Information Theory and Mixing Least-Squares Regressions

Gilbert Leung, *Member, IEEE*, and Andrew R. Barron, *Senior Member, IEEE*

***Abstract*—For Gaussian regression, we develop and analyze methods for combining estimators from various models. For squared-error loss, an unbiased estimator of the risk of the mixture of general estimators is developed. Special attention is given to the case that the component estimators are least-squares projections into arbitrary linear subspaces, such as those spanned by subsets of explanatory variables in a given design. We relate the unbiased estimate of the risk of the mixture estimator to estimates of the risks achieved by the components. This results in simple and accurate bounds on the risk and its estimate, in the form of sharp and exact oracle inequalities. That is, without advance knowledge of which model is best, the resulting performance is comparable to or perhaps even superior to what is achieved by the best of the individual models. Furthermore, in the case that the unknown parameter has a sparse representation, our mixture estimator adapts to the underlying sparsity. Simulations show that the performance of these mixture estimators is better than that of a related model-selection estimator which picks a model with the highest weight. Also, the connection between our mixtures with Bayes procedures is discussed.**

***Index Terms*—Bayes mixtures, combining least-squares regressions, complexity, model adaptation, model selection target, oracle inequalities, resolvability, sparsity, unbiased risk estimate.**

I. Introduction

REGRESSION problems in statistics concern estimating some functional relation between a response variable and explanatory variables. Often there are multiple models describing such a relation. It is common to employ a two-stage practice which first examines the data and picks a best model based on some model assessment criterion, and then uses an appropriate regression estimator for that model. This is useful when a parsimonious model for explaining the response is desired. However, 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 the estimator for the chosen model can be overly optimistic as model uncertainty from the selection procedure is often neglected.

Combining estimators from different models is an alternative to model selection-based estimation. We may take each component model to be a linear subspace of the full model space and the corresponding estimator to be the least-squares projection of the observations into that subspace. The combined estimator consists of a convex mixture of the component estimators with weights that may depend on the data. In this paper, we study properties of the statistical risk (mean-squared error) of the combined estimator. An information-theoretic characterization of an unbiased estimate of its risk is provided. Furthermore, the risk of the resulting mixture is not much more than an idealized target defined by the minimum of risks achieved by the various estimators (one for each model considered). This is what Yang [1] calls combining for adaptation and the risk target is termed the model selection target by Tsybakov [2] since it lower-bounds the risks of all model selection-based estimators. The general sharp risk bounds, or oracle inequalities, shown in this paper are obtained by choosing certain types of weights that adapt to the data. Moreover, the resulting mixture estimator often performs better in simulations than a related model-selection estimator, which picks the estimate corresponding to the highest weight model.

A primary motivation behind mixing estimators is that it often improves the risk in regression estimation, as "betting" on multiple models provides a type of insurance against a singly selected model being poor. Another motivation comes from consideration of Bayes procedures which are known to possess desirable properties in any statistical decision problem. Indeed, Bayes procedures minimize the average case risk with respect to the prior. With squared-error loss, a Bayes estimator is a convex combination of estimators weighted by the corresponding models' posterior probability (see Hoeting *et al.* [3] and the references cited therein).

A key tool in our analysis is the unbiased estimate of risk by Stein [4], [5]. We adapt it to provide risk assessment for mixtures of general estimators and to produce risk bounds for the mixture of least-squares estimators in linear models. Mixtures of shrinkage estimators, which are nonlinear, are analyzed in the thesis [6].

A. Overview

In regression and function estimation problems with fixed design, one has observations Y_i of response values μ_i plus independent Gaussian noise, for indices $i = 1, 2, \ldots, n$. These response values μ_i may be equal to the values of an unknown function $f(\underline{x}_i)$ where the given $\underline{x}_i$ are vectors of explanatory variables. We also have functional models available for our consideration that may or may not approximate such a true f well.

Manuscript received August 10, 2004; revised March 27, 2006. The material in this paper was presented in part at the Conference on Information Sciences and Systems, Princeton, NJ, March 2004 and the IEEE International Symposium on Information Theory, Adelaide, Australia, September 2005.

G. Leung is with Qualcomm Inc., Campbell, CA 95008 USA (e-mail: gleung@qualcomm.com).

A. R. Barron is with the Statistics Department, Yale University, New Haven, CT 06511 USA (e-mail: andrew.barron@yale.edu).

Coommunicated by P. L. Bartlett, Associate Editor for Pattern Recognition, Statistical Learning and Inference.

Digital Object Identifier 10.1109/TIT.2006.878172

In choosing a procedure for estimating the response values μ_i, we recognize that unobserved or hidden variables may also contribute to the true μ_i. Thus, we adopt the general setting that $Y = \mu + \epsilon$ with μ in $\mathbb{R}^n$ and errors ϵ distributed as $\mathcal{N}ormal(0, \sigma^2 I)$, where for simplicity σ^2 is known. Though the true μ is allowed to be general, our estimators will be constructed from (usually linear) functional models and combinations thereof. We shall use squared-error loss $\sum_{i=1}^n (\mu_i - \hat{\mu}_i)^2$ and its expectation, the mean-squared error, as the risk in assessing the performance of our estimators.

