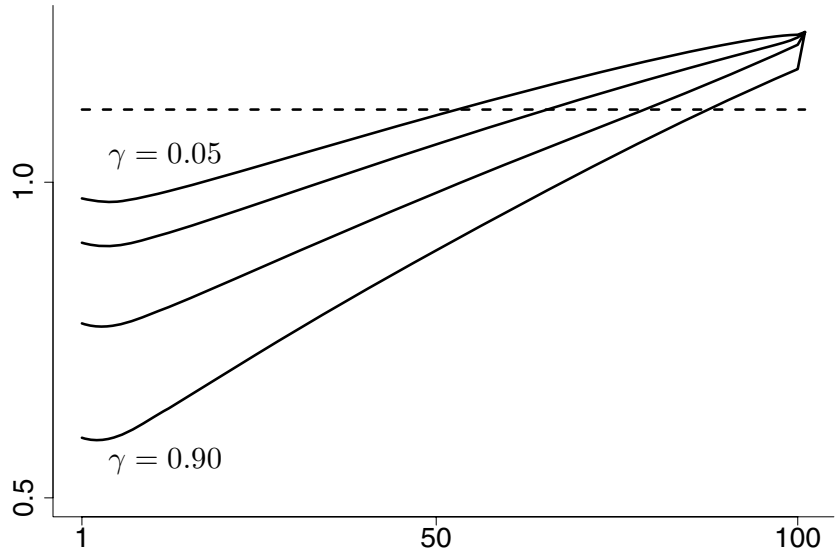

FIGURE 7.2. Constants ρ_S from Example 7.69. The horizontal axis is the dimension of the submodel. The four curves show ρ_S for $\gamma = 0.05, 0.15, 0.50, 0.90$. The highest curve corresponds to $\gamma = 0.05$ and the curves get lower as γ increases. The dotted line is the radius of the χ^2 ball.

Method	Coverage	Radius
χ^2	0.950	1.115
Baraud ($\gamma = 0.90$)	1.000	0.973
($\gamma = 0.50$)	1.000	0.904
($\gamma = 0.15$)	1.000	0.779
($\gamma = 0.05$)	0.996	0.605
Pivotal	0.998	0.582

TABLE 7.1. Simulation results from Example 7.69 based on 1000 simulations.

s_n then $E_\theta(s_n) \geq C_1\sigma_n n^{1/4}$ for **every** θ and $E_\theta(s_n) \geq C_2\sigma_n n^{1/2}$ for **some** θ. Here, C_1 and C_2 are positive constants. The χ^2 ball has radius $\sigma_n n^{1/2}$ for all θ. This suggests that the χ^2 ball can be improved upon, and indeed, the Baraud confidence ball can achieve the faster $\sigma_n n^{1/4}$ rate at some points in the parameter space. We will provide some of the details in this section. But first, let us compare this with point estimation.

From Theorem 7.32, the optimal rate of convergence of a point estimator over a Sobolev space of order m is $n^{-2m/(2m+1)}$. According to Theorem 7.50, we can construct estimators that achieve this rate, without prior knowledge of m. This raises the following questions: Can we construct confidence balls that adaptively achieve this optimal rate? The short answer is no. Robins and van der Vaart (2005), Juditsky and Lambert-Lacroix (2003), and Cai and Low (2005) show that some degree of adaptivity is possible for confidence sets but the amount of adaptivity is quite restricted. Without any smoothness assumptions, we see from our comments above that the fastest rate of convergence one can attain is $\sigma_n n^{1/4}$ which is of order $O(n^{-1/4})$ when $\sigma_n = \sigma/\sqrt{n}$.

Turning to the details, we begin with the following Theorem due to Li (1989).

7.71 Theorem (Li 1989). *Let $\mathcal{B}_n = \{\theta^n \in \mathbb{R}^n : ||\widehat{\theta}^n - \theta^n|| \leq s_n\}$ where $\widehat{\theta}^n$ is any estimator of θ^n and $s_n = s_n(Z^n)$ is the radius of the ball. Suppose that*

$$\liminf_{n\to\infty} \inf_{\theta^n \in \mathbb{R}^n} \mathbb{P}_{\theta^n}(\theta^n \in \mathcal{B}_n) \geq 1 - \alpha. \tag{7.72}$$

Then for any sequence θ^n and any $c_n \to 0$,

$$\limsup_{n\to\infty} \mathbb{P}_{\theta^n}(s_n \leq c_n\sigma_n n^{1/4}) \leq \alpha. \tag{7.73}$$

Finite sample results are available from Baraud (2004) and Cai and Low (2005). For example, we have the following result, whose proof is in the appendix.

7.74 Theorem (Cai and Low 2004). *Assume the model (7.1). Fix $0 < \alpha < 1/2$. Let $\mathcal{B}_n = \{\theta : ||\widehat{\theta} - \theta|| \leq s_n\}$ be such that*

$$\inf_{\theta\in\mathbb{R}^n} \mathbb{P}_\theta(\theta \in \mathcal{B}_n) \geq 1 - \alpha.$$

Then, for every $0 < \epsilon < (1/2) - \alpha$,

$$\inf_{\theta\in\mathbb{R}^n} \mathbb{E}_\theta(s_n) \geq \sigma_n(1 - 2\alpha - 2\epsilon)n^{1/4}(\log(1+\epsilon^2))^{1/4}. \tag{7.75}$$

In particular, if $\sigma_n = \sigma/\sqrt{n}$*, then*

$$\inf_{\theta \in \mathbb{R}^n} \mathbb{E}_\theta(s_n) \geq \frac{C}{n^{1/4}} \tag{7.76}$$

where $C = \sigma(1 - 2\alpha - 2\epsilon)(\log(1 + \epsilon^2))^{1/4}$.

