

Improving Regression through Model Mixing

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Gilbert Leung

Dissertation Director: Andrew R. Barron
Department of Statistics

May, 2004

© 2004 by Gilbert Leung
All rights reserved.

Abstract

In problems of regression, techniques of model selection and model mixing are often used to produce a combined procedure (without advance knowledge of which model is best) for which one would hope that the resulting performance (in square-error loss) would be comparable or perhaps even superior to what is achieved by the best of the individual models.

For Gaussian regression, we examine two major cases and develop methods for model mixing (convexly combining) and analysing the risks of the mixture estimators. In either case, our component models arise from using arbitrary subsets of available regressors, which also includes models of the common leading-term type. The first case involves mixing ordinary least-squares estimators on chosen subsets (the coefficients outside the chosen subset are estimated by zero). In the second case, we mix estimators that apply two positive-part James-Stein shrinkage estimators: one on the chosen subset and the other on the subset's complement. (This is done because of the shrinkage estimator's nice risk properties, uniform in the unknown parameter.) In both cases, we consider model weights related to the risk estimate for each individual estimator, as this can sometimes arise when the weights are determined from Bayes posterior probabilities. We analyse an unbiased estimate of the risk of the averaged estimator and relate it to estimates of the risk achieved by estimators for the individual models. Our analysis provides simple and accurate bounds on the risks, in the form of sharp and exact oracle inequalities.

Furthermore, we provide simulation results for the leading-term models. They show that performance of these mixture estimators is always better than that suggested by the analytical risk bounds. Also, our mixture always performs better than the model selection estimators using Akaike's information criterion.

Contents

1	Overview	1
1.1	Problem Statement	1
1.2	Background	2
1.2.1	Least-Squares (“All-or-Nothing”) Estimators	2
1.2.2	Shrinkage Estimators	4
1.3	Outline of Thesis	4
2	Mixture Estimators for the Multivariate Normal Mean	6
2.1	General Mixture	6
2.1.1	Introduction and Examples	6
2.1.2	Risk Estimate	7
2.2	Mixing Least-Squares Estimators	9
2.2.1	Risk Estimate for Mixture	9
2.2.2	An Upper-bound for the Combined Risk Estimate	9
2.2.3	Risk Bound	11
2.2.4	Simulations	12
2.3	Mixing Shrinkage Estimators	16
2.3.1	Risk Estimates for Positive-Part James-Stein	16
2.3.2	Two-Set Shrinkage Risk Estimates	16
2.3.3	Risk Estimate for Mixture	17
2.3.4	Risk Bound	20
2.3.5	Risk Targets: A Comparison with Ideal Linear Estimation	20
2.3.6	Simulations	22
3	Controlling Model Complexity	25
3.1	Complexity of Subset Models	25
3.2	Complexity of Two-Set Shrinkage Models	26
4	Discussion	27
4.1	Bayesian Considerations	27
4.1.1	Form of Pseudo-Bayes Estimator	27
4.1.2	Least-Squares Estimators for Subset Models	27
4.1.3	Two-Set Shrinkage Estimators	28
4.2	Concluding Remarks	28

A	Appendix	29
A.1	From Function Estimation to Canonical Regression	29
A.2	Incomplete Gamma function, g_L , and Gamma distribution	30
B	Bibliography	31

List of Theorems and Definitions

Theorem 2.8 (Stein)	7
Theorem 2.9 (Unbiased Risk Estimate for Mixture)	7
Lemma 2.10 (Orthogonality)	8
Definition 2.11	8
Proposition 2.15	9
Definition 2.17	10
Definition 2.18	10
Lemma 2.20	10
Corollary 2.25	11
Theorem 2.26 (Least-Squares Mixture Risk Bound)	11
Lemma 2.30	16
Definition 2.31 (Continuous risk estimate for positive James-Stein)	16
Definition 2.37	17
Lemma 2.38	17
Definition 2.42	18
Lemma 2.43	18
Corollary 2.44	18
Proposition 2.45	19
Lemma 2.47	19
Corollary 2.48	19
Theorem 2.49 (Two-Set Shrinkage Mixture Risk Bound)	20
Definition 2.53	21
Proposition 2.54	21
Lemma 2.55	21
Proposition 2.56	22
Proposition 3.3	25
Lemma 3.4	26
Theorem 3.6	26
Lemma 3.8	26
Theorem 3.9	26

List of Examples

Example 2.2 (Block Models)	6
Example 2.3 (General Subset Models)	6
Example 2.4 (Two-Block Models)	7
Example 2.5 (Two-Set Models)	7
Example 2.27 (Leading-Term)	11
Example 2.28 (Block Models)	12
Example 2.29 (All-Subset)	12
Example 2.50 (Two-Block Shrinkage)	20
Example 3.1 (Complexity-regularized Subsets)	25

List of Figures

2.1	Risks and Target: Constant One-Block. $\theta_i^2 \propto \mathbb{1}_{\{i \leq 10\}}$	13
2.2	Risks and Target: Gradual Decay. $\theta_i^2 \propto 1/i$	14
2.3	Risks and Target: Odd or Even Function. $\theta_i^2 \propto \mathbb{1}_{\{i \text{ odd}\}}$	14
2.4	Risks and Target: Ramp-Up with Cut-off. $\theta_i^2 \propto (i-1)\mathbb{1}_{\{i \leq 15\}}$	15
2.5	Risk and Targets: Constant One-Block. $\theta_i^2 \propto \mathbb{1}_{\{i \leq 10\}}$	23
2.6	Risk and Targets: Gradual Decay. $\theta_i^2 \propto 1/i$	23
2.7	Risk and Targets: Odd or Even Function. $\theta_i^2 \propto \mathbb{1}_{\{i \text{ odd}\}}$	24
2.8	Risk and Targets: Ramp-Up with Cut-off. $\theta_i^2 \propto (i-1)\mathbb{1}_{\{i \leq 15\}}$	24

Chapter 1

Overview

Often in regression, a common practice is to select a subset of available regressors, and to use the least-squares estimate on these selected variables to fit the response. This is useful especially when a parsimonious model for explanation of the response is desired. However, it is well-known that model selection procedures can be unstable, as small changes in the data often lead to a significant change in model choice. Moreover, the inference done with least-squares on the chosen model does not account for model uncertainty from the selection procedure, and therefore can be overly optimistic.

In this thesis, we demonstrate a scheme for convexly combining regression procedures for which the risk, under squared-error loss, is not much more than an idealized target defined by the risk achieved by the best of all the models considered. (This is what Yang (2004) calls *combining for adaptation* and the risk target is termed the *model selection target* by Tsybakov (2003) since a best model is being selected from the class.) One motivation for such mixtures comes from consideration of Bayes procedures (these and their limits are the admissible procedures in any statistical decision problem). With squared-error loss, the Bayes procedures are posterior weighted averages of the estimators for each model (see the references in Hoeting et al. (1999)), not model selections. A secondary motivation is that mixing effectively expands the model class to the class comprising all convex combinations of the models in the considered class.

We achieve this goal here by certain choices of the weights that adapt to the data. Our theoretical results are the sharp risk bounds (also called oracle inequalities) mentioned above. Moreover, in simulations, the resulting mixture estimator often performs better than a selection based estimator for a good part of the parameter space.

We will present our results in mixing two types of subset regression estimators: least-squares and positive-part James-Stein shrinkage estimators (or simply shrinkage estimators henceforth). The former refers to choosing a subset of regressors and simply using the least-squares estimate for the regression coefficients (thus, setting the coefficients outside the subset to zero). These are also nicknamed the “all-or-nothing” estimators. The latter refers to using two separate applications of the positive-part James-Stein estimators on the chosen subset and its complement (with each being the least-squares estimate on each part shrunk by a data-determined fraction between 0 and 1). This is desirable because it is well-known that such a shrinkage estimator has uniformly lower risk than least-squares. In each case, the mixture estimator is a convex combination of the described component estimators

(indexed by the different chosen subsets).

1.1 Problem Statement

Consider a Gaussian regression problem in which we have n observations and a design of rank $d \leq n$. For mathematical convenience, we assume that the variances of our observation errors are known so that we can always rescale our variables to arrive at a regression canonical form. That is, we observe $X \in \mathbb{R}^d$, normally distributed with mean θ ,

$$X \sim N_d(\theta, \sigma_n^2 I),$$

where each θ_i is the unknown coefficient for regressor i , which can be a combination of multiple explanatory variables (and transforms thereof). We will put

$$\sigma_n^2 = 1/n$$

here as it is the case for a typical regression problem in which better accuracy for observing and estimating the responses is obtained as the sample size n increases. However, n is fixed in this thesis: even though our analysis is non-asymptotic, n is retained in the scaling here for a rough comparison of our risk bounds with those in the literature which use n (often asymptotic).

Thus we have a multivariate normal location problem under square-error loss. In this setting, X is both the least-squares estimator and the maximum-likelihood estimator. One may also think of the θ_i as the coefficients in the representation of a response function in an orthonormalized basis, with dimension d that may be as large as n . Please see Appendix 1 for more discussion of the relationship of function estimation with the canonical regression problem.

Let m be a subset of $\{1, 2, \dots, d\}$, representing a particular subset of available regressors and $\hat{\theta}^m$ is an estimator for the model supported by this subset. For example, we can simply use the least-squares estimator for the regressors included in m , and estimate the coefficients of the other regressors by 0. This way,

$$\hat{\theta}_i^m = X_i \mathbb{1}_{\{i \in m\}}, \quad i = 1, \dots, d.$$

We will also consider positive-part James-Stein shrinkage estimators in lieu of least-squares

$$\hat{\theta}_i^m = (c_m \mathbb{1}_{\{i \in m\}} + c_m^c \mathbb{1}_{\{i \notin m\}}) X_i, \quad i = 1, \dots, d$$

where $c_m \in [0, 1]$ depends only on $\{X_i : i \in m\}$, and *mutatis mutandis* for c_m^c where m^c denotes $\{1, \dots, d\} \setminus m$. We form a convex combination of a finite class \mathcal{M} of these estimators

$$\hat{\theta} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{\theta}^m, \quad (1.1)$$

where the weights $\{\hat{\rho}_m : m \in \mathcal{M}\}$ (chosen later) sum to one and depend on the data. In general, we use a prior model probability of π_m for model m (with $\sum_{m \in \mathcal{M}} \pi_m = 1$), which can be simplified to the uniform distribution on \mathcal{M} if no prior knowledge is available. For the mixture of least-squares estimators, we will show the following risk bound (Theorem 3.6)

$$\mathbb{E} |\hat{\theta} - \theta|^2 \leq \inf_{m \in \mathcal{M}} \left[\mathbb{E} |\hat{\theta}^m - \theta|^2 + 4\sigma_n^2 \log \frac{1}{\pi_m} \right], \quad (1.2)$$

where $|\cdot|$ is the Euclidean norm; and for the mixture of shrinkage estimators, we will show (Theorem 3.9)

$$\mathbb{E} |\hat{\theta} - \theta|^2 \leq \inf_{m \in \mathcal{M}} \left\{ \mathbb{E} |\hat{\theta}^m - \theta|^2 + \sigma_n^2 \left[1 + 8 \log \frac{1}{\pi_m} \right] \right\}. \quad (1.3)$$

Recall that $\sigma_n^2 = 1/n$ for a canonical regression problem with n observations. When uniform priors $\pi_m = 1/M$ are used, where $M = \#\mathcal{M}$ is the cardinality of \mathcal{M} , then the above results become

$$\mathbb{E} |\hat{\theta} - \theta|^2 \leq \inf_{m \in \mathcal{M}} \mathbb{E} |\hat{\theta}^m - \theta|^2 + \frac{4}{n} \log M \quad (1.2')$$

for least-squares and

$$\mathbb{E} |\hat{\theta} - \theta|^2 \leq \inf_{m \in \mathcal{M}} \mathbb{E} |\hat{\theta}^m - \theta|^2 + \frac{1 + 8 \log M}{n} \quad (1.3')$$

for shrinkage estimators. Tighter bounds are obtained for the uniform model prior in Theorem 2.26 (least-squares) and Theorem 2.49 (shrinkage). In both (1.2') and (1.3'), we may define the risk target as the minimum of the risks of the estimators mixed, i.e.

$$r_* = r_*(\mathcal{M}) = \inf_{m \in \mathcal{M}} \mathbb{E} |\hat{\theta}^m - \theta|^2. \quad (1.4)$$

In both cases, the excess beyond the targets in the upper-bounds are of order $1/n \log M$. When the model classes \mathcal{M} are the same for both mixture estimators, the risk target is lower in the shrinkage case than in the least-squares case, as the individual shrinkage estimators have lower risks than those of least-squares, and the shrinkage risk target is much lower when $|\theta|$ is small. This shrinkage risk advantage offsets the larger coefficient of 8, as opposed to 4 in least-squares for the log-cardinality term in the risk bounds.

A theme common to both cases is to use weights $\hat{\rho}_m$ that are decreasing functions of risk estimates of the individual models, thus providing a focus on models assessed to have smaller risks. Specifically, we use

$$\hat{\rho}_m = \hat{\rho}_m(\beta) = \frac{\exp(-\beta \hat{r}_m / \sigma_n^2)}{\sum_{m'} \exp(-\beta \hat{r}_{m'} / \sigma_n^2)}, \quad \beta > 0, \quad (1.5)$$

for model m , where \hat{r}_m is a risk estimate for model m because such a choice yields the above bounds on the risk of the resulting mixture. We will discuss choices for β in the next chapter.

A notable feature of our work is that in each of the two cases, we compute an *unbiased estimate of risk* for the mixture estimator based on Stein (1973, 1981), which can be expressed in terms of the combination of the risk estimates for the component estimators under the individual models. This is useful for not only the resulting risk bound, but also in real practice when one wish to assess the quality of the mixture estimator.

1.2 Background

When least-squares estimators are used, our weights (1.5) at $\beta = 1/2$ are similar to those proposed in Buckland et al. (1997) for mixing (arbitrary) estimators, where only numerical analysis is provided. The exponential form of weights (with an arbitrary β) was also used in section 2.6 of Yang (2004) for prediction oracle inequalities, but it is closely related to the Cesaro mean of densities used in Barron (1987), Yang (2000, 2003), and Catoni (1997), when applied to Gaussian errors. In both Catoni (1999) (an extended version of Catoni (1997)) and Yang (2004), oracle inequalities similar to ours were obtained for the mean-squared error for prediction via mixing arbitrary bounded regression functions. However, their log M terms have coefficients depending on the assumptions of the problems, and are larger than ours even in the simplest Gaussian setting. In most of the work by Yang and Catoni, they also split the data into two sets, one for setting the weights, and the other for forming the estimates $\hat{\theta}^m$. In contrast, the analysis technique employed here allows use of all the data, and all at once in constructing both the weights and the estimates.

In the case of mixing shrinkage estimators, even though our weights retain the exponential form, the use of risk estimates takes us quite afar from the contexts considered in the aforementioned papers.

In both cases, the log M terms in (1.2'), (1.3') are reminiscent of that in the oracle inequality with the model selection target in Tsybakov (2003), although the setting there is in function estimation restricted to certain situations.

We will discuss these two cases separately in relation to existing work in the literature, but our techniques used to arrive at the risk bounds in these cases are similar. The least-squares estimators are discussed in greater details in this overview for their simplicity — the details of the shrinkage case will be deferred to the next chapter. In particular, because they are widely studied in the literature, and also because of their simplicity, we will discuss mixing estimators under the leading-term models in greater depth. This also serves well as an introduction. But we emphasize that our results hold for arbitrary subset regression models.

1.2.1 Least-Squares (“All-or-Nothing”) Estimators

Leading-Term Models

One focus is the case when the regressors are ordered in some natural way. We will examine the case when the regressors are arranged in decreasing relevance. For example, the regressors may be of decreasing importance judged by experts in a scientific field, or simply of

increasing complexity (e.g. polynomial functions of increasing degrees), which limits the suitability or the convincing power of their use. Another interpretation is that θ represents the coefficients of Fourier basis functions (in ascending frequency) of a frequency band-limited signal, corrupted by additive white Gaussian noise. In any case, if the response we are modelling has some regularity (of unknown extent), we anticipate that the tail sums of squares of the coefficients become small at some point, where it would be appropriate to zero out the estimated coefficients from then on. If we adhered to the previous notation, a model m here is the set of contiguous initial coordinate indices ending at some number no larger than d . For notational convenience, we let m be such an ending coordinate with $m \leq d$, so the subset of regressors in consideration is $\{1, 2, \dots, m\}$. Indeed, consider for $m = 0, 1, \dots, d$ the estimators

$$\hat{\theta}^m = (X_1, \dots, X_m, 0, \dots, 0),$$

which project X onto the spaces spanned by the first m elements in the standard basis. Such an estimator is sensible when the observations left out, X_{m+1}, \dots, X_n are small (comparable to the noise level). The risk of $\hat{\theta}^m$ is the mean-squared error

$$\begin{aligned} r_m &= \mathbb{E} |\hat{\theta}^m - \theta|^2 = \sum_{i \leq m} \mathbb{E} (X_i - \theta_i)^2 + \sum_{i > m} \theta_i^2 \\ &= \frac{m}{n} + \sum_{i > m} \theta_i^2, \end{aligned}$$

which has a decomposition into variance (or estimation error) m/n and bias (or approximation error) $\sum_{i > m} \theta_i^2$. Thus the risk depends on the tail sum of the excluded θ_i^2 as well as the number of terms included, relative to the sample size. In cases with the true coefficients being zero past some point i , then, the best balance in bias and variance occurs at some $m = k$, and mean-squared error would then be of the order k/n . Perhaps more common are cases in which the tail sum of the coefficients tapers less dramatically, e.g., like a polynomial in $1/i$, for which the best balance occurs for m that grows as a fractional power of n , and correspondingly the mean-squared error is some fractional power of $1/n$. It is known (see Yang and Barron (1998)) that for various such rates of tail sum decay, the minimax rates of statistical risk are achieved by using such $\hat{\theta}^m$ of a suitable dimension m . Nonetheless, it is not wise to presume in advance a particular order of regularity of the tail sum.

In our analysis, the model $m^* = m_n^*$ which achieves the best balance $r_{m^*} = \min_m r_m$, provides a target level of performance $r_{m^*} = r_*$ as in (1.4). A selection based estimator attempts to find an \hat{m} whose estimator $\hat{\theta}^{\hat{m}}$ has a risk that approaches this target. For each m , an unbiased estimate of the risk r_m is

$$\hat{r}_m = \sum_{i > m} X_i^2 + \frac{2m}{n} - \frac{d}{n} \quad (1.6)$$

which can be obtained either from Stein's unbiased risk estimator or from Akaike's (1970; 1973) information criterion (AIC). AIC is motivated by this unbiasedness of \hat{r}_m for each model m . It picks \hat{m} to achieve $\hat{r}_{\hat{m}} = \min_m \hat{r}_m$. However, in so doing, the unbiasedness property is lost due to the selection. Indeed, $\hat{r}_{\hat{m}}$ is an under-estimate of risk with expected value $\mathbb{E} \min_m \hat{r}_m \leq \min_m r_m$. The actual risk of any such selection estimator $\mathbb{E} |\hat{\theta}^{\hat{m}} - \theta|^2$ is larger than r_* .

There are asymptotic and finite sample analyses of AIC and other estimators based on selection, such as in Shibata (1981); Li (1987); Birgé and Massart (1997, 1998); Barron et al. (1999); Yang (1999); Baraud (1999); Kabaila (2002). Some of these asymptotic results show in the present setting that the factor by which the risk of such selections exceeds r_* tends to 1, as sample size gets large. However, the available finite sample bounds require a coefficient of r_* greater than 1, and if in these bounds one wants it to be close to 1, one incurs the cost of a large term added to the risk beyond r_* (where the term added can tend to infinity as the multiplier tends to 1). Also the indicated asymptotic conclusion requires the assumption (as may indeed be appropriate) that there is no k for which the coefficients are all zero past k . Some attempts to adjust the criterion (e.g. by using $m(\log n)/n$ instead of the $2m/n$) possibly motivated by Bayes or description length considerations (as in Schwarz (1978) or Rissanen (1978)) have risk that exceeds r_* by multiplicative factors of order $\log n$ asymptotically.

The individual coordinates of the mixture (1.1) are $\hat{\theta}_i = \hat{c}_i X_i$, where the filtering coefficients $\hat{c}_i = \sum_{m \geq i} \hat{\rho}_m$ form a decreasing sequence in i . For other approaches to obtaining decreasing multiplier sequences, see Pinsker (1980) who determines the asymptotically minimax solution for particular ellipsoidal classes (with bounded $\sum_i i^{2s} \theta_i^2$ and known index of regularity s), and Beran and Dümbgen (1998) who use empirically optimized decreasing \hat{c}_i , with the aim to approach a risk target smaller than ours considered here (namely the minimal risk among all monotone decreasing multiplier sequences) but at a greater added cost of order at least $n^{-1/2}$. In contrast, for our less ambitious target r_* (minimal risk among all leading-term models), our added cost will be only of order $1/n \log n$.