If the estimate is not constrained to live in any of the models of interest, the simple estimator Y can be obtained by maximizing likelihood or by least-squares, and has a mean-squared error of $n\sigma^2$. Least-squares regression into a lower dimensional space has risk that can potentially be much smaller, depending on how close the true μ is to that space. Here we combine least-squares regressions into various linear subspaces, and provide accurate upper bounds for the risks of the mixtures which can be vastly reduced from $n\sigma^2$ due to model adaptation. Again, our aim is to have the risk of the combined estimator close to the minimum of the risks of the individual estimators.

A linear regression model m is a d_m-dimensional linear subspace of $\mathbb{R}^n$ in which the mean vector μ may reside. We consider classes $\mathcal{M}$ of linear models m. Typically, each m in $\mathcal{M}$ is spanned by subsets of columns of a design matrix of predictors. For each model m, there is a basis of d_m columns denoted by X_m for which the mean μ is modeled as $X_m\theta$ for some unknown $\theta \in \mathbb{R}^{d_m}$. Let $\hat{\mu}^m = \hat{\mu}^m(Y)$ be the least-squares projection of the observed Y for each model m. Its risk can be decomposed into squared bias and variance via the Pythagorean identity

$$r_m = \mathbb{E}\,\|\hat{\mu}^m - \mu\|^2 = \|\mu^m - \mu\|^2 + d_m\sigma^2 \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm, μ^m is the projection of the true mean μ into the subspace m, and the expectation is taken with respect to the sampling distribution of Y given μ. Thus, if μ is close to the subspace m with d_m small compared to n, then the projection estimator $\hat{\mu}^m$ will have a small risk, perhaps much smaller than $n\sigma^2$.

We now propose a convex combination of these estimators

$$\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}^m$$

where the data-determined weights $w_m = w_m(Y)$ are chosen to give emphasis to models assessed to be better. In particular, for each model m, let $\hat{r}_m$ be an unbiased estimate of the risk of $\hat{\mu}^m$ given by

$$\hat{r}_m = \|Y - \hat{\mu}^m\|^2 + \sigma^2(2d_m - n) \tag{2}$$

in accordance with Akaike [7], [8], Mallows [9], or Stein [4], [5], which means that $\mathbb{E}\,\hat{r}_m = \mathbb{E}\,\|\hat{\mu}^m - \mu\|^2$ for each μ in $\mathbb{R}^n$. Then we define the weights to be

$$w_m \propto \exp\Big[-\beta\frac{\hat{r}_m}{2\sigma^2}\Big], \qquad \beta > 0 \tag{3}$$

normalized to have unit sum over $m \in \mathcal{M}$. The tuning parameter β adjusts the degree of concentration of the weights on the models with small risk estimates. The two extremes are $\beta \to 0$, which gives the uniform distribution on $\mathcal{M}$, and $\beta \to \infty$, which assigns nonzero weights to only the models with minimal estimated risk. Typical values are $\beta = 1$, which gives the weighted mixture a Bayes interpretation, and $\beta = 1/2$, which leads to the main risk bounds.

We will show that the w-averaged risk estimate $\sum_m w_m \hat{r}_m$ is a crucial part of an unbiased risk estimate $\hat{r}$ of the mixture $\hat{\mu}$. In fact, for $\beta \leq 1/2$, it is an upper bound

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

with equality when $\beta = 1/2$. Let $\hat{m}$ be a minimizer of risk estimates satisfying $\hat{r}_{\hat{m}} = \min_m \hat{r}_m$. We will show that the average risk estimate admits the representation

$$\sum_{m \in \mathcal{M}} w_m \hat{r}_m = \hat{r}_{\hat{m}} + \frac{2\sigma^2}{\beta}\big[H(w) + \log w_{\hat{m}}\big] \tag{4}$$

where $H(w) = -\sum_m w_m \log w_m$ is the entropy of the weights w. This identity shows how the interplay between the positive $H(w)$ and the negative $\log w_{\hat{m}}$ terms characterizes how close the w-averaged risk estimate is to the minimal risk estimate $\hat{r}_{\hat{m}}$. In any case, $H(w)$ is upper-bounded by $\log M$ where M is the cardinality of $\mathcal{M}$. In particular, when $\beta = 1/2$

$$\hat{r} < \min_{m \in \mathcal{M}} \hat{r}_m + 4\sigma^2 \log M. \tag{5}$$

Taking expectation, we show that the risk satisfies

$$\mathbb{E}\,\|\hat{\mu} - \mu\|^2 \leq \min_{m \in \mathcal{M}} \mathbb{E}\,\|\hat{\mu}^m - \mu\|^2 + 4\sigma^2 \log M. \tag{6}$$

We have used a risk target

$$r_* = \min_{m \in \mathcal{M}} r_m = \min_{m \in \mathcal{M}} \mathbb{E}\,\|\hat{\mu}^m - \mu\|^2 \tag{7}$$

which corresponds to a model with optimal bias and variance tradeoff. This r_* is the main term in the bound (6) for the risk of the combined estimator. Indeed, the first term on the right-hand side of (1), $\|\mu^m - \mu\|^2 = \sum_{i=1}^n (\mu_i^m - \mu_i)^2$ is a sum of n terms, so typically r_* is much larger than the $\log M$ term in (6) (unless one has the surprisingly good fortune that μ is close to one of the subspaces considered with dimension lower than $\log M$).