The lower bound in the above theorem cannot be attained everywhere, as the next result shows.

7.77 Theorem (Cai and Low 2004). *Assume the model (7.1). Fix* $0 < \alpha < 1/2$*. Let* $\mathcal{B}_n = \{\theta : ||\widehat{\theta} - \theta|| \leq s_n\}$ *be such that*

$$\inf_{\theta \in \mathbb{R}^n} \mathbb{P}_\theta(\theta \in \mathcal{B}_n) \geq 1 - \alpha.$$

Then, for every $0 < \epsilon < (1/2) - \alpha$*,*

$$\sup_{\theta \in \mathbb{R}^n} \mathbb{E}_\theta(s_n) \geq \epsilon\, \sigma_n z_{\alpha+2\epsilon} \sqrt{n} \sqrt{\frac{\epsilon}{1 - \alpha - \epsilon}}. \tag{7.78}$$

In particular, if $\sigma_n = \sigma/\sqrt{n}$*, then*

$$\sup_{\theta \in \mathbb{R}^n} \mathbb{E}_\theta(s_n) \geq C \tag{7.79}$$

where $C = \epsilon z_{\alpha+2\epsilon} \sqrt{\epsilon/(1 - \alpha - \epsilon)}$.

Despite these pessimistic sounding results, there is some potential for adaptation since the infimum in Theorem 7.74 is smaller than the supremum in Theorem 7.77. For example, the χ^2 ball has radius $O(\sigma_n\sqrt{n})$ but the lower bound in the above theorem is $O(\sigma_n n^{1/4})$ suggesting that we can do better than the χ^2 ball. This was the motivation for the Baraud and pivotal confidence sets. The Baraud confidence set does have a certain type of adaptivity: if $\theta \in S$ then $\widehat{\rho} \leq \rho_S$ with high probability. This follows easily from the way that the ball is defined. Let us formalize this as a lemma.

7.80 Lemma. *Define* $\mathcal{S}$*,* α*,* γ *and* $(\rho_S : S \in \mathcal{S})$ *as in Theorem 7.59. For each* $S \in \mathcal{S}$*,*

$$\inf_{\theta \in S} \mathbb{P}_\theta(\widehat{\rho} \leq \rho_S) \geq 1 - \gamma. \tag{7.81}$$

Baraud also gives the following results which show that his construction is essentially optimal. The first result gives a lower bound on any adaptive confidence ball. The result after that shows that the radius ρ_S of his confidence set essentially achieves this lower bound.

7.82 Theorem (Baraud 2004). *Suppose that $\widehat{\theta} = \widehat{\theta}(Z)$ and $r = r(Z)$ are such that $\mathcal{B} = \{\theta : \ ||\theta - \widehat{\theta}||^2 \leq r^2\}$ is a $1-\alpha$ confidence ball. Also suppose that $2\alpha + \gamma < 1 - e^{-1/36}$ and that $d(S) \leq n/2$. If*

$$\inf_{\theta \in S} \mathbb{P}_\theta(r \leq r_S) \geq 1 - \gamma \tag{7.83}$$

then, for some $C = C(\alpha, \gamma) > 0$,

$$r_S^2 \geq C\sigma_n^2 \max\Big\{d(S), \sqrt{n}\Big\}. \tag{7.84}$$

Taking S to consist of a single point yields the same result as Theorem 7.74 and taking $S = \mathbb{R}^n$ yields the same result as Theorem 7.77.

7.85 Theorem (Baraud 2004). *Define $\mathcal{S}$, α, γ and $(\rho_S : \ S \in \mathcal{S})$ as in Theorem 7.59. Assume that $d(S) \leq n/2$ for every $S \in \mathcal{S}$ except for $S = \mathbb{R}^n$. There exists a universal constant $C > 0$ such that*

$$\rho_S^2 \leq C\,\sigma_n^2\, \max\Big\{d(S), \sqrt{n\log(1/\alpha_S)}, \log(1/\alpha_S)\Big\}. \tag{7.86}$$

When σ_n is only known to lie in an interval $I = [\sqrt{1-\eta_n}\tau_n, \tau_n]$, Baraud shows that the lower bound (7.84) becomes

$$r_S^2 \geq C\tau_n^2 \max\Big\{\eta_n n/2, d(S)(1-\eta_n), \sqrt{n - d(S)}(1-\eta_n)\Big\} \tag{7.87}$$

which shows that information about σ is crucial. Indeed, the best we realistically could hope for is to know σ^2 up to order $\eta_n = O(n^{-1/2})$ in which case the lower bound is of order $\max\{\sqrt{n}, d(S)\}$.

7.10 Random Radius Bands?

We have seen that random radius confidence balls can be adaptive in the sense that they can be smaller than fixed radius confidence balls at some points in the parameter space. Is the same true for confidence bands? The answer is no, as follows from results in Low (1997). Actually, Low considers estimating a density f at a single point x but essentially the same results apply to regression and to confidence bands. He shows that any random radius confidence interval for $f(x)$ must have expected width at least as large as a fixed width confidence interval. Thus, there is a qualitative difference between constructing a confidence ball versus a confidence band.

