Droge, Bernd

Working Paper

On the minimax regret estimation of a restricted normal mean, and implications

SFB 373 Discussion Paper, No. 2002,81

Provided in Cooperation with:
Collaborative Research Center 373: Quantification and Simulation of Economic Processes, Humboldt University Berlin

Suggested Citation: Droge, Bernd (2002) : On the minimax regret estimation of a restricted normal mean, and implications, SFB 373 Discussion Paper, No. 2002,81, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification and Simulation of Economic Processes, Berlin, https://nbn-resolving.de/urn:nbn:de:kobv:11-10049549

This Version is available at:
https://hdl.handle.net/10419/65311

On the Minimax Regret Estimation of a Restricted Normal Mean, and Implications

Bernd Droge

Sonderforschungsbereich 373, Humboldt University,

Unter den Linden 6, 10099 Berlin, Germany

Summary. Consider estimating the mean of a normal distribution with known variance, when that mean is known to lie in a bounded interval. In a decision-theoretic framework we study finite sample properties of a class of nonlinear estimators. These estimators are based on thresholding techniques which have become very popular in the context of wavelet estimation. Under squared error loss we show that there exists a unique minimax regret solution for the problem of selecting the threshold. For comparison, the behaviour of linear shrinkers is also investigated.

In special cases we illustrate the implications of our results for the problem of estimating the regression function in a nonparametric situation. This is possible since, as usual, a coordinatewise application of the scalar results leads immediately to results for multivariate (sequence space) problems. Then it is well known that orthogonal transformations can be employed to turn statements about estimation over coefficient bodies in sequence space into statements about estimation over classes of smooth functions in noisy data. The performance of the proposed minimax regret optimal curve etimator is demonstrated by simulated data examples.

AMS 1991 subject classifications: Primary 62C99; secondary 62G07.

Key words: Bounded normal mean, soft thresholding, minimax regret decision theory, nonlinear estimation; nonparametric regression; orthogonal series estimation.

The research on this paper was carried out within the Sonderforschungsbereich 373 at Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

1 Introduction

Suppose we observe a normally distributed random variable

$$y \sim N(\theta, \sigma^2) \ , \tag{1}$$

where $\sigma^2 > 0$ is known. We wish to estimate the mean θ when prior knowledge specifies that this mean lies in a bounded interval, say,

$$\theta \in [-c, c] \quad \text{for some} \quad c > 0 \ . \tag{2}$$

The performance of an estimator $\hat{\theta}$ of θ may be measured by a squared-error loss, $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|^2$, and the associated risk is then the expected loss (mean squared error)

$$R(\hat{\theta}, \theta) = EL(\hat{\theta}, \theta) \ .$$

In the absence of prior information on θ, the usual (e.g. maximum likelihood) estimator $\hat{\theta}_0 = y$ would be admissible and minimax, but both properties are lost under (2).

The problem of estimating a bounded normal mean has been investigated by a number of authors. For example, Casella & Strawderman (1981) and Bickel (1981) focused on the minimax approach, whereas Gatsonis et al. (1987) compared the Bayes estimator corresponding to the uniform prior on $[-c, c]$ with the maximum likelihood estimator and the estimators proposed by Casella & Strawderman and Bickel.

In the last decade, wavelet methods have been applied to various curve estimation problems, since they can effectively compress signals with possible irregularities. One feature of a variety of such methods is that they act coordinatewise. Hence, many decision-theoretic results in this field can be obtained by studying scalar statistical problems.

In this paper we apply the minimax regret approach to a class of linear and nonlinear shrunk estimators. The minimax regret principle as well as the class of estimators are introduced in the next section. The nonlinear estimators are based on the soft thresholding technique which has become very popular in the context of wavelet estimation. We show that there exists a unique minimax regret solution for the problem of selecting the threshold. This optimal threshold is a monotonically decreasing function of the parameter bound c, and turns out to be independent of the prior constraint for rather large values of c. On the other hand, for very precise prior constraints it is found that linear shrinkage may outperform the soft thresholding technique in the minimax regret sense. A disadvantage of linear shrinkers is,

however, that their risk function is generally unbounded when θ varies on the whole real line. This property is shared by the maximum likelihood estimator.

In Section 3 we illustrate how our results can be applied to the problem of estimating the regression function in a nonparametric situation. Under specific prior constraints on the regression function it turns out that the scalar results imply immediately optimal curve estimates. The considered methods are applied to some simulated data examples, which suggest to use the minimax regret soft thresholding rule to avoid serious bias problems in cases of possibly misspecified prior constraints. All proofs are deferred to the Appendix.

2 Minimax regret shrinkage

2.1 Some shrunk estimators

In the literature, different classes of estimators of θ have been considered. Many of them may be seen as shrinkers. A simple class is given by the set of linear shrinkers

$$\hat{\Theta}^L = \{\hat{\theta}^L_\lambda \mid \hat{\theta}^L_\lambda(y) = \lambda y \ , \ \lambda \in [0,1]\}. \tag{3}$$

But there exists also a variety of nonlinear estimators which shrink y towards 0 in a data-dependent way. Here we consider only one of the various forms of thresholding estimators:

$$\hat{\Theta}^S = \{\hat{\theta}^S_\lambda \mid \hat{\theta}^S_\lambda(y) = \mathrm{sgn}(y)(|y| - \sigma\lambda)_+ \ , \ \lambda \in [0,\infty]\} \ , \tag{4}$$

where $a_+ = \max(a, 0)$. These estimators are known as soft threshold rules in the context of wavelet estimation, compare e.g. Donoho & Johnstone (1994). A key feature of the estimators is that inside a threshold zone $[-\lambda\sigma, \lambda\sigma]$ the data is interpreted as noise. In nonparametric regression problems, coordinatewise thresholding of empirical (wavelet) coefficients lead to estimators of the regression function that possess several optimal and near-optimal properties over a wide class of (smooth) function spaces. Details and an extensive discussion may be found in Donoho et al. (1995). This motivates the consideration of the class (4), since the study of questions about estimation over classes of smooth functions in noisy data can often be reduced to the study of univariate Gaussian problems.

In what follows, we will need the exact risk of the considered estimators. Let $\tau = \theta/\sigma$ and $\hat{\tau} = y/\sigma$. Then we have

$$R(\hat{\theta}^L_\lambda, \theta) = \sigma^2 \rho_L(\lambda, \tau) \ , \quad R(\hat{\theta}^S_\lambda, \theta) = \sigma^2 \rho_S(\lambda, \tau) \ ,$$

where the normalized risks ρ_L, ρ_S are given by

$$\begin{aligned} \rho_L(\lambda,\tau) &= E|\lambda\hat{\tau}-\tau|^2 = \lambda^2+(1-\lambda)^2\tau^2 \quad \text{and} \qquad (5) \\ \rho_S(\lambda,\tau) &= E\{\mathrm{sgn}(\hat{\tau})(|\hat{\tau}|-\lambda)_+ - \tau\}^2 \\ &= 1+\lambda^2+(\tau^2-\lambda^2-1)\{\Phi(\lambda-\tau)-\Phi(-\lambda-\tau)\} \\ &\quad -(\lambda+\tau)\phi(\lambda-\tau)-(\lambda-\tau)\phi(\lambda+\tau)\ , \qquad (6) \end{aligned}$$

see, for example, Donoho & Johnstone (1994) and Droge (1998). Here, Φ and ϕ denote the distribution function and the density of the standard normal law, respectively. Clearly, $\rho_L(\lambda,\tau)$ is unbounded unless $\lambda = 1$, whereas, for fixed threshold λ, $\rho_S(\lambda,\tau)$ is monotonically increasing and approaches $1+\lambda^2$ as $\tau \to \infty$.

2.2 Minimax regret principle

The minimax principle tries to protect against the worst possible magnitude of the risk. Hence it ignores the fact that there is often a lower risk bound which cannot be improved by any permissible estimator. An obvious way to remedy the situation is to consider the so-called regret risk of an estimator $\hat{\theta}$,

$$reg(\hat{\theta},\theta;\hat{\Theta}) = R(\hat{\theta},\theta) - \inf_{\tilde{\theta}\in\hat{\Theta}} R(\tilde{\theta},\theta), \qquad (7)$$

where $\hat{\Theta}$ denotes some class of estimators which is of interest in the problem under consideration. Although the lower risk bound, $\inf_{\tilde{\theta}\in\hat{\Theta}} R(\tilde{\theta},\theta)$, cannot be attained for all θ by any estimator $\hat{\theta} \in \hat{\Theta}$, the aim is to be as close as possible to this ideal bound. Consequently, given the underlying mean θ, the regret risk evaluates estimators on the basis of how they compare with the best possible one within $\hat{\Theta}$. A minimax regret estimator $\hat{\theta}^*$ of θ is then defined by being minimax with respect to the regret risk (Berger, 1985, § 5.5.5), that is

$$\hat{\theta}^* \in \arg\inf_{\hat{\theta}\in\hat{\Theta}} \sup_{|\theta|\le c} reg(\hat{\theta},\theta;\hat{\Theta}). \qquad (8)$$

If necessary, $\hat{\theta}^*$ defined by (8) will be called $\hat{\Theta}$-minimax regret estimator since, to be precise, it is $\hat{\Theta}$-minimax (restricted minimax) w.r.t. the regret risk.

Droge (1993) and Droge & Georg (1995) have applied the minimax regret principle to the problem of model selection in linear regression when the regressors are orthogonal. The first paper deals with the case of a known error variance, whereas the latter covers the case when this variance is unknown. The selection estimators considered in both papers may be seen as so-called hard threshold rules in the context of wavelet estimation. Model selection based on minimizing Mallows' (1973) C_p

criterion turns out to be nearly optimal in the minimax regret sense. In the problem of orthogonal series regression estimation, hard and soft thresholding techniques have been compared by Droge (1998). The use of the minimax regret approach shows the superiority of soft thresholding over hard thresholding and provides thus an additional motivation for dealing just with the class (4) of estimators. All three papers are concerned with estimating the regression function when no prior information is available about its shape. The key step in deriving the optimal rules in all three papers is therefore to solve the related univariate Gaussian problem without prior constraints on the parameter. In contrast, the next subsection is devoted to the scalar minimax regret problem under the prior information (2).