It is sometimes useful to incorporate a deterministic factor π_m in the model weight w_m to account for model complexity or model preference, in a manner that facilitates desirable risk properties. Suppose such factors π_m are assigned, where expressing them in the form $\pi_m = \exp(-C_m)$ and requiring that they sum to at most one endows model m with an interpretation of having descriptive complexity C_m. Thus, low-complexity models are favored. The new weights become

$$w_m \propto \pi_m \exp\Big[-\beta\frac{\hat{r}_m}{2\sigma^2}\Big] \tag{3a}$$

where these combined weights w are again normalized to have unit sum. As before, the choice $\beta = 1$ has a Bayes interpretation, and $\beta = 1/2$ leads to the main risk bounds.

As in the case without π, the w-averaged risk estimate is an upper bound for the unbiased risk estimate $\hat{r}$ of this mixture $\hat{\mu}$, formed with the new weights (3a), when $\beta \le 1/2$, with equality at $\beta = 1/2$.

Information theory also elucidates the risk analysis of mixing with this more general form of weights. The average risk estimate admits the following analogous representation:

$$\sum_{m\in\mathcal{M}} w_m \hat{r}_m = \hat{r}_{\hat{m}} + \frac{2\sigma^2}{\beta}\left[C_{\hat{m}} - D(w\|\pi) + \log w_{\hat{m}}\right] \quad (4a)$$

where now

$$\hat{m} = \arg\min_{m\in\mathcal{M}} \left\{\hat{r}_m + \frac{2\sigma^2}{\beta}C_m\right\}$$

is the model with the highest weight, and

$$D(w\|\pi) = \sum_m w_m \log(w_m/\pi_m)$$

is the information divergence between the weights w and π. The $\log w_{\hat{m}} - D(w\|\pi)$ terms in (4a) gauge how close the average risk estimate is to the minimum risk estimate plus complexity $\hat{r}_{\hat{m}} + 2\sigma^2 C_{\hat{m}}/\beta$. When $\beta = 1/2$

$$\hat{r} < \min_{m\in\mathcal{M}} \left\{\hat{r}_m + 4\sigma^2 C_m\right\}. \quad (5a)$$

Moreover, the following risk bound is shown to hold:

$$\mathbb{E}\,\|\hat{\mu} - \mu\|^2 \le \min_{m\in\mathcal{M}} \left\{\,\mathbb{E}\,\|\hat{\mu}^m - \mu\|^2 + 4\sigma^2 C_m\right\}. \quad (6a)$$

The right side, expressed via (1) as

$$\min_{m\in\mathcal{M}} \left\{\,\|\mu^m - \mu\|^2 + d_m\sigma^2 + 4\sigma^2 C_m\right\}$$

is an index of resolvability of μ by the model class $\mathcal{M}$ which calibrates the mixture estimator by the best tradeoff in approximation, dimension, and complexity (corresponding to the three terms, respectively) among the models in $\mathcal{M}$.

Note that the $\log M$ terms in (5) and (6) (excess beyond the minimum) are now subsumed under the C_m terms in (5a) and (6a). Not surprisingly, the latter recovers the former when the uniform model weights $\pi_m = 1/M$ are used, but in this case, we will show tighter bounds in Section IV due to a technical refinement.

B. Background

Essential to the concept of Bayes mixtures are Bayesian interpretations of individual least-squares regressions, which date back to ideas of Bayes, Laplace, and Gauss. In particular, the linear least-squares projections in Gaussian models arise as the Bayes estimators with (improper) uniform prior on the coefficients of linear combinations. Each associated posterior weight for such a model is proportional to $\exp\{-\|Y - \hat{\mu}^m\|^2/(2\sigma^2)\}$, times a function of the model dimension d_m. The heights of the uniform priors (with infinite total mass) are arbitrary. These heights do not affect the individual Bayes estimators, but they do lead to ambiguous posterior weights. To resolve this ambiguity, Hartigan [10] assigns these prior weights based on hypothesis testing interpretations and arranged the posterior weights to be $\exp[-\hat{r}_m/(2\sigma^2)]$ (normalized to have unit sum), favoring the models with lower risk assessment $\hat{r}_m$. See also Buckland *et al.* [11] for numerical evaluations with these weights.