Similar comments apply for other norms. The L_p norm is defined by

$$||\theta||_p = \begin{cases} (\sum_i |\theta_i|^p)^{1/p} & p < \infty \\ \max_i |\theta_i| & p = \infty. \end{cases}$$

Confidence bands can be thought of as L_∞ confidence balls. It can be shown that confidence balls in the L_p norm with $2 < p < \infty$ fall in between the two extremes of L_2 and L_∞ in the sense that they have some adaptivity, but not as much as in the L_2 norm. Similar comments apply to hypothesis testing; see Ingster and Suslina (2003).

7.11 Penalization, Oracles and Sparsity

Consider again the many Normal means problem

$$Z_i \sim \theta_i + \sigma_n \epsilon_i, \quad i = 1, \ldots, n.$$

If we choose $\widehat{\theta}$ to minimize the sums of squares $\sum_{i=1}^n (Z_i - \widehat{\theta}_i)^2$, we get the MLE $\widehat{\theta} = Z = (Z_1, \ldots, Z_n)$. If instead we minimize a penalized sums of squares, we get different estimators.

7.88 Theorem. *Let $J : \mathbb{R}^n \to [0, \infty)$, $\lambda \geq 0$ and define the* **penalized sums of squares**

$$M = \sum_{i=1}^n (Z_i - \theta_i)^2 + \lambda J(\theta).$$

Let $\widehat{\theta}$ minimize M. If $\lambda = 0$ then $\widehat{\theta} = Z$. If $J(\theta) = \sum_{i=1}^n \theta_i^2$ then $\widehat{\theta}_i = Z_i/(1+\lambda)$ which is a linear shrinkage estimator. If $J(\theta) = \sum_{i=1}^n |\theta_i|$ then $\widehat{\theta}$ is the soft-thresholding estimator (7.21). If $J(\theta) = \#\{\theta_i : \ \theta_i \neq 0\}$ then $\widehat{\theta}$ is the hard-thresholding estimator (7.23).

Thus we see that linear shrinkage, soft thresholding and hard thresholding are all special cases of one general approach. The case of the L_1 penalty $\sum_{i=1}^n |\theta_i|$ is especially interesting. According to Theorem 7.88, the estimator that minimizes

$$\sum_{i=1}^n (Z_i - \widehat{\theta}_i)^2 + \lambda \sum_{i=1}^n |\theta_i| \tag{7.89}$$

is the soft-threshold estimator $\widehat{\theta}_\lambda = (\widehat{\theta}_{\lambda,1}, \ldots, \widehat{\theta}_{\lambda,n})$ where

$$\widehat{\theta}_{i,\lambda} = \text{sign}(Z_i)(|Z_i| - \lambda)_+.$$

The criterion (7.89) arises in variable selection for linear regression under the name lasso (Tibshirani (1996)) and in signal processing under the name basis pursuit (Chen et al. (1998)). We will see in Chapter 9 that soft thresholding also plays an important role in wavelet methods.

To get more insight on soft thresholding, we consider a result from Donoho and Johnstone (1994). Consider estimating θ_i and suppose we use either Z_i or 0 as an estimator. Such an estimator might be appropriate if we think the vector θ is sparse in the sense that it has many zeroes. The risk of Z_i is σ_n^2 and the risk of 0 is θ_i^2. Imagine an **oracle** that knows when Z_i has better risk and when 0 has better risk. The risk of the oracles estimator is $\min\{\sigma_n^2, \theta_i^2\}$. The risk for estimating the whole vector θ is

$$R_{\text{oracle}} = \sum_{i=1}^{n} \min\{\sigma_n^2, \theta_i^2\}.$$

Donoho and Johnstone (1994) showed that soft thresholding gives an estimator that comes close to the oracle.

7.90 Theorem (Donoho and Johnstone 1994). *Let $\lambda = \sigma_n\sqrt{2\log n}$. Then, for every $\theta \in \mathbb{R}^n$,*

$$\mathbb{E}_\theta||\widehat{\theta}_\lambda - \theta||^2 \leq (2\log n + 1)(\sigma_n^2 + R_{\text{oracle}}).$$

Moreover, no estimator can get substantially closer to the oracle in the sense that, as $n \to \infty$,

$$\inf_{\widehat{\theta}} \sup_{\theta\in\mathbb{R}^n} \frac{\mathbb{E}_\theta||\widehat{\theta} - \theta||^2}{\sigma_n^2 + R_{\text{oracle}}} \sim 2\log n. \tag{7.91}$$

Consider now a sparse vector θ that is 0 except for k large components, where $k << n$. Then, $R_{\text{oracle}} = k\sigma_n^2$. In function estimation problems, we will see in the next chapter that $\sigma_n^2 = O(1/n)$ and hence $R_{\text{oracle}} = O(k/n)$ which is small in sparse cases (k small).

7.12 Bibliographic Remarks

The idea of reducing nonparametric models to Normal means models (or the white noise model in the appendix) dates back at least to Ibragimov and Has'minskii (1977), Efromovich and Pinsker (1982), and others. See Brown and Low (1996), Nussbaum (1996a) for examples of recent results in this area. A thorough treatment of Normal decision theory and its relation to nonparametric problems is contained in Johnstone (2003). There is also a substantial literature on hypothesis testing in this framework. Many of the results are due to Ingster and are summarized in Ingster and Suslina (2003).