2.3 Minimax regret linear and soft threshold estimation

We aim at finding the minimax regret solution within the class

$$\hat{\Theta} = \hat{\Theta}_L \cup \hat{\Theta}_S \tag{9}$$

of linear and soft threshold estimators. First, we need a lower risk bound. Minimizing the risk in $\hat{\Theta}_L$ over $\lambda \in [0,1]$ based on full knowledge of θ provides the ideal linear shrunk 'estimator',

$$\hat{\theta}^L_{\lambda^*} = \lambda^* y \ , \quad \lambda^* = \lambda^*(\tau) = \frac{\tau^2}{1+\tau^2} \ ; \tag{10}$$

see e.g. Donoho & Johnstone (1994). Moreover, Droge (1998) has verified that the risk of (10) gives, for each θ, a lower bound for the risk of all shrunk estimators in $\hat{\Theta}$:

$$\inf_{\tilde{\theta} \in \hat{\Theta}} R(\tilde{\theta}, \theta) = \lambda^* \sigma^2 = R(\hat{\theta}^L_{\lambda^*}, \theta) \ . \tag{11}$$

We remark that (11) remains true if $\hat{\Theta}$ is replaced by the larger set of estimators

$$\tilde{\Theta} = \{\hat{\theta} \mid \hat{\theta}(y) = \xi(|\hat{\tau}|)y, \ \ \xi : [0,\infty) \to [0,1] \ \text{ nondecreasing and continuous a.e.}\} \ .$$

Consequently, we compare all estimators $\hat{\theta}$ by their regret risk

$$reg(\hat{\theta}, \theta; \hat{\Theta}) = R(\hat{\theta}, \theta) - \lambda^* \sigma^2 \ . \tag{12}$$

Let us start with deriving the minimax regret solution within the class of linear estimators $\hat{\Theta}_L$. Straightforward optimization (see Appendix) provides the following.

Theorem 1 . *The minimax regret linear estimator (MMRLE) is given by* $\hat{\theta}^L_{\lambda_L} = \lambda_L y$, *where*

$$\lambda_L = \lambda_L(c/\sigma) = 1 - \frac{\sigma}{\sqrt{\sigma^2 + c^2}} \ ,$$

leading to the maximal regret risk $\sup_{|\theta| \le c} reg(\hat{\theta}^L_{\lambda_L}, \theta; \hat{\Theta}) = \lambda_L^2 \sigma^2$.

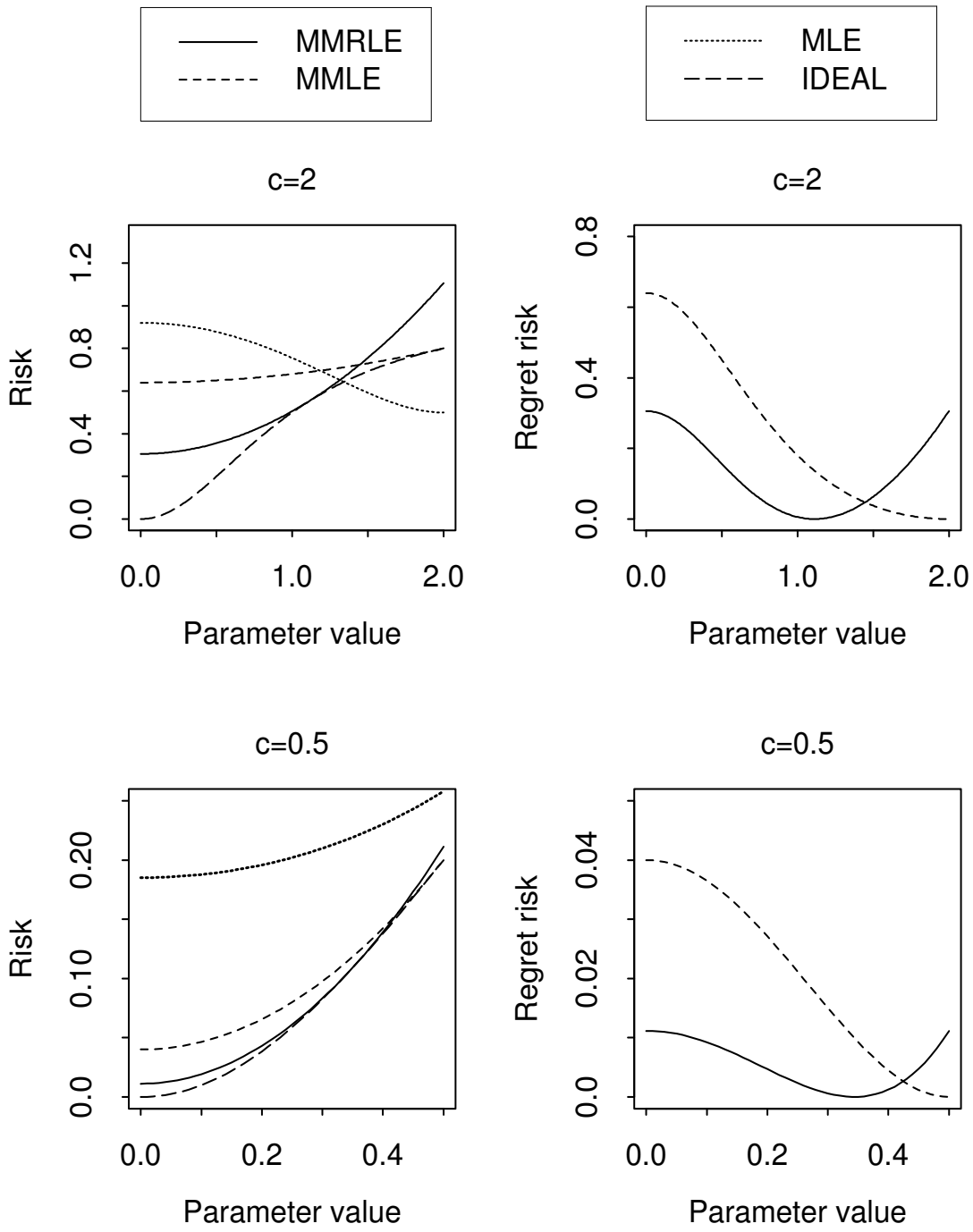

Figure 1: *Comparison of the MMRLE, $\hat{\theta}^L_{\lambda_L}$, and the MMLE, $\hat{\theta}^L_{\lambda_M}$. For $\sigma = 1$ and for some values of c, the left-hand and right-hand panels depict their risk and regret risk, respectively, as function of the parameter $|\theta|$. In the left-hand panel, the risk of both the ideal linear shrinker, $\hat{\theta}^L_{\lambda^*_L}$, and the MLE, $\hat{\theta}_{ML}$, is additionally included.*

Figure 1 displays, for some values of the upper bound c of the parameter space, the behaviour of the MMRLE, $\hat{\theta}^L_{\lambda_L}$, in comparison with the ideal (but impractical) linear shrinker, $\hat{\theta}^L_{\lambda^*_L}$, and the minimax linear estimator (MMLE),

$$\hat{\theta}^L_{\lambda_M} = \lambda_M y \ , \quad \lambda_M = \lambda_M(c/\sigma) = \lambda^*(c/\sigma) = \frac{c^2}{\sigma^2 + c^2} = 1 - \frac{\sigma^2}{\sigma^2 + c^2} \ . \tag{13}$$

For illustrational purposes, the maximum likelihood estimator (MLE),

$$\hat{\theta}_{ML} = \hat{\theta}_{ML}(y, c) = yI\{|y| \le c\} + \mathrm{sgn}(y)cI\{|y| > c\} \ , \tag{14}$$

is also included in the figure. We note, however, that this estimator does even not belong to the class $\tilde{\Theta}$, so that (11) provides no lower risk bound for the MLE, as it is visible in the case $c = 2$.

We observe furthermore that $\lambda_M = \lambda^*(c/\sigma)$, $\lambda_L < \lambda_M$ and, as shown in the Appendix, $\sup_{|\theta| \le c} reg(\hat{\theta}^L_{\lambda_M}, \theta; \hat{\Theta}) = R(\hat{\theta}^L_{\lambda_M}, 0) = \lambda_M^2 \sigma^2$, which is, of course, larger than the maximal regret risk of $\hat{\theta}^L_{\lambda_L}$. Clearly, the maximal regret risk of both estimators $\hat{\theta}^L_{\lambda_L}$ and $\hat{\theta}^L_{\lambda_M}$ increase strictly from 0 to σ^2 as c varies from 0 to $+\infty$. For the presented examples, the minimax regret linear estimator dominates the minimax linear estimator over a large part of the parameter interval. Only for parameters near the endpoints of the interval, the minimax linear estimator is superior. It is easy to verify (see Appendix), that this is the case if and only if

$$\tau^2 > \frac{2\sigma^{-1}(\sigma^2 + c^2)^{3/2} - 3c^2 - 2\sigma^2}{c^2} \ . \tag{15}$$

Before stating the minimax regret result for soft thresholding, we summarize some properties of the corresponding normalized regret risk

$$r(\lambda, \tau) := \sigma^{-2} reg(\hat{\theta}^S_\lambda, \theta; \hat{\Theta}) = \rho_S(\lambda, \tau) - \lambda^*(\tau) \ . \tag{16}$$

Obviously, $r(\lambda, \tau)$ is a continuous function of both variables λ and τ. Further properties are stated in the next lemma.

Lemma 1 .

(i) For fixed λ, $r(\lambda, \tau)$ is symmetric about $\tau = 0$, i.e. $r(\lambda, -\tau) = r(\lambda, \tau)$, and has a local maximum at the point $\tau = 0$.

(ii) There is a λ_0, with $\lambda_0 \approx 0.338$, such that, for all $\lambda \le \lambda_0$, $r(\lambda, \tau)$ is monotonically decreasing in $\tau \ge 0$.