Demonstration of detailed risk properties of weighted regressions has been challenging. Analogous information-theoretic bounds for Bayes predictive density estimation (or Cesaro averages thereof) have been developed by Barron [12], [13], Catoni [14], and Yang [15], [16]. We call attention to Yang [1, Sec. 2.6] where he gives an exponential form of weights (with arbitrary β), which, when his theory is specialized to Gaussian errors, produces the weights $\exp[-\beta\hat{r}_m/(2\sigma^2)]$ we use here. Catoni [17] and Yang [1] give oracle inequalities similar to ours for prediction mean-squared error 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 $\hat{\mu}^m$. In contrast, the analysis technique employed in this paper allows use of all the data, and all at once in constructing both the weights and the estimates.

To achieve such bounds, we give an unbiased risk assessment of the combined estimator with weights $\exp[-\beta\hat{r}_m/(2\sigma^2)]$ for arbitrary $\beta > 0$. The choice $\beta = 1$ produces Bayes procedures. The best bounds via our technique occur with $\beta = 1/2$.

George [18], [19] also studied mixing estimators, with emphasis on Stein's shrinkage estimators, which are nonlinear, and provided an expression for the risk estimate of the mixture using Stein's result [5]. The form we give here has an explicit interpretability that leads to risk bounds for the applications to mixing least-squares estimators. Mixtures of shrinkage estimators using similar techniques are also analyzed in [6].

II. Unbiased Risk Assessment

As above, we have $Y \sim \mathcal{N}ormal(\mu, \sigma^2 I)$ in $\mathbb{R}^n$ and for each model $m \in \mathcal{M}$, we have an estimator $\hat{\mu}^m = \hat{\mu}^m(Y)$. Typically, each estimator is tied to various explanatory variables given in a design matrix via a functional model. In Section II-A, we give expressions for the risk estimates of general mixture estimators composed of arbitrary estimators (not necessarily linear). We propose a special form of weights that simplify the expression for the mixture risk estimate in Section II-B. Finally, we will apply the general risk estimate results to the case of linear models and least-squares in Section II-C.

An important realization is that, unlike Akaike's information criterion (AIC) [8] which gives an unbiased risk estimate only for each model separately, Stein's identity [4], [5] can be applied more generally to provide an unbiased estimator of the risk of a mixture estimator.

We shall use

$$\sigma^2 = 1$$

for Sections II–VI for notational simplicity.

A. Risk Assessment for General Mixture

We use the notation $a \centerdot b = \sum_{i=1}^n a_i b_i$ for the inner product of vectors a and b and ∇ for the gradient where $\nabla_i = \partial/\partial Y_i$.

Suppose for each m, the estimator $\hat{\mu}^m$ is almost differentiable in Y (that is, its coordinates can be represented by well-defined integrals of its almost-everywhere derivatives $\nabla_i \hat{\mu}_i^m$, which is implied by continuity together with piecewise differentiability) and that $\nabla_i \hat{\mu}_i^m$ have finite first moments. Then Stein [4], [5] gives an unbiased estimate $\hat{r}_m$ for the risk $r_m = \mathbb{E}\, \|\hat{\mu}^m - \mu\|^2$, i.e., $\mathbb{E}\, \hat{r}_m = r_m$ for each μ.

Our goal is to give an unbiased risk estimate for the mixture

$$\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}^m$$

where the weights $w_m(Y)$ are nonnegative, sum to one, and almost differentiable. We further assume that $\mathbb{E}\, |(\nabla_i w_m) \hat{\mu}_i^m|$ are finite. We also suppose $\mathcal{M}$ is finite (though under mild conditions, the conclusions can be extended for infinite $\mathcal{M}$). The following theorem relates the unbiased assessment of the risk of $\hat{\mu}$ to unbiased assessments of the risks of the individual estimators $\hat{\mu}^m$.

Theorem 1: With the above assumptions, an unbiased estimate of the risk $r = \mathbb{E}\, \|\hat{\mu} - \mu\|^2$ of the mixture $\hat{\mu} = \sum_m w_m \hat{\mu}^m$ is given by

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_m - \|\hat{\mu}^m - \hat{\mu}\|^2 + 2(\nabla \log w_m) \cdot (\hat{\mu}^m - \hat{\mu}) \Big]. \quad (8)$$

In addition, if

$$w_m(Y) = \frac{\exp(-\rho_m)\pi_m}{\sum_{m'} \exp(-\rho_{m'})\pi_{m'}} \quad (9)$$

for almost differentiable $\rho_m = \rho_m(Y)$ and arbitrary constants π_m, then

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_m - \|\hat{\mu}^m - \hat{\mu}\|^2 + 2(\nabla \rho_m) \cdot (\hat{\mu} - \hat{\mu}^m) \Big]. \quad (10)$$

This unbiased estimate of risk (8) has three terms. The principal term, $\sum_m w_m \hat{r}_m$, is the weighted average of the individual risk estimates. This average is a crude risk assessment, possibly biased. However, with suitable design of the weights, we will show that it becomes an upper bound for the unbiased risk assessment $\hat{r}$ for the mixture of least-squares regressions. Also, an information-theoretic representation of this term yields the conclusion that it is not much larger than $\min_m \hat{r}_m$.