7.13 Appendix

The White Noise Model. Regression is also connected with the **white noise model**. Here is a brief description. Recall that a standard Brownian motion $W(t)$, $0 \leq t \leq 1$ is a random function such that $W(0) = 0$, $W(s+t) - W(s) \sim N(0,t)$ and, $W(v) - W(u)$ is independent of $W(t) - W(s)$ for $0 \leq u \leq v \leq s \leq t$. You can think of W as a continuous version of a random walk. Let $Z_i = f(i/n) + \sigma\epsilon_i$ with $\epsilon_i \sim N(0,1)$. For $0 \leq t \leq 1$, define

$$Z_n(t) = \frac{1}{n}\sum_{i=1}^{[nt]} Z_i = \frac{1}{n}\sum_{i=1}^{[nt]} f(i/n) + \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{i=1}^{[nt]} Z_i.$$

The term $\frac{1}{n}\sum_{i=1}^{[nt]} f(i/n)$ converges to $\int_0^t f(s)ds$ as $n \to \infty$. The term $n^{-1/2}$ $\sum_{i=1}^{[nt]} Z_i$ converges to a standard Brownian motion. (For any fixed t, this is just an application of the central limit theorem.) Thus, asymptotically we can write

$$Z(t) = \int_0^t f(s)ds + \frac{\sigma}{\sqrt{n}}W(t).$$

This is called the **standard white noise model**, often written in differential form as

$$dZ(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW(t) \tag{7.92}$$

where $dW(t)$ is the white noise process.[6]

Let $\phi_1, \phi_2, \ldots$ be an orthonormal basis for $L_2(0,1)$ and write $f(x) = \sum_{i=1}^{\infty}\theta_i\phi_i(x)$ where $\theta_i = \int f(x)\phi_i(x)dx$. Multiply (7.92) by ϕ_j and integrate. This yields $Z_i = \theta_i + (\sigma/\sqrt{n})\epsilon_i$ where $Z_i = \int \phi_i(t)dZ(t)$ and $\epsilon_i = \int \phi_i(t)dW(t) \sim N(0,1)$. We are back to the Normal means problem. A more complicated argument can be used to relate density estimation to the white noise model as in Nussbaum (1996a).

Weak Differentiability. Let f be integrable on every bounded interval. Then f is **weakly differentiable** if there exists a function f' that is integrable on every bounded interval, such that

$$\int_x^y f'(s)ds = f(y) - f(x)$$

whenever $x \leq y$. We call f' the weak derivative of f. An equivalent condition is that for every ϕ that is compactly supported and infinitely differentiable,

$$\int f(s)\phi'(s)ds = -\int f'(s)\phi(s)ds.$$

[6] Intuitively, think of $dW(t)$ as a vector of Normals on a very fine grid.

See Härdle et al. (1998), page 72.

Proof of Pinsker's Theorem (Theorem 7.28). (Following Nussbaum (1996b).) We will need to use Bayes estimators, which we now review. Let π_n be a prior for θ^n. The **integrated risk** is defined to be $B(\widehat{\theta}, \pi_n) = \int R(\widehat{\theta}^n, \theta^n) d\pi_n(\theta^n) = \mathbb{E}_{\pi_n} \mathbb{E}_\theta L(\widehat{\theta}, \theta)$. The **Bayes estimator** $\widehat{\theta}_{\pi_n}$ minimizes the Bayes risk:

$$B(\pi_n) = \inf_{\widehat{\theta}} B(\widehat{\theta}^n, \pi_n). \tag{7.93}$$

An explicit formula for the Bayes estimator is

$$\widehat{\theta}_{\pi_n}(y) = \text{argmin}_a \mathbb{E}\big(L(a, \theta) \mid Z^n\big).$$

In the case of squared error loss $L(a, \theta) = ||a - \theta||_n^2$, the Bayes estimator is $\widehat{\theta}_{\pi_n}(y) = \mathbb{E}(\theta | Z^n)$.

Let $\Theta_n = \Theta_n(c)$. Let

$$R_n = \inf_{\widehat{\theta}} \sup_{\theta \in \Theta_n} R(\widehat{\theta}, \theta)$$

denote the minimax risk. We will find an upper bound and a lower bound on the risk.

Upper Bound. Let $\widehat{\theta}_j = c^2 Z_j/(\sigma^2 + c^2)$. The bias of this estimator is

$$\mathbb{E}_\theta(\widehat{\theta}_j) - \theta_j = -\frac{\sigma^2 \theta_j}{\sigma^2 + c^2}$$

and the variance is

$$\mathbb{V}_\theta(\widehat{\theta}_j) = \left(\frac{c^2}{c^2 + \sigma^2}\right)^2 \sigma_n^2 = \left(\frac{c^2}{c^2 + \sigma^2}\right)^2 \frac{\sigma^2}{n}$$

and hence the risk is

$$\begin{aligned}
\mathbb{E}_\theta ||\widehat{\theta} - \theta||^2 &= \sum_{j=1}^n \left[\left(\frac{\sigma^2 \theta_j}{\sigma^2 + c^2}\right)^2 + \left(\frac{c^2}{c^2 + \sigma^2}\right)^2 \left(\frac{\sigma^2}{n}\right) \right] \\
&= \left(\frac{\sigma^2}{\sigma^2 + c^2}\right)^2 \sum_{j=1}^n \theta_j^2 + \sigma^2 \left(\frac{\sigma^2}{\sigma^2 + c^2}\right)^2 \\
&\leq c^2 \left(\frac{\sigma^2}{\sigma^2 + c^2}\right)^2 + \sigma^2 \left(\frac{\sigma^2}{\sigma^2 + c^2}\right)^2 \\
&= \frac{\sigma^2 c^2}{\sigma^2 + c^2}.
\end{aligned}$$

Hence,

$$R_n \leq \frac{c^2 \sigma^2}{c^2 + \sigma^2}$$

for all n.