General Subset Models

An extreme of this subset model class entails considering all subsets. The corresponding model selection problem for such a model class is widely studied as estimators based on thresholding individual coefficients. That is, if the size of a particular coefficient is smaller than some threshold, set it to zero; otherwise, leave it alone. For example, Donoho and Johnstone (1995) and Johnstone (1998) apply such estimators to wavelet coefficients and show that they have good asymptotic properties over many classes of function spaces (the context in which they consider such a canonical Gaussian sequence model is function estimation.)

However, when such a large model class (exponential in the number of coefficients) is considered, mixing across models often incur a large penalty because there is simply not enough data to compute weights accurately. However, we will provide a way to penalize the complex models via prior weights. Our theory will show that with such complexity regularization, our mixing procedure performs well with respect to the best procedure plus the penalty of its associated complexity.

Nevertheless, our theory holds for the general subset model classes. That is, we provide the generalization for (1.5), (1.6) for arbitrary subsets of $\{1, 2, \dots, d\}$. Provided that d is large enough such that the underlying regression function is well-approximated by the regression design, our work corresponds to mixing arbitrary regression functions.

1.2.2 Shrinkage Estimators

The James-Stein (1961) shrinkage estimator was shown to have uniformly lower mean-squared error than the least-squares estimator for $d \geq 3$. And the positive-part James-Stein (estimator) has uniform risk improvement over the James-Stein, so we shall use it for risk reduction. Both estimators shrink X toward 0 if $|X|^2$ is small but otherwise do little.

Two-Block (Leading and Trailing Terms) Shrinkage

For $d \geq 6$, we can exploit more shrinkage opportunities if we divide the d coefficients into two contiguous blocks, each of size at least three coordinates, and employ two positive-part James-Stein estimators independently. This way, if the coefficients in one block is small while the other large, the small one is shrunk and the large one is almost left intact. Each coordinate where the first block may end represents a particular model with the corresponding two-block shrinkage estimator just described. Since we do not know *a priori* which one of these would provide for the most shrinkage, we want to mix across these two-block shrinkage estimators.

Blockwise James-Stein procedures have been widely used in to achieve adaptive estimators in the wavelet context, e.g. Donoho and Johnstone (1995); Johnstone (1998). In particular, asymptotic analysis by Cavalier and Tsybakov (2001) (see also Goldenshluger and Tsybakov (2001)) shows that it yields an exact oracle inequality for the class of estimators $(c_i X_i)$ with monotone non-increasing multiplier sequences c_i mentioned above (studied by Beran and Dümbgen (1998)) and is sharp minimax over general ℓ_2 ellipsoids. Our two-block shrinkage mixture (finite-dimensional) approaches a lower risk target than the risk of the best of least-squares for the leading-term models, although higher than the risk target for the monotone class. This is in part due to the risk overhead (of about 1.2 per block) introduced by the positive-part James-Stein estimator even when the underlying $\theta = 0$. But such overhead becomes negligible in the asymptotic regime.

General Two-Set Shrinkage

As the “all-or-nothing” least-squares estimators can be applied to general subset models as a generalization of the leading-term ones, we can also generalize the two-block shrinkage estimators to arbitrary subsets. George (1986a) examined such multiple shrinkage estimators in this multivariate normal mean problem, and the framework started in George (1986b) includes shrinkage estimators that project to lower-dimensional spaces. His method is based on pseudo-Bayes interpretation of the positive-part James-Stein estimator. However, such a method, as with many other similar ones, often has arbitrary scales associated with the preference for the models. He tried to address this via the Bayesian perspective by using calibration weights on these models that act like prior probabilities.

The weights we choose for mixing these two-part shrinkage models are different from George’s, because we retain the exponential form (1.5) using risk estimates for the individual models. We are motivated by the goal of having a mixture procedure that yields an oracle inequality upper-bounding its risk (compared to the best model risk target), rather than a pseudo-Bayes interpretation.

Bayes Estimators

Strawderman (1971) proposed a prior that induces a Bayes estimator similar to the positive-part James-Stein estimator, and showed that for dimension $d \geq 5$, the prior is proper and that the resulting estimator is minimax and admissible. One could apply this estimator to the two-block shrinkage case here, with the weighting being the posterior probabilities of the models. Such an estimator should have desired risk properties, as well as the adaptability to the models.

One problem is that the estimators under the individual models, as well as the Bayes factors for weighting them are complicated. Though they are relatively costly to compute, the bigger difficulty is that the Bayes mixture’s risk is almost intractable to analyse. Thus, usually only simulation results can be used for comparison. The requirement that each block must have cardinality at least 5 is a nuisance, though in practice, it does not pose a large risk penalty due to lack of adaptability to the models with small blocks.

For Bayesian procedures for model selection and averaging, see George and McCulloch (1997); Berger and Pericchi (1996); Kass and Raftery (1995).

1.3 Outline of Thesis

In Chapter 2, we develop our theory of mixing estimators and analyse the mixture estimator’s risk in details. The chapter starts with using Stein’s unbiased estimates for the risks of component estimators to compute the risk estimate for the mixture, and eventually leads to oracle inequalities for the mixture, with an idealized risk target based on the minimum risk of all the models considered. The mixture of least-squares estimator and mixture of shrinkage estimators on subset models are discussed separately, although the analyses techniques are more or less the same. Simulation results for both mixture estimators are shown for the leading-term models.

The theory in Chapter 2 is developed without a model prior probability, which essentially corresponds to using a uniform model prior. This gives sharp oracle inequalities for mixing. In Chapter 3, we discuss the use of a model prior for complexity regularization. In short, weights are employed to favour the simpler models and resembles using a prior probability for the models. The risk bounds in this chapter are generally not as tight as those in Chapter 2, but they may be better suited to practical cases when the models should not be considered with equal importance. Some discussions are given in Chapter 4.

Having completed the comparison between the work in this thesis and those in the literature with n , for the rest of the thesis, we will work with the canonical regression problem with

$$\sigma_n^2 = 1$$

unless stated otherwise. This unclutters the analysis symbolically. If desired, such a simple scaling by n can be easily replenished.

A Note on the Notation

Since there are only so many letters I could use, I reuse $r, \hat{\theta}, \hat{\theta}^m$ for both the least-squares and the shrinkage cases. It should be clear from the context which estimators and risks I am talking about. And when the two are compared, I sometimes employ superscripts to distinguish among them, and sometimes just use more words to avoid any possible confusion at the risk of being overly verbose.

I am fond of the de Finetti notation of writing indicators by identifying them with their sets. For example,

$$\mathbb{I}_{\{\gamma=1\}} \stackrel{\text{def}}{=} \{\gamma = 1\}.$$

But I use them both depending on their functional purposes and their visual effect — the former is nice when you wish to emphasize the indicator as a function of a variable of choice when multiple variables are involved (usually the case), e.g.

$$\mathbb{I}_{\{1\}}(\gamma) = \mathbb{I}_{\{\gamma(x)=1\}} = \mathbb{I}_{\{\gamma=1\}}(x);$$

the latter comes in handy when such a distinction is unnecessary and when it appears in superscripts or subscripts so that one does not have to squint the eyes to read the sub- or super-subscripts, e.g.

$$a^{\{\gamma=1\}} = \begin{cases} a & \text{if } \gamma = 1 \\ 1 & \text{if } \gamma \neq 1 \end{cases}.$$

A vector X in \mathbb{R}^d can be written in terms of its ordered members

$$X = (X_i)_{i \leq d}.$$

Minima and maxima can be denoted by

$$a \wedge b = \min \{a, b\}, \quad a \vee b = \max \{a, b\}.$$

Expectation over the data, say X , taken with respect to the sampling distribution parameterized by θ , is denoted by either \mathbb{E} or \mathbb{E}_θ with the latter emphasizing the dependence on θ .

Chapter 2

Mixture Estimators for the Multivariate Normal Mean

2.1 General Mixture

2.1.1 Introduction and Examples

In a canonical regression problem, we want to estimate the unknown mean of a d -variate normal X which is understood to be the data to be fitted to models comprising at most d regressors. In our simplified setting, the variance is assumed to be known, such that we scale our variables X_i to be identically distributed and have unit variance. That is, we assume $X \sim \text{Normal}_d(\theta, I)$ and we want to estimate θ under squared-error loss

$$\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|^2$$

where $\|\cdot\|$ is the Euclidean norm. And the criterion for evaluating an estimator $\hat{\theta}$ will be its risk, or expected loss

$$r(\theta, \hat{\theta}) = \mathbb{E} \|\theta - \hat{\theta}\|^2.$$

We study estimators of the form

$$\hat{\theta} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{\theta}^m \quad \text{with} \quad \sum_{m \in \mathcal{M}} \hat{\rho}_m = 1. \quad (2.1)$$

That is, the estimator is a convex combination of estimators $\hat{\theta}^m$ over a finite collection \mathcal{M} , each with normalized weights $\hat{\rho}_m \in [0, 1]$ that could be either fixed or determined by the data X .

Least-Squares Estimators

[2.2] **EXAMPLE (Block Models).** The leading-term models mentioned in the introduction can be generalized to block models. We use

$$\hat{\theta}^m = (0, \dots, 0, X_{k_1}, \dots, X_{k_2}, 0, \dots, 0)' \quad \text{where } 1 \leq k_1 \leq k_2 \leq d$$

to estimate θ . Notationally, we can denote each model m by (k_1, k_2) , and the model class by

$$\mathcal{M} = \{(k_1, k_2) : 1 \leq k_1 \leq k_2 \leq d\},$$

such that the cardinality of \mathcal{M} is $\#\mathcal{M} = d(d+1)/2$. If θ_i are (generalized) Fourier coefficients, just as a leading-term model gives rise to an estimator $\hat{\theta}^m$ termed a *low-pass filter* analogously, this can also be called a *band-pass filter*. ||

EXAMPLE (General Subset Models). Here each model m is a subset of $\{1, 2, \dots, d\}$, [2.3] representing the coordinate indices of X for inclusion. That is, the resulting estimator is

$$\hat{\theta}_i^m = X_i \mathbb{1}_{\{i \in m\}}, \quad i = 1, \dots, d.$$

On one extreme, m can range over all subsets of the indices, which will give $\#\{0, 1\}^d = 2^d$ models. For such a large number of models, we generally would not mix across all models without ways to penalize the more complex ones. We will say more of this later.

Observe that this is a generalization of the leading-term models (Section 1.2) and block models above. For example, the leading-term models can be represented by $m = \{1, \dots, k\}$ for $0 \leq k \leq d$ (with the convention that m represents the empty set for $k = 0$). ||

In general, the least-squares estimators under the subset models can be nicknamed “**all-or-nothing**” estimators as they multiply the data X_i by ones or zeros.

Two-Set Shrinkage Estimators

Instead of using only ones and zeros, we could also use a shrinkage coefficient between zero and one to multiply the data to yield an estimator. To estimate such a coefficient, the positive part James-Stein estimator is the tool that we employ as a building block under each model. First, the basic estimator (one-block, applied to all $d \geq 3$ coefficients) is

$$\begin{aligned} & \left(1 - \frac{d-2}{|X|^2}\right)_+ X, \quad \text{where } a_+ = a \vee 0 \\ & = X - \gamma X, \quad \gamma = \gamma(X) \stackrel{\text{def}}{=} 1 \wedge \frac{d-2}{|X|^2} \end{aligned}$$

where γ is the data-determined shrinkage coefficient. It shrinks X toward 0 if $|X|$ is small but otherwise does little. This estimator was shown by James and Stein (1961) to dominate

the maximum likelihood X under squared-error loss. The risk reduction can be up to about $d - 1$ if $\theta = 0$.

[2.4] **EXAMPLE (Two-Block Models).** For $d \geq 6$, we can exploit more shrinkage opportunities if we divide the d coefficients into two contiguous blocks, each of size at least three coordinates, and employ two positive-part James-Stein estimators independently. This way, if the coefficients in one block are small while the other large, the former are shrunk and the latter are almost left intact. Let the first block end at coordinate m (inclusively) with $m \in \mathcal{M} = \{3, \dots, d-3\}$ with $\#\mathcal{M} = d-5$. Then, we can define

$$\hat{\theta}_i^m = (1 - a_i^m)X_i, \quad \text{where } a_i^m = a_i^m(X) = \begin{cases} 1 \wedge \frac{m-2}{\sum_{j \leq m} X_j^2}, & \text{if } i \leq m \\ 1 \wedge \frac{d-m-2}{\sum_{j > m} X_j^2}, & \text{if } i > m \end{cases}.$$

Since we do not know *a priori* which boundary m would yield the smallest risk, we want to mix over \mathcal{M} . ||

[2.5] **EXAMPLE (Two-Set Models).** The previous example is the shrinkage analogue of least-squares for the leading-term models. Now we present the generalization of this for the general subset models. Let each model $m \in \mathcal{M}$ be a subset of $\{1, 2, \dots, d\}$ such that $\#m \in \{3, 4, \dots, d-3\}$. Denote the shrinkage coefficient for the coordinates within and without the set m by

$$\gamma_m = \gamma_m(X) = 1 \wedge \frac{\#m-2}{\sum_{j \in m} X_j^2}, \quad \text{and} \quad \gamma_{m^c} = \gamma_{m^c}(X) = 1 \wedge \frac{d-\#m-2}{\sum_{j \notin m} X_j^2} \quad (2.6)$$

respectively, where the complement of m is taken with respect to $\{1, \dots, d\}$. Finally, we define for $i = 1, \dots, d$

$$\hat{\theta}_i^m = (1 - a_i^m)X_i, \quad \text{where } a_i^m = a_i^m(X) = \gamma_m(X) \mathbb{1}_{\{i \in m\}} + \gamma_{m^c}(X) \mathbb{1}_{\{i \notin m\}}. \quad (2.7)$$

As before, we want to mix across $m \in \mathcal{M}$. ||

Remark: The theory we are going to develop does not require that the models in \mathcal{M} be distinct, though there is no reason for including repeated models. For instance, since we treat m and its complement m^c symmetrically in Example 2.5, the model m is equivalent to the model $m' = m^c$, and they can coexist in \mathcal{M} . However, the multiplicity of a model will be reflected in an apparent increase in the weight for that model by the respective multiplicity. ◁

2.1.2 Risk Estimate

From now on, we assume that the estimator under each model is square-integrable (and hence, so is the mixture estimator) such that it is meaningful to examine the risks under squared-error loss. The key tool in this setting is Stein's (1973; 1981) unbiased estimator of risk. For us, an important realization is that, unlike AIC (Akaike (1973)) which gives

unbiased risk estimators only for each model separately, Stein's identity can be applied more generally to provide an unbiased estimator of the risk of a mixture estimator.

We restate Theorem 1 from Stein (1981) below. We write $a \cdot b = \sum_{i=1}^d a_i b_i$ for the inner product and ∇ the gradient (∇_i) where $\nabla_i = \partial/\partial X_i$, and hence

$$\nabla \cdot h = \sum_{i=1}^d \frac{\partial}{\partial X_i} h_i(X).$$

THEOREM (Stein). Consider the estimator $\delta(X) = X - h(X)$ for θ such that $h : \mathbb{R}^d \mapsto \mathbb{R}^d$ is an almost differentiable¹ function for which

$$\mathbb{E}_\theta |\nabla_i h_i(X)| < \infty, \quad \text{for each } i = 1, \dots, d.$$

Then an unbiased estimate of the risk of δ is

$$\hat{r}_\delta(X) = d + |h(X)|^2 - 2\nabla \cdot h(X),$$

meaning $\mathbb{E} |\theta - \delta(X)|^2 = \mathbb{E}_\theta \hat{r}_\delta(X)$ for each θ .

Proof: Essentially integration by parts using the normal density. See also Corollary 7.2 on p. 273 of Lehmann and Casella (1998). ◻

THEOREM (Unbiased Risk Estimate for Mixture). Assume that $\hat{\rho}_m(X)$ is almost differentiable for each model m in a finite collection \mathcal{M} , and that each $\hat{\theta}^m$ satisfies the condition (for δ) of Theorem 2.8 such that an unbiased estimate of its risk exists and equals

$$\hat{r}_m \stackrel{\text{def}}{=} d + |X - \hat{\theta}^m|^2 - 2\nabla \cdot (X - \hat{\theta}^m).$$

Then an unbiased estimate \hat{r} of the risk $\mathbb{E} |\theta - \hat{\theta}|^2$ of the mixture estimator $\hat{\theta} = \sum_m \hat{\rho}_m \hat{\theta}^m$ is

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[\hat{r}_m - |\hat{\theta} - \hat{\theta}^m|^2 - 2(\nabla \log \hat{\rho}_m) \cdot (\hat{\theta} - \hat{\theta}^m) \right].$$

Remark: If $\hat{\rho}_m \equiv 0$, we mean $\hat{\rho}_m \nabla \log \hat{\rho}_m = \nabla \hat{\rho}_m = 0$. ◁

Proof: Since $\hat{\rho}_m \in [0, 1]$ for each $m \in \mathcal{M}$, it is clear that $\hat{\theta} = \sum_m \hat{\rho}_m \hat{\theta}^m$ satisfies the condition (for δ) of Theorem 2.8. We use Stein's identity to obtain an unbiased estimate \hat{r}

¹A function is **almost differentiable** if each of its coordinates can be represented by a directional integral. That is, for $x, z \in \mathbb{R}^d$ and $i = 1, \dots, d$,

$$h_i(x+z) - h_i(z) = \int_0^1 z \cdot \nabla h_i(x+tz) dt.$$

And $\nabla_j h_i$ is naturally called the **derivative** of h with respect to x_j . If a function is continuous and piecewise differentiable, then it is almost differentiable.

of the risk $r(\theta, \hat{\theta}) = \mathbb{E} |\hat{\theta} - \theta|^2$ such that $r = \mathbb{E}_\theta \hat{r}$ for each θ . Let $g^m(X) = X - \hat{\theta}^m$ and $g(X) = X - \hat{\theta} = \sum_m \hat{\rho}_m g^m(X)$. Then the risk estimate is

$$\begin{aligned} \hat{r} &= d - 2 \sum_{i=1}^d \nabla_i g_i + |g|^2, \quad \nabla_i = \frac{\partial}{\partial x_i} \\ &= d - 2 \sum_{i=1}^d \sum_{m \in \mathcal{M}} \nabla_i (\hat{\rho}_m g^m)_i + |g|^2 \\ &= d - 2 \sum_{i=1}^d \sum_{m \in \mathcal{M}} [(\nabla_i \hat{\rho}_m) g_i^m + \hat{\rho}_m (\nabla_i g_i^m)] + |g|^2 \\ &= \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[d - 2 \sum_{i=1}^d \nabla_i g_i^m \right] - 2 \sum_{i=1}^d \sum_{m \in \mathcal{M}} (\nabla_i \hat{\rho}_m) g_i^m + |g|^2 \\ &= \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[d - 2 \sum_{i=1}^d \nabla_i g_i^m + |g^m|^2 \right] - 2 \sum_{m \in \mathcal{M}} \sum_{i=1}^d (\nabla_i \hat{\rho}_m) g_i^m + |g|^2 - \sum_{m \in \mathcal{M}} \hat{\rho}_m |g^m|^2. \end{aligned}$$

Here we have exchanged the order of summation in the double sum because \mathcal{M} is finite, and we have added and subtracted $\sum_{m \in \mathcal{M}} \hat{\rho}_m |g^m|^2$ because it is finite (for almost every X). Then

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[\hat{r}_m - |g - g^m|^2 \right] - 2 \sum_{m \in \mathcal{M}} \sum_{i=1}^d (\nabla_i \hat{\rho}_m) g_i^m,$$

where

$$\hat{r}_m \stackrel{\text{def}}{=} d - 2 \sum_{i=1}^d \nabla_i g_i^m + |g^m|^2$$

is the unbiased estimate of risk of the component estimator $\hat{\theta}^m(X) = X - g^m(X)$. Use the variance calculation $\mathbb{E} (Z - \mathbb{E} Z)^2 = \mathbb{E} Z^2 - (\mathbb{E} Z)^2$ over each coordinate i to obtain

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m |g^m|^2 - |g|^2 = \sum_{m \in \mathcal{M}} \hat{\rho}_m |g^m - g|^2 = \sum_{m \in \mathcal{M}} \hat{\rho}_m |\hat{\theta}^m - \hat{\theta}|^2 \geq 0,$$

We have arrived at

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[\hat{r}_m - |\hat{\theta}^m - \hat{\theta}|^2 \right] - 2 \sum_{m \in \mathcal{M}} (\nabla \hat{\rho}_m) \cdot (X - \hat{\theta}^m).$$

Now the rest of the proof is a direct application of the following technical lemma, which although seems trivial, enunciates a structure in mixing estimators. \square

[2.10] LEMMA (Orthogonality). *Assume that the derivative $\nabla_i \hat{\rho}_m$ is finite for each i and m . Any random vector $h(X) \in \mathbb{R}^d$ not a function of m will have 0 as its inner product with $\nabla \hat{\rho}_m$.*

$$\sum_{m \in \mathcal{M}} (\nabla \hat{\rho}_m) \cdot h(X) = 0.$$

Moreover, if $(Z^m)_{m \in \mathcal{M}}$ are any collection of vectors in \mathbb{R}^d that has the null vector 0 as the mean when mixed with $\hat{\rho}$, i.e.