The second term, $-\sum_m w_m \|\hat{\mu} - \hat{\mu}^m\|^2$, wonderfully illustrates an advantage of mixing estimators. If the estimates $\hat{\mu}^m$ vary with m, then combining them reduces the unbiased risk assessment by the weighted average of the squared distances of the $\hat{\mu}^m$ from their centroid $\hat{\mu}$. The unbiased risk estimate for the mixture (8) intuitively reveals this reduction based on variability of estimates among a model class (as m varies for a given sample), rather than based on the variance of the estimators (as the sample varies for a fixed model m), which is a motivation for resampling-type estimators.

The third term, $2\sum_m (\nabla w_m) \cdot (\hat{\mu}^m - \hat{\mu})$, quantifies the effect of the data sensitivity of the weights via their gradients with respect to the data Y. Constant weights would make this term zero, but would not permit means to adapt the fit to the models that have smaller $\hat{r}_m$. Finally, the exponential form of weights (9) gives a particularly clean mixture risk estimate (10) that depends on the weights via the gradient of the exponents in the relative weighting only and not the normalization.

If our weights focus on models assessed to be good, then our intuition says that the third term quantifies the price one pays for making the mixture estimator adaptive, so it should have a positive expectation (otherwise, mixing offers a "free lunch"). However, in the corollary in the next section, we will show how to design weights such that this third term can be canceled with the second.

Proof of Theorem 1: According to [4], [5], an unbiased estimate of the risk of any estimator $\hat{\mu}$ is given by

$$\hat{r} = \|\hat{\mu} - Y\|^2 + 2\sum_{i=1}^n \nabla_i \hat{\mu}_i - n \quad (11)$$

as long as each $\nabla_i \hat{\mu}_i$ has finite absolute expectation, but our assumptions are sufficient to ensure this. Now with a variance calculation using the weights w as a distribution on $\mathcal{M}$, summing over each of the coordinates, we rewrite the first term above as

$$\|\hat{\mu} - Y\|^2 = \sum_{m \in \mathcal{M}} w_m \Big[\|\hat{\mu}^m - Y\|^2 - \|\hat{\mu}^m - \hat{\mu}\|^2 \Big].$$

The second term can be expanded via differentiation under the summation sign

$$\nabla_i \sum_{m \in \mathcal{M}} w_m \hat{\mu}_i^m = \sum_{m \in \mathcal{M}} w_m \nabla_i \hat{\mu}_i^m + \sum_{m \in \mathcal{M}} (\nabla_i w_m) \hat{\mu}_i^m$$

and we recognize in these components the terms of

$$\hat{r}_m = \|\hat{\mu}^m - Y\|^2 + 2\sum_{i=1}^n \nabla_i \hat{\mu}_i^m - n. \quad (12)$$

such that

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_m - \|\hat{\mu}^m - \hat{\mu}\|^2 \Big] + 2\sum_{i=1}^n \sum_{m \in \mathcal{M}} (\nabla_i w_m) \hat{\mu}_i^m$$

after exchanging the order of summation over m and i. The last term here is the same as

$$2\sum_{i=1}^n \sum_{m \in \mathcal{M}} (\nabla_i w_m)(\hat{\mu}_i^m - \hat{\mu}_i)$$

because $\sum_m (\nabla_i w_m) \hat{\mu}_i = \big[\nabla_i \big(\sum_m w_m\big)\big] \hat{\mu}_i = 0$ (as the weights w_m sum to a constant). The above display equals $2(\nabla w_m) \cdot (\hat{\mu}^m - \hat{\mu})$ by exchanging the order of summation again and the first claim (8) follows.

For the second claim, $\nabla \log w_m(Y)$ equals $-\nabla \rho_m(Y)$ minus a function (the gradient of $\log \sum_k \exp(-\rho_k)\pi_k$) which does not depend on m. Now since $\hat{\mu} - \hat{\mu}^m$ has w-average being the null vector $\underline{0}$, its inner product with a quantity not depending on m averages to 0 under the weights w, so that we are left with the $\nabla \rho_m(Y)$ term. This proves (10). □

Remark: One can adjust $\rho_m(Y)$ by adding any function of Y that does not depend on m without changing either the value of w_m or the validity of (10).