(iii) For all fixed $\lambda > \lambda_0$, there is one and only one positive local maximizer, $\bar{\tau} = \bar{\tau}(\lambda)$, say, of $r(\lambda, \tau)$ with $\bar{\tau}^2(\lambda) > \max(3, \lambda^2)$. $\bar{\tau}(\lambda)$ and $r(\lambda, \bar{\tau}(\lambda))$ are continuous and monotonically increasing functions of $\lambda > \lambda_0$. Moreover, $r(\lambda, \bar{\tau}(\lambda))$ approaches 0.2347 and ∞ as λ tends to λ_0 and ∞, respectively.

(iv) $r(\lambda,0) = \rho_S(\lambda,0)$ *is monotonically decreasing in* λ *with* $r(0,0) = 1$ *and* $r(\lambda,0) \to 0$ *as* $\lambda \to \infty$.

(v) For fixed $\tau > 0$, *there is one and only one stationary point,* $\bar{\lambda} = \bar{\lambda}(\tau)$, *say, of* $r(\lambda,\tau)$ *which is a minimizer and which is a continuous and monotonically decreasing function of* τ *with* $\bar{\lambda}(\tau) \to 0$ *as* $\tau \to \infty$ *and* $\bar{\lambda}(\tau) \to \infty$ *as* $\tau \to 0$.

The last statement is shown in the Appendix, whereas the remaining results may be found in Droge (1998).

Without the prior knowledge (2), the minimax regret optimal threshold is approximately $\lambda_S(\infty) = 0.545$, see Droge (1998). With this value we obtain $\bar{\tau}_\infty := \bar{\tau}(\lambda_S(\infty)) \approx 2.73$ (cf. part (iii) of Lemma 1) and $r(\bar{\tau}_\infty, \lambda_S(\infty)) \approx 0.387$. Employing the above lemma, we establish in the Appendix the following result on the existence of a unique minimax regret soft threshold estimator under (2).

Theorem 2 .

(i) Within the class (4) of soft threshold estimators there is a minimax regret solution $\hat{\theta}^S_{\lambda_S}$ *with respect to the regret function (12), where, for given* c *and* σ, $\lambda_S = \lambda_S(c/\sigma)$ *is the unique minimizer of the maximal regret risk,* $\sup_{|\theta|\le c} reg(\hat{\theta}^S_\lambda, \theta; \hat{\Theta})$, *over the threshold* λ.

(ii) It holds $\lambda_S(c/\sigma) = \lambda_S(\infty)$ *if* $c/\sigma \ge \bar{\tau}_\infty$, *and* $\lambda_S(c/\sigma) \to \infty$ *as* $c/\sigma \to 0$. *Moreover, as* c/σ *varies from 0 to* $\bar{\tau}_\infty$, *the optimal threshold* $\lambda_S(c/\sigma)$ *is strictly decreasing, whereas the resulting maximal regret risk of* $\hat{\theta}^S_{\lambda_S}$ *is strictly increasing.*

Using the properties of Theorem 2 and Lemma 1, it is easy to calculate numerically the optimal thresholds and the corresponding maximal regret risks for all "signal to noise ratios" $c/\sigma < \bar{\tau}_\infty$. Figure 2 shows the results. For example, we obtain $\lambda_S(1) \approx 0.85, \lambda_S(0.3) \approx 3.77$ and $\lambda_S(0.1) \approx 8$, which has to be compared with $\lambda_S \approx 0.545$ for all $c/\sigma > 2.73$.

The right-hand panel of Figure 2 provides also a comparison of the (normalized) maximal regret risk of the optimal soft threshold estimator $\hat{\theta}^S_{\lambda_S}$ with those of the minimax linear estimator and the minimax regret linear estimator, showing that the latter dominates even the optimal nonlinear estimator $\hat{\theta}^S_{\lambda_S}$ in cases of rather small values of the signal to noise ratio c/σ. More precisely, one finds $\sup_{|\theta|\le c} reg(\hat{\theta}^L_{\lambda_L}, \theta; \hat{\Theta}) < \sup_{|\theta|\le c} reg(\hat{\theta}^S_{\lambda_S}, \theta; \hat{\Theta})$, that is that linear shrinkage outperforms nonlinear shrinkage by soft thresholding in terms of the maximal regret risk, if $c \le c^*\sigma$ with $c^* \approx 2.42$. This is not surprising since the lower risk bound (11) is in general not attainable by the ideal soft threshold 'estimator', $\hat{\theta}^S_{\bar{\lambda}}$, where $\bar{\lambda} = \bar{\lambda}(\tau)$ is defined as minimizer of $R(\hat{\theta}^S_\lambda, \theta)$ over λ based on full knowledge of θ. This fact is

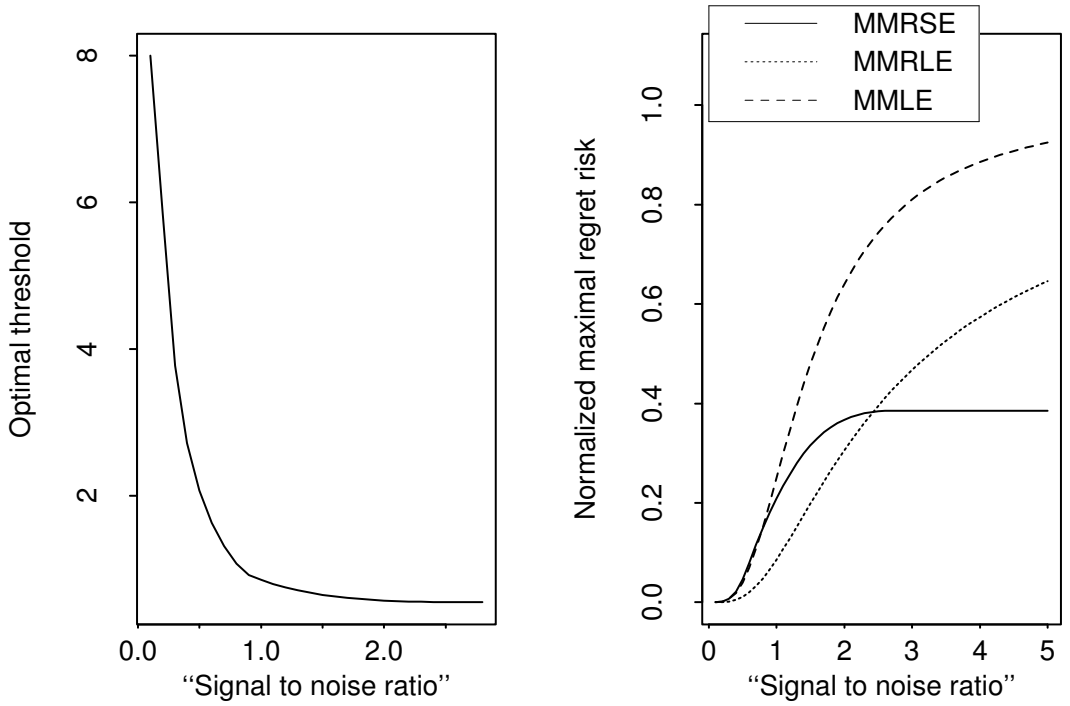

Figure 2: *The left-hand panel displays the minimax regret optimal soft threshold λ_S in dependence of the "signal to noise ratio" c/σ. The right-hand panel shows the normalized maximal regret risk, $\sigma^{-2}\sup_{|\theta|\leq c} reg(\hat{\theta},\theta;\hat{\Theta})$, of the optimal soft threshold estimator $\hat{\theta}^S_{\lambda_S}$ in comparison with those of the MMRLE $\hat{\theta}^L_{\lambda_L}$ and the MMLE $\hat{\theta}^M_{\lambda_M}$.*

visible in Figure 3. For the maximal risk difference between both ideal estimators we find $\sup_\tau r(\bar{\lambda}(\tau),\tau) \approx 0.24$, and this maximum is attained at $\tau \approx 1.4$. Since minimax regret soft thresholding is inferior to minimax regret linear shrinking unless $c > c^*\sigma$, Figure 3 gives a risk comparison of both estimators only for the case $c = 3\sigma$.

Due to the above discussion, $reg(\hat{\theta},\theta;\hat{\Theta})$ is different from $reg(\hat{\theta},\theta;\hat{\Theta}_S)$. Hence, $\hat{\theta}^S_{\lambda_S}$ is not exactly a minimax regret solution within the class $\hat{\Theta}_S$. But in Theorem 2 we have selected the threshold on the basis of the regret function (12), because this allows a direct comparison with the linear shrinkage approach. The consequences may be summarized as follows.