LOWER BOUND. Fix $0 < \delta < 1$. Let π_n be a Normal prior for which $\theta_1, \ldots, \theta_n$ are IID $N(0, c^2\delta^2/n)$. Let $B(\pi_n)$ denote the Bayes risk. Recall that $B(\pi_n)$ minimizes the integrated risk $B(\widehat{\theta}, \pi_n)$ over all estimators. The minimum is obtained by taking $\widehat{\theta}$ to be the posterior mean which has coordinates $\widehat{\theta}_j = c^2\delta^2 Z_j/(c^2\delta^2 + \sigma^2)$ with risk

$$R(\theta, \widehat{\theta}) = \sum_{i=1}^{n} \left[\theta_i^2 \left(\frac{\sigma_n^2}{\frac{c^2\delta^2}{n} + \sigma_n^2} \right)^2 + \sigma^2 \left(\frac{\frac{c^2\delta^2}{n}}{\frac{c^2\delta^2}{n} + \sigma_n^2} \right)^2 \right].$$

The Bayes risk is

$$B(\pi_n) = \int R(\theta, \widehat{\theta}) d\pi_n(\theta) = \frac{\sigma^2\delta^2c^2}{\sigma^2 + \delta^2c^2}.$$

So, for any estimator $\widehat{\theta}$,

$$\begin{aligned} B(\pi_n) &\leq B(\widehat{\theta}, \pi_n) \\ &= \int_{\Theta_n} R(\theta, \widehat{\theta}) d\pi_n + \int_{\Theta_n^c} R(\theta, \widehat{\theta}) d\pi_n \\ &\leq \sup_{\theta \in \Theta_n} R(\theta, \widehat{\theta}) + \int_{\Theta_n^c} R(\theta, \widehat{\theta}) d\pi_n \\ &\leq \sup_{\theta \in \Theta_n} R(\theta, \widehat{\theta}) + \sup_{\widehat{\theta}} \int_{\Theta_n^c} R(\theta, \widehat{\theta}) d\pi_n. \end{aligned}$$

Taking the infimum over all estimators that take values in Θ_n yields

$$B(\pi_n) \leq R_n + \sup_{\widehat{\theta}} \int_{\Theta_n^c} R(\theta, \widehat{\theta}) d\pi_n.$$

Hence,

$$\begin{aligned} R_n &\geq B(\pi_n) - \sup_{\widehat{\theta}} \int_{\Theta_n^c} R(\theta, \widehat{\theta}) d\pi_n \\ &= \frac{\sigma^2\delta^2c^2}{\delta^2c^2 + \sigma^2} - \sup_{\widehat{\theta}} \int_{\Theta_n^c} R(\theta, \widehat{\theta}) d\pi_n. \end{aligned}$$

Now, using the fact that $||a + b||^2 \leq 2(||a||^2 + ||b||^2)$, and the Cauchy–Schwartz inequality,

$$\begin{aligned} \sup_{\widehat{\theta}} \int_{\Theta_n^c} R(\theta, \widehat{\theta}) d\pi_n &\leq 2 \int_{\Theta_n^c} ||\theta||^2 d\pi_n + 2 \sup_{\widehat{\theta}} \int_{\Theta_n^c} \mathbb{E}_\theta ||\widehat{\theta}||^2 d\pi_n \\ &\leq 2\sqrt{\pi_n(\Theta_n^c)} \sqrt{\mathbb{E}_{\pi_n} \left(\sum_j \theta_j^2 \right)^2} + 2c^2 \pi_n(\Theta_n^c). \end{aligned}$$

Thus,

$$R_n \geq \frac{\sigma^2\delta^2c^2}{\sigma^2+\delta^2c^2} - 2\sqrt{\pi_n(\Theta_n^c)}\sqrt{\mathbb{E}_{\pi_n}\left(\sum_j \theta_j^2\right)^2} - 2c^2\pi_n(\Theta_n^c). \tag{7.94}$$

We now bound the last two terms in (7.94).

We shall make use of the following large deviation inequality: if $Z_1, \ldots, Z_n \sim N(0,1)$ and $0 < t < 1$, then

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_j (Z_j^2 - 1)\right| > t\right) \leq 2e^{-nt^2/8}.$$

Let $Z_j = \sqrt{n}\theta_j/(c\delta)$ and let $t = (1-\delta^2)/\delta^2$. Then,

$$\begin{aligned}
\pi_n(\Theta_n^c) &= \mathbb{P}\left(\sum_{j=1}^n \theta_j^2 > c^2\right) = \mathbb{P}\left(\frac{1}{n}\sum_{j=1}^n (Z_j^2-1) > t\right) \\
&\leq \mathbb{P}\left(\left|\frac{1}{n}\sum_j (Z_j^2-1)\right| > t\right) \leq 2e^{-nt^2/8}.
\end{aligned}$$

Next, we note that

$$\begin{aligned}
\mathbb{E}_{\pi_n}\left(\sum_j \theta_j^2\right)^2 &= \sum_{i=1}^n \mathbb{E}_{\pi_n}(\theta_i^4) + \sum_{i=1}^n \sum_{j\neq i}^n \mathbb{E}_{\pi_n}(\theta_i^2)\mathbb{E}_{\pi_n}(\theta_j^2) \\
&= \frac{c^4\delta^4\mathbb{E}(Z_1^4)}{n} + \binom{n}{2}\frac{c^4\delta^4}{n^2} = O(1).
\end{aligned}$$