Remark: The above risk estimate formulae hold coordinatewise. That is, putting $\hat{\mu} = (Y_1, \ldots, Y_{i-1}, \hat{\mu}_i^m, Y_{i+1}, \ldots, Y_n)'$ in (11) yields an unbiased risk estimate for $\hat{\mu}_i^m$

$$\hat{r}_{m,i} = (\hat{\mu}_i^m - Y_i)^2 + 2\nabla_i \hat{\mu}_i^m - 1,$$

such that $\mathbb{E}\, \hat{r}_{m,i} = (\hat{\mu}_i^m - \mu_i)^2$ for each μ_i. Then an unbiased risk estimate for the mixture $\hat{\mu}_i = \sum_m w_m \hat{\mu}_i^m$ is given by

$$\sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_{m,i} - (\hat{\mu}_i^m - \hat{\mu}_i)^2 + 2(\nabla_i \log w_m)(\hat{\mu}_i^m - \hat{\mu}_i) \Big].$$

With weights (9), we can further simplify this to

$$\sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_{m,i} - \|\hat{\mu}_i^m - \hat{\mu}_i\|^2 + 2(\nabla_i \, \rho_m)(\hat{\mu}_i - \hat{\mu}_i^m) \Big]. \quad \square$$

Given a collection of models and its corresponding estimators, we can use Theorem 1 to design data-determined weights w_m that make the unbiased estimate of risk (8) for the mixture small. The weights (9) offer a tractable start, and we can further simplify (10) in certain cases laid out in Section II-B. Our risk bounds developed later is one such application.

A second application of the theorem is evaluation of model classes and their respective mixture estimators, as there can be multiple model classes that meaningfully decompose a common parameter space into various scientifically reasonable models (linear and curved). Provided that we have the component estimators in each model class and weight them appropriately, we can evaluate how effectively each model class explains the data using (8). One can go further with this for model class design. For instance, a goal may be to heuristically choose a collection of models rich enough to cover the considered parameter space, and yet the models are different enough to provide enough variability in their corresponding estimates such that the second term in the right-hand side of (8) offers a large reduction in the unbiased risk estimate (while the third term is controlled).

B. Special Forms of Weights and a Bayesian Interpretation

A special form of weights (9) allows further simplification of the mixture's unbiased risk estimate.

Corollary 2: If the weight exponent $\rho_m(Y)$ has gradient $\beta(Y - \hat{\mu}^m)$ for all $m \in \mathcal{M}$ and some fixed $\beta \geq 0$, then

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_m - (1 - 2\beta)\|\hat{\mu}^m - \hat{\mu}\|^2 \Big]. \tag{13}$$

In addition, if $\beta \leq 1/2$, the risk estimate can be bounded by

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

with equality when $\beta = 1/2$.

Proof: From the stated assumption of the form of $\rho_m(Y)$, we see that after adding a function not depending on m, $\nabla \rho_m(Y)$ matches a multiple of $\hat{\mu} - \hat{\mu}^m$ so the first claim follows from (10) and the first remark in Section II-A. Choosing $\beta = 1/2$ or smaller eliminates the second term. $\square$

We turn our attention to Bayes procedures (strictly speaking, posterior Bayes). Possibly improper prior measures λ_m for μ in $\mathbb{R}^n$ are said to produce proper posterior distributions if the integral of the Gaussian likelihood

$$\int (2\pi)^{-n/2} e^{-\|Y - \mu\|^2/2} \, d\lambda_m(\mu) \tag{14}$$

is finite for each Y and m. In that case, expression (14) is called the marginal density of Y (also known as Bayes factor for m) and is denoted by $p(Y \,|\, m)$; and w_m, proportional to $p(Y \,|\, m)\, \pi_m$, is the posterior probability of model m. Moreover, $\hat{\mu} = \mathbb{E}\,[\mu \,|\, Y] = \sum_m w_m \hat{\mu}^m$ is the Bayes mixture of the individual Bayes estimators $\hat{\mu}^m = \mathbb{E}\,[\mu \,|\, Y, m]$.

Corollary 3: For a Bayes mixture, the unbiased risk estimate (13) holds with $\beta = 1$. That is,

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \Big[\hat{r}_m + \|\hat{\mu}^m - \hat{\mu}\|^2 \Big].$$

Proof: For each fixed m, the (posterior) Bayes estimator satisfies [20, Ch. 4, Theorem 3.2]

$$\begin{aligned} \hat{\mu}^m = \mathbb{E}\,[\mu \,|\, Y, m] &= Y + \mathbb{E}\,[\mu - Y \,|\, Y, m] \\ &= Y + \nabla \log p(Y \,|\, m). \end{aligned} \tag{15}$$

Indeed, having assumed that $p(Y \,|\, m)$ is finite for all Y, differentiation of it under the integration sign (14) is justified for the Gaussian likelihood (cf. [21, Ch. 2, Theorem 9] for a more general result about exponential families) and this permits us to rewrite the posterior expectation of $\mu - Y$ as $\nabla p(Y \,|\, m)/p(Y \,|\, m)$, yielding the last equality in (15). Thus, $\rho_m(Y) = -\log p(Y \,|\, m)$ has gradient $Y - \hat{\mu}^m$ so that (13) holds with $\beta = 1$ by Corollary 2. $\square$

Alternatively, we can heuristically apply Theorem 1 to weights emphasizing models with small risk estimates $\hat{r}_m$

$$w_m = \frac{\pi_m \exp(-\beta \hat{r}_m/2)}{\sum_{m'} \pi_{m'} \exp(-\beta \hat{r}_{m'}/2)}, \qquad \beta > 0 \tag{16}$$

where the positive constants π_m are a mechanism for assigning model preference. That is, we take $\rho_m = \beta \hat{r}_m/2$ in (9). The parameter β controls the relative importance of averaging across models (small β) and picking out the one that is empirically best (large β). The two extremes are $\beta \to 0$, which ignores the observations Y and weights the models by π_m only, and $\beta \to \infty$, which uses only the model(s) with minimal estimated risk.