Corollary 1 . *For given c and σ, the unique minimax regret estimator (MMRE) within the class $\hat{\Theta}$ of linear and soft threshold estimators is given by*

$$\hat{\theta}^* = \hat{\theta}^*(y, c/\sigma) = \begin{cases} \hat{\theta}^S_{\lambda_S} & if \quad c > c^*\sigma \\ \hat{\theta}^L_{\lambda_L} & if \quad c \leq c^*\sigma \ . \end{cases}$$

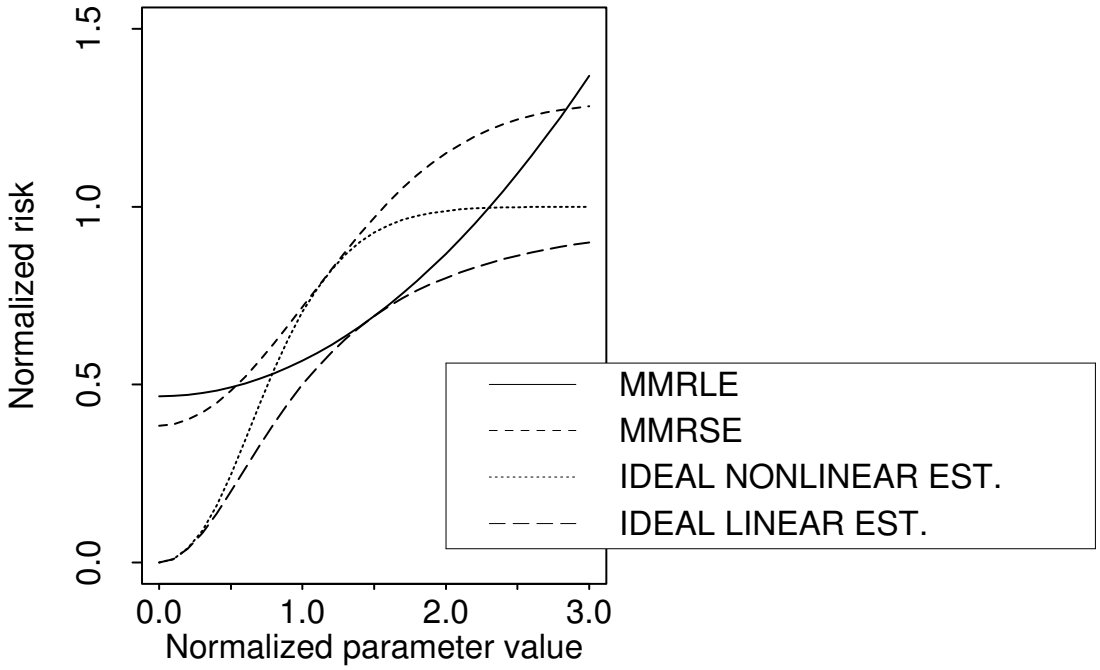

Figure 3: *For $c = 3\sigma$, normalized risk of some estimators as function of the normalized parameter τ. The risks of the ideal linear estimator, $\hat{\theta}^L_{\lambda*}$, and the ideal nonlinear estimator, $\hat{\theta}^S_{\lambda}$, are included.*

2.4 Discussion

The estimators introduced in this section have been derived under the prior constraint (2) that the parameter θ lies in a bounded interval. Nevertheless, their risk could be considered as function of θ varying on the whole real line. Then it turns out that all linear estimators, except for the naive one $\hat{\theta}_0 = y$, have an unbounded risk due to their biases. This property is shared by the MLE. On the other hand, the risk of the soft threshold (nonlinear) estimators is always bounded, which is a very desirable property when misspecifications of the prior information on the parameter cannot be excluded. This may happen, for example, in applications of the sequence space model of the next section, where the hyperrectangle constraint requires the specification of bounds for many coefficients. Consequently, curve estimates based on soft thresholding of coefficients according to Theorem 2 may even be preferable to those which use the optimal result of Corollary 1.

The realizations of the MLE lie completely in the parameter space $[-c, c]$, since it is just the projection of the naive estimator onto this interval. In contrast, all estimators of the class $\hat{\Theta}$ may have realizations on the whole real line. Therefore, a natural idea consists in projecting these estimators onto the parameter space.

The resulting estimators improve the original ones with respect to the risk over the interval $[-c, c]$. However, outside this interval their bias effects become more serious and even the projections of the soft threshold estimators are unbounded in risk. Having in mind our earlier comments, this is the reason for not investigating this type of seemingly improved versions of our estimators.

3 Application to nonparametric regression

3.1 Model and methods

In this section we illustrate the implications of our results for the problem of estimating the regression function in a nonparametric regression situation. This is possible since, as usual, a coordinatewise application of the scalar results leads immediately to results for multivariate (sequence space) problems. Then orthogonal transformations can be employed to turn statements about estimation over coefficient bodies in sequence space into statements about estimation over classes of smooth functions.

Suppose we observe responses $y_1, \ldots, y_n$, at nonrandom (distinct) design points $x_1, \ldots, x_n$, which follow the model

$$y_i = f(x_i) + \varepsilon_i \quad (i = 1, \ldots, n), \tag{17}$$

where the ε_i are independent and identically distributed as $N(0, \sigma^2)$. For simplicity, we assume that σ^2 is known. The goal is to use the data, y, to estimate the unknown real-valued regression function f. The performance of an estimator $\hat{f}(\cdot)$ of the regression function may be assessed by a normalized squared-error loss at the design points,

$$L(\hat{f}, f) = \frac{1}{n\sigma^2} \sum_{i=1}^{n} \{f(x_i) - \hat{f}(x_i)\}^2, \tag{18}$$

or by the corresponding normalized risk, $R(f, \hat{f}) = EL(\hat{f}, f)$.

Here we focus on orthogonal series regression, which is a widely used technique for nonparametric regression and closely related to other methods such as smoothing splines and kernel estimation. Let $\{\psi_j\}_{j=1}^n$ be a sequence of given basis functions which are orthonormal with respect to the counting measure on the design points $x_1, \ldots, x_n$, that is

$$\sum_{i=1}^{n} \psi_j(x_i)\psi_k(x_i) = \delta_{jk},$$

where δ_{jk} denotes the Kronecker symbol. Then, for $i = 1, \ldots, n$,

$$f(x_i) = \sum_{j=1}^{n} \theta_j \psi_j(x_i), \tag{19}$$

where $\theta_j = \sum_{k=1}^n f(x_k)\psi_j(x_k)$. In the traditional orthogonal series estimation approach, f is estimated by truncating (19) in an appropriate way and estimating the coefficients involved by their empirical versions or, equivalently, by least squares:

$$\hat{\theta}_j = \sum_{k=1}^{n} y_k \psi_j(x_k) \qquad (j = 1, \ldots, n).$$

In contrast, the approach of this paper is to apply the methods of Section 2 coordinatewise to the estimated coefficients. This leads to the consideration of the following classes of nonlinear shrinkers and diagonal linear shrinkers, repectively:

$$\mathcal{S} = \{\hat{f}_\Lambda^S \mid \hat{f}_\Lambda^S(x) = \sum_{i=1}^{n} \mathrm{sgn}(\hat{\theta}_i)(|\hat{\theta}_i| - \sigma\lambda_i)_+ \psi_i(x), \Lambda = (\lambda_1, \ldots, \lambda_n), \lambda_i \in [0, \infty]\} \ ,$$

$$\mathcal{L} = \{\hat{f}_\Lambda^L \mid \hat{f}_\Lambda^L(x) = \sum_{i=1}^{n} \lambda_i \hat{\theta}_i \psi_i(x), \ \ \Lambda = (\lambda_1, \ldots, \lambda_n), \ \ \lambda_i \in [0, 1]\} \ .$$

Let $\mathcal{S}^* = \mathcal{S} \cup \mathcal{L}$, $\tau_i = \theta_i/\sigma$ and $\hat{\tau}_i = \hat{\theta}_i/\sigma$. Within the class $\mathcal{S}^*$ we seek then for minimax regret optimal estimators with respect to the regret risk

$$\mathrm{Reg}(\hat{f}, f; \mathcal{S}^*) = R(\hat{f}, f) - \inf_{\tilde{f} \in \mathcal{S}^*} R(\tilde{f}, f) \ .$$

Droge (1998) has shown that

$$\inf_{\tilde{f} \in \mathcal{S}^*} R(\tilde{f}, f) = \tfrac{1}{n} \sum_1^n \lambda^*(\tau_i) = R(\hat{f}_{ideal}^L, f) \ ,$$

where $\hat{f}_{ideal}^L = \sum_{i=1}^n \lambda^*(\tau_i)\hat{\theta}_i\psi_i$ is the ideal linear shrunk estimator, compare (10). Consequently, we get for the soft threshold estimators

$$\mathrm{Reg}(\hat{f}_\Lambda^S, f; \mathcal{S}^*) = \frac{1}{n} \sum_i r(\lambda_i, \tau_i) \ ,$$

where $r(\lambda, \tau)$ is defined by (16). An analogous result holds for the diagonal linear shrinkers $\hat{f}_\Lambda^L$, replacing in (16) ρ_S by ρ_L.

In contrast to Droge (1998), however, we assume here to have some prior information on the smoothness of $f \in \mathcal{F}$, which leads typically to constraints on the Fourier coefficients θ_i. In particular we assume that in the corresponding sequence space model

$$\hat{\theta}_i = \theta_i + e_i \ , \quad e_i \sim N(0, \sigma^2) \quad (i = 1, \ldots, n)$$

the vector of coefficients $\theta = (\theta_1, \ldots, \theta_n)^T$ lies in an n-dimensional hyperrectangle, that is,

$$f \in \mathcal{F} \Leftrightarrow \theta \in \Theta := \{\theta \ : \ |\theta_i| \le c_i \ , \ i = 1, \ldots, n\} \ . \tag{20}$$

Of particular interest are hyperrectangles with

$$c_i = c i^{-\alpha} \ , i = 1, \ldots, n \ . \tag{21}$$

For more information about smoothness constraints of this type we refer to Donoho et al. (1990) and the references quoted therein. Under (20), we have

$$\sup_{f\in\mathcal{F}} \mathrm{Reg}(\hat{f}^S_\Lambda, f; \mathcal{S}^*) = \frac{1}{n}\sum_i \sup_{|\tau_i|\le c_i/\sigma} r(\lambda_i, \tau_i) \ ,$$

as well as an analogous relation for diagonal linear shrinkage. Consequently, the minimax regret problem can be reduced to the scalar problems solved in the previous section.

Corollary 2 . *Suppose (20). Then, for given σ, the unique minimax regret estimators of f within the classes $\mathcal{L}$, $\mathcal{S}$ and $\mathcal{S}^*$ are given by*

$$\begin{aligned}
\hat{f}_L(x) &= \hat{f}^L_{\Lambda_L}(x) \quad \textit{with } \Lambda_L = (\lambda_L(c_1/\sigma), \ldots, \lambda_L(c_n/\sigma)) \ , \\
\hat{f}^S(x) &= \hat{f}^S_{\Lambda_S}(x) \quad \textit{with } \Lambda_S = (\lambda_S(c_1/\sigma), \ldots, \lambda_S(c_n/\sigma)) \ , \textit{and} \\
\hat{f}^*(x) &= \sum_{i=1}^n \hat{\theta}^*_i \psi_i(x) \quad \textit{with } \hat{\theta}^*_i = \hat{\theta}^*(\hat{\theta}_i, c_i/\sigma) \ ,
\end{aligned}$$

respectively, where the scalar functions $\lambda_L(\cdot)$, $\lambda_S(\cdot)$ and $\hat{\theta}^*(\cdot,\cdot)$ are defined as in Theorems 1, 2 and Corollary 1.