Therefore, from (7.94),

$$R_n \geq \frac{\sigma^2\delta^2c^2}{\sigma^2+\delta^2c^2} - 2\sqrt{2}e^{-nt^2/16}O(1) - 2c^2e^{-nt^2/8}.$$

Hence,

$$\liminf_{n\to\infty} R_n \geq \frac{\sigma^2\delta^2c^2}{\sigma^2+\delta^2c^2}.$$

The conclusion follows by letting $\delta \uparrow 1$. ∎

Proof of Theorem 7.74. Let

$$a = \frac{\sigma_n}{n^{1/4}}(\log(1+\epsilon^2))^{1/4}$$

and define

$$\Omega = \Big\{\theta = (\theta_1, \ldots, \theta_n) : \ |\theta_i| = a, \quad i = 1, \ldots, n\Big\}.$$

Note that Ω contains 2^n elements. Let f_θ denote the density of a multivariate Normal with mean θ and covariance $\sigma_n^2 I$ where I is the identity matrix. Define the mixture

$$q(y) = \frac{1}{2^n} \sum_{\theta \in \Omega} f_\theta(y).$$

Let f_0 denote the density of a multivariate Normal with mean $(0, \ldots, 0)$ and covariance $\sigma_n^2 I$. Then,

$$\begin{aligned} \int |f_0(x) - g(x)| dx &= \int \frac{|f_0(x) - g(x)|}{\sqrt{f_0(x)}} \sqrt{f_0(x)} dx \\ &\leq \sqrt{\int \frac{(f_0(x) - g(x))^2}{f_0(x)} dx} \\ &= \sqrt{\int \frac{g^2(x)}{f_0(x)} dx - 1}. \end{aligned}$$

Now,

$$\begin{aligned} \int \frac{q^2(x)}{f_0(x)} dx &= \int \left(\frac{q(x)}{f_0(x)}\right)^2 f_0(x) dx = \mathbb{E}_0 \left(\frac{q(x)}{f_0(x)}\right)^2 \\ &= \left(\frac{1}{2^n}\right)^2 \sum_{\theta,\nu \in \Omega} \mathbb{E}_0 \left(\frac{f_\theta(x) f_\nu(x)}{f_0^2(x)}\right) \\ &= \left(\frac{1}{2^n}\right)^2 \sum_{\theta,\nu \in \Omega} \exp\left\{-\frac{1}{2\sigma_n^2}(||\theta||^2 + ||\nu||^2)\right\} \mathbb{E}_0 \left(\exp\left\{\epsilon^T(\theta + \nu)/\sigma_n^2\right\}\right) \\ &= \left(\frac{1}{2^n}\right)^2 \sum_{\theta,\nu \in \Omega} \exp\left\{-\frac{1}{2\sigma_n^2}(||\theta||^2 + ||\nu||^2)\right\} \exp\left\{\sum_{i=1}^n (\theta_i + \nu_i)^2/(2\sigma_n^2)\right\} \\ &= \left(\frac{1}{2^n}\right)^2 \sum_{\theta,\nu \in \Omega} \exp\left\{\frac{\langle \theta, \nu \rangle}{\sigma_n^2}\right\}. \end{aligned}$$

The latter is equal to the mean of $\exp(\langle \theta, \nu \rangle / \sigma_n^2)$ when drawing two vectors θ and ν at random from Ω. And this, in turn, is equal to

$$\mathbb{E} \exp\left\{\frac{a^2 \sum_{i=1}^n E_i}{\sigma_n^2}\right\}$$

where $E_1, \ldots, E_n$ are independent and $\mathbb{P}(E_i = 1) = \mathbb{P}(E_i = -1) = 1/2$. Moreover,

$$\begin{aligned}
\mathbb{E} \exp\left\{ \frac{a^2 \sum_{i=1}^n E_i}{\sigma_n^2} \right\} &= \prod_{i=1}^n \mathbb{E} \exp\left\{ \frac{a^2 E_i}{\sigma_n^2} \right\} \\
&= \left(\mathbb{E} \exp\left\{ \frac{a^2 E_1}{\sigma_n^2} \right\} \right)^n \\
&= \left(\cosh\left(\frac{a^2}{\sigma_n^2} \right) \right)^n
\end{aligned}$$

where $\cosh(y) = (e^y + e^{-y})/2$. Thus,

$$\int \frac{q^2(x)}{f_0(x)} dx = \left(\cosh\left(\frac{a^2}{\sigma_n^2} \right) \right)^n \leq e^{a^4 n/\sigma_n^4}$$

where we have used the fact that $\cosh(y) \leq e^{y^2}$. Thus,

$$\int |f_0(x) - q(x)| dx \leq \sqrt{e^{a^4 n/\sigma_n^4} - 1} = \epsilon.$$

So, if Q denotes the probability measure with density q, we have, for any event A,

$$\begin{aligned}
Q(A) &= \int_A q(x)dx = \int_A f_0(x)dx + \int_A (q(x) - f_0(x))dx \\
&\geq \mathbb{P}_0(A) - \int_A |q(x) - f_0(x)| dx \geq \mathbb{P}_0(A) - \epsilon. \qquad (7.95)
\end{aligned}$$

Define two events, $A = \{(0, \ldots, 0) \in \mathcal{B}_n\}$ and $B = \{\Omega \bigcap \mathcal{B}_n \neq \emptyset\}$. Every $\theta \in \Omega$ has norm

$$||\theta|| = \sqrt{na^2} = \sigma_n n^{1/4} (\log(1 + \epsilon^2))^{1/4} \equiv c_n.$$