Intuitive appeal aside, an important motivation for these weights is that, in the case of using least-squares estimators $\hat{\mu}^m$ for linear models m (explored in the next subsection), weights (16) yield further simplification of (10) via Corollary 2. In particular, linear least-squares coincide with (posterior) Bayes estimators (15) when one chooses a prior uniform over (and restricted to) the linear subspace m for each model m. In this case, the posterior probability takes the form of (16) with $\beta = 1$ when prior densities for μ under m (with respect to the Lebesgue measure on m) have relative heights $1/(\sqrt{2\pi e})^{d_m}$ and the prior probabilities for model m are π_m.

Remark: One can also think of the parameter $1/\beta$ as a tuning coefficient for inflating the error variance $\sigma^2 = 1$. We will show that mixing estimators with $\beta = 1/2$, a conservative approach regarding the noise to have twice its actual variance, achieves the best risk bound.

Remark: Mixtures composed of positive-part James–Stein shrinkage estimators using the heuristic weights (16) also prove to have low risks, as shown in [6].

C. Linear Least-Squares

Now we specialize to the case that each model $m \in \mathcal{M}$ is a linear subspace of $\mathbb{R}^n$. The estimator $\hat{\mu}^m$ under such a model is the least-squares projection of the observations Y into the d_m-dimensional linear space, the column space of a design matrix X_m of a subset of explanatory variables. This can be accomplished by Gram–Schmidt procedures, or explicitly via the projection matrix $\mathcal{P}_m = X_m(X_m'X_m)^{-1}X_m'$ such that $\hat{\mu}^m = \mathcal{P}_m Y$.

In essence, combining these least-squares projections produces a shrinkage estimator which draws the observations Y toward the linear models in $\mathcal{M}$. The closer Y seems to be to a certain model m (as assessed by the unbiased estimates of risks of the individual estimators), the more the shrinkage, since the weight w_m for the projection $\hat{\mu}^m$ would be large, drawing the mixture closer to m.

Lemma 4: For each linear model m, the expression assigned to $\hat{r}_m$ in (2)

$$\hat{r}_m = \|Y - \hat{\mu}^m\|^2 + 2d_m - n$$

is an unbiased risk estimate for $\hat{\mu}^m$. Moreover, $\hat{r}_m$ has gradient

$$\nabla \hat{r}_m = 2(Y - \hat{\mu}^m).$$

Proof: It is fruitful to consider an orthonormal basis for $\mathbb{R}^n$ for which the first d_m elements of this basis spans m. A point Y in $\mathbb{R}^n$ can be represented by a linear combination of these basis elements, whose coefficients are obtained by inner products with Y. In other words, there exists an orthonormal matrix Q, a function of m, whose first d_m columns span m. Then Y has a representation QZ, with coefficients obtained as $Z = Q'Y$. Moreover, $Z \sim \mathcal{N}ormal(\theta, I)$, with $\theta = Q'\mu$ and

$$\hat{\theta}^m = Q'\hat{\mu}^m = (Z_1, \ldots, Z_{d_m}, \ 0, \ldots, 0)'$$

is the corresponding least-squares projection in the new coordinate system which simply retains the first d_m elements of Z. Similarly, the projection μ^m of μ has the representation

$$(\theta_1, \ldots, \theta_{d_m}, \ 0, \ldots, 0)'$$

in this system. Then, since the norm is preserved by orthonormal transformations, the risk of $\hat{\mu}^m$ is

$$\begin{aligned} r_m(\mu) &\overset{\text{def}}{=} \mathbb{E}\,\|\hat{\mu}^m - \mu\|^2 \\ &= \mathbb{E}\,\|\hat{\theta}^m - \theta\|^2 \\ &= \sum_{k > d_m} \theta_k^2 + d_m. \end{aligned} \tag{17}$$

With θ^m as the projection of θ into m, the sum above equals

$$\sum_{k > d_m} \theta_k^2 = \|\theta - \theta^m\|^2 = \|\mu - \mu^m\|^2.$$

Thus, we have re-established the Pythagorean identity (1) for the risk

$$r_m = \|\mu^m - \mu\|^2 + d_m. \tag{18}$$

The unbiased risk estimate $\hat{r}_m$ is easily computed in the new coordinate system. From (17), and the unbiasedness of $Z_k^2 - 1$ for θ_k^2 for each k, we deduce that the following is an unbiased estimate for r_m:

$$\hat{r}_m = \sum_{k > d_m} Z_k^2 + 2d_m - n = \|Z - \hat{\theta}^m\|^2 + 2d_m - n.$$