In the following, $\hat{f}_L$, $\hat{f}_S$ and $\hat{f}^*$ will be called minimax regret linear estimator (MMRLE), minimax regret soft threshold estimator (MMRSE) and minimax regret estimator (MMRE), respectively, of the regression function f. The maximum likelihood estimator (MLE) of f under (20) is easily seen to be

$$\hat{f}_{ML} = \sum_{i=1}^n \hat{\theta}^{ML}_i \psi_i(x) \ \text{ with } \ \hat{\theta}^{ML}_i = \hat{\theta}_{ML}(\hat{\theta}_i, c_i) \ ,$$

where $\hat{\theta}_{ML}(\cdot,\cdot)$ is defined in (14). Similarly, a coordinatewise application of (13) provides the minimax linear estimator (MMLE) of f within (3):

$$\hat{f}_{MM} = \sum_{i=1}^n \hat{\theta}^{MM}_i \psi_i(x) \ \text{ with } \ \hat{\theta}^{MM}_i = \hat{\theta}^L_{\lambda_M(c_i/\sigma)}(\hat{\theta}_i) \ ,$$

compare (13) and (3) for the definition of the scalar estimator.

3.2 Some simulated data examples

The considered methods were applied to several simulated data examples. We simulated, according to (17), three sets of $n = 101$ observations at uniformly spaced design points $x_i = (i-1)/n$, for $i = 1, \ldots, n$. The true regression function f and standard deviation σ were chosen as follows:

(I) $f_1(x) = 4\sin(3\pi x) + 2\cos(5\pi x) + 2\cos(16\pi x)$, $\sigma = 1$

(II) f_1, $\sigma = 0.4$

(III) $f_2(x) = \sin(2\pi x^3)^3$, $\sigma = 0.1$.

Table 1 summarizes some features of these examples.

Table 1: *Some characteristics of Examples (I)-(III).*

Example	$\|f\|_n^2$	$\|y\|_n^2$	$sd(f)/\sigma$	$\inf_{\tilde{f}\in S^*} R(\tilde{f}, f)$	$\inf_{\tilde{f}\in S} R(\tilde{f}, f)$
(I)	11.953	13.103	3.363	0.192	0.222
(II)	11.953	12.406	8.408	0.375	0.448
(III)	0.199	0.230	4.362	0.163	0.181

The trigonometric functions,

$$\psi_j(x) = \begin{cases} n^{-1/2} & if \quad j = 1 \\ (2/n)^{1/2}\cos(2\pi kx) & if \quad j = 2k \geq 2 \text{ is even} \\ (2/n)^{1/2}\sin(2\pi kx) & if \quad j = 2k+1 \geq 3 \text{ is odd,} \end{cases}$$

served as basis, which fulfills the required orthonormality condition under the above design.

We remark that any available information on f would naturally have an impact on choosing an appropriate basis. For example, the trigonometric basis may be used when the regression curve is known to be smooth and periodic, since then most of the information about it is contained in the lower frequency terms; that is the coefficients θ_j for large j will be negligible compared with those for small j. Generally, a desirable property of a basis is for most of the signal to be concentrated in few coefficients. But the issue of choosing an appropriate basis is beyond the scope of this paper.

The parameter constraints in (20) were chosen according to (21). We considered five different values of α, including even the case $\alpha = 0$ where the same bound is used for all coefficients θ_i. Due to Parseval's identity we started with a bound, C say, for

$$\|f\|_n^2 := n^{-1}\sum_{i=1}^{n} f(x_i)^2 \ ,$$

and calculated the values c_i for the parameter contraints by $\sum_i c_i^2 = nC$.

Under a variety of parameter constraints, we calculated the normalized risk of all estimators introduced in Subsection 3.1. The results for Examples (I)-(III) are presented in Tables 2-4, where the minimum-risk-values have been underlined. Additionally, the minimal values for the estimators, $\tilde{f}$ say, derived by the minimax regret approach have been marked by a star. Both values differ only in Example (III), where the best performance is given by some MMLE.

The normalized risks in the tables should be compared with the lower risk bounds for linear shinking and soft thresholding, which are given in the last two columns of Table 1. For instance, even if we knew the true coefficients in Example (I), the normalized risk of any soft threshold estimator would be at least 0.222, while the MMRSE calculated under $C = 50$ and $\alpha = 1.5$ yields the risk 0.268.

Notice that the naive estimator, $\hat{f}_0(x) = \sum_{i=1}^{n} \hat{\theta}_i \psi_i(x)$, coincides with the MLE, MMLE and MMRLE in case of no prior constraints ($C = \infty$), and has risk one. Obviously, the MLE provides unacceptable risks in all considered cases. As discussed in Subsection 2.4, this is a consequence of the serious bias effects which may result when the empirical coefficients are projected onto some possibly misspecified intervals.

Many papers on wavelet based estimation such as Donoho et al. (1995) advocate the universal threshold parameter $\lambda_n = \sqrt{2 \log n}$ for all coefficients. However, as pointed out by Droge (1998), shrinkage with large thresholds may incur large biases, especially when some Fourier coefficients are large. This is the case in our Examples (I)-(III), where the normalized risk of the associated soft threshold estimator amounts to 1.057, 1.945 and 1.27, respectively, and exceeds even the risk for the naive estimator.

In some cases we observe that the prior constraint has no influence on the optimal soft thresholding method. This fact is in accordance with our Theorem 2, and happens, of course, if $C = \infty$ as well as for all values of C when $\alpha = 0$, with the only exception of $C = 5$ in Example (I). Even in one case of unequal bounds for the coefficients, i.e. for $\alpha = 0.5$, this occurs for all values of C in Example (II) and for some values of C in the other examples.

Generally, the optimal soft threshold estimator, MMRSE, is less sensitive to variation of the assumptions than are the linear estimators, MMLE, MMRLE and partly MMRE. This may be explained by their unbounded risk property, compare again the discussion in Subsection 2.4. In particular the specification of α would cause some problems in applications, so that it seems preferable to use always the minimax regret soft thresholding method. Note that a rule of thumb like taking the MMRSE with $C \approx \|y\|_n^2$ and $\alpha = 1$ would provide reasonable results in all three examples. Because of $E\|y\|_n^2 = \|f\|_n^2 + \sigma^2$, the choice of C could slightly be changed,

Table 2: *Normalized risk of some estimators for Example (I).*

		Prior constraint C for $\|f\|_n^2$					
Esimator	α	∞	50	20	12	10	5
MLE	0.0	1.000	4.403	6.620	7.691	8.035	9.138
	0.5	1.000	1.367	2.740	3.896	4.313	5.779
	1.0	1.000	1.738	2.611	3.620	4.017	5.525
	1.5	1.000	2.831	4.336	5.406	5.779	7.081
	2.0	1.000	4.609	6.276	7.158	7.447	8.406
MMLE	0.0	1.000	0.966	0.934	0.923	0.925	1.026
	0.5	1.000	0.906	0.797	0.708	0.673	0.548
	1.0	1.000	0.429	0.351	0.379	0.408	0.636
	1.5	1.000	0.689	1.237	1.581	1.704	2.185
	2.0	1.000	2.129	2.595	2.908	3.031	3.572
MMRLE	0.0	1.000	0.974	1.180	1.442	1.575	2.342
	0.5	1.000	0.698	0.645	0.660	0.679	0.860
	1.0	1.000	0.453	0.645	0.865	0.967	1.472
	1.5	1.000	1.324	1.973	2.368	2.516	3.127
	2.0	1.000	2.727	3.348	3.762	3.922	4.589
MMRSE	0.0	0.533	0.533	0.533	0.533	0.533	0.528
	0.5	0.533	0.533	0.529	0.514	0.504	0.452
	1.0	0.533	0.382	0.310	0.286	0.280	0.270
	1.5	0.533	0.268*	0.295	0.330	0.348	0.453
	2.0	0.533	0.507	0.867	1.065	1.089	2.502
MMRE	0.0	0.533	0.533	0.533	0.533	0.533	2.342
	0.5	0.533	0.533	0.513	0.460	0.437	0.343
	1.0	0.533	0.313	0.587	0.739	0.809	1.144
	1.5	0.533	1.220	1.838	2.148	2.251	2.753
	2.0	0.533	2.540	3.050	3.706	3.853	4.438

Table 3: *Normalized risk of some estimators for Example (II).*

		Prior constraint C for $\|f\|_n^2$					
Esimator	α	∞	50	20	12	10	5
MLE	0.0	1.000	22.558	36.477	43.223	45.416	52.713
	0.5	1.000	3.402	12.263	19.876	22.697	32.921
	1.0	1.000	8.140	14.680	21.395	23.988	33.748
	1.5	1.000	17.007	26.674	33.433	35.791	43.992
	2.0	1.000	28.568	39.038	44.569	46.384	52.414
MMLE	0.0	1.000	0.994	0.989	0.987	0.987	1.011
	0.5	1.000	0.984	0.961	0.937	0.926	0.869
	1.0	1.000	0.767	0.624	0.566	0.556	0.616
	1.5	1.000	0.669	1.468	2.478	2.966	5.387
	2.0	1.000	6.699	10.103	11.744	12.300	14.439
MMRLE	0.0	1.000	1.129	1.423	1.767	1.941	2.995
	0.5	1.000	0.906	0.925	0.977	1.009	1.242
	1.0	1.000	0.739	0.955	1.257	1.412	2.3278
	1.5	1.000	2.334	4.489	6.239	6.949	9.946
	2.0	1.000	10.318	13.532	15.315	15.977	18.727
MMRSE	0.0	0.670	0.670	0.670	0.670	0.670	0.670
	0.5	0.670	0.670	0.670	0.670	0.670	0.670
	1.0	0.670	0.656	0.618	0.587	0.576	0.546
	1.5	0.670	0.547	0.580	0.612	0.626	0.698
	2.0	0.670	0.763	0.932	1.074	1.136	1.458
MMRE	0.0	0.670	0.670	0.670	0.670	0.670	0.670
	0.5	0.670	0.670	0.670	0.670	0.670	0.670
	1.0	0.670	0.600	0.521	0.489	0.483*	0.488
	1.5	0.670	0.542	4.079	5.587	6.165	8.607
	2.0	0.670	9.618	12.922	14.434	14.927	17.419