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m Z^m = 0,$$

then any such h (independent of m) and $(Z^m)_{m \in \mathcal{M}}$ are orthogonal under $\hat{\rho}$,

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m [Z^m \cdot h(X)] = 0.$$

In particular, $\nabla \log \hat{\rho}_m$ has $\hat{\rho}$ -mean 0 and is orthogonal to h under $\hat{\rho}$.

Proof: Observe that $\sum_m \hat{\rho}_m = \text{constant}$ implies $0 = \nabla \sum_m \hat{\rho}_m = \sum_m \nabla \hat{\rho}_m$ by the finiteness of $\nabla_i \hat{\rho}_m$. The fact that h is not a function of m means that we can exchange the order of summation: the one over m and that over the coordinates for the inner product. For the second statement, since the inner product is bilinear, we have $(\sum_m \hat{\rho}_m Z^m) \cdot h(X) = 0 \cdot h = 0$. The last statement is obvious by combining the previous two. \square

This unbiased assessment of risk has three terms. The primary term $\sum_m \hat{\rho}_m \hat{r}_m$ is the weighted average of the individual risk estimates. With suitable design of the weights, this average will not be much larger than $\min_m \hat{r}_m$, as we shall see. In this respect, this term is analogous to what appears in the AIC analysis, except that it is here appearing in an unbiased estimate of the risk of the combined model, not only for the individual models.

The second term $- \sum_m \hat{\rho}_m |\hat{\theta} - \hat{\theta}^m|^2$ wonderfully illustrates an advantage of model mixing. If the estimates $\hat{\theta}^m$ vary with m (that is, if the fits are different for different m), then averaging them (with weights $\hat{\rho}_m$) leads to a reduction in the unbiased risk assessment given by the weighted average of the squared distance of the $\hat{\theta}^m$ from their centroid $\hat{\theta}$. A nice feature of the unbiased risk estimator is that it cleanly reveals this reduction based on variability of estimates (as m varies), rather than based on the classical variance of the estimators (which addresses variability with changes in the sample, not changes in the estimates with m for a given sample).

The third term $-2 \sum_m \hat{\rho}_m (\nabla \log \hat{\rho}_m) \cdot (\hat{\theta} - \hat{\theta}^m)$ quantifies the effect of the data-sensitivity of the weights (through their gradients with respect to the data X). Constant weights would make this term zero, but would not permit means to adapt the fit to the models that have smaller \hat{r}_m .

The following weights are a reasonable way to form a mixture by emphasizing the component estimators assessed to have low risks. Indeed, for mixing least-squares estimators, this form of weights will yield the desired oracle inequality (1.2') in the introductory chapter.

DEFINITION. For $\beta \geq 0$, define

$$\hat{\rho}_m = \frac{\exp(-\beta \hat{r}_m)}{\sum_{m' \in \mathcal{M}} \exp(-\beta \hat{r}_{m'})}. \quad \triangleleft$$

Note that $\hat{\rho}_m$ is strictly positive for each m .

This parameterization of weights allows us to control the degree of concentration in the model with low risk estimates. Indeed, when $\beta > 0$, we put smaller weights on the models

with higher risk estimates. The higher the β , the stronger the concentration is on the models with low risk estimates. On one extreme, if $\beta = 0$, the weights are uniform, ignoring \hat{r}_m altogether. When $\beta \rightarrow \infty$, the mixture $\hat{\theta}$ consists of only the models with minimal risk estimates $\{m : \hat{r}_m = \min_m \hat{r}_m\}$ with uniform weights. When the minimum is unique, then $\hat{\theta}$ is just a model selection estimator based on the model $m = \hat{m}$ minimizing \hat{r}_m .

Now write the normalization constant in $\hat{\rho}_m$ as $C = \sum_m \exp(-\beta \hat{r}_m)$. Then

$$\nabla_i \hat{\rho}_m = C^{-1} (-\beta \nabla_i \hat{r}_m) \exp(-\beta \hat{r}_m) - C^{-2} \exp(-\beta \hat{r}_m) \sum_{k \in \mathcal{M}} (-\beta \nabla_i \hat{r}_k) \exp(-\beta \hat{r}_k).$$

And hence,

$$\nabla \hat{\rho}_m = -\beta \hat{\rho}_m \left[\nabla \hat{r}_m - \sum_k \hat{\rho}_k \nabla \hat{r}_k \right].$$

Thus, we can apply Lemma 2.10 with $Z^m = \hat{\theta} - \hat{\theta}^m$ and $h(X) = \sum_k \hat{\rho}_k \nabla \hat{r}_k$ to obtain

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[\hat{r}_m - |\hat{\theta} - \hat{\theta}^m|^2 + 2\beta \nabla \hat{r}_m \cdot (\hat{\theta} - \hat{\theta}^m) \right]. \quad (2.12)$$

2.2 Mixing Least-Squares Estimators

2.2.1 Risk Estimate for Mixture

In this section, we specialize the theory to the least-squares estimators in Example 2.3. Recall that each model $m \in \mathcal{M}$ is a subset² of $\{1, \dots, d\}$, with the estimator

$$\hat{\theta}^m = (X_i \mathbb{1}_{\{i \in m\}})_{i \leq d}.$$

Its risk is

$$\begin{aligned} r_m(\theta) &\stackrel{\text{def}}{=} \mathbb{E} |\hat{\theta}^m - \theta|^2 = \sum_{i \in m} \mathbb{E} (X_i - \theta_i)^2 + \sum_{i \notin m} \theta_i^2 \\ &= \#m + \sum_{i \notin m} \theta_i^2. \end{aligned} \quad (2.13)$$

An unbiased estimate of risk for this estimator is

$$\hat{r}_m = \sum_{i \notin m} X_i^2 + 2\#m - d,$$

which can be obtained from Theorem 2.8 or directly knowing that $X_i^2 - 1$ is unbiased for θ_i^2 . This is exactly the (**stepwise**) AIC quantity (when restricted to **nested models**) to be minimized for model selection over the leading-term estimators $\hat{\theta}^k = (X_1, X_2, \dots, X_k, 0, \dots, 0)'$ where

$$\text{AIC}_k = \sum_{i > k} X_i^2 + 2k,$$

²For the leading-term models, m is an initial set.

up to a constant offset d not depending on the model. The crucial fact that leads to much simplification in the formula for the unbiased estimate of the subset mixture estimator is as follows.

$$\nabla \hat{r}_m = 2(X_i \mathbb{1}_{\{i \notin m\}})_{i \leq d}.$$

In other words,

$$\nabla \hat{r}_m = 2(X - \hat{\theta}^m). \quad (2.14)$$

And applying Lemma 2.10 with $h = X - \hat{\theta}$ and $Z = \hat{\theta} - \hat{\theta}^m$ yields

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m [\nabla \hat{r}_m \cdot (\hat{\theta} - \hat{\theta}^m)] = 2 \sum_{m \in \mathcal{M}} \hat{\rho}_m |\hat{\theta} - \hat{\theta}^m|^2.$$

PROPOSITION. *The mixture (with weights above) of subset estimators has an unbiased estimator of risk* [2.15]

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[\hat{r}_m - (1 - 4\beta) |\hat{\theta}^m - \hat{\theta}|^2 \right].$$

Furthermore, for $0 \leq \beta \leq 1/4$, the risk estimate can be bounded by

$$\hat{r} \leq \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m,$$

with equality when $\beta = 1/4$.

Proof: The third term of the unbiased estimate of risk for $\hat{\theta}$ in (2.12) is now proportional to the second and we can combine them. Since $(1 - 4\beta) |\hat{\theta}^m - \hat{\theta}|^2 \geq 0$ for $\beta \leq 1/4$, the upper bound follows. \square

The parameter β in the weights $\hat{\rho}$, controls the relative importance of averaging across models (small β) and picking out the one that was empirically best (large β). When β is strictly less than $1/4$, we continue to see overall improvement due to averaging — the risk is strictly less than $\sum_m \hat{\rho}_m \hat{r}_m$.

We now focus on the first term since it is the upper-bound, and in particular, when $\beta = 1/4$ the two other terms in the risk estimate mentioned above cancel each other.

2.2.2 An Upper-bound for the Combined Risk Estimate

To further the story, we compare the unbiased risk estimate of this model-averaged estimator $\hat{\theta}$ to the best of the unbiased risk estimates of the models,

$$\hat{r}_* \stackrel{\text{def}}{=} \min_{m \in \mathcal{M}} \hat{r}_m \stackrel{\text{def}}{=} \hat{r}_{\hat{m}} \quad \text{for some } \hat{m} \in \mathcal{M}.$$

This is useful because it is related to the risk target

$$r_* = r_*(\theta) = \min_{m \in \mathcal{M}} r_m(\theta),$$

where r_m was defined in (2.13), by virtue of

$$\mathbb{E}_\theta \hat{r}_* = \mathbb{E}_\theta \min_m \hat{r}_m \leq \min_m \mathbb{E}_\theta \hat{r}_m = \min_m r_m = r_*. \quad (2.16)$$

In this section, we assume $\beta > 0$.

[2.17] DEFINITION. The discrete **entropy** for the probability vector w over a space \mathcal{M} is

$$H(w) = \sum_{m \in \mathcal{M}} w_m \log \frac{1}{w_m}. \quad \triangleleft$$

It is well-known that H is concave and bounded in the interval $[0, \log(\#\mathcal{M})]$ (See Cover and Thomas, 1991, Chapter 2).

[2.18] DEFINITION. Let $\psi = \psi(\#\mathcal{M})$ be a constant defined by the solution to

$$\psi = \log \frac{\#\mathcal{M} - 1}{\psi} - 1. \quad \triangleleft$$

It is clear that ψ is increasing in $\#\mathcal{M}$ and $\psi < \log(\#\mathcal{M})$. Also, for each $K > 0$,

$$\psi \leq \max \left\{ K, \log \frac{\#\mathcal{M} - 1}{K} - 1 \right\}, \quad (2.19)$$

by considering separately whether $\psi \leq K$.

[2.20] LEMMA. With $\hat{r}_{\hat{m}} = \min_m \hat{r}_m = \hat{r}_*$,

(a) we have

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m = \hat{r}_* + \frac{1}{\beta} \left[H(\hat{\rho}) + \log \hat{\rho}_{\hat{m}} \right]. \quad (2.21)$$

An immediate upper-bound is thus obtained:

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m - \hat{r}_* < \frac{1}{\beta} H(\hat{\rho}) \leq \frac{1}{\beta} \log(\#\mathcal{M}). \quad (2.22)$$

(b) For ψ in Definition 2.18,

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m \leq \hat{r}_* + \frac{\psi(\#\mathcal{M})}{\beta}.$$

Proof: (a) It is easy to check that

$$\hat{r}_m = \frac{1}{\beta} \left[\log \frac{1}{\hat{\rho}_m} - \log \sum_{k \in \mathcal{M}} \exp(-\beta \hat{r}_k) \right]. \quad (2.23)$$

Then by adding and subtracting the minimal risk estimate \hat{r}_* ,

$$\begin{aligned} \hat{r}_m &= \hat{r}_* + \frac{1}{\beta} \left[\log \frac{1}{\hat{\rho}_m} - \log \sum_{k \in \mathcal{M}} \exp(-\beta \hat{r}_k) - \beta \hat{r}_* \right] \\ &= \hat{r}_* + \frac{1}{\beta} \left[\log \frac{1}{\hat{\rho}_m} + \log \hat{\rho}_{\hat{m}} \right]. \end{aligned}$$

Now average with respect to weights $\hat{\rho}_m$ to obtain the first statement. Since $\hat{\rho}_{\hat{m}} < 1$, its log is strictly negative. Hence the first inequality in (2.22) follows. The second follows from the fact that the entropy is bounded by the log-cardinality of the space.

(b) If we consider m as a random variable on the space \mathcal{M} with probability $\hat{\rho}$ (given X), then

we manipulate with conditional entropy (Cover and Thomas, 1991, Chapter 2), depending on whether $m = \hat{m}$, to obtain the identity

$$H(\hat{\rho}) = (1 - \hat{\rho}_{\hat{m}})H(\tilde{\rho}) + H(\hat{\rho}_{\hat{m}}),$$

where $\{\tilde{\rho}_m : m \neq \hat{m}\}$ are the the renormalized weights on $\mathcal{M} \setminus \{\hat{m}\}$ and $H(\hat{\rho}_{\hat{m}})$ is the binary entropy. (Cf. Fano's inequality.) Thus, (2.21) becomes³

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m - \hat{r}_* = \frac{1}{\beta} \left[(1 - \hat{\rho}_{\hat{m}})H(\tilde{\rho}) + H(\hat{\rho}_{\hat{m}}) + \log \hat{\rho}_{\hat{m}} \right].$$

Hence, the bracketed terms on the right is upper-bounded by

$$(1 - \hat{\rho}_{\hat{m}}) \log(\#\mathcal{M} - 1) + H(\hat{\rho}_{\hat{m}}) + \log \hat{\rho}_{\hat{m}} = (1 - \hat{\rho}_{\hat{m}}) \left[\log(\#\mathcal{M} - 1) - \log \frac{1 - \hat{\rho}_{\hat{m}}}{\hat{\rho}_{\hat{m}}} \right], \quad (2.24)$$

which is clearly concave in $\hat{\rho}_{\hat{m}}$. Setting to zero the first derivative of (2.24) with respect to $\hat{\rho}_{\hat{m}}$, we see that the maximum of the bound occurs at $\hat{\rho}_{\hat{m}} = \rho_{\dagger}$ satisfying

$$\log(\#\mathcal{M} - 1) - \log \frac{1 - \rho_{\dagger}}{\rho_{\dagger}} = \frac{1}{\rho_{\dagger}}.$$

In terms of the odds of “error” (the event $m \neq \hat{m}$), the maximum of the bound occurs at

$$O_{\dagger} \stackrel{\text{def}}{=} \frac{1 - \rho_{\dagger}}{\rho_{\dagger}} = \frac{1}{\rho_{\dagger}} - 1 = \log \frac{\#\mathcal{M} - 1}{O_{\dagger}} - 1.$$

³A proof by direct computation is as follows. Define for each $m \neq \hat{m}$

$$\tilde{\rho}_m = \frac{\exp(-\beta \hat{r}_m)}{\sum_{k \neq \hat{m}} \exp(-\beta \hat{r}_k)} = \frac{\exp[-\beta(\hat{r}_m - \hat{r}_*)]}{\sum_{k \neq \hat{m}} \exp[-\beta(\hat{r}_k - \hat{r}_*)]}.$$

It is easy to show that

$$\sum_{m \neq \hat{m}} \exp[-\beta(\hat{r}_k - \hat{r}_*)] = \frac{1 - \hat{\rho}_{\hat{m}}}{\hat{\rho}_{\hat{m}}},$$

such that

$$\hat{r}_m - \hat{r}_* = \frac{1}{\beta} \left[\log \frac{1}{\hat{\rho}_m} - \log \frac{1 - \hat{\rho}_{\hat{m}}}{\hat{\rho}_{\hat{m}}} \right].$$

Then,

$$\begin{aligned} \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m - \hat{r}_* &= (1 - \hat{\rho}_{\hat{m}}) \left[\sum_{m \neq \hat{m}} \tilde{\rho}_m \hat{r}_m - \hat{r}_* \right] \\ &= \frac{1 - \hat{\rho}_{\hat{m}}}{\beta} \left[\sum_{m \neq \hat{m}} \tilde{\rho}_m \log \frac{1}{\hat{\rho}_m} - \log \frac{1 - \hat{\rho}_{\hat{m}}}{\hat{\rho}_{\hat{m}}} \right] \\ &= \frac{1}{\beta} \left[(1 - \hat{\rho}_{\hat{m}})H(\tilde{\rho}) + H(\hat{\rho}_{\hat{m}}) + \log \hat{\rho}_{\hat{m}} \right] \end{aligned}$$

because

$$-(1 - \hat{\rho}_{\hat{m}}) \log \frac{1 - \hat{\rho}_{\hat{m}}}{\hat{\rho}_{\hat{m}}} = H(\hat{\rho}_{\hat{m}}) + \log \hat{\rho}_{\hat{m}}.$$

Substituting this back in (2.24) yields the desired bound

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m - \hat{r}_* \leq \frac{O_\dagger}{\beta} = \frac{\psi(\#\mathcal{M})}{\beta}. \quad \square$$

The combined risk estimate on the left of (2.22) is bounded by the minimum of the individual risk estimates plus a price due to the mixing, which is a function of mixing weights $\hat{\rho}$. According to the rest of (2.22), if the distribution $\hat{\rho}$ is concentrated on mostly one model \hat{m} , then $H(\hat{\rho})$ is close to zero and the combined risk estimate is very close to the minimum \hat{r}_* . Moreover, if in the distribution $\hat{\rho}$ there are several, say J , values of m with nearly minimal values among \hat{r}_m , then accounting for those J values in the sum on the right side of (2.23), one has a further reduction of about $1/\beta \log J$, which aids in further quantifying the advantage of the mixture. In any case, since H is less than the log-cardinality, the combined risk estimate cannot exceed \hat{r}_* by more than a relatively small amount $1/\beta \log(\#\mathcal{M})$.

The second part of the lemma gives a better bound. A tight bound for ψ is obtained if we take K large in (2.19) when $\#\mathcal{M}$ is large. One numerical method to obtain this is by solving for the fixed point. Put $K_0 = \log(\#\mathcal{M} - 1) - 1$ and iterate

$$K_{\text{new}} = \log \frac{\#\mathcal{M} - 1}{K_{\text{old}}} - 1.$$

Now we have the corresponding version of the bound by taking expectation \mathbb{E}_θ with respect to the sampling distribution, X given θ .

[2.25] **COROLLARY.** *With $\rho = \rho(\beta) = \mathbb{E}_\theta \hat{\rho}$ and $\rho_* = \mathbb{E}_\theta \hat{\rho}_{\hat{m}}$, the expected value of the combined risk estimate*

$$\mathbb{E}_\theta \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m$$

can be bounded by any of the following

- (a) $r_* + 1/\beta [H(\rho) + \log \rho_*]$;
- (b) $1/\beta [H(\rho) - \log \sum_{m \in \mathcal{M}} \exp(-\beta r_m)]$;
- (c) $r_* + 1/\beta [(1 - \rho_*) \log(\#\mathcal{M} - 1) + H(\rho_*) + \log \rho_*]$;
- (d) $r_* + \psi(\#\mathcal{M})/\beta$.

Proof: Recall (2.16). For parts (a), (c) and (d), both H and the logarithm are concave so we can apply Jensen's inequality. Part (b) follows from mixing (2.23) with $\hat{\rho}$ and the concavity of the function $(\hat{r}_m)_{m \in \mathcal{M}} \mapsto -\log \sum_{m \in \mathcal{M}} \exp(-\beta \hat{r}_m)$, which is well-known but a proof is as follows.

Since $\beta > 0$, by dilation, it suffices to show that the function

$$(x_m)_{m \in \mathcal{M}} \mapsto \log \sum_{m \in \mathcal{M}} \exp(-x_m)$$

is convex. That is, for $\lambda \in (0, 1)$, $\bar{\lambda} = 1 - \lambda$, and $(x_m), (y_m) \in \mathbb{R}^{\#\mathcal{M}}$ we want to show

$$\log \sum_{m \in \mathcal{M}} e^{-(\lambda x_m + \bar{\lambda} y_m)} \leq \lambda \log \sum_{m \in \mathcal{M}} e^{-x_m} + \bar{\lambda} \log \sum_{m \in \mathcal{M}} e^{-y_m},$$

which is equivalent to showing

$$\sum_m e^{-(\lambda x_m + \bar{\lambda} y_m)} \leq \left[\sum_m e^{-x_m} \right]^\lambda \left[\sum_m e^{-y_m} \right]^{\bar{\lambda}}.$$

But this is a direct application of Hölder's inequality for

$$\begin{aligned} f : \mathcal{M} &\mapsto \mathbb{R}_+, & f(m) &= e^{-\lambda x_m} \\ g : \mathcal{M} &\mapsto \mathbb{R}_+, & g(m) &= e^{-\bar{\lambda} y_m} \end{aligned}$$

by identifying the above inequality with

$$\|fg\|_1 \leq \|f\|_{1/\lambda} \|g\|_{1/\bar{\lambda}}$$

where the norms are the L_p norms over the counting measure on \mathcal{M} . □

2.2.3 Risk Bound

Each part of Corollary 2.25 and Proposition 2.15 provides an immediate risk bound for $\hat{\theta}$. That is, for each $\beta > 0$, Corollary 2.25a, for instance, yields

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq r_* + \frac{1}{\beta} H(\rho) + \log \rho_* - (1 - 4\beta) \mathbb{E}_\theta \sum_{m \in \mathcal{M}} \hat{\rho}_m |\hat{\theta} - \hat{\theta}^m|^2.$$

But this is difficult to use because we cannot obtain the expectations of the quantities $\hat{\rho}$, $\min_m \hat{\rho}_m$, and $\sum_m \hat{\rho}_m |\hat{\theta} - \hat{\theta}^m|^2$ in closed form. Corollary 2.25d gives the most useful bound. We also restrict $\beta \leq 1/4$ to get rid of the variation term $(1 - 4\beta) \mathbb{E} \sum_m \hat{\rho}_m |\hat{\theta} - \hat{\theta}^m|^2$.