Hence, $A \bigcap B \subset \{s_n \geq c_n\}$. Since $\mathbb{P}_\theta(\theta \in \mathcal{B}_n) \geq 1 - \alpha$ for all θ, it follows that $\mathbb{P}_\theta(B) \geq 1 - \alpha$ for all $\theta \in \Omega$. Hence, $Q(B) \geq 1 - \alpha$. From (7.95),

$$\begin{aligned}
\mathbb{P}_0(s_n \geq c_n) &\geq \mathbb{P}_0(A \bigcap B) \geq Q(A \bigcap B) - \epsilon \\
&= Q(A) + Q(B) - Q(A \bigcup B) - \epsilon \\
&\geq Q(A) + Q(B) - 1 - \epsilon \\
&\geq Q(A) + (1 - \alpha) - 1 - \epsilon \\
&\geq \mathbb{P}_0(A) + (1 - \alpha) - 1 - 2\epsilon \\
&\geq (1 - \alpha) + (1 - \alpha) - 1 - 2\epsilon \\
&= 1 - 2\alpha - 2\epsilon.
\end{aligned}$$

So, $\mathbb{E}_0(s_n) \geq (1-2\alpha-2\epsilon)c_n$. It is easy to see that the same argument can be used for any $\theta \in \mathbb{R}^n$ and hence $\mathbb{E}_\theta(s_n) \geq (1-2\alpha-2\epsilon)c_n$ for every $\theta \in \mathbb{R}^n$. ∎

Proof of Theorem 7.77. Let $a = \sigma_n z_{\alpha+2\epsilon}$ where $0 < \epsilon < (1/2)(1/2-\alpha)$ and define

$$\Omega = \Big\{\theta = (\theta_1, \ldots, \theta_n) : \ |\theta_i| = a, \quad i = 1, \ldots, n\Big\}.$$

Define the loss function $L = L(\widehat{\theta}, \theta) = \sum_{i=1}^n I(|\widehat{\theta}_i - \theta_i| \geq a)$. Let π be the uniform prior on Ω. The posterior mass function over Ω is $p(\theta|y) = \prod_{i=1}^n p(\theta_i|y_i)$ where

$$p(\theta_i|y_i) = \frac{e^{2ay_i/\sigma_n^2}}{1+e^{2ay_i/\sigma_n^2}} I(\theta_i = a) + \frac{1}{1+e^{2ay_i/\sigma_n^2}} I(\theta_i = -a).$$

The posterior risk is

$$\mathbb{E}(L(\widehat{\theta}, \theta)|y) = \sum_{i=1}^n \mathbb{P}(|\widehat{\theta}_i - \theta_i| \geq a|y_i)$$

which is minimized by taking $\widehat{\theta}_i = a$ if $y_i \geq 0$ and $\widehat{\theta}_i = -a$ if $y_i < 0$. The risk of this estimator is

$$\begin{aligned}
&\sum_{i=1}^n \left(\mathbb{P}(Y_i < 0|\theta_i = a) I(\theta_i = a) + \mathbb{P}(Y_i > 0|\theta_i = -a) I(\theta_i = -a) \right) \\
&\quad = n\Phi(-a/\sigma_n) = n(\alpha + 2\epsilon).
\end{aligned}$$

Since this risk is constant, it is the minimax risk. Therefore,

$$\begin{aligned}
\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^n} \sum_{i=1}^n \mathbb{P}_\theta(|\widehat{\theta}_i - \theta_i| \geq a) &\geq \inf_{\widehat{\theta}} \sup_{\theta \in \Omega} \sum_{i=1}^n \mathbb{P}_\theta(|\widehat{\theta}_i - \theta_i| \geq a) \\
&= n(\alpha + 2\epsilon).
\end{aligned}$$

Let $\gamma = \epsilon/(1-\alpha-\epsilon)$. Given any estimator $\widehat{\theta}$,

$$\gamma n \mathbb{P}_\theta(L < \gamma n) + n\mathbb{P}_\theta(L \geq \gamma n) \geq L$$

and so

$$\sup_\theta \big(\gamma n \mathbb{P}_\theta(L < \gamma n) + n\mathbb{P}_\theta(L \geq \gamma n)\big) \geq \sup_\theta \mathbb{E}_\theta(L) \geq n(\alpha + 2\epsilon).$$

This inequality, together with the fact that $\mathbb{P}_\theta(L < \gamma n) + \mathbb{P}_\theta(L \geq \gamma n) = 1$ implies that

$$\sup_\theta \mathbb{P}_\theta(L \geq \gamma n) \geq \alpha + \epsilon.$$

Thus,

$$\sup_\theta \mathbb{P}_\theta(||\widehat{\theta} - \theta||^2 \geq \gamma n a^2) \geq \sup_\theta \mathbb{P}_\theta(L \geq \gamma n) \geq \alpha + \epsilon.$$

Therefore,

$$\begin{aligned}
\sup_\theta \mathbb{P}_\theta(s_n^2 \geq \gamma n a^2) &\geq \sup_\theta \mathbb{P}_\theta(s_n^2 \geq ||\widehat{\theta} - \theta||^2 \geq \gamma n a^2) \\
&= \sup_\theta \mathbb{P}_\theta(s_n^2 \geq ||\widehat{\theta} - \theta||^2) + \sup_\theta \mathbb{P}_\theta(||\widehat{\theta} - \theta||^2 \geq \gamma n a^2) - 1 \\
&\geq \alpha + \epsilon + 1 - \alpha - 1 = \epsilon.
\end{aligned}$$