Table 4: *Normalized risk of some estimators for Example (III).*

		Prior constraint C for $\|f\|_n^2$					
Esimator	α	∞	2	1	0.5	0.2	0.1
MLE	0.0	1.000	4.863	7.491	9.882	12.781	14.629
	0.5	1.000	0.996	1.549	3.208	6.557	9.040
	1.0	1.000	1.115	2.045	3.210	6.068	8.624
	1.5	1.000	2.903	4.302	5.960	9.045	11.274
	2.0	1.000	6.057	7.277	9.215	11.910	13.639
MMLE	0.0	1.000	0.9906	0.982	0.969	0.952	0.991
	0.5	1.000	0.974	0.950	0.905	0.795	0.667
	1.0	1.000	0.677	0.539	0.410	0.288	0.264
	1.5	1.000	0.220	0.219	0.284	0.545	0.933
	2.0	1.000	0.535	0.829	1.224	2.012	3.031
MMRLE	0.0	1.000	0.963	1.008	1.131	1.561	2.301
	0.5	1.000	0.819	0.766	0.712	0.681	0.754
	1.0	1.000	0.451	0.389	0.381	0.518	0.808
	1.5	1.000	0.355	0.521	0.822	1.517	2.347
	2.0	1.000	1.180	1.715	2.445	3.838	5.257
MMRSE	0.0	0.530	0.530	0.530	0.530	0.530	0.530
	0.5	0.530	0.530	0.530	0.530	0.526	0.500
	1.0	0.530	0.491	0.442	0.370	0.290	0.255
	1.5	0.530	0.248	0.236*	0.239	0.286	0.380
	2.0	0.530	0.314	0.416	0.568	0.839	1.168
MMRE	0.0	0.530	0.530	0.530	0.530	0.530	0.530
	0.5	0.530	0.530	0.530	0.530	0.513	0.441
	1.0	0.530	0.436	0.371	0.311	0.269	0.361
	1.5	0.530	0.251	0.337	0.584	0.908	1.321
	2.0	0.530	0.913	1.176	1.560	3.756	5.081

but the choice of α remains hard.

Figure 4 illustrates the behaviour of the best minimax regret estimators $\tilde{f}$ in the three examples. The approximation of the true regression function by $\tilde{f}$ is in general quite satisfactory. The left-hand panel depicts possible misspecifications of the prior constraints. The horizontal (dashed) line shows, however, that serious misspecifications of the bounds for the coefficients occur mainly above the value $\bar{\tau}_\infty * \sigma$ and have thus no influence on the optimal thresholds.

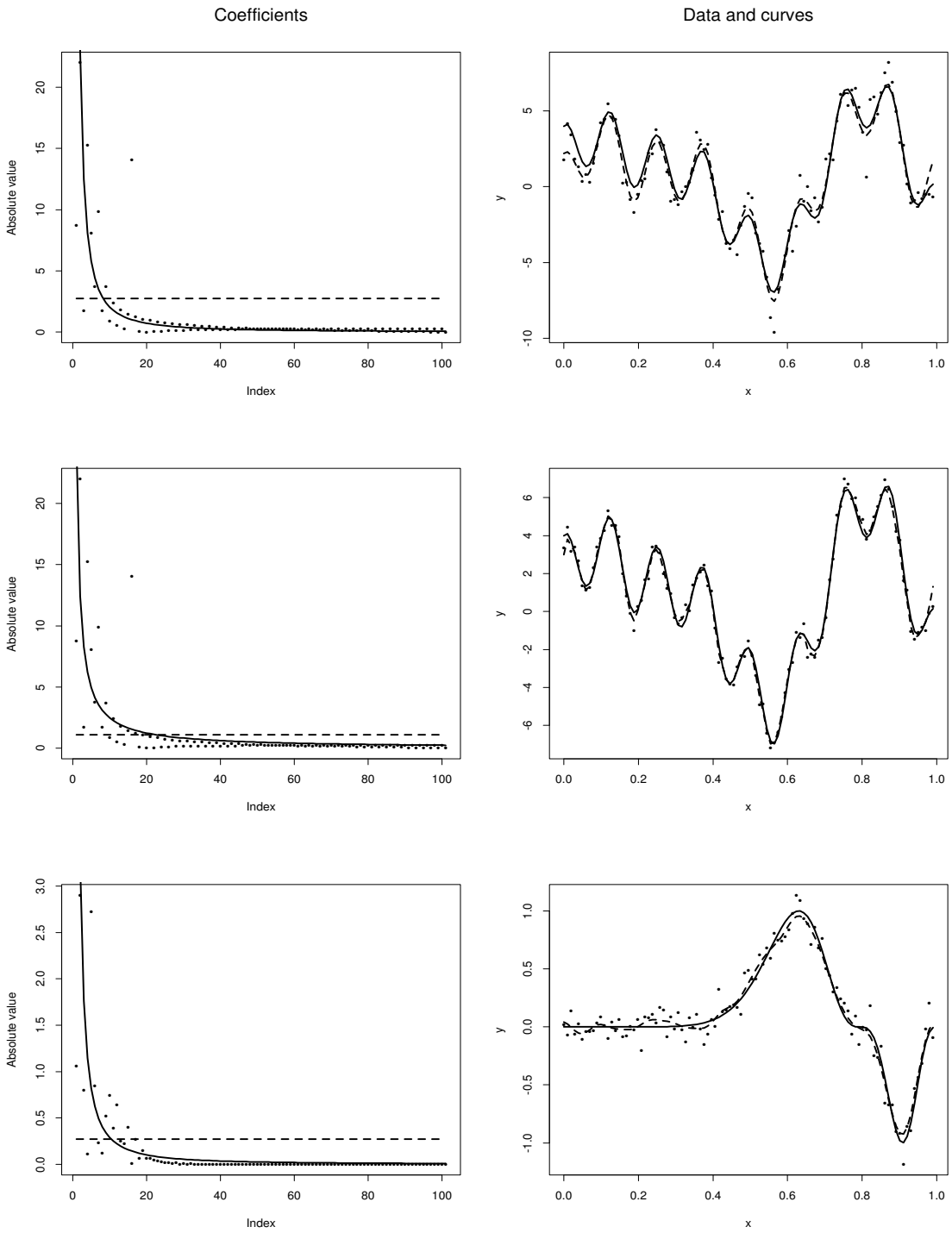

Figure 4: *From top to bottom: Illustration of examples (I)–(III). The right-hand panel shows the data (dots) together with the true regression function f (solid curve) and its best minimax regret estimate $\tilde{f}$ (dashed curve). The left-hand panel displays the true coefficients $|\theta_i|$ (dots), the bounds c_i (solid line) used for $\tilde{f}$, and the value $\bar{\tau}_\infty * \sigma$ (dashed line), above which the optimal thresholds do not depend on the c_i.*

Appendix

Proof of Theorem 1. On account of (5), (10) and (12), we obtain

$$r_L(\lambda,\tau) := \sigma^{-2} reg(\hat{\theta}^L_\lambda, \theta; \hat{\Theta}) = \lambda^2 + (1-\lambda)^2\tau^2 - \frac{\tau^2}{1+\tau^2} \ ,$$

which is, for fixed λ, obviously symmetric about $\tau = 0$, i.e., $r_L(\lambda,-\tau) = r_L(\lambda,\tau)$. Hence, it is sufficient to restrict the considerations to $\tau \geq 0$.

For $\lambda = 0$ or $\lambda = 1$, $r_L(\lambda,\tau)$ has a unique stationary point at $\tau = 0$, which is a maximum if $\lambda = 1$ and a minimum if $\lambda = 0$. For fixed $\lambda \in (0,1)$, $r_L(\lambda,\tau)$ has a unique (local) maximum at $\tau = 0$ and a unique positive minimum at $\tau = \tau_{min} = \sqrt{\frac{\lambda}{1-\lambda}}$ (with $r_L(\lambda,\tau_{min}) = 0$) . This leads to the maximal normalized regret risk

$$\sup_{\tau \leq c/\sigma} r_L(\lambda,\tau) = \max\{r_L(\lambda,0), r_L(\lambda,c/\sigma)\} \ , \tag{A1}$$

which has to be minimized over $\lambda \in [0,1]$.

Clearly, $r_L(\lambda,0) = \lambda^2$ is monotonically increasing in λ, whereas straightforward algebra provides that $r_L(\lambda,c/\sigma)$ is monotonically decreasing if and only if

$$\lambda < \frac{c^2}{\sigma^2 + c^2} = \lambda_M \ ,$$

compare (13). Furthermore, one verifies that

$$r_L(\lambda,0) < r_L(\lambda,c/\sigma) \quad \text{if and only if} \quad \lambda < 1 - \frac{1}{\sqrt{1+(c/\sigma)^2}} = \lambda_L \ .$$

Because of $\lambda_L \leq \lambda_M$, the unique minimum of (A1) is therefore attained at $\lambda = \lambda_L$, and the corresponding value of the maximal normalized regret risk is

$$r_L(\lambda_L,0) = r_L(\lambda_L,c/\sigma) = \lambda_L^2 \ , \tag{A2}$$

which completes the proof. □

Maximal regret risk of minimax linear estimator. If $\lambda = \lambda_M$, then we have $\tau_{min} = c/\sigma$, see proof of Theorem 1. Consequently, $r_L(\lambda_L,\tau)$ is monotonically decreasing in $|\tau| \leq c/\sigma$, leading to the desired result:

$$\sup_{|\theta|\leq c} reg(\hat{\theta}^L_{\lambda_M}, \theta; \hat{\Theta}) = \sigma^2 \sup_{|\tau|\leq c/\sigma} r_L(\lambda_M,\tau) = \sigma^2 r_L(\lambda_M,0) = R(\hat{\theta}^L_{\lambda_M},0) = \lambda_M^2\sigma^2 \ .$$

□

Proof of (15). The risk formula (5) gives $R(\hat{\theta}^L_{\lambda_M},\theta) < R(\hat{\theta}^L_{\lambda_L},\theta)$ iff

$$\begin{aligned} \lambda_M^2 + (1-\lambda_M)^2\tau^2 &< \lambda_L^2 + (1-\lambda_L)^2\tau^2 \quad \text{iff} \\ \lambda_M^2 - \lambda_L^2 &< \tau^2(\lambda_M - \lambda_L)(2 - \lambda_M - \lambda_L) \quad \text{iff} \\ \tau^2 &> \frac{\lambda_M + \lambda_L}{2 - \lambda_M - \lambda_L} \ . \end{aligned}$$

The result follows by inserting the expressions for λ_M and λ_L (cp. (13) and Theorem 1) on the right had side of the last line. □

Proof of Lemma 1, item (v).