THEOREM (Least-Squares Mixture Risk Bound). *For the mixture of least-squares estimator $\hat{\theta}$, when $\beta \leq 1/4$ and for each $K > 0$, we have* [2.26]

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq r_* + \frac{\psi(\#\mathcal{M})}{\beta} \leq r_* + \frac{1}{\beta} \max \left\{ K, \log \frac{\#\mathcal{M} - 1}{K} - 1 \right\}. \quad \square$$

EXAMPLE (Leading-Term). For the leading-term models discussed in Section 1.2.1, [2.27] we have $\mathcal{M} = \{0, 1, \dots, d\}$, with each model m representing the ending coordinate of an initial set of regressors. Then using weights $\hat{\rho}$ [Definition 2.11] with $\beta = 1/4$ gives an unbiased estimate of the risk of $\hat{\theta}$,

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m = \hat{r}_* + 4[H(\hat{\rho}) + \log \hat{\rho}_{\hat{m}}].$$

In addition, since $\mathcal{M} = \{0, 1, \dots, d\}$ with $\#\mathcal{M} = d + 1$, this can be bounded by

$$\hat{r} < \hat{r}_* + 4[1 \vee (\log d - 1)].$$

The bound implied by (2.22)

$$H(\hat{\rho}) + \log \hat{\rho}_{\hat{m}} < \log d$$

is tight in the sense that there exists a sequence of $X = X^{(d)}$ such that

$$\lim_{d \rightarrow \infty} \frac{H(\hat{\rho}) + \log \hat{\rho}_{\hat{m}}}{\log d} \rightarrow 1.$$

Indeed, for each d , we pick a worst-case X . Let $k > 0$ and

$$|X_i|^2 = \begin{cases} k + 2 & \text{if } i = d \\ 2 & \text{if } i < d \end{cases}.$$

Then

$$\hat{r}_m = \begin{cases} 2d - d = d & \text{if } m = d \\ 2(d - 1) + (k + 2) - d = d + k & \text{if } m < d \end{cases}.$$

Then $\hat{r}_* = \hat{r}_d = d$ and $\hat{\rho}_m$ is roughly uniform over the first d models ($m < d$) with $H(\hat{\rho}) \sim \log d$ for d large. As the difference $k = \hat{r}_1 - \hat{r}_d$ between the risk estimates for the minimal model $m = d$ and other models grows large, the weights $\hat{\rho}$ concentrates more at $m = d$. We can put $k = d$ such that $\hat{\rho}_{\hat{m}}$ tends to 1 and $\log \hat{\rho}_{\hat{m}}$ tends to 0. But such a sequence of X is highly unlikely so that the risk bound with $H(\hat{\rho})$ replaced by $\log d$ is not tight.

Then we have

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq r_* + 4\psi(d + 1) \leq r_* + 4 \max\{1, \log d - 1\},$$

with a relatively small additive constant of $4\psi(d + 1)$, which is strictly smaller than d for dimension as small as 2, as tabulated below.

d	2	5	10	20	40	100
$4 \max\{1, \log d - 1\}$	4	4	5.2	8	10.8	14.4
$4\psi(d + 1)$	1.9	3.3	4.6	6.2	8.0	10.5

In our simulations (shown in the next subsection) using $\beta = 1/4$, the computed risk for various choices of θ suggests that the excess beyond r_* is usually less than $\log d$ rather than the upper bound of $4(\log d - 1)$.

This estimator is *not minimax* (for all $\theta \in \mathbb{R}^d$). The X above provides evidence for this because $\hat{r}_m > d$ for all $m < d$ and $\hat{r}_d = d$ such that for any $\beta \geq 1/4$,

$$\hat{r} \geq \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m > d.$$

This happens because the nested subsets favour keeping the leading coordinates to the detriment of the trailing ones. Since our mixture with $\beta = 1/2$ is the improper Bayes procedure examined in Hartigan (2002), his study shows such an estimator in the one-dimensional case with $\beta = 1/2$ is not minimax, i.e. its risk exceeds 1 for some range of θ .

In general, if the trailing observations are large, as X_d is in this toy example, the fact that it is left out under most of the leading-term models contribute to the overall high risk. \parallel

Remark: Minimaxy is a global attribute invariant to the ordering of the coefficients. The leading-term mixture fails to be minimax in part because the subsets under consideration are unbalanced (favouring initial coordinates). But if sufficient number of subsets considered are symmetric in the coordinates, then the mixture estimator would have a shot at achieving minimaxy. \triangleleft

EXAMPLE (Block Models). For the block models in Example 2.2, we have $\#\mathcal{M} = d(d + 1)/2 \leq d^2$. Thus, using $\beta = 1/4$ gives the inequality $\hat{r} < \hat{r}_* + 4 \log d^2$ and hence, the oracle inequality [2.28]

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq r_* + 8 \log d. \quad \parallel$$

EXAMPLE (All-Subset). For the all-subset case mentioned in Example 2.3, we have \mathcal{M} is the power set of $\{1, \dots, d\}$ such that $\#\mathcal{M} = 2^d$. Then the above theory for subset mixture estimators yields the risk bound [2.29]

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq r_* + 4d \log 2,$$

which has the undesirably large additive constant as a multiple of d because the number of subsets we consider is simply too large. Since the least-square estimator for θ has risk only d , it says that mixing all subsets in the simple fashion above will probably yield poor risk performance, as reflected by this bound. In the next chapter, we will provide a method for accounting the complexity of each subset model. This offers better control of the risk bound when mixing over many subsets. \parallel

2.2.4 Simulations

In this subsection, we will show simulations based on the leading-term models for $d = 20$. Besides the mixture estimator proposed, we will examine the AIC estimator in this **nested model** setting as well for comparison.

To learn something from the simulations, we impose structure on θ , and the function estimation context is chosen for such considerations. That is, the θ_i is considered as the coefficient of an orthonormal basis expansion of the function, or signal, to be estimated. Here we have simulation results divided into four scenarios for the underlying signal θ .

Constant One-Block The first case describes θ when it is indeed a low-pass signal.

Gradual Decay The second is probably quite typical in reality when θ_i^2 decays as the reciprocal of i .

Odd or Even Function The third scenario describes an odd (or even) function parameterized by the Fourier basis, so only every other coordinate of θ is non-zero.

Ramp-Up with Cut-off In the last case, θ_i^2 increases linearly in i but is cut off to zero after $i = 15$. It is somewhat academic, designed as a hard example which the convex combination of the leading-term models (monotone decreasing θ_i^2) cannot represent

well. The sharp artificial cut-off presents more difficulty: the monotone decreasing weights of our mixture estimator assigns to the coordinates cannot get down to zero quickly enough beyond the cut-off point. Indeed, since the signal is not significantly strong in the leading terms, the risk estimates for the models with extraneous terms beyond $i = 15$ may not be high and the weights assigned to these models ought not be negligible.

In all of the following plots, we have taken the variance in each dimension σ_n^2 to be 1. The least-square estimator has a risk of $d = 20$. Our bound is $4 \log d$ in excess beyond the risk target. But the simulations below show that the true cost for mixing is probably around $\log d \approx 3$.

In this set of simulation results, we have used $\beta = 1/2$ in our mixture estimator, because this corresponds to a Bayes procedure (see the Discussion chapter). But we have also tried $\beta = 1/4$ and $\beta = 1$ — the performance of the mixture estimator is not very sensitive to the choice of β in this region.

Note that we are using $|\theta|$ (without squaring) in all the x -axes to better illustrate the behaviour of the estimators in small θ settings. Also, be careful that the y -axes of the following plots do not have a common scale.

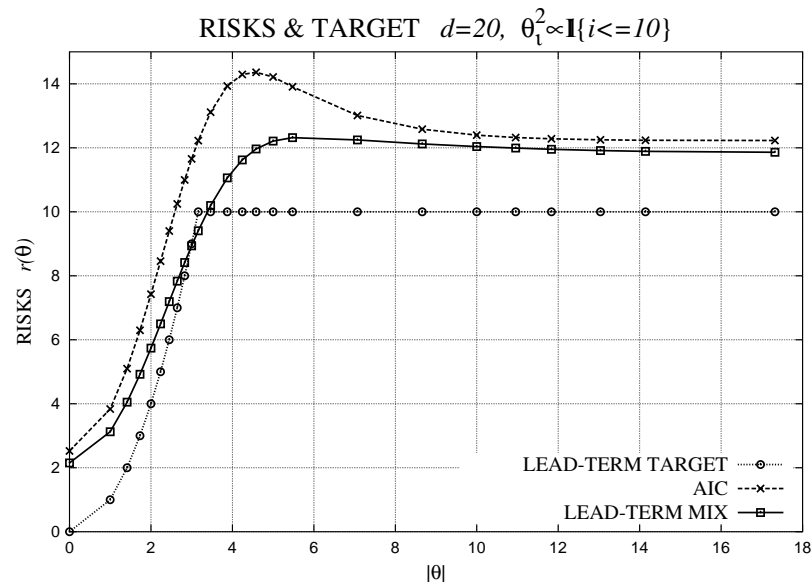


Figure 2.1: Risks and Target: Constant One-Block. $\theta_i^2 \propto \mathbb{I}_{\{i \leq 10\}}$

The leading-term risk target is

$$r_*(\theta) = |\theta|^2 \wedge 10.$$

First, it is clear that the target excludes the trailing ten coefficients, since they are in fact zero. So we only need to consider the first ten coordinates of θ_i^2 together since they are constant. If $|\theta|^2 < 10$, we are better off leaving them out since the bias so incurred is less than the variance of 10 if we include them. The kink of the target above at $|\theta|^2 = 10$ or $|\theta| \approx 3.16$ is due to this minimum operation.

The plot says that the our leading-term mixture estimator performs only about 2 worse than the risk target at small and large θ , but matches, and even beats the target around $|\theta|^2 = 10$. And the AIC leading-term selection estimator is uniformly worse than our mixture: but these two are close when $|\theta|^2$ is large. This is expected since the θ in this case is represented exactly by the model $m = 10$ such that AIC picks it correctly when the “signal-to-noise” ratio is high; the adaptive weights in our mixture give strong emphasis on the right model as well.

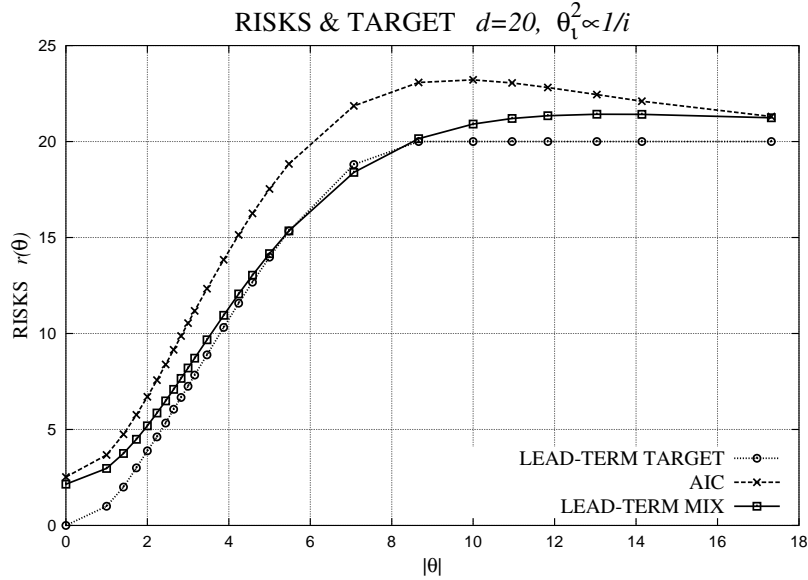


Figure 2.2: Risks and Target: Gradual Decay. $\theta_i^2 \propto 1/i$

For this scenario, an integral approximation shows that the tail sum

$$\sum_{i>m} \theta_i^2 \approx |\theta|^2 \left[1 - \frac{\log m}{\log d} \right]$$

and the minimum of r_m occurs at roughly $m \approx d \wedge [|\theta|^2 / \log d]$. Note that now not any one of the component estimators considered in our model class truly describes the underlying θ . Also, the risk target does increase up to d as the total $|\theta|^2$ increases.

The plot says that our leading-term mixture estimator tracks the risk target very well, a remarkable fact since the θ is not one of the leading-term type. In fact, for moderately sized θ , the mixture performs slightly better than the risk target. The AIC selection estimator, however, starts out being roughly the same as the mixture when $\theta = 0$, but ultimately performs worse uniformly by 2 or 3 as θ is non-zero.

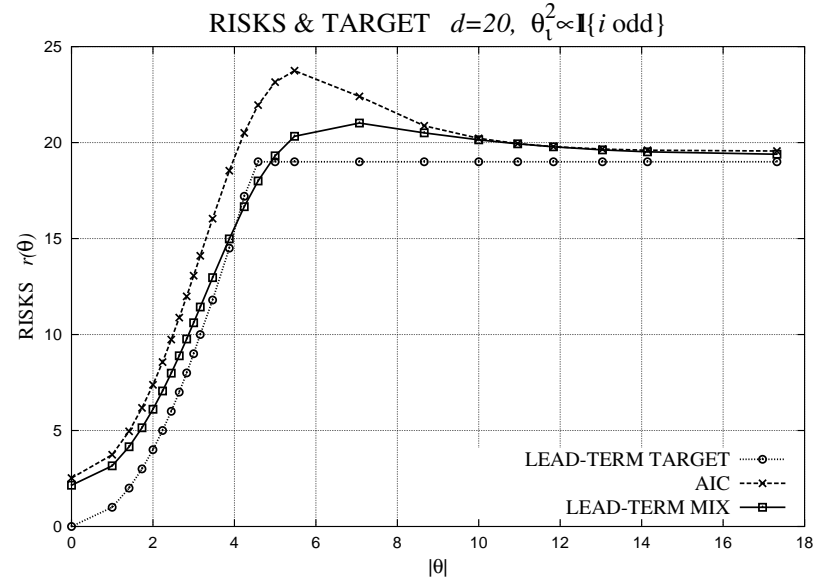


Figure 2.3: Risks and Target: Odd or Even Function. $\theta_i^2 \propto \mathbb{1}_{\{i \text{ odd}\}}$

The leading-term risk target for this case is

$$r_* = |\theta|^2 \wedge 19,$$

and it switches from the quadratic to constant at $|\theta| \approx 4.36$. Since the leading-term models cannot exclude individual coordinates in a leading block, this kind of θ essentially resembles a single constant block to the target, and hence the form of r_* analogous to that in the constant one-block case, with the only change of the saturation point of 19 instead of 10 for the cardinality of the block.

As before, the leading-term mixture has an overhead of 2 at $\theta = 0$ but then it matches the target risk at around the saturation point of $|\theta|^2 = 19$. Its risk creeps beyond $d = 20$ a little right after $|\theta|^2$ reaches 19 before it tapers down to almost the risk target (up to an offset of 0.5) as θ becomes large.

The behaviour of the AIC estimator starts off being roughly the same as that of our mixture when θ is small, but then it overshoots the target, by quite a bit more, at around $|\theta|^2 = 19$; then it asymptotes to the risk of the mixture again as θ becomes large.

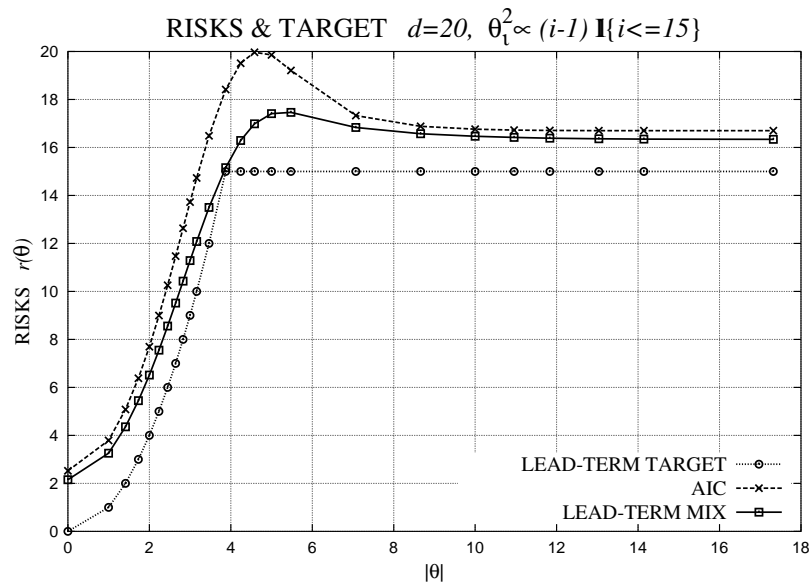


Figure 2.4: Risks and Target: Ramp-Up with Cut-off. $\theta_i^2 \propto (i - 1) \mathbb{I}\{i \leq 15\}$

The leading-term risk target for this case is

$$r_* = |\theta|^2 \wedge 15,$$

as it the ramp-up of θ_i^2 up to $i = 15$ acts like a single block to the models we consider.

Here, the mixture estimator's risk over-shoots the risk target by more than the previous case, probably due to the difficulty of modelling such a θ as described above, but it eventually tapers down almost to the target again, staying at about 2 above it.

The behaviour of the AIC estimator starts off being roughly the same as that of our mixture when θ is small, but then its risk overshoots the target, by quite a bit more than that of the mixture, probably due to the same difficulty; then it approaches the target from above again as θ becomes large, staying at about 2.5 above it.

2.3 Mixing Shrinkage Estimators

2.3.1 Risk Estimates for Positive-Part James-Stein

We now investigate the two-block shrinkage in Example 2.4 using the positive-part James-Stein estimator as a building block. Recall the basic one-block estimator on all d coefficients is

$$\hat{\theta}_{+S}(X) = \left(1 - \frac{d-2}{|X|^2}\right)_+ X = X - \gamma X, \quad \gamma = 1 \wedge \frac{d-2}{|X|^2},$$

where γ is the shrinkage coefficient. When the James-Stein estimator is involved, the constant $d-2$ appears excessively where d is the dimension of the shrinkage block. To avoid writing -2 *ad nauseam*, we therefore write

$$\tilde{d} \stackrel{\text{def}}{=} d - 2,$$

not just for d but possibly for other integers whenever it is clear.

The following result is a corollary of Theorem 2.8. It is framed in a lemma because the author has not seen it in the literature.

[2.30] **LEMMA.** *The positive-part James-Stein estimator has the following unbiased estimate of risk*

$$\hat{r}_{+S}(X) = \begin{cases} d - \frac{\tilde{d}^2}{|X|^2} & \text{if } |X|^2 > \tilde{d} \\ |X|^2 - d & \text{if } |X|^2 \leq \tilde{d} \end{cases}, \quad d = \dim X$$

Moreover, $|\hat{r}_{+S}| \leq d$ and \hat{r}_{+S} is strictly increasing in $|X|^2$. \square

Observe that \hat{r}_{+S} depends on X through $|X|^2$ only. It has a positive piece (for $|X|^2 \geq \tilde{d}$) and a negative piece ($|X|^2 < \tilde{d}$) separated by 4 at the discontinuity at $|X|^2 = \tilde{d}$. In practice, this discontinuity does not pose *any* problem in implementing the mixture scheme as before by forming weights according to [2.11]. But we want to analyse the mixture of shrinkage estimators with similar techniques, such as Stein's unbiased risk estimate, which requires that the weights be almost differentiable.

Thus, we want to replace \hat{r}_m by risk estimates continuous in X . For the one-block shrinkage estimator, this is denoted by \hat{r}_{+S}^c : we perturb \hat{r}_{+S} by ± 2 — pulling both the positive and the negative pieces toward 0. Note that \hat{r}_{+S}^c will no longer be unbiased.

[2.31] **DEFINITION** (Continuous risk estimate for positive James-Stein). *Let*

$$\hat{r}_{+S}^c(X) = \begin{cases} \tilde{d} - \frac{\tilde{d}^2}{|X|^2}, & \text{if } |X|^2 > \tilde{d} \\ |X|^2 - \tilde{d} & \text{if } |X|^2 \leq \tilde{d} \end{cases} = (|X|^2 - \tilde{d})\gamma,$$

where $\gamma = 1 \wedge \tilde{d}/|X|^2$ as defined before.

Note that this risk estimate \hat{r}_{+S}^c is continuous at $|X|^2 = \tilde{d}$ with value 0, and bounded $|\hat{r}_{+S}^c| \leq \tilde{d}$, and strictly increasing in $|X|^2$. In addition, $|\hat{r}_{+S}^c| < |\hat{r}_{+S}|$ for any X .