Thus, $\sup_\theta E_\theta(s_n) \geq \epsilon a\sqrt{\gamma n}$. ∎

7.14 Exercises

1. Let $\theta_i = 1/i^2$ for $i = 1, \ldots, n$. Take $n = 1000$. Let $Z_i \sim N(\theta_i, 1)$ for $i = 1, \ldots, n$. Compute the risk of the MLE. Compute the risk of the estimator $\widetilde{\theta} = (bZ_1, bZ_2, \ldots, bZ_n)$. Plot this risk as a function of b. Find the optimal value b_*. Now conduct a simulation. For each run of the simulation, find the (modified) James–Stein estimator $\widehat{b}Z$ where

$$\widehat{b} = \left[1 - \frac{n}{\sum_i Z_i^2}\right]^+.$$

   You will get one $\widehat{b}$ for each simulation. Compare the simulated values of $\widehat{b}$ to b_*. Also, compare the risk of the MLE and the James–Stein estimator (the latter obtained by simulation) to the Pinsker bound.

2. For the Normal means problem, consider the following *curved soft threshold estimator*:

$$\widehat{\theta}_i = \begin{cases} -(Z_i + \lambda)^2 & Z_i < -\lambda \\ 0 & -\lambda \leq Z_i \leq \lambda \\ (Z_i - \lambda)^2 & Z_i > \lambda \end{cases}$$

   where $\lambda > 0$ is some fixed constant.

   (a) Find the risk of this estimator. *Hint*: $R = \mathbb{E}(\text{SURE})$.

   (b) Consider problem (1). Use your estimator from (2a) with λ chosen from the data using SURE. Compare the risk to the risk of the James–Stein estimator. Now repeat the comparison for

$$\theta = (\overbrace{10, \ldots, 10}^{10 \text{ times}}, \overbrace{0, \ldots, 0}^{990 \text{ times}}).$$

3. Let $J = J_n$ be such that $J_n \to \infty$ and $n \to \infty$. Let

$$\widehat{\sigma}^2 = \frac{n}{J} \sum_{i=n-J+1}^{n} Z_i^2$$

where $Z_i \sim N(\theta_i, \sigma^2/n)$. Show that if $\theta = (\theta_1, \theta_2, \ldots)$ belongs to a Sobolev body of order $m > 1/2$ then $\widehat{\sigma}^2$ is a uniformly consistent estimator of σ^2 in the Normal means model.

4. Prove Stein's lemma: if $X \sim N(\mu, \sigma^2)$ then $\mathbb{E}(g(X)(X-\mu)) = \sigma^2 \mathbb{E} g'(X)$.

5. Verify equation 7.22.

6. Show that the hard threshold estimator defined in (7.23) is not weakly differentiable.

7. Compute the risk functions for the soft threshold estimator (7.21) and the hard threshold estimator (7.23).

8. Generate $Z_i \sim N(\theta_i, 1)$, $i = 1, \ldots, 100$, where $\theta_i = 1/i$. Compute a 95 percent confidence ball using: (i) the χ^2 confidence ball, (ii) the Baraud method, (iii) the pivotal method. Repeat 1000 times and compare the radii of the balls.

9. Let $||a-b||_\infty = \sup_j |a_j - b_j|$. Construct a confidence set B_n of the form $B_n = \{\theta \in \mathbb{R}^n : \ ||\theta - Z^n||_\infty \leq c_n\}$ such that $\mathbb{P}_\theta(\theta \in B_n) \geq 1-\alpha$ for all $\theta \in \mathbb{R}^n$ under model (7.1) with $\sigma_n = \sigma/\sqrt{n}$. Find the expected diameter of your confidence set.

10. Consider Example 7.24. Define

$$\delta = \max_{S \in \mathcal{S}} \sup_{\theta \in \mathbb{R}^n} |\widehat{R}_S - R(\widehat{\theta}_S, \theta)|.$$

Try to bound δ in the following three cases: (i) $\mathcal{S}$ consists of a single model S; (ii) nested model selection; (iii) all subsets selection.

11. Consider Example 7.24. Another method for choosing a model is to use penalized likelihood. In particular, some well-known penalization model selection methods are **AIC** (Akaike's Information Criterion), Akaike (1973), **Mallows'** C_p, Mallows (1973), and **BIC** (Bayesian Information Criterion), Schwarz (1978). In the Normal means model, minimizing

SURE, AIC and C_p are equivalent. But BIC leads to a different model selection procedure. Specifically,

$$\text{BIC}_B = \ell_B - \frac{|B|}{2} \log n$$

where ℓ_B is the log-likelihood of the submodel B evaluated at its maximum likelihood estimator. Find an explicit expression for BIC_B. Suppose we choose B by maximizing BIC_B over $\mathcal{B}$. Investigate the properties of this model selection procedure and compare it to selecting a model by minimizing SURE. In particular, compare the risk of the resulting estimators. Also, assuming there is a "true" submodel (that is, $\theta_i \neq 0$ if and only if $i \in B$), compare the probability of selecting the true submodel under each procedure. In general, estimating θ accurately and finding the true submodel are not the same. See Wasserman (2000).

12. By approximating the noncentral χ^2 with a Normal, find a large sample approximation to for ρ_0 and ρ_n in Example 7.69. Then prove equation (7.70).