(a) Some properties of the noncentral χ^2-distribution. We start with some useful facts on the noncentral χ^2-distribution. Let $F_k(\cdot;\alpha)$ and $f_k(\cdot;\alpha)$ denote the distribution function and the probability density function, respectively, of a noncentral χ^2-distributed random variable with k degrees of freedom and noncentrality parameter α.Then we have, for $x > 0$, (see e.g. (A1), (A2) and (A8) of Droge, 1993)

$$f_k(x;\alpha) = \frac{1}{2}\exp\left(-\frac{\alpha+x}{2}\right)\sum_{j=0}^{\infty}\frac{(\frac{\alpha}{2})^j(\frac{x}{2})^{\frac{k}{2}+j-1}}{j!\Gamma(\frac{k}{2}+j)} \quad . \tag{A3}$$

Differentiating this w.r.t. x provides

$$\frac{\partial}{\partial x}f_k(x;\alpha) = \frac{1}{2}[f_{k-2}(x;\alpha) - f_k(x;\alpha)] \quad . \tag{A4}$$

Note that relation (A4) holds for any integer k if we define the functions f_k in (A3) formally also for negative integers k (so that f_k is not necessarily a density), since the Γ-function fulfills $\Gamma(z+1) = z\Gamma(z)$ for all real numbers $z \neq 0, -1, -2, \ldots$.

By arguments similar to those leading to (A4) we get (see also (A10) and (A11) of Droge, 1993)

$$\begin{aligned}\frac{\partial}{\partial\alpha}F_k(x;\alpha) &= \frac{1}{2}[F_{k+2}(x;\alpha) - F_k(x;\alpha)] &\text{(A5)}\\ &= -f_{k+2}(x;\alpha) &\text{(A6)}\end{aligned}$$

and

$$\frac{\partial}{\partial\alpha}f_k(x;\alpha) = \frac{1}{2}[f_{k+2}(x;\alpha) - f_k(x;\alpha)] \quad . \tag{A7}$$

(b) Uniqueness of local minimizer of $r(\lambda,\tau)$. To find the stationary points of (16) for fixed $\tau > 0$, we differentiate the normalized regret risk w.r.t. λ. Using $\phi'(y) = -y\phi(y)$, this yields

$$\begin{aligned}\frac{\partial}{\partial\lambda}r(\lambda,\tau) &= \frac{\partial}{\partial\lambda}\rho_S(\lambda,\tau)\\ &= -2\lambda[\Phi(\lambda-\tau) - \Phi(-\lambda-\tau) - 1] - 2[\phi(\lambda-\tau) + \phi(\lambda+\tau)] &\text{(A8)}\\ &= 2\lambda[1 - F_1(\lambda^2;\tau^2) - 2f_1(\lambda^2;\tau^2)] \quad , &\text{(A9)}\end{aligned}$$

where the last equality follows by

$$\Phi(\lambda-\tau) - \Phi(-\lambda-\tau) = F_1(\lambda^2;\tau^2) \text{ and } \phi(\lambda-\tau) + \phi(\lambda+\tau) = 2\lambda f_1(\lambda^2;\tau^2) \quad .$$

In view of (A9), setting $x := \lambda^2$ and $\alpha := \tau^2$, we have to verify that, for fixed $\alpha > 0$, there is a unique (positive) root of

$$h(x,\alpha) = 1 - F_1(x;\alpha) - 2f_1(x;\alpha) \quad , \tag{A10}$$

which corresponds to a minimum of $r(\lambda,\tau)$. To accomplish this, we first observe that, on account of (A4) and (A3),

$$\frac{\partial}{\partial x}h(x,\alpha) = -f_1(x;\alpha) - 2\frac{\partial}{\partial x}f_1(x;\alpha) = -f_{-1}(x;\alpha) = s_\alpha(x)\sum_{j=0}^{\infty}\left(\frac{x}{2}\right)^j a_j(\alpha) \ , \tag{A11}$$

where

$$s_\alpha(x) = \frac{1}{2}\left(\frac{x}{2}\right)^{-3/2}\exp\left(-\frac{\alpha+x}{2}\right) \quad \text{and} \quad a_j(\alpha) = -\frac{\left(\frac{\alpha}{2}\right)^j}{j!\Gamma(j-\frac{1}{2})} \ .$$

Clearly, $s_\alpha(x) > 0$ for all $x > 0$ and $\alpha \geq 0$. Moreover, $a_0(\alpha) = -\frac{1}{\Gamma(-1/2)} = \frac{1}{2\Gamma(1/2)} > 0$ and $a_j(\alpha) < 0$ for all $j \geq 1$, so that the sequence $\{a_j(\alpha)\}_{j=0}^{\infty}$ has one and only one sign change. Thus, Lemma B2 of Droge (1993) implies the existence of a unique positive root, $x_0 = x_0(\alpha)$, say, of (A11), that is, of a unique stationary point of $h(x,\alpha)$, which must provide a maximum due to part (ii) of that lemma. Furthermore, $h(x,\alpha)$ tends to $-\infty$ and 0 as x approaches 0 and ∞, respectively. This and the strict unimodality of $h(x,\alpha)$ establish the existence of a unique positive root, $\bar{x} = \bar{x}(\alpha) < x_0$, of $h(x,\alpha)$. Consequently, $\bar{\lambda} = \bar{\lambda}(\tau) := \sqrt{\bar{x}(\tau^2)}$ is the unique stationary point of $r(\lambda,\tau)$, which turns out to be a minimum since $h(x,\alpha)$ is increasing at $\bar{x}$ and, hence,

$$\frac{\partial^2}{\partial\lambda^2}r(\lambda,\tau)|_{\lambda=\bar{\lambda}} = 4\bar{\lambda}^2\frac{\partial}{\partial x}h(x,\tau^2)|_{x=\bar{\lambda}^2} > 0 \ ,$$

compare (A9).

(c) Behaviour of $\bar{\lambda}(\tau)$. First recall that

$$\frac{\partial}{\partial x}h(x,\alpha)|_{x=\bar{x}(\alpha)} > 0 \ ,$$

since $\bar{x}(\alpha) < x_0(\alpha)$ and $h(\cdot,\alpha)$ is monotonically increasing on the interval $(0, x_0(\alpha))$. Moreover, from (A6) and (A7) we deduce, for all $x > 0$,

$$\frac{\partial}{\partial\alpha}h(x,\alpha) = f_1(x,\alpha) > 0 \ .$$

Consequently, the formula for the derivative of an implicit function yields

$$\bar{x}\,'(\alpha) = -\left[\frac{\partial h(x,\alpha)}{\partial\alpha} \Big/ \frac{\partial h(x,\alpha)}{\partial x}\right]_{x=\bar{x}(\alpha)} < 0 \ , \tag{A12}$$

so that $\bar{\lambda}(\tau) = \sqrt{\bar{x}(\tau^2)}$ is a monotonically decreasing function of τ.

This monotonicity result has the immediate consequence that $\bar{x}(\alpha)$ converges as $\alpha \to \infty$, since $\bar{x}(\alpha) \geq 0$. It remains therefore to prove that the corresponding limit is 0. Given any $\alpha_0 > 0$, we have, for all $\alpha > \alpha_0$, $\bar{x}(\alpha) \leq \bar{x}(\alpha_0) =: \bar{x}_0$. Thus, for all $\alpha > \alpha_0$,

$$F_1(\bar{x}(\alpha),\alpha) \leq F_1(\bar{x}_0,\alpha) \ ,$$

which converges to 0 as $\alpha \to \infty$ since then, for any fixed c, $F_k(c,\alpha) \to 0$. Because of $h(\bar{x}(\alpha),\alpha) = 0$, this implies

$$f_1(\bar{x}(\alpha),\alpha) \to \frac{1}{2} \quad \text{as} \quad \alpha \to \infty \ ,$$

compare (A10). Observing $\lim_{\alpha\to\infty} f_1(x,\alpha) = 0$ for any fixed $x > 0$, we conclude $\lim_{\alpha\to\infty} \bar{x}(\alpha) = 0$ and hence $\bar{\lambda}(\tau) \to 0$ as $\tau \to \infty$.

Finally, $\bar{\lambda}(\tau) \to \infty$ as $\tau \to 0$ follows from part (iv) of Lemma 1 by a continuity argument.