Note that both of these risk estimates are defined by cases, and the conditions under the two cases can be more succinctly expressed using the indicators $\{\gamma < 1\}$ and $\{\gamma = 1\}$,

signifying whether X is large and small, respectively. That is, “ X large” means “shrink some” and “ X small” means “shrink everything”.

$$\begin{aligned} \hat{r}_{+S} &= (|X|^2 - \tilde{d})\gamma + 2(-1)^{\{\gamma=1\}} \\ &= \hat{r}_{+S}^c + 2(-1)^{\{\gamma=1\}} \end{aligned} \quad (2.32)$$

It is now obvious that $|\hat{r}_{+S}^c - \hat{r}_{+S}| \leq 2$. Moreover, \hat{r}_{+S}^c is almost differentiable, with

$$\nabla \hat{r}_{+S}^c(X) = \begin{cases} \frac{\tilde{d}^2}{|X|^4} 2X & \text{if } |X|^2 > \tilde{d} \\ 2X & \text{if } |X|^2 \leq \tilde{d} \end{cases} = 2\gamma^2 X = 2\gamma(X - \hat{\theta}_{+S}). \quad (2.33)$$

Remark: There are many other ways to adjust the unbiased risk estimate to make it continuous. The only crucial fact about our continuous estimate is that it preserves the gradient of the original unbiased estimate. This is what we care about because it is used in our mixture estimator's risk estimate, and such a gradient has special properties quintessential to later analysis. Indeed, the unbiased estimate \hat{r}_{+S} is negative for $|X|^2 \leq \tilde{d}$, which means it always underestimates the risk for small X when the risk ought to be non-negative. We see that perhaps a more accurate risk estimate than ours would be $\hat{r}_{+S}^c + 2$, which moves the negative piece up by 4 and leaves the positive piece unchanged. However, since the sole purpose of this continuous adjustment to the risk estimate is for forming the model weights, we see that any constant offset will be immaterial after the normalization of the weights due to its exponentiating form. In other words, we are only interested in comparing the model's risk estimate relative to those of other models in \mathcal{M} .

Our continuous modification, however, facilitates the language for the analysis later somewhat because, just as this is the case with \hat{r}_{+S} , we can still say that $\hat{r}_{+S} - \hat{r}_{+S}^c$ is non-negative iff $|X|^2 > \tilde{d}$, and vice versa. \triangleleft

2.3.2 Two-Set Shrinkage Risk Estimates

Here we expound on the general setting used in Example 2.5 where we partition the coordinates into two sets and apply independent positive-part James-Stein estimators. Each $m \in \mathcal{M}$ is such that $\#m \in \{3, 4, \dots, d-3\}$.

Remark: Such a restriction on the cardinality of each m is merely to ensure that the positive-part James-Stein estimators be well-defined. However, this is unnecessary if the least-squares estimator (and its unbiased risk estimate) is used instead of the shrinkage estimator for a set with cardinality of 1 or 2. Our ultimate bounds for the unbiased estimate of risk and the risk of the mixture estimator still hold. The theory is developed here with the restriction for simplicity. \triangleleft

Recall the component estimator under model m is

$$\hat{\theta}^m = X - (a_i^m X_i)_{i \leq d}$$

where as in (2.6, 2.7),

$$a_i^m = a_i^m(X) = \gamma_m(X) \mathbb{1}_{\{i \in m\}} + \gamma_{m^c}(X) \mathbb{1}_{\{i \notin m\}},$$

with

$$\gamma_m(X) \stackrel{\text{def}}{=} 1 \wedge \frac{\#m - 2}{\sum_{i \in m} X_i^2}.$$

To unclutter the algebra by suppressing the coordinate index i , we also use matrix notation to write

$$\hat{\theta}^m = (I - A_m)X,$$

where

$$A_m = \text{diag}(a^m)$$

is a $d \times d$ diagonal matrix consisting of the elements of a^m . Write

$$X_{\in m} = \{X_i : i \in m\} \quad \text{and} \quad X_{\notin m} = \{X_i : i \notin m\}.$$

Then, because the shrinkage coefficients for the two sets

$$\gamma_m = \gamma_m(X_{\in m}) \quad \text{and} \quad \gamma_{m^c} = \gamma_{m^c}(X_{\notin m})$$

are independent, the unbiased estimate of risk for $\hat{\theta}^m$ is simply the sum of those for the shrinkage estimators on the individual sets,

$$\begin{aligned} \hat{r}_m &= \hat{r}_{+S}(X_{\in m}) + \hat{r}_{+S}(X_{\notin m}) \\ &= \left\{ \begin{array}{ll} \#m - \frac{(\#m - 2)^2}{|X_{\in m}|^2} & \text{if } \gamma_m < 1 \\ |X_{\in m}|^2 - \#m & \text{if } \gamma_m = 1 \end{array} \right\} + \left\{ \begin{array}{ll} \#m^c - \frac{(\#m^c - 2)^2}{|X_{\notin m}|^2} & \text{if } \gamma_{m^c} < 1 \\ |X_{\notin m}|^2 - \#m^c & \text{if } \gamma_{m^c} = 1 \end{array} \right\}. \end{aligned}$$

Analogous to [2.31], we define a continuous (but biased) risk estimate for the estimator $\hat{\theta}^m$.

$$\begin{aligned} \hat{r}_m^c &= \hat{r}_{+S}^c(X_{\in m}) + \hat{r}_{+S}^c(X_{\notin m}) \\ &= (|X_{\in m}|^2 - [\#m - 2])\gamma_m + (|X_{\notin m}|^2 - [\#m^c - 2])\gamma_{m^c}. \end{aligned} \quad (2.34)$$

Using (2.32), we rewrite the unbiased risk estimate as

$$\hat{r}_m = \hat{r}_m^c + 2(-1)^{\{\gamma_m=1\}} + 2(-1)^{\{\gamma_{m^c}=1\}}.$$

Then it is clear that

$$\sup_{m \in \mathcal{M}} |\hat{r}_m - \hat{r}_m^c| \leq 4. \quad (2.35)$$

In fact, the adjustment to the unbiased risk estimates can only take on values 0 and ± 4 . In particular,

(a) $\hat{r}_m - \hat{r}_m^c = 4$ for some m implies the positivity of both \hat{r}_m and \hat{r}_m^c (equivalently, $\gamma_m < 1$ and $\gamma_{m^c} < 1$); and $|X|^2 > d - 4$.

(b) Likewise, $\hat{r}_m - \hat{r}_m^c = -4$ for some m implies the negativity of both \hat{r}_m and \hat{r}_m^c (that is, $\gamma_m = 1 = \gamma_{m^c}$); and $|X|^2 \leq d - 4$.

And these two cases are mutually exclusive.

2.3.3 Risk Estimate for Mixture

The time is now ripe for a little abstraction that was unnecessary when the method and analysis of mixing least-squares estimators were presented. Let our weights for mixing estimators be of the form

$$\hat{w}_m = \hat{w}_m(X) = \frac{\exp(-\hat{\ell}_m)}{\sum_{k \in \mathcal{M}} \exp(-\hat{\ell}_k)}, \quad m \in \mathcal{M} \quad (2.36)$$

where $\hat{\ell}_m = \hat{\ell}_m(X)$ is almost differentiable in X for each m . Even though they are not needed for forming the mixture estimator, we find that our analysis is facilitated by the old form of weights using the unbiased risk estimates,

$$\hat{\rho}_m = \frac{\exp(-\beta \hat{r}_m)}{\sum_{k \in \mathcal{M}} \exp(-\beta \hat{r}_k)}, \quad \beta > 0, \quad m \in \mathcal{M}$$

because they are tied in the same fashion to the minimum of the risk estimates, and in turn, to the risk target (after taking expectation).

DEFINITION. *The discrete **Kullback-Leibler divergence** between the probability p and the non-negative vector q , both indexed by the same space \mathcal{M} , is* [2.37]

$$D(p||q) = \sum_{m \in \mathcal{M}} p_m \log \frac{p_m}{q_m}. \quad \triangleleft$$

It is well-known that D is convex in the pair (p, q) (See Cover and Thomas, 1991, Chapter 2), and clear that D is monotonically decreasing in q . It is easy to show that $D(p||q) \geq -\log \sum_m q_m$. If q is a **sub-probability**, i.e. $\sum_m q_m \leq 1$, then $D \geq 0$, with equality iff $p_m = q_m$ for each m (which implies that q is a probability).

LEMMA. *The mixture estimator $\hat{\theta} = \sum_{m \in \mathcal{M}} \hat{w}_m \hat{\theta}^m$ (with \hat{w} defined in (2.36)) has the following unbiased estimate of risk,* [2.38]

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{w}_m \left[\hat{r}_m - |\hat{\theta} - \hat{\theta}^m|^2 + 2\nabla \hat{\ell}_m \bullet (\hat{\theta} - \hat{\theta}^m) \right]. \quad (2.39)$$

Write $\hat{r}_{\hat{m}} = \min_m \hat{r}_m = \hat{r}_*$. Then the first term above can be written as

$$\sum_{m \in \mathcal{M}} \hat{w}_m \hat{r}_m = \hat{r}_* + \frac{1}{\beta} [D(\hat{w}||\hat{\rho}) + H(\hat{w}) + \log \hat{\rho}_{\hat{m}}]$$

In particular, with ψ defined in Definition 2.18, the following inequality holds.

$$\sum_{m \in \mathcal{M}} \hat{w}_m \hat{r}_m < \hat{r}_* + \frac{1}{\beta} [D(\hat{w}||\hat{\rho}) + \psi(\#\mathcal{M}) + \log \frac{\hat{\rho}_{\hat{m}}}{\hat{w}_{\hat{m}}}] \quad (2.40)$$

Proof: The first expression for the risk estimate can be obtained in the same fashion we

obtained (2.12). In a line similar to the least-squares case in Lemma 2.20a,

$$\begin{aligned}\hat{r}_m &= \frac{1}{\beta} \left[\log \frac{1}{\hat{\rho}_m} - \log \sum_m \exp(-\beta \hat{r}_m) \right] \\ &= \frac{1}{\beta} \left[\log \frac{\hat{w}_m}{\hat{\rho}_m} + \log \frac{1}{\hat{w}_m} - \log \sum_m \exp(-\beta \hat{r}_m) \right] \\ &= \hat{r}_* + \frac{1}{\beta} \left[\log \frac{\hat{w}_m}{\hat{\rho}_m} + \log \frac{1}{\hat{w}_m} + \log \hat{\rho}_{\hat{m}} \right]\end{aligned}\quad (2.41)$$

Now take average with respect to the weights \hat{w} to obtain the second statement. The inequality follows from the proof of Lemma 2.20b. \square

[2.42] DEFINITION. For $\beta > 0$, the almost differentiable weights for mixing shrinkage estimators are

$$\hat{\rho}_m^c = \hat{\rho}_m^c(\beta) = \frac{\exp(-\beta \hat{r}_m^c)}{\sum_{k \in \mathcal{M}} \exp(-\beta \hat{r}_k^c)}, \quad m \in \mathcal{M}. \quad \triangleleft$$

Thus, the mixture estimator is formed:

$$\hat{\theta} = \sum_{m \in \mathcal{M}} \hat{\rho}_m^c \hat{\theta}^m.$$

We proceed to bound the risk estimate using Lemma 2.38. But first we will have a technical result bounding the Kullback-Leibler divergence.

[2.43] LEMMA. For each $m \in \mathcal{M}$, let

$$p_m = \exp(-\ell_m^p) / \sum_{k \in \mathcal{M}} \exp(-\ell_k^p)$$

and

$$q_m = \exp(-\ell_m^q) / \sum_{k \in \mathcal{M}} \exp(-\ell_k^q).$$

Suppose

$$\sup_{m \in \mathcal{M}} |\ell_m^p - \ell_m^q| \leq K,$$

then $D(p \| q) \leq K^2/2$. Equality holds if (a) $K = 0$, or (b) $\ell_m^p - \ell_m^q = \text{constant}$ for all $m \in \mathcal{M}$.

Proof: We will prove a weaker version of the lemma for the purposes here only. The general proof is similar. That is, we $D(\hat{\rho}^c \| \hat{\rho}) \leq 8\beta^2$ holds for $\beta \geq 0$.

First, if $\beta = 0$, then it is trivial that $D(\hat{\rho}^c \| \hat{\rho}) = 0$ since both $\hat{\rho}^c$ and $\hat{\rho}$ become the uniform distribution on \mathcal{M} . Assume that $\beta > 0$. In light of (2.35) and the remark below it, we prove this by cases. Assume that $|X|^2 > d - 4$ such that $\hat{r}_m - \hat{r}_m^c \in \{0, 4\}$ for each m . Let

$$\mathcal{M}^+ = \{m \in \mathcal{M} : \hat{r}_m - \hat{r}_m^c = 4\}$$

Then,

$$\begin{aligned}\log \frac{\hat{\rho}_m^c}{\hat{\rho}_m} &= \beta(\hat{r}_m - \hat{r}_m^c) + \log \frac{\sum_{m \in \mathcal{M}} \exp(-\beta \hat{r}_m)}{\sum_{m \in \mathcal{M}} \exp(-\beta \hat{r}_m^c)} \\ &= 4\beta \mathbb{1}_{\{m \in \mathcal{M}^+\}} + \log \frac{\sum_{m \in \mathcal{M}} \exp(-\beta \hat{r}_m^c) e^{-4\beta \mathbb{1}_{\{m \in \mathcal{M}^+\}}}}{\sum_{m \in \mathcal{M}} \exp(-\beta \hat{r}_m^c)} \\ &= 4\beta \mathbb{1}_{\{m \in \mathcal{M}^+\}} + \log \left(1 + (e^{-4\beta} - 1) \sum_{m \in \mathcal{M}^+} \hat{\rho}_m^c \right)\end{aligned}$$

Denote

$$\hat{\rho}_+^c \stackrel{\text{def}}{=} \sum_{m \in \mathcal{M}^+} \hat{\rho}_m^c \leq 1,$$

and observe that if either $\mathcal{M}^+ = \emptyset$ or $\mathcal{M}^+ = \mathcal{M}$ (such that the constant offset of 4 uniformly over \mathcal{M} gets normalized out), then $\hat{\rho}^c = \hat{\rho}$ and $D = 0$. So, we may assume that \mathcal{M}^+ is a proper subset of \mathcal{M} so that $0 < \hat{\rho}_+^c < 1$. Sum the above expression over \mathcal{M} with weights $\hat{\rho}_m^c$ to obtain

$$\begin{aligned}D(\hat{\rho}^c \| \hat{\rho}) &= 4\beta \hat{\rho}_+^c + \log(1 + (e^{-4\beta} - 1)\hat{\rho}_+^c) \\ &\leq 4\beta \hat{\rho}_+^c + (e^{-4\beta} - 1)\hat{\rho}_+^c \\ &= \hat{\rho}_+^c (4\beta + e^{-4\beta} - 1) \\ &< \hat{\rho}_+^c (4\beta)^2 / 2 \\ &< 8\beta^2.\end{aligned}$$

The second case $|X|^2 \leq d - 4$ is similar, except that we work with the definitions

$$\mathcal{M}^o = \{m \in \mathcal{M} : \hat{r}_m - \hat{r}_m^c = 0\} \quad \text{and} \quad \hat{\rho}_o^c \stackrel{\text{def}}{=} \sum_{m \in \mathcal{M}^o} \hat{r}_m^c \leq 1,$$

which play the roles of \mathcal{M}^+ and $\hat{\rho}_+^c$ in the proof respectively. \square

COROLLARY. The first term of the unbiased risk estimate satisfies

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m^c \hat{r}_m \leq \hat{r}_* + 8\beta + \frac{1}{\beta} [H(\hat{\rho}^c) + \log \hat{\rho}_{\hat{m}}] \leq \hat{r}_* + 8\beta + 4 + \frac{\psi(\#\mathcal{M})}{\beta}. \quad [2.44]$$

Furthermore, with $\rho^c = \mathbb{E}_\theta \hat{\rho}^c$ and $\rho_* = \mathbb{E}_\theta \hat{\rho}_{\hat{m}}$ and $\rho_*^c = \mathbb{E}_\theta \hat{\rho}_m^c$,

$$\mathbb{E}_\theta \sum_{m \in \mathcal{M}} \hat{\rho}_m^c \hat{r}_m$$

can be upper-bounded by any of the following.

- (a) $r_* + 8\beta + 1/\beta [H(\rho^c) + \log \rho_*]$;
- (b) $8\beta + 1/\beta [H(\rho^c) - \log \sum_{m \in \mathcal{M}} \exp(-\beta r_m)]$;
- (c) $r_* + 8\beta + 4 + 1/\beta [(1 - \rho_*^c) \log(\#\mathcal{M} - 1) + H(\rho_*^c)]$;
- (d) $r_* + 8\beta + 4 + \psi(\#\mathcal{M})/\beta$.

Proof: Put $\hat{w} = \hat{\rho}^c$ in (2.40). It is clear that

$$\log \frac{\hat{\rho}_m^c}{\hat{\rho}_m^c} \leq 4\beta.$$

Also, $D(\hat{\rho}^c \parallel \rho^c) \leq 8\beta^2$ from the previous lemma. Combining these bounds yields the results for parts (a), (c) and (d). Part (b) follows similarly from mixing (2.41) with $\hat{\rho}^c$. \square

The analysis of its unbiased estimate of risk, according to the above lemma, requires calculating

$$\nabla \hat{r}_m^c(X) = 2A_m^2 X = 2A_m(X - \hat{\theta}^m),$$

[Cf. (2.33)]. Compare this with the $\nabla \hat{r}_m$ for least-squares (2.14) – there is an extra A_m factor here. This turns out to cost us a factor of two in the upper-bound for the risk estimate.

[2.45] **PROPOSITION.** *The two-set shrinkage mixture estimator has an unbiased estimate of risk upper-bounded by*

$$\hat{r} \leq \sum_{m \in \mathcal{M}} \hat{\rho}_m^c \left[\hat{r}_m - (1 - 8\beta) |\hat{\theta} - \hat{\theta}^m|^2 \right].$$

Proof: From (2.39), it suffices to show

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m^c \nabla \hat{r}_m^c \cdot (\hat{\theta} - \hat{\theta}^m) \leq 2(1 + \max_j a_j) \sum_m \hat{\rho}_m^c |\hat{\theta} - \hat{\theta}^m|^2, \quad (2.46)$$

where $2(1 + \max_j a_j) \leq 4$. Since

$$\begin{aligned} 1/2 \nabla \hat{r}_m^c &= A_m X - A_m \hat{\theta}^m \\ &= X - \hat{\theta}^m - A_m \hat{\theta}^m, \end{aligned}$$

and since $X - \hat{\theta}$ is not a function of m , we apply Lemma 2.10 to obtain

$$\begin{aligned} \sum_{m \in \mathcal{M}} \hat{\rho}_m^c \nabla \hat{r}_m^c \cdot (\hat{\theta} - \hat{\theta}^m) &= 2 \sum_{m \in \mathcal{M}} \hat{\rho}_m^c (\hat{\theta} - \hat{\theta}^m - A_m \hat{\theta}^m) \cdot (\hat{\theta} - \hat{\theta}^m) \\ &= 2 \left[\sum_m \hat{\rho}_m^c |\hat{\theta} - \hat{\theta}^m|^2 - \sum_m \hat{\rho}_m^c A_m \hat{\theta}^m \cdot (\hat{\theta} - \hat{\theta}^m) \right] \end{aligned}$$

Thus, we want to bound the second quantity in brackets on the right side. $-\sum_m \hat{\rho}_m^c (A_m \hat{\theta}^m) \cdot (\hat{\theta} - \hat{\theta}^m)$. Let

$$A = \sum_m \hat{\rho}_m^c A_m, \quad a_i = \sum_m \hat{\rho}_m^c a_i^m$$

i.e. A a diagonal matrix with the mixed shrinkage weights $\{a_i, 1 \leq i \leq d\}$ on the diagonal, and again a^m is the shrinkage weights under model m . Then the left side of (*) becomes

$$\begin{aligned} - \sum_m \hat{\rho}_m^c [A_m(I - A_m)X] \cdot [(A_m - A)X] &= \sum_m \hat{\rho}_m^c \sum_{i=1}^d a_i^m (1 - a_i^m) X_i^2 (a_i - a_i^m) \\ &= \sum_i X_i^2 \sum_m \hat{\rho}_m^c a_i^m (1 - a_i^m) (a_i - a_i^m) \end{aligned}$$

We now apply the following result to show

$$\sum_m \hat{\rho}_m^c a_i^m (1 - a_i^m) (a_i - a_i^m) \leq a_i \sum_m \hat{\rho}_m^c (a_i - a_i^m)^2. \quad (2.47)$$

LEMMA. *Let Y be a real random variable with finite mean μ . Then*

(a) *for any real number λ , we have*

$$\text{var } Y = \mathbb{E}[(\lambda - Y)(\mu - Y)].$$

(b) *If Y is a non-negative random variable with $Y \leq \lambda$, then*

$$\mathbb{E}[Y(\lambda - Y)(\mu - Y)] \leq \mu \text{var } Y.$$

Proof: (a) Take expectation of $(\lambda - Y)(\mu - Y) = (\mu - Y)^2 + (\lambda - \mu)(\mu - Y)$.

(b) We may assume that Y has finite variance for the inequality holds trivially otherwise. Since $\lambda - Y \geq 0$, we have

$$\begin{aligned} Y(\lambda - Y)(\mu - Y)\{\mu \geq Y\} &\leq \mu(\lambda - Y)(\mu - Y)\{\mu \geq Y\} \\ Y(\lambda - Y)(\mu - Y)\{\mu < Y\} &< \mu(\lambda - Y)(\mu - Y)\{\mu < Y\} \end{aligned}$$

Combining the two inequalities yields the following, which we can integrate and apply part (a) to obtain the result.

$$Y(\lambda - Y)(\mu - Y) \leq \mu(\lambda - Y)(\mu - Y) \quad \square$$

We can now continue by identifying mixing with $\hat{\rho}^c$ over \mathcal{M} as an expectation. Apply the lemma with $\lambda = 1$, $Y = a_i^m$, $\mu = a_i$,

$$\begin{aligned} \sum_i X_i^2 \sum_m \hat{\rho}_m^c a_i^m (1 - a_i^m) (a_i - a_i^m) &\leq \sum_i X_i^2 a_i \sum_m \hat{\rho}_m^c (a_i - a_i^m)^2 \\ &\leq \sum_i X_i^2 (\max_j a_j) \sum_m \hat{\rho}_m^c (a_i - a_i^m)^2 \\ &= (\max_j a_j) \sum_m \hat{\rho}_m^c \sum_i X_i^2 (a_i - a_i^m)^2 \\ &= (\max_j a_j) \sum_m \hat{\rho}_m^c |\hat{\theta} - \hat{\theta}^m|^2 \end{aligned}$$

by noting that both $a_i^m, a_i \in [0, 1]$ by definition. Finally, combine with the above to give (2.46). Then the proposition follows immediately. \square

Note that the factor of 4 in Proposition 2.15 has turned to 8 here, and the equality is now demoted to an inequality. But this would still permit us to upper-bound the risk below. Here is an interesting by-product.

COROLLARY. *The two-set shrinkage mixture estimator with each $\beta \leq 1/8$ is minimax.* [2.48]

Proof: An unbiased estimate of risk for $\hat{\theta}$ satisfies

$$\hat{r} \leq \sum_{m \in \mathcal{M}} \hat{\rho}_m^c \hat{r}_m \leq d. \quad \square$$

Remark: The mixture estimator is probably minimax for a larger β as well because all the component estimators being mixed are minimax. The weights being dependent on the data presents some difficulty in the proof. The above is a simple way around it. \triangleleft

2.3.4 Risk Bound

We now have a new risk target

$$r_* = r_*^{+S}(\mathcal{M}) = \min_{m \in \mathcal{M}} r_m^{+S} = \min_{m \in \mathcal{M}} \mathbb{E} |\hat{\theta}^m - \theta|^2$$

against which we gauge the performance of our mixture estimator. We restate (2.16) here to display the relations between the risk-related quantities in the new two-set shrinkage case.

$$\mathbb{E}_\theta \hat{r}_* = \mathbb{E}_\theta \min_m \hat{r}_m \leq \min_m \mathbb{E}_\theta \hat{r}_m = \min_m r_m = r_*.$$

[2.49] **THEOREM (Two-Set Shrinkage Mixture Risk Bound).** *For the mixture of two-set shrinkage estimator $\hat{\theta}$, when $\beta \leq 1/8$, we have*

$$\hat{r} < \hat{r}_* + 8\beta + 4 + \psi(\#\mathcal{M})/\beta$$

and hence,

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq r_* + 8\beta + 4 + \psi(\#\mathcal{M})/\beta,$$

where for each $K > 0$,

$$\psi(\#\mathcal{M}) \leq \max \left\{ K, \log \frac{\#\mathcal{M} - 1}{K} - 1 \right\}. \quad \square$$

Observe that the additive penalty beyond r_*^{+S} is larger than that in the least-squares case. But the new risk target r_*^{+S} is usually lower than that for the former – the only exception is that there is a shrinkage overhead of about one per set of coordinates when the true parameter θ within and without the set m are *both small*. Indeed, the corresponding risk target for the subset models can get down all the way to 0 but the shrinkage risk target cannot.

[2.50] **EXAMPLE (Two-Block Shrinkage).** Recall Example 2.4 in which the models are written as $\mathcal{M} = \{3, \dots, d-3\}$. We use the convenient notation below to write the coefficient for first block, ending at coordinate $m \in \mathcal{M}$, and that for the second block as

$$X_{\leq m} = (X_1, \dots, X_m) \quad \text{and} \quad X_{> m} = (X_{m+1}, \dots, X_d), \quad (2.51)$$

respectively. Then the estimator under model m is

$$\hat{\theta}_i^m(X) = \begin{cases} X_i - \gamma(X_{\leq m})X_i & \text{if } i \leq m \\ X_i - \gamma(X_{> m})X_i & \text{if } i > m \end{cases}.$$

The continuous risk estimate for this is

$$\hat{r}_m^c = (|X_{\leq m}|^2 - [m-2])\gamma(X_{\leq m}) + (|X_{> m}|^2 - [d-m-2])\gamma(X_{> m}).$$

Mixing these estimators with weights $\hat{\rho}_m^c \propto \exp(-1/8 \hat{r}_m^c)$ gives an unbiased risk estimate that can be bounded by

$$\hat{r} \leq \sum_{m \in \mathcal{M}} \hat{\rho}_m^c \hat{r}_m \leq \hat{r}_* + 5 + 8\psi(\#\mathcal{M}) \leq \hat{r}_* + 1 + \log(\#\mathcal{M}),$$

where $\hat{r}_* = \min_m \hat{r}_m$. In addition, if $d \geq 13$, then $\#\mathcal{M} \geq 9$ and the first bound above can be loosened to

$$\hat{r} \leq \hat{r}_* - 3 + 8 \log(d-6).$$

The risk can be bounded by taking expectation,

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq r_* - 3 + 8 \log(d-6).$$

The bound is useful for moderate dimensionality, say $d \geq 20$ or so, because the excess beyond the target is less than d , the risk of the least-square procedure. For instance,

d	7	10	20	40	100	200
$5 + 8[1 \vee \log(d-6) - 1]$	13	13	18.1	25.2	33.3	39.1
$5 + 8\psi(\#\mathcal{M})$	7.2	10.7	15.7	20.1	25.7	30.0

However, in simulation, the performance of our mixture estimator is very good, and the excess beyond the risk target is merely about $\log d$. In any case, our oracle inequality has a smaller additive constant than those obtained in the literature; and we stay clear from any multiplicative constant. \parallel

One implication of this two-set shrinkage risk bound is that it encourages us to mix across many models, because we know that the “penalty” enters into the risk at by most its logarithm. Thus, in a canonical model with n data points such that $\sigma_n^2 = 1/n$, this bound suggests that we can mix across a number of models sub-geometric in n . This is useful when θ_i^2 tapers very slowly in i . In the function estimation setting, this means that any good representation of the signal requires many basis functions.

Before we discuss simulation results for the two-set shrinkage mixture estimators, we want to emphasize that the risk target here is often much lower than that used in the subset regression examples (although there is an overhead of about 2.5 when θ is small because each application of the positive-part James-Stein estimator contributes about 1.25 in risk even when the underlying θ is 0. We shall now segue into an interlude about our new risk target r_*^{+S} , by analysing related risk quantities.

2.3.5 Risk Targets: A Comparison with Ideal Linear Estimation

In this section, we explore the fact that shrinkage estimation using the (regular) James-Stein estimator is close to ideal linear estimation where we estimate θ using a linear function of the observation X with the coefficient between 0 and 1. Not surprisingly, the risk of the regular James-Stein estimator is extremely similar that of the positive-part James-Stein (estimator). Therefore, one can characterize the latter risk by the (unachievable) risk of the ideal linear estimator.

We first examine the risk of the original one-block James-Stein estimator

$$\hat{\theta}_S = \left(1 - \frac{\tilde{d}}{|X|^2}\right)X.$$

(Recall our notation $\tilde{d} = d - 2$.) It is well-known (e.g. Lehmann and Casella, 1998, Chapter 4) that the regular James-Stein estimator has the following unbiased estimate of risk

$$\hat{r}_S = d - \frac{\tilde{d}^2}{|X|^2}. \quad (2.52)$$

[2.53] DEFINITION. For $L > 0$, let

$$g_L(\phi) = e^{-\phi/2} \sum_{k=0}^{\infty} \frac{(\phi/2)^k}{k!} \frac{L}{L+2k}, \quad \phi \geq 0$$

with $g_L(0) = 1$ and $\lim_{\phi \rightarrow \infty} g_L(\phi) = 0$ for each L .

[2.54] PROPOSITION. The risk of the James-Stein estimator has this closed form expression.

$$\mathbb{E} |\theta - \hat{\theta}_S|^2 = d - \tilde{d} g_{\tilde{d}}(|\theta|^2).$$

Furthermore, it satisfies

$$2 \leq \mathbb{E} |\theta - \hat{\theta}_S|^2 \leq d - u_d(|\theta|^2)$$

with

$$u_d(\phi) = \tilde{d} \left(e^{-\phi/d} \vee \frac{\tilde{d}}{\tilde{d} + \phi} \right),$$

where the minimum $\mathbb{E}_0 |\hat{\theta}_S|^2 = 2$ is achieved at $\theta = 0$. It is a concave increasing function in $|\theta|^2$, and

$$\lim_{|\theta|^2 \rightarrow \infty} \mathbb{E} |\theta - \hat{\theta}_S|^2 = d.$$

Proof: It is well-known that the non-central chi-square random quantity $|Z|^2$ can be decomposed into a Poisson mixture of central chi-square random variables. (See for example, Lehmann (1986), p. 428.) That is, $|X|^2 \sim \chi_d^2(\phi)$, where $\phi = |\theta|^2$ is the non-centrality parameter, and $|X|^2$ has the Lebesgue density

$$e^{-\phi/2} \sum_{i=0}^{\infty} \frac{(\phi/2)^i}{i!} f_{d+2i}(\cdot), \quad \phi = |\theta|^2$$

where f_k is the Lebesgue density of a (central) χ_k^2 random variable. Hence,

$$\begin{aligned} \mathbb{E}_{\theta} \frac{1}{|X|^2} &= \mathbb{E}_{K \sim \text{Poi}(\phi/2)} \mathbb{E}_{V_K \sim \chi_{d+2K}^2} \left[\frac{1}{V_K} \mid K \right], \quad \phi = |\theta|^2 \\ &= \sum_{k=0}^{\infty} e^{-\phi/2} \frac{(\phi/2)^k}{k!} \frac{1}{d-2+2k} \\ &= g_{\tilde{d}}(\phi) / \tilde{d}. \end{aligned}$$

Then the risk expression follows from (2.52) and its upper- and lower-bounds follow from the following technical lemma about the function g_L . \square

[2.55] LEMMA. For each $L > 0$, the non-negative function $g_L : \mathbb{R}^+ \mapsto \mathbb{R}^+$

(a) satisfies

$$e^{-\phi/L} \leq g_{L-2}(\phi) \leq \frac{L}{L+\phi} \leq g_L(\phi) \leq 1;$$

(b) is convex and decreasing in ϕ , with maximum at $g_L(0) = 1$. Also $g_L'(0) = -\frac{1}{L+2}$.

Proof: (a) The upper bound 1 is obtained by dropping the k in the summation. For the exponential lower bound, it suffices to show that

$$\sum_{k=0}^{\infty} \frac{(\phi/2)^k}{k!} \frac{L}{L+2k} \geq e^{\phi/2} \exp\left(-\frac{\phi}{L+2}\right) = \exp\left(\frac{\phi}{2} \frac{L}{L+2}\right).$$

But this is immediate from the Taylor expansion of the right hand side. Namely, for each $L > 0$,

$$\left(\frac{L}{L+2}\right)^k = \frac{L^k}{L^k + \binom{k}{1} 2L^{k-1} + \dots} \leq \frac{L^k}{L^k + 2kL^{k-1}} = \frac{L}{L+2k} \quad \text{for all } k = 0, 1, \dots$$

To show $g_{L-2}(\phi) \leq \frac{L}{L+\phi} \leq g_L(\phi)$, we compare $g_L(\phi) \frac{L+\phi}{L}$ and $g_L(\phi) \frac{L+2+\phi}{L+2}$ with 1 for $L > 0$. That is, let $\alpha = 0, 2$,

$$\begin{aligned} g_L(\phi) \frac{L+\alpha+\phi}{L+\alpha} &= e^{-\phi/2} \sum_{k=0}^{\infty} \frac{(\phi/2)^k}{k!} \frac{L}{L+2k} \left(1 + \frac{\phi}{L+\alpha}\right) \\ &= e^{-\phi/2} \sum_{k=0}^{\infty} \frac{(\phi/2)^k}{k!} \left(\frac{L}{L+2k} + \frac{2k}{L+\alpha} \frac{L}{L+2(k-1)}\right) \end{aligned}$$

and it suffices to observe that the parenthesized coefficients on the right,

$$\alpha = 0 : \quad \frac{L}{L+2k} + \frac{2k}{L+2(k-1)} \geq \frac{L}{L+2k} + \frac{2k}{L+2k} = 1,$$

$$\alpha = 2 : \quad \frac{L}{L+2k} + \frac{2k}{L+2} \frac{L}{L+2(k-1)} = \frac{L}{L+2k} + \frac{2k}{L+2k+4(k-1)\{k \geq 1\}/L} \leq 1.$$

Thus, $g_{L-2}(\phi) \leq \frac{L}{L+\phi} \leq g_L(\phi)$ for $L > 2$.

(b) The following shows that $g_L' < 0$ and $g_L'' > 0$.

$$\begin{aligned} g_L'(\phi) &= -e^{-\phi/2} \sum_{k=0}^{\infty} \frac{(\phi/2)^k}{k!} \frac{L}{(L+2k)(L+2k+2)} \\ g_L''(\phi) &= 2e^{-\phi/2} \sum_{k=0}^{\infty} \frac{(\phi/2)^k}{k!} \frac{L}{(L+2k)(L+2k+2)(L+2k+4)}. \quad \square \end{aligned}$$

Remark: See Appendix 2 for more interesting facts about the g_L function. \triangleleft

Pinsker (1980) and Johnstone (1998) examined the idealized best linear “estimator” of the form cX , for a deterministic $c = c(\theta)$ which minimizes the risk $\mathbb{E}_{\theta} |cX - \theta|^2$:

$$c = c_d = \frac{|\theta|^2}{d + |\theta|^2}$$

So $0 \leq c_d(\theta) < 1$ and this estimator could be termed the idealized linear shrinkage estimator, as this is neither a statistic nor a *bona fide* estimator since it requires the knowledge of the unknown θ . The minimum risk is the harmonic mean of the squared norm of θ and the total variance:

$$\mathbb{E} |\theta - \hat{\theta}_{\text{il}}|^2 = \frac{d|\theta|^2}{d + |\theta|^2} = c_d d.$$

Observe that the idealized risk has value 0 when $\theta = 0$ and asymptotes to d as $|\theta|^2 \rightarrow \infty$. It is bounded by

$$\mathbb{E} |\theta - \hat{\theta}_{\text{il}}|^2 \leq d \wedge |\theta|^2.$$

Moreover, it is very close to the risk of the James-Stein estimator.

[2.56] **PROPOSITION.** *The risks of the James-Stein estimator $\hat{\theta}_S$ and the idealized linear shrinkage estimator $\hat{\theta}_{\text{il}}$ are tightly coupled.*

$$\mathbb{E} |\theta - \hat{\theta}_{\text{il}}|^2 \leq \mathbb{E} |\theta - \hat{\theta}_S|^2 \leq 2 + \mathbb{E} |\theta - \hat{\theta}_{\text{il}}|^2.$$

Proof: This is a direct consequence of the previous proposition and Lemma 2.55a. Let $\phi = |\theta|^2$, then

$$\begin{aligned} d - \tilde{d}g_{\tilde{d}}(\phi) &\geq d[1 - g_{\tilde{d}}(\phi)] \geq \frac{d\phi}{d + \phi}, \\ d - \tilde{d}g_{\tilde{d}}(\phi) &\leq d - \frac{\tilde{d}^2}{\tilde{d} + \phi} \leq 2 + \frac{\tilde{d}\phi}{\tilde{d} + \phi} \leq 2 + \frac{d\phi}{d + \phi}. \end{aligned}$$

And risk bounds follow. \square

Remark: Perhaps a more remarkable way to tell the story is in the canonical form for regression with a sample size n such that $\sigma_n^2 = 1/n$. In this setting, the two risks, when normalized, are asymptotically equal as $n \rightarrow \infty$.

$$\mathbb{E} |\theta - \hat{\theta}_S|^2 = \mathbb{E} |\theta - \hat{\theta}_{\text{il}}|^2 + O(1/n), \quad n \rightarrow \infty. \quad \triangleleft$$

Therefore, one can interpret that the regular James-Stein estimator provides a way to empirically estimate the ideal linear coefficient of X and it does very well in matching the idealized linear estimation risk target. In fact, the positive-part James-Stein estimator does even better. Unfortunately, its risk cannot be expressed in closed form. But we know that it is uniformly lower than that of the regular James-Stein estimator (see Lehmann and Casella, 1998, Chapter 4). However, we know that the overhead of the positive-part James-Stein estimator at $\theta = 0$ is only about 1.2, as opposed to 2 for the original James-Stein estimator, say, from their respective unbiased risk estimates,

$$\mathbb{E}_0 |\hat{\theta}_S|^2 - \mathbb{E}_0 |\hat{\theta}_{+S}|^2 = 2d \mathbb{P}(\chi_d^2 \leq \tilde{d}) + \mathbb{E} \left[\left(\chi_d^2 - \frac{\tilde{d}^2}{\chi_d^2} \right) \{ \chi_d^2 \leq \tilde{d} \} \right] \approx 0.8,$$

where χ_d^2 is a (central) chi-squared random variable with d degrees of freedom. However, the two risks are very close for large $|\theta|^2$ (as this can be gleaned from the fact that the two estimators shrink X by the same portion for large $|X|^2$).

Now we consider two-set shrinkage. Given a model $m \subseteq \{1, 2, \dots, d\}$, we first write

$$\theta_{\in m} = (\theta_i)_{i \in m} \quad \text{and} \quad \theta_{\notin m} = (\theta_i)_{i \notin m}.$$

We then would like to consider the idealized risk formed by independently estimating $\theta_{\in m}$ and $\theta_{\notin m}$ for the sets m and m^c by

$$\check{\theta}^m = (c_i X_i)_{i \leq d}, \quad c_i = c_m \mathbb{1}_{\{i \in m\}} + c_{m^c} \mathbb{1}_{\{i \in m^c\}}$$

where c_m and c_{m^c} are allowed to depend on θ . Due to the independence of the two sets of coefficients, it is clear that

$$c_m = \frac{|\theta_{\in m}|^2}{|\theta_{\in m}|^2 + \#m} \quad \text{and} \quad c_{m^c} = \frac{|\theta_{\notin m}|^2}{|\theta_{\notin m}|^2 + \#m^c}.$$

The risk of this two-block ideal linear estimation risk, given m , is simply the sum of the ideal linear estimation risks for the individual sets,

$$r_m^{\text{il}} = \mathbb{E} |\theta - \check{\theta}^m|^2 = \frac{|\theta_{\in m}|^2 \#m}{|\theta_{\in m}|^2 + \#m} + \frac{|\theta_{\notin m}|^2 \#m^c}{|\theta_{\notin m}|^2 + \#m^c}.$$

We are interested in forming a risk target that represents that of the best estimator among $m \in \mathcal{M}$.

$$r_*^{\text{il}} = \min_{m \in \mathcal{M}} r_m^{\text{il}}.$$

And it is clear from the one-block shrinkage results that this risk target is closely coupled with our two-block positive-part James-Stein risk target r_*^{+S} . Hence, when we discuss in the next subsection the simulations we conduct for two-set shrinkage mixture estimator, we will compare its risks against both targets r_*^{+S} and r_*^{il} .

2.3.6 Simulations

As in Section 2.2.4, we have simulations in four scenarios:

1. Constant One-Block,
2. Gradual Decay
3. Odd or Even Function
4. Ramp-Up with Cut-off

Please refer to the descriptions in Section 2.2.4 for details.

In this set of simulation results, we have used $\beta = 1/4$ in our mixture estimator. But we have also tried $\beta = 1/8$ and $\beta = 1/2$ — the performance of the mixture estimator is not very sensitive to the choice of β in this region.