(d) More precise asymptotics for $\bar{\lambda}(\tau)$. Analogously to (A12), a formula for $\bar{\lambda}\,'(\tau)$ can be derived by setting (A8) equal to 0. This leads to

$$\bar{\lambda}\,'(\tau) = -\frac{\bar{\lambda}(\tau)\tau}{1 - \bar{\lambda}(\tau)\tau \tanh(\bar{\lambda}(\tau)\tau)} \ , \tag{A13}$$

where, because of $\bar{\lambda}\,'(\tau) < 0$, the denominator must be positive, which is equivalent to $0 < \lambda(\tau)\tau < \kappa$ with $\kappa \approx 1.2$. Continuity arguments together with the results proved above under (c) imply now that $\bar{\lambda}\,'(\tau)$ appoaches 0 and $-\infty$ as τ tends to ∞ and 0, respectively. In view of (A13), $\lambda(\tau)\tau$ must therefore converge to 0 and κ as τ tends to ∞ and 0, respectively. Taking $\delta(\tau) = \bar{\lambda}(\tau)\tau/\kappa$, $\bar{\lambda}(\tau)$ can be represented as

$$\bar{\lambda}(\tau) = \frac{\kappa}{\tau}\delta(\tau) \ ,$$

where $\delta : (0,\infty) \to (0,1)$ is a continuous function converging to 0 and 1, respectively, as $\tau \to \infty$ and $\tau \to 0$. □

Proof of Theorem 2. *(i).* Using (16), the maximal regret risk of a soft threshold estimator $\hat{\theta}^S_\lambda$ may be rewritten as

$$\sup_{|\theta|\le c} reg(\hat{\theta}^S_\lambda, \theta; \hat{\Theta}) = \sigma^2 \sup_{|\tau|\le c/\sigma} r(\lambda,\tau) \ .$$

Without the prior knowledge (2), the minimax regret optimal threshold, $\lambda_S(\infty)$, say, is given as unique solution of $r(\lambda,0) = r(\lambda,\bar{\tau}(\lambda))$, where $\bar{\tau}(\lambda)$ is defined in part (iii) of Lemma 1, compare Theorem 3 and its proof in Droge (1998). The associated estimator $\hat{\theta}^S_{\lambda_S(\infty)}$ provides thus the (unconstraint) normalized minimax regret risk

$$\inf_\lambda \sup_\tau r(\lambda,\tau) = r(\lambda_S(\infty),0) = r(\lambda_S(\infty),\bar{\tau}_\infty) \ , \tag{A14}$$

where $\bar{\tau}_\infty = \bar{\tau}(\lambda_S(\infty))$.

Under the constraint (2), the minimax regret risk (A14) can only be reduced by choosing a threshold larger than $\lambda_S(\infty)$, since $\tau = 0$ belongs always to the parameter space and $r(\lambda,0)$ is monotonically decreasing in λ, see Lemma 1, (iv). Consequently, we have for all c

$$\lambda_S(c/\sigma) \ge \lambda_S(\infty) \ . \tag{A15}$$

On the other hand, part (v) of the same lemma provides that, for fixed $\tau > 0$, $r(\lambda, \tau)$ is monotonically increasing in λ if and only if $\lambda > \bar{\lambda}(\tau)$. Numerical approximations, employing properties established in Lemma 1, lead now to $\lambda_S(\infty) \approx 0.545$, $\bar{\tau}_\infty \approx 2.73$ and $\bar{\lambda}(\bar{\tau}_\infty) \approx 0.0316$. Because of $\lambda_S(\infty) > \bar{\lambda}(\bar{\tau}_\infty)$, an increasing value of λ would therefore increase the (normalized) minimax regret risk (A14) as long as $\bar{\tau}_\infty \sigma$ belongs to the parameter space. This implies

$$\lambda_S(c/\sigma) = \lambda_S(\infty) \approx 0.545 \quad \text{if} \quad c/\sigma \geq \bar{\tau}_\infty \approx 2.73 \ . \tag{A16}$$

Consider now the case $c/\sigma < \bar{\tau}_\infty$. In view of (A15), we can confine our considerations to the case $\lambda \geq \lambda_S(\infty)$. Statements (i)–(iii) of Lemma 1, in particular the monotonicity of $\bar{\tau}(\lambda)$, lead then to

$$\sup_{|\tau| \leq c/\sigma} r(\lambda, \tau) = \max\{r(\lambda, 0), r(\lambda, c/\sigma)\} \ \text{ for all } \ \lambda \geq \lambda_S(\infty) \ .$$

Again, in view of items (iv) and (v) of Lemma 1, $r(\lambda, 0)$ is monotonically decreasing in λ, whereas $r(\lambda, c/\sigma)$ is first decreasing in $\lambda \leq \bar{\lambda}(c/\sigma)$ and then increasing in $\lambda \geq \bar{\lambda}(c/\sigma)$. Because of the strict monotonicity in Lemma 1 (compare the above proof), the existence of a unique minimax regret rule follows. More precisely, we have $\lambda_S(c/\sigma) = \bar{\lambda}(c/\sigma)$ if

$$r(\bar{\lambda}(c/\sigma), c/\sigma) > r(\bar{\lambda}(c/\sigma), 0) \ . \tag{A17}$$

If (A17) is not fulfilled, then there is a unique threshold $\lambda_S(c/\sigma) \geq \bar{\lambda}(c/\sigma)$ with

$$r(\lambda_S(c/\sigma), c/\sigma) = r(\lambda_S(c/\sigma), 0) \ ,$$

which provides the optimal solution. The maximal regret risk of the corresponding estimator $\hat{\theta}^S_{\lambda_S}$ (i.e., the minimax regret risk within $\hat{\Theta}_S$) is $\sigma^2 r(\lambda_S(c/\sigma), c/\sigma)$, which coincides with $\sigma^2 r(\lambda_S(c/\sigma), 0)$ unless (A17) holds.

(ii). Because of (A16) it suffices to consider the case $c/\sigma \leq \bar{\tau}_\infty$. For such values of c, we know from the above proof of item (i) that (A15),

$$\lambda_S(c/\sigma) \geq \bar{\lambda}(c/\sigma) \ , \tag{A18}$$

and

$$\inf_{\lambda} \sup_{|\tau| \leq c/\sigma} r(\lambda, \tau) = r(\lambda_S(c/\sigma), c/\sigma) \geq r(\lambda_S(c/\sigma), 0) \ , \tag{A19}$$

where equality holds if $\lambda_S(c/\sigma) > \bar{\lambda}(c/\sigma)$. Moreover, item (iii) of Lemma 1 and its proof in Droge (1998) yield $\bar{\tau}(\lambda_S(c/\sigma)) \geq \bar{\tau}_\infty$ and, hence,

$$r(\lambda_S(c/\sigma), c/\sigma) > r(\lambda_S(c/\sigma), \tau) \quad \text{for all} \quad \tau \in (0, c/\sigma) \ , \tag{A20}$$

since there is only one stationary point of $r(\lambda, \tau)$ on the interval $(0, \bar{\tau}(\lambda))$ which must provide a minimum.

From inequality (A18) we obtain the first asymptotic result, $\lambda_S(c/\sigma) \to \infty$ as $c/\sigma \to 0$, since $\bar{\lambda}(\tau) \to \infty$ as $\tau \to 0$ due to item (v) of Lemma 1.

Let $c_1 < c_2 \leq \bar{\tau}_\infty \sigma$ and, for $i = 1, 2$, $\lambda_i = \lambda_S(c_i/\sigma)$ as well as $\bar{\lambda}_i = \bar{\lambda}(c_i/\sigma)$. Then it remains to show that (a) $\lambda_1 > \lambda_2$ and (b) $r(\lambda_1, c_1/\sigma) < r(\lambda_2, c_2/\sigma)$.

Ad (a). If $\lambda_2 = \bar{\lambda}_2$, then (A18) and the strict monotonicity of $\bar{\lambda}$ in Lemma 1, (v) give immediately the result: $\lambda_1 \geq \bar{\lambda}_1 > \bar{\lambda}_2 = \lambda_2$. On the other hand, $\lambda_2 \neq \bar{\lambda}_2$ implies $\lambda_2 > \bar{\lambda}_2$ and thus $r(\lambda_2, c_2/\sigma) = r(\lambda_2, 0)$, compare (A19). From this the result follows, since $\lambda_1 \geq \bar{\lambda}_1$ by (A18) and the assumption $\lambda_1 \in [\bar{\lambda}_1, \lambda_2]$ would provide a contradiction to (A19):

$$r(\lambda_1, c_1/\sigma) \leq r(\lambda_2, c_1/\sigma) < r(\lambda_2, c_2/\sigma) = r(\lambda_2, 0) \leq r(\lambda_1, 0) \ ,$$

recall that $r(\lambda, c_1/\sigma)$ increases in $\lambda \geq \bar{\lambda}_1$ as well as (A20) and Lemma 1, (iv).

Ad (b). Suppose first that $\lambda_1 = \bar{\lambda}_1$. Then the the minimizing property of $\bar{\lambda}_1$ and (A20) provide $r(\lambda_1, c_1/\sigma) \leq r(\lambda_2, c_1/\sigma) < r(\lambda_2, c_2/\sigma)$. Finally, for $\lambda_1 > \bar{\lambda}_1$ we have equality in (A19) so that we conclude from (a), the strict monotonicity in part (iv) of Lemma 1 and again (A19)

$$r(\lambda_1, c_1/\sigma) = r(\lambda_1, 0) < r(\lambda_2, 0) \leq r(\lambda_2, c_2/\sigma) \ ,$$

completing the proof. □

References

Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis.* 2nd ed., Springer-Verlag, New York.

Bickel, P.J. (1981). Minimax estimation of the mean of a normal distribution when the parameter space is restricted. *Ann. Statist.* **9**, 1301-1309.

Casella,L.D. and Strawderman, W. (1981). Estimating a bounded normal mean. *Ann. Statist.* **9**, 868-876.

Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: Asymptopia ? (With discussion) *J. Roy. Statist. Soc.* **B 57**, 301-369.

DONOHO, D.L., LIU, R.C. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *Ann. Statist.* **18**, 1416-1437.

DROGE, B. (1993). On finite sample properties of adaptive least squares regression estimates. *Statistics* **24**, 181-203.

DROGE, B. (1998). Minimax regret analysis of orthogonal series regression estimation: Selection versus shrinkage. *Biometrika* **85**, 631-643.

DROGE, B. and GEORG, T. (1995). On selecting the smoothing parameter of least squares regression estimates using the minimax regret approach. *Statistics & Decisions* **13**, 1-20.

GATSONIS, C., MACGIBBON, B. and STRAWDERMAN, W. (1987). On the estimation of a restricted normal mean. *Statist. Probab. Lett.* **6**, 21-30.

MALLOWS, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661-75.