Also, be careful that the y -axes of the following plots do not have a common scale.

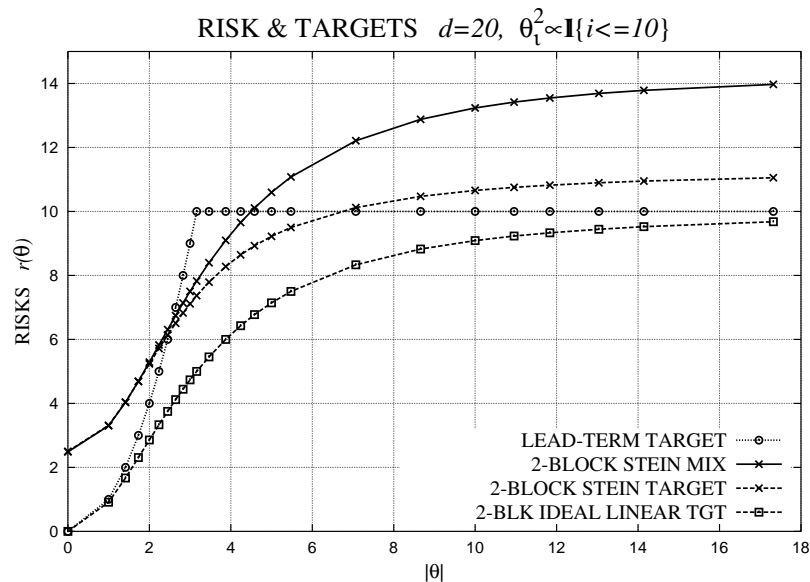


Figure 2.5: Risk and Targets: Constant One-Block. $\theta_i^2 \propto \mathbb{1}_{\{i \leq 10\}}$

We plotted the leading-term target (circle) here $|\theta|^2 \wedge 10$ for reference, which stays above the better target of the ideal two-block linear target (dashed square). The latter is just the ideal linear target on the leading-terms

$$r_*^{i1} = \frac{10|\theta|^2}{10 + |\theta|^2}.$$

The two-block positive-part James-Stein target (dashed cross) stays roughly 2 to 2.5 above that.

Our 2-block shrinkage mixture matches the James-Stein target for small θ but is worse by $3 \approx \log d$ for large θ . It outperforms the leading-term target for $|\theta|$ between 2.4 and 4.6.

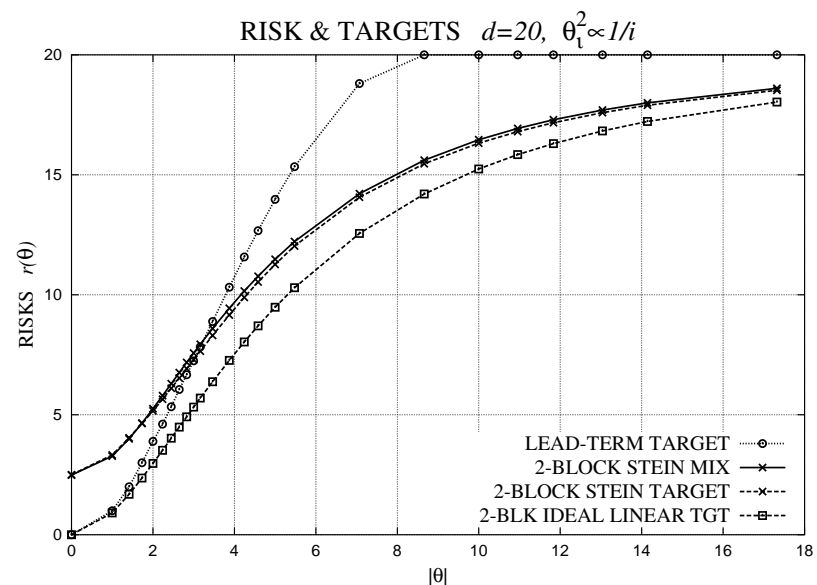


Figure 2.6: Risk and Targets: Gradual Decay. $\theta_i^2 \propto 1/i$

Except for the overhead of 2.5 for small θ , the James-Stein target stays below the leading-term target.

Our 2-block shrinkage mixture is almost on top of the James-Stein target, which in turn is not much worse than the two-block ideal linear target.

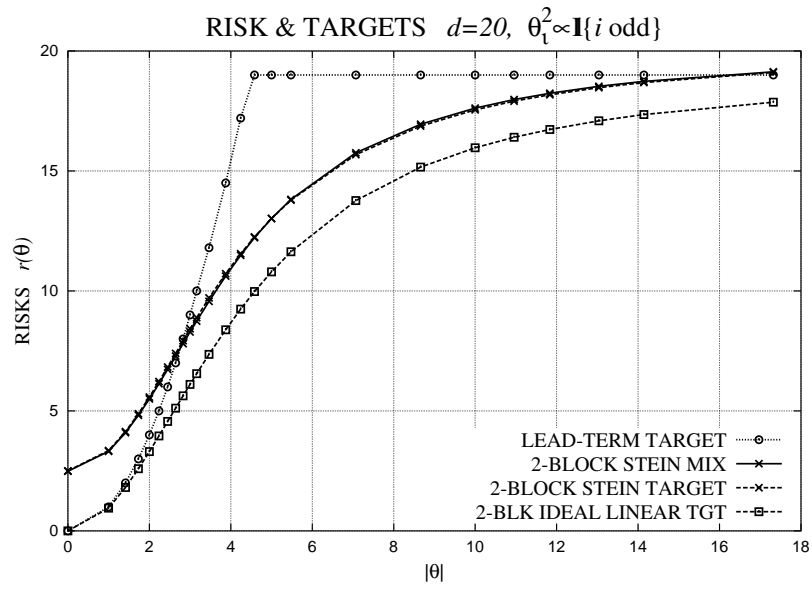


Figure 2.7: Risk and Targets: Odd or Even Function. $\theta_i^2 \propto \mathbb{1}_{\{i \text{ odd}\}}$

The conclusions here are similar to the previous case.

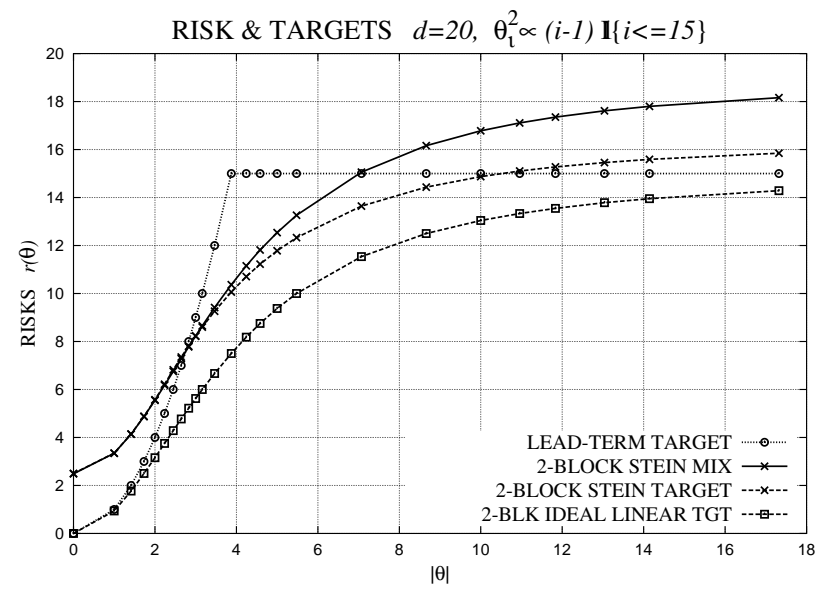


Figure 2.8: Risk and Targets: Ramp-Up with Cut-off. $\theta_i^2 \propto (i-1) \mathbb{1}_{\{i \leq 15\}}$

The conclusions here are similar to the one-block constant case.

Chapter 3

Controlling Model Complexity

As we saw in Example 2.29 that the previously developed theory yields a poor oracle inequality for the all-subset case because of the excess number of models considered, we would like to have a way of mixing a large number of subset models, favouring the models with lower complexities.

Related to coding in information theory is the assignment of a weight of the form $\exp(-C_m)$ for model m , where $C_m > 0$ is the **complexity** for model m . This is a number depending on m only (and not on the observations) deduced from *a priori* assumptions that could be subjective. We will give an example later. The Bayesian interpretation of this is that we employ the prior probability $\pi_m \propto \exp(-C_m)$ for model m . In information theory, the complexity is interpreted as the length of the codeword in a uniquely decodable code. Thus, the higher the complexity of a particular model, the more coding units it takes a code to describe the model, and the larger C_m is. In this regard, these complexity numbers must satisfy **Kraft's inequality** (see Cover and Thomas, 1991, Chapter 5)

$$\sum_{m \in \mathcal{M}} \exp(-C_m) \leq 1$$

so as to guarantee a uniquely decodable code. We require this here also for the validity of later analyses. But this is not necessary for the well-definedness of the weights with these **complexity regularization** in place that we are going to employ. Indeed, it is the relative sizes of e^{-C_m} that control the relative importance of the models.

[3.1] **EXAMPLE (Complexity-regularized Subsets).** We would like to model our prior knowledge that the coordinates at which θ is nonzero exhibit the following patterns with the associated complexities, which are expressed in the coding unit “nats”, for its relation to the use of the natural logarithm (as supposed to “bits” for binary digits when expressed in base-2 logarithm).

Nonzero coordinates ($m =$)	Complexity C_m
\emptyset	$\log 4 + 0$
$\{1, 2, \dots, k\}, \quad k > 0$	$\log 4 + \log d$
$\{k_1, \dots, k_2\}, \quad 0 < k_1 < k_2 < d$	$\log 4 + \log \binom{d}{2}$
any other m with $\#m = k, \quad k = 1, 2, \dots, d$	$\log 4 + \log d + \log \binom{d}{k}$

The complexities are chosen for the following properties. First, the above four cases have equal probabilities of $1/4$. And no coding is needed to describe the empty set, such that its complexity is just $\log 4$. Second, it takes $\log d$ nats to describe one out of d coordinates, and it takes $\log \binom{d}{k}$ nats to describe any k distinct coordinates, and $\log d$ nats to describe $k \leq d$ in the last case.

It is easy to check that these complexities assigned to the models satisfy Kraft's inequality $\sum_m \exp(-C_m) \leq 1$. ||

3.1 Complexity of Subset Models

We modify the weights of our models to reflect their respective complexities. Now, each is proportional to $\exp(-\beta \hat{r}_m - C_m)$. That is

$$\hat{\rho}_m = \frac{\exp(-\beta \hat{r}_m - C_m)}{\sum_{m'} \exp(-\beta \hat{r}_{m'} - C_{m'})} \quad (3.2)$$

To continue the Bayesian analogy, now $\hat{\rho}_m$ has the interpretation of the posterior probability of model m given the data X , because $\exp(-\beta \hat{r}_m)$ acts as the “sampling” probability density given model m , and hence the usual normalization appears in the denominator as in the Bayes formula.

It is straightforward to check that the new weights still satisfy the equation in (2.12) because C_m does not depend on X such that $\nabla C_m = 0$ for each m . Hence, the statements of Proposition 2.15 also holds for these new weights.

PROPOSITION. *Mixing subset estimators with complexity-regularizing weights in (3.2) yields $\hat{\theta}$ with an unbiased estimator of risk* [3.3]

$$\hat{r} = \sum_{m \in \mathcal{M}} \hat{\rho}_m \left[\hat{r}_m - (1 - 4\beta) |\hat{\theta}^m - \hat{\theta}|^2 \right].$$

Furthermore, for $0 \leq \beta \leq 1/4$, the risk estimate can be bounded by

$$\hat{r} \leq \sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m,$$

with equality when $\beta = 1/4$. \square

Denote the ‘‘prior’’ for model m as $\pi_m = \exp(-C_m)$. Now the analogue of Lemma 2.20 follows. This quantifies the combined risk estimate using the minimum of complexity-inflated risk $\hat{r}_{\hat{m}} + 1/\beta C_{\hat{m}}$ (up to a factor of β).

[3.4] LEMMA. *With \hat{m} defined by $\hat{r}_{\hat{m}} + 1/\beta C_{\hat{m}} = \min_m [\hat{r}_m + 1/\beta C_m]$, the combined risk estimates admits this representation*

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m = \hat{r}_{\hat{m}} + \frac{1}{\beta} \left[C_{\hat{m}} - D(\hat{\rho} \parallel \pi) + \log \hat{\rho}_{\hat{m}} \right] \quad (3.5)$$

In particular,

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m \hat{r}_m < \min_m \left[\hat{r}_m + \frac{1}{\beta} C_m \right].$$

Proof: First,

$$\hat{r}_m = \frac{1}{\beta} \left[\log \frac{\pi_m}{\hat{\rho}_m} - \log \sum_{m \in \mathcal{M}} \exp(-\beta \hat{r}_m - C_m) \right].$$

Now by adding and subtracting $\hat{r}_{\hat{m}} + 1/\beta C_{\hat{m}}$, we have

$$\hat{r}_m = \hat{r}_{\hat{m}} + \frac{1}{\beta} \left[C_{\hat{m}} + \log \frac{\pi_m}{\hat{\rho}_m} + \log \hat{\rho}_{\hat{m}} \right].$$

Now average with respect to the weights $\hat{\rho}_m$ to obtain (3.5). For the second statement, the definition¹ of D and Kraft’s inequality implies that

$$D(\hat{\rho} \parallel \pi) \geq -\log \sum_m \pi_m \geq 0.$$

Therefore, the negativity of the last two terms in (3.5) yields the bound. \square

This first expression in the lemma says that the combined risk estimate is simply the minimum of the complexity-inflated risk $\hat{r}_{\hat{m}} + 1/\beta C_{\hat{m}}$ (up to the scale factor β) minus some adjustments. Thus it is upper-bounded by the quantity without the adjustments. The Kullback divergence term D would be small if knowing the data X does not make the posterior weights $\hat{\rho}$ differ much from the prior weights π . At any rate, D is non-negative because π is a sub-probability by Kraft’s inequality. And just as before, if \hat{m} is a strong minimizer with few competing models m , the bound $\log \hat{\rho}_{\hat{m}} \leq 0$ would be tight because $\hat{\rho}_{\hat{m}}$ is close to 1. Otherwise, there are further risk savings due to this term.

Note that the log-cardinality additive price before is now subsumed in the C_m/β term, which would have been about $1/\beta \log(\#\mathcal{M})$ as before if the models were equally considered.

[3.6] THEOREM. *For the mixture of least-squares estimator $\hat{\theta}$ with complexity-regularized weights, when $\beta = 1/4$, we have*

$$\mathbb{E} |\theta - \hat{\theta}|^2 < \min_m [r_m + 4C_m]. \quad \square$$

¹Or simply, let π' be a normalized derivation of π such that $\pi'_m = \pi_m / \sum_m \pi_m$ for each m . Then $D(\hat{\rho} \parallel \pi) \geq D(\hat{\rho} \parallel \pi') \geq 0$.

See section 2.6 of Yang (2004) and section 10 of Catoni (1999) for similar oracle inequalities for prediction for mixing arbitrary regression functions. They yield an additive cost in excess of the best model risk of the same order as ours, though their constant depends on the assumptions of the error and the uniform upper-bound for the regression functions, and can be quite large.

3.2 Complexity of Two-Set Shrinkage Models

There is a parallel story about controlling complexity in the Two-Set Shrinkage Models. Please note that all the risk and weight-related quantities, $r, \hat{r}, \hat{\rho}, C$ (and the subscripted versions) refer to the shrinkage case in this section.

Recall the definition for the continuous risk estimates $\{\hat{r}_m^c\}$ (2.34). We now define the complexity-regularizing weights as,

$$\hat{\rho}_m^c = \frac{\exp(-\beta \hat{r}_m^c - C_m)}{\sum_{m' \in \mathcal{M}} \exp(-\beta \hat{r}_{m'}^c - C_{m'})} \quad (3.7)$$

which are almost differentiable, and we shall use them for mixing our estimators $\{\hat{\theta}^m\}$.

LEMMA. *Define \hat{m} by $\hat{r}_{\hat{m}} + 1/\beta C_{\hat{m}} = \min_m [\hat{r}_m + 1/\beta C_m]$, the combined risk estimates admits this representation* [3.8]

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m^c \hat{r}_m = \hat{r}_{\hat{m}} + \frac{1}{\beta} \left[C_{\hat{m}} - D(\hat{\rho}^c \parallel \pi) + D(\hat{\rho}^c \parallel \hat{\rho}) + \log \hat{\rho}_{\hat{m}} \right]$$

In particular, $D(\hat{\rho}^c \parallel \hat{\rho}) \leq 8\beta$ and

$$\sum_{m \in \mathcal{M}} \hat{\rho}_m^c \hat{r}_m < \min_m \left[\hat{r}_m + \frac{1}{\beta} C_m \right] + 8\beta.$$

Proof: The proof is completely analogous to the least-squares case. Refer to Lemma 3.4 and Lemma 2.38. Also, Lemma 2.43 still holds with the complexity-regularizing weights, which yields the required bound $D \leq 8\beta^2$. \square

THEOREM. *For the mixture of two-set shrinkage estimator $\hat{\theta}$ with complexity-regularized weights (3.7) with $\beta = 1/8$, we have* [3.9]

$$\mathbb{E} |\theta - \hat{\theta}|^2 \leq \min_m [r_m + 1 + 8C_m]. \quad \square$$

Chapter 4

Discussion

4.1 Bayesian Considerations

We mentioned that the data-driven weights $\hat{\rho}$ for the models are analogous to the Bayesian posterior probability for weighting the estimators under the models. Here we clarify this connection for the mixture of least-squares estimator, noting that the unbiased risk estimate for the least-square estimator under each model is related to AIC. However, when we examine the shrink-two-set estimators, we are deriving an extension that is non-Bayesian.

Nevertheless, our justifications for using the weights in [2.11] are rooted in neither AIC nor Bayes, but rather that the unbiasedness of the risk estimates \hat{r}_m provides a common ground for model comparison, and the models should be weighted according to some decreasing function of these risk estimates. The use of the exponential function is vindicated because it yields the clean risk bounds with simple yet direct applications of information-theoretic tools.

4.1.1 Form of Pseudo-Bayes Estimator

The crux of the Bayesian interpretation lies in the fact that a Bayes estimator in our canonical multivariate normal mean problem (under squared-error loss) can always be written as

$$X + \nabla \log p_w(X) \quad (4.1)$$

where

$$p_w(x) = \int_{\mathbb{R}^d} \phi(x - \theta) dw(\theta)$$

is the marginal density induced by the prior w for θ , and ϕ the standard d -variate normal density. This is because the Bayes estimator is the posterior mean

$$\mathbb{E}[\Theta | X] = X + \mathbb{E}[\Theta - X | X]$$

where the posterior expectation on the right can be written as

$$\nabla \log p_w(X) = \frac{\nabla p_w(X)}{p_w(X)} = \frac{\int (\theta - X) \phi(X - \theta) dw(\theta)}{\int \phi(X - \theta) dw(\theta)}$$

using Bayes formula and integration by parts with the normal density (Note that $p_w(x) > 0$ for any x by the normality of $X | \theta$.) Many estimators, even if they are not formally derived

from a prior on θ in a Bayesian framework, can be written in such a way with some other function q in lieu of p_w , and Brown (1971) showed that all admissible estimators for this problem have to be in the form (4.1). And the function q is called **pseudo-marginal density** because it needs not integrate to one (any constant scaling will vanish under the gradient) and it needs not come from integrating out θ with respect to a prior measure.

What remains are the questions

1. Can our mixture estimator $\hat{\theta}$ and its component estimators $\hat{\theta}^m$ can be written in such a form (4.1)? What are their respective pseudo-marginal densities q and q_m ?
2. Are there prior distributions for θ that would induce these pseudo-marginal densities?

We answer these separately for the subset models and for the two-block shrinkage models.

4.1.2 Least-Squares Estimators for Subset Models

The first question has an easy answer for the least-squares estimators on subset models. Conditional on model m , we take

$$q_m(X) = \exp(-1/2 \hat{r}_m) = \exp\left[-\frac{1}{2} \sum_{i \notin m} X_i^2 - \#m + \frac{d}{2}\right],$$

which does give $\hat{\theta}^m = (X_i \mathbb{I}_{\{i \in m\}})_{i \leq d}$ with $\nabla \log q_m = -(X_i \mathbb{I}_{\{i \neq m\}})$ by putting $\beta = 1/2$ in our weights.

It is common to use model selection to choose one appropriate subset model, and then use the least-square estimator for each subset model. The problem of figuring out the prior distribution for θ that would give rise to a Bayes procedure mimicking this, in the sense of choosing a model with the maximum posterior weight, has been studied. See Hartigan (2002) for a good survey. It is well-known that it requires a uniform improper prior on the parameter in the subset model to obtain a Bayes procedure that coincides with the least-squares estimator for that model. The difficulty involved is the arbitrariness of the height of the uniform prior, because any scaling has no effect on the resulting estimator (least-squares) under the model but does affect the posterior probabilities of the models. In other words,

one needs to specify the preference between any two models m and m' . If we assume that this only depends on the difference between the model dimensions, then this can be cast as the ratio of prior weights (the concept of probability is lost due to the lack of an absolute scale) between two models of dimensions, say k and $k + 1$,

$$\frac{\text{weight}(\#m = k + 1)}{\text{weight}(\#m = k)} \stackrel{\text{def}}{=} w$$

with $0 < w < 1$ since the prior must give preference to the models with lower dimensions to counteract the fact that the larger models give better fits. Hartigan examines this (for the one-dimensional case $X \sim N(\theta, 1)$) such that there are only two models, $\theta = 0$ and $\theta \in \mathbb{R}$). His method for selecting w is based on specifying the desired level of significance in testing whether the extra variable should be included. He further shows that the choice

$$w = \frac{1}{\sqrt{2\pi}} e^{-1} \approx 0.147$$

coincides with AIC for selecting the model with the highest posterior weight. This is exactly what we use, as is apparent from the fact that the Stein's unbiased risk estimate for model m is exactly AIC's criterion up to an additive constant d . Alternatively, we equate the posterior weight of $\theta = 0$ in Hartigan with our $\hat{\rho}_0$ for the leading-term case with $d = 1$

$$\frac{\phi(x)}{\phi(x) + w} = \frac{\exp(-x^2/2)}{\exp(-x^2/2) + e^{-1}} \stackrel{\text{def}}{=} \hat{\rho}_0$$

to obtain the w above. In the general subset case, it is equivalent to using a prior of

$$w^{\#m} = (\sqrt{2\pi}e)^{-\#m}$$

for model m .

Note that the above discussion is only limited to the case $\beta = 1/2$. However, using other values of β can be interpreted as using a modelling density with a different variance other than 1. For example, when $\beta = 1/4$, as required for our cleanest risk bounds, the effect of this discrepancy is to make the form of the dependence of the weights on the sum of squares of the left out coefficients $\sum_{i \notin m} X_i^2$ more diffuse (that is longer-tailed, though still with exponential decay) than they would have been with $\beta = 1/2$.

George (1986b) provides the methodology for producing the pseudo-posterior probabilities for mixing estimators under different models, and this applies as long as the component estimators can be written in the form (4.1) (without considering any prior distribution.) He also provided a formula for Stein's unbiased estimate of risk of such a convex combination of estimators. But he focused on component estimators with superharmonic pseudo-marginals, because they are minimax; and he further showed that in this case the resulting mixture estimator will also be minimax. But he did not examine the subset model estimators in particular, probably because their pseudo-marginals are not superharmonic. Indeed, if we view this problem in his framework, it implies $\beta = 1/2$, as a true (but improper) Bayes procedure should. Therefore, with $q_m = \exp(-1/2 \hat{r}_m)$, its Laplacian

$$\nabla^2 q_m(X) = \sum_{i \notin m} (X_i^2 - 1) q_m(X) = \left(\sum_{i \notin m} X_i^2 + \#m - d \right) q_m(X)$$

may not be negative.

George studied positive-part James-Stein estimator closely because it is minimax.

4.1.3 Two-Set Shrinkage Estimators

The (one-block) positive-part James-Stein shrinkage estimator on d coefficients can be easily seen to in the form of (4.1) with

$$\nabla \log q^d(X) = -\gamma X, \quad \text{where } \gamma = \left(1 \wedge \frac{\tilde{d}}{|X|^2} \right),$$

where we recall that $\tilde{d} = d - 2$. George (1986b,a) examined this case closely, and his pseudo-marginal is

$$q^d(X) = \left\{ \begin{array}{ll} \left(\frac{\tilde{d}}{e|X|^2} \right)^{\tilde{d}/2}, & \text{if } |X|^2 > \tilde{d} \\ \exp(-|X|^2/2), & \text{if } |X|^2 \leq \tilde{d} \end{array} \right\} = (e/\gamma)^{-\gamma|X|^2/2}.$$

This can be shown to be superharmonic, and thus, the resulting positive-part James-Stein estimator is minimax. However, this only agrees with our pseudo-marginal

$$q^d(X) = \exp(-\beta \hat{r})$$

for $\beta = 1/2$ and $|X|^2 \leq \tilde{d}$, but they otherwise differ. Moreover, even in this regime, George's pseudo-marginal will not agree with our q using the continuous risk estimate \hat{r}_{+S}^c .

Using George (1986a), the pseudo-marginal for the two-part shrinkage estimator (on the sets m and m^c) is simply the product of the pseudo-marginals for the individual blocks.

$$q_m(X) = q^m(X_{\in m}) q^{d-m}(X_{\notin m})$$

George did provide a way of putting prior weights for models with different dimensions, what he called calibration, but he did not provide any theoretical justification for setting such weights.

4.2 Concluding Remarks

The sharp and exact oracle inequalities for finite-dimensional regression presented in this thesis are appealing. One by-product is that is that the two-set shrinkage estimator is minimax (it dominates least-squares, which has risk d), and in general can be used instead of least-squares if the mean squared-error is the only concern in a regression problem.

Future research possibilities include letting d grow with n and examine whether any blocking or partitioning regression scheme can achieve other oracle inequalities or minimax risk targets. With some mild regular conditions, our theory can be made to accommodate a countable number of models (when the cardinality of the model class grows with dimension $d \rightarrow \infty$).

Chapter A

Appendix

A.1 From Function Estimation to Canonical Regression

In this appendix, we want to show how function estimation is related to linear regression by using a linear parameterization of a function via a basis.

Linear Parametrization of a Function via a Basis

We observe $Y_j \in \mathbb{R}$ through the following model:

$$Y_j = f(X_j) + \epsilon_j, \quad j = 1, 2, \dots, n$$

where $X_j \in \mathbb{R}^d$ are distributed as μ and ϵ_j are independent errors distributed as $N(0, \sigma^2)$ and independent of all X_j . In particular, $\mathbb{E}[Y_j | X_j] = f(X_j)$. The unknown f is a real $L_2(\mu)$ function we want to estimate. For the moment, we assume σ^2 known – we leave it in a general form for its flexibility in our analyses.

Suppose the function takes the form

$$f(x) = \sum_{i=1}^{\infty} \beta_i \varphi_i(x), \quad \beta_i \in \mathbb{R}$$

where $\{\varphi_i\}_{i \in \mathbb{N}}$ are given linear independent functions spanning $L_2(\mu)$. It is often believed that the function is approximated well by the leading terms

$$\sum_{i=1}^m \beta_i \varphi_i(x)$$

although the decay rate of the approximation error with m is not known. If μ is given, we may assume that the basis $\{\varphi_i\}_{i \in \mathbb{N}}$ is orthonormal by a procedure like Gram-Schmidt that preserves the leading-term models. Then

$$\beta_i = \int f(x) \varphi_i(x) d\mu(x)$$

and in vector notation,

$$f(x) = \Phi(x)\beta$$

where $\Phi(x) = (\varphi_1(x), \varphi_2(x), \dots)$ and $\beta = (\beta_1, \beta_2, \dots)^t$. Consequently, estimating f amounts to estimating β

$$\hat{f}(x) = \Phi(x)\hat{\beta} = \sum_{i=1}^{\infty} \varphi_i(x) \hat{\beta}_i(Y), \quad Y = (Y_1, Y_2, \dots, Y_n).$$

The square loss

$$\|f - \hat{f}\|_{\mu}^2 = \int |f(x) - \hat{f}(x)|^2 d\mu(x),$$

can then be computed using

$$\|f - \hat{f}\|_{\mu}^2 = \sum_{i=1}^{\infty} (\beta_i - \hat{\beta}_i)^2 = \|\beta - \hat{\beta}\|^2$$

by the distance-preserving $L_2 - \ell_2$ isometry.

Common procedures consider estimates that use the first m basis functions and then select m . For any given m , the risk so obtained is

$$r_n(m) := \sum_{i=1}^m \mathbb{E}(\hat{\beta}_i - \beta_i)^2 + \sum_{i \geq m}^{\infty} \beta_i^2$$

which can be interpreted as the tradeoff of variance and bias. For instance, one may use

$$\hat{\beta}_i = \frac{1}{n} \sum_{j=1}^n \varphi_i(X_j) Y_j$$

which produces an unbiased estimate of β_i of variance of order $1/n$. There is a best $m_n^* = m_n^*(f, \sigma^2)$ that achieves the best bias-and-variance tradeoff

$$r_n^* = r_n(m_n^*) = \min_m \left\{ \sum_{i=1}^m \mathbb{E}(\hat{\beta}_i - \beta_i)^2 + \sum_{i \geq m}^{\infty} \beta_i^2 \right\}$$

A natural question is whether there is an estimate of f (without knowledge of m_n^* and f) that has a risk nearly as small as r_n^* .

A Simplified Problem using the Canonical Linear Model

Some aspects of the problem are simplified by considering a fixed design, or conditioning on $X = (X_1, X_2, \dots, X_n)$. Let $\Phi^{(k)}(x) = (\varphi_1(x), \varphi_2(x), \dots, \varphi_k(x))$ be the k -truncated basis, where $k \leq n$ is the largest size model that will be considered. Denote the $n \times k$ matrix

$$A = \Phi^{(k)}(X) = \begin{pmatrix} \Phi^{(k)}(X_1) \\ \Phi^{(k)}(X_2) \\ \vdots \\ \Phi^{(k)}(X_n) \end{pmatrix} = \begin{pmatrix} \varphi_1(X_1) & \varphi_2(X_1) & \cdots & \varphi_k(X_1) \\ \varphi_1(X_2) & \varphi_2(X_2) & \cdots & \varphi_k(X_2) \\ \vdots & \vdots & \cdots & \vdots \\ \varphi_1(X_n) & \varphi_2(X_n) & \cdots & \varphi_k(X_n) \end{pmatrix}.$$

For simplicity, take $k = n$ and assume that A has full rank. Then our problem is transformed to the following approximate model,

$$Y = A\beta + \epsilon \quad (\text{A.1})$$

where $Y = (Y_1, Y_2, \dots, Y_n)$, $\beta = \beta(k) = (\beta_1, \beta_2, \dots, \beta_k)^t$, $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^t$. The approximate aspect of this model is the neglecting of the bias from omission of basis functions $\{\varphi_i\}_{i>n}$ in the representation of f . For simplicity, we assume here the validity of (A.1) with ϵ_j iid $N(0, \sigma^2)$.

Let $S^t S = (1/n A^t A)^{-1}$, where the square-root matrix S may be taken to be upper triangular for the ease of computation.¹ Then an equivalent simplified model is

$$\begin{aligned} Z &:= \frac{1}{n} S A^t Y = \frac{1}{n} S A^t A \beta + \frac{1}{n} S A^t \epsilon \\ &= S^{-t} \beta + \tilde{\epsilon} \\ &= \theta + \tilde{\epsilon} \end{aligned}$$

such that, conditional on X , the new response

$$Z \sim \text{Normal}_k(\theta, \sigma_n^2 I),$$

with unknown mean $\theta = S^{-t} \beta$ and known variance $\sigma_n^2 = \sigma^2/n = 1/n$ by letting σ^2 to be unity.²

¹If the basis $\{\varphi_i\}_{i \in \mathbb{N}}$ is orthonormal in $L_2(\mu)$, then $\frac{1}{n} A^t A$ is almost the identity matrix for large n , whence S is also close to the identity, and our new parameter θ is close to the original $\beta = S^t \theta$.

²We basically have reduced the problem to a white noise observation model,

$$Z_i = \theta_i + \sigma_n \tilde{\epsilon}_i, \quad i = 1, 2, \dots, k$$

where θ_i are unknown, and $\tilde{\epsilon}_i$ are iid $N(0, 1)$. The per-dimension error variance $\sigma_n^2 = \sigma^2/n$ has a scale of $1/n$ chosen to emphasize the regression framework.

Alternatively, one could start with a Gramm-Schmidt procedure to orthogonalize the design such that $QR = A$ where Q is an $n \times k$ matrix with orthonormal columns and R is $k \times k$ upper triangular. Then we work with the model

$$Y = Q\tilde{\beta}^{(k)} + \epsilon,$$

where $\tilde{\beta}^{(k)} = R\beta$. An equivalent model is then

$$\tilde{Z} := n^{-1/2} Q^t Y = n^{-1/2} \tilde{\beta}^{(k)} + n^{-1/2} Q^t \epsilon = \theta + \tilde{\epsilon}$$

However, in the literature as well as in this thesis, we often take $\sigma_n^2 = 1$ for convenience, especially when the problem at hand deals only with a finite sample. We also take $k = d$ as the (truncated) dimension of our problem as an approximation to the original one.

A.2 Incomplete Gamma function, g_L , and Gamma distribution

The function g_L was defined for evaluating the risk of the James-Stein estimator. The complementary incomplete Gamma function was used for evaluating the form of the Bayes estimator using the Strawderman prior. The two are closely related: they are essentially conveniently standardized version of each other, but extended for a negative argument when the other function is only defined for a non-negative domain.

Recall our definition of g_L conveniently standardized for the James-Stein risk evaluation:

$$g_L(x) = e^{-x/2} \sum_{k=0}^{\infty} \frac{(x/2)^k}{k!} \frac{L}{L+2k}, \quad x \geq 0, \quad L = 1, 2, \dots$$

Using the complementary incomplete Gamma function

$$\gamma(r, x) = \int_0^x t^{r-1} e^{-t} dt = x^r \sum_{k=0}^{\infty} \frac{(-x)^k}{k!} \frac{1}{r+k}, \quad r > 0, \quad x \geq 0$$

we could write g_L as

$$g_L(x) = e^{-x/2} \frac{L}{2} \left(-\frac{x}{2}\right)^{-L/2} \gamma\left(\frac{L}{2}, -\frac{x}{2}\right).$$

by extending the definition of γ for $x \in \mathbb{R}$. This is justified by an application of the ratio test, showing that the related series defining both functions converge for each $x \in \mathbb{C}$, $r > 0$, $L > 0$. Thus, g_L is also defined for $L > 0$, $x \in \mathbb{R}$. So,

$$\gamma(r, x) = g_{2r}(-2x) x^r r^{-1} e^{-x}, \quad x \in \mathbb{R}$$

When $x \geq 0$, $\gamma(r, x)$ also defines the cumulative distribution function G_r of a Gamma($r, 1$) random variable

$$\gamma(r, x) = \Gamma(r) G_r(x) \quad r > 0, \quad x \geq 0.$$

In other words, one could more succinctly write

$$g_{2r}(-x) = \frac{G_r(x/2)}{G'_{r+1}(x/2)} = \frac{2r}{x} \frac{G_r(x/2)}{G'_r(x/2)}, \quad x \geq 0, \quad r > 0,$$

where $G'_r(x) = x^{r-1} e^{-x} / \Gamma(r)$ is the density of a Gamma($r, 1$) random variable.

where $\theta = n^{-1/2} R\beta$. The errors $\tilde{\epsilon} = n^{-1/2} Q^t \epsilon$, conditional on X , are distributed as $\text{Normal}_k(0, \sigma_n^2 I_k)$ for $\sigma_n^2 = \sigma^2/n$. We estimate β from the inverse transform of an estimate of θ

$$\hat{\beta}^{(k)} = \sqrt{n}(R'R)^{-1} R^t \hat{\theta}.$$

When r is a positive integer, the function admits a finite sum representation

$$\gamma(r, x) = (r-1)!e^{-x} \left[e^x - \sum_{k=0}^{r-1} \frac{x^k}{k!} \right] = \Gamma(r) \left[1 - e^{-x} \sum_{k=0}^{r-1} \frac{x^k}{k!} \right]$$

because $\mathbb{P}\{\text{Gamma}(r) \leq x\} = \mathbb{P}\{\text{Poisson}(x) \geq r\}$, and this expression holds also for the extension to $x \in \mathbb{R}$ because of the uniform convergence (over \mathbb{C}) of the series stated above.

Thus, when L is even, g_L also admits the following finite sum representation using its relation to $\gamma(r, x)$:

$$g_{2r}(x) = \frac{r!}{(-x/2)^r} \left[e^{-x/2} - \sum_{k=0}^{r-1} \frac{(-x/2)^k}{k!} \right], \quad r \in \mathbb{N}, x \in \mathbb{R}.$$

Bibliography

- Akaike, H. (1970). Statistical Predictor Identification. *Annals of Institute of Statistical Mathematics*, **22**:203–217.
- Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In Petrov, B. N. and Csáki, F., editors, *Second International Symposium on Information Theory*, 267–281, Budapest: Akadémia Kiado.
- Baraud, Y. (1999). Model Selection for Regression on a Fixed Design. *Probability Theory and Related Fields*, **117**:467–493.
- Barron, A. R. (1987). Are Bayes rules consistent in information? Springer-Verlag.
- Barron, A. R., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, **113**:301–413.
- Beran, R. and Dümbgen, L. (1998). Modulation of estimators and confidence sets. *Annals of Statistics*, **26**(5):1826–1856.
- Berger, J. O. and Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**:109–122.
- Birgé, L. and Massart, P. (1997). From model selection to adaptive estimation. In Pollard, D., Torgersen, E., and Yang, G., editors, *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, 55–87. Springer-Verlag, New York.
- Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, **4**(3):329–375.
- Brown, L. D. (1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Annals of Mathematical Statistics*, **42**:855–903.
- Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model Selection: An Integral Part of Inference. *Biometrics*, **53**:603–618.
- Catoni, O. (1997). Mixture approach to universal model selection. Preprint 30, Laboratoire de Mathématiques de l’Ecole Normale Supérieure, Paris.
- Catoni, O. (1999). “Universal” aggregation rules with exact bias bounds”. Preprint 510, Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris.
- Cavalier, L. and Tsybakov, A. B. (2001). Penalized blockwise Stein’s method, monotone oracles and sharp adaptive estimation. *Mathematical Methods of Statistics*, **10**(3):247–282.
- Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.
- Donoho, D. L. and Johnstone, I. M. (1995). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Journal of the American Statistical Association*, **90**(432):1200–1224.
- George, E. I. (1986a). Combining minimax shrinkage estimators. *Journal of the American Statistical Association*, **81**:431–445.
- George, E. I. (1986b). Minimax multiple shrinkage estimation. *Annals of Statistics*, **14**:188–205.
- George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, **7**:339–373.
- Goldenshluger, A. and Tsybakov, A. B. (2001). Adaptive Prediction and Estimation in Linear Regression with Infinitely Many Parameters. *Annals of Statistics*, **29**(6):1601–1619.
- Hartigan, J. A. (2002). Bayesian Regression Using Akaike Priors. Technical report, Yale University.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussions). *Statistical Science*, **14**:382–417.
- James, W. and Stein, C. (1961). Estimation with quadratic loss. In University of California Press, editor, *Proceedings of Fourth Berkeley Symposium of Mathematical Statistics and Probability*, **1**:361–380.
- Johnstone, I. M. (1998). *Function Estimation in Gaussian Noise: Sequence Models*. Available at www-stat.stanford.edu.
- Kabaila, P. (2002). On variable selection in linear regression. *Econometric Theory*, **18**:913–925.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**:773–795.

- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*. Springer-Verlag, New York, second edition.
- Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer-Verlag, New York, second edition. (originally by E.L. Lehmann, 1983).
- Leung, G. and Barron, A. R. (2004). Information Theory, Model Selection and Model Mixing for Regression. In *Proceedings of the Conference on Information Sciences and Systems*, Princeton.
- Li, K. C. (1987). Asymptotic optimality for C_p , C_L , cross-validation and generalized cross-validation: discrete index set. *Annals of Statistics*, **15**(1):958–975.
- Pinsker, M. S. (1980). Optimal Filtering of Square Integrable Signals in Gaussian White Noise. *Problems in Information Transmission*, **16**:120–133. translated from Russian.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, **14**:465–471.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2):461–464.
- Shibata, R. (1981). An optimal selection of regression variables. *Biometrika*, **68**(1):45–54.
- Stein, C. (1973). Estimation of the mean of a multivariate normal distribution. In *Proceedings of the Prague Symposium in Asymptotic Statistics*, 345–381.
- Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, **9**:1135–1151.
- Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics*, **42**(1):385–388.
- Tsybakov, A. B. (2003). Optimal rates of aggregation. In B.Scholkopf and M.Warmuth, editors, *Computational Learning Theory and Kernel Machines: Lecture Notes in Artificial Intelligence*, **2777**:303–313, Heidelberg. Springer.
- Yang, Y. (1999). Model selection for nonparametric regression. *Statistica Sinica*, **9**:475–499.
- Yang, Y. (2000). Combining different regression procedures for adaptive regression. *Journal of Multivariate Analysis*, **74**:135–161.
- Yang, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, **13**:783–809.
- Yang, Y. (2004). Combining forecasting procedures: some theoretical results. *Econometric Theory*, **20**:176–222.
- Yang, Y. and Barron, A. R. (1998). An asymptotic property of model selection criteria. *IEEE Transactions in Information Theory*, **44**:95–116.