Since Q is a norm-preserving transformation, this shows the first claim, and yields a simple expression for the gradient $\nabla_Z \hat{r}_m$ of $\hat{r}_m$ with respect to Z because

$$\frac{d\hat{r}_m}{dZ_k} = 2Z_k \mathbb{1}_{\{k > d_m\}} = 2(Z_k - \hat{\theta}_k^m),$$

where $\mathbb{1}_{\{k > d_m\}} = 1$ if $k > d_m$ and 0 otherwise. Since the elements Q_{ik} of Q are exactly the derivatives dZ_k/dY_i, applying the multivariate chain rule gives

$$\nabla_Y \hat{r}_m = Q \nabla_Z \hat{r}_m = 2Q(Z - \hat{\theta}^m) = 2(Y - \hat{\mu}^m)$$

and the second claim follows. □

Remark: An alternative proof is to use Stein's identity (12), together with the fact that $\operatorname{tr} \mathcal{P}_m = d_m$ to show that $\hat{r}_m$ is unbiased. Then write

$$\|Y - \hat{\mu}^m\|^2 = Y'(I - \mathcal{P}_m)'(I - \mathcal{P}_m)Y = Y'(I - \mathcal{P}_m)Y$$

where the last equality follows from the fact that $I - \mathcal{P}_m$ is symmetric and also a projection (onto the orthogonal space of m). Then the gradient of (2) is $2(I - \mathcal{P}_m)Y = 2(Y - \hat{\mu}^m)$.

Thus, for linear least-squares estimators, by choosing w_m proportional to $\pi_m \exp(-\beta \hat{r}_m/2)$, the condition for Corollary 2 is satisfied. With these weights at $\beta = 1/2$, the resulting expression in (10) is only the w-average of the unbiased risk estimates $\hat{r}_m$ of the individual models.

This puts us in a setting where we can give simple information-theoretic characterization of the risk assessment for the mixture $\hat{\mu}$.

III. Information-Theoretic Characterization of Risk Assessment

We analyze the average risk estimate $\sum_m w_m \hat{r}_m$ in this section. It is the primary term in the estimate for the risk of the mixture $\hat{\mu}$; and for $\beta \le 1/2$, it is a tight upper bound of the unbiased risk estimate $\hat{r}$ as concluded by Corollary 2.

Remark: When the unknown mean μ can be well-approximated by multiple models m, the resulting risk of the mixture at μ would not be very sensitive to the choice of β around the values of interest at 1 (Bayes) and $1/2$ (clean bound). See Section VI for numerical results.

Since the choice $\beta = 1/2$ makes this average risk estimate unbiased for the risk of $\hat{\mu}$, we will set it so in this section for a brisk exposition. The generalization to any $\beta > 0$ can be obtained by replacing 4 with $2/\beta$, though the average risk estimate will no longer be unbiased when $\beta \neq 1/2$. This will be explicitly done in the next section where a tighter bound is proven for the case with weights (3).

A. Sharp Bounds on Risk Estimate of Mixture

The following enunciates the relationship between the average risk estimate $\sum_m w_m \hat{r}_m$ and the minimum. From now on, let $M = \#\mathcal{M}$ be the cardinality of $\mathcal{M}$.

Theorem 5: (a) For each $m \in \mathcal{M}$, let

$$w_m = \frac{\exp(-\hat{r}_m/4)}{\sum_{m'} \exp(-\hat{r}_{m'}/4)} \tag{19}$$

then with $\hat{m}$ being any model achieving $\hat{r}_* = \min_m \hat{r}_m$, the unbiased risk estimate for $\hat{\mu} = \sum_m w_m \hat{\mu}^m$ satisfies

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \hat{r}_m = \hat{r}_* + 4\Big[H(w) + \log w_{\hat{m}}\Big] \tag{20}$$

$$< \hat{r}_* + 4 \log M. \tag{21}$$

(b) More generally, for each m, let

$$w_m = \frac{\pi_m \exp(-\hat{r}_m/4)}{\sum_{m'} \pi_{m'} \exp(-\hat{r}_{m'}/4)} \tag{22}$$

where $\pi_m = \exp(-C_m)$ and $\sum_m \pi_m \leq 1$. Then here, with $\hat{m}$ being any model attaining $\min_m \{\hat{r}_m + 4C_m\}$ the unbiased risk estimate for $\hat{\mu}$ satisfies

$$\hat{r} = \sum_{m \in \mathcal{M}} w_m \hat{r}_m = \hat{r}_{\hat{m}} + 4\Big[C_{\hat{m}} - D(w\|\pi) + \log w_{\hat{m}}\Big]$$

$$< \min_{m \in \mathcal{M}} \{\hat{r}_m + 4C_m\}. \tag{23}$$

Proof: Part (a) is a special case of part (b) with $\pi_m = 1/M$. For part (b), observe that

$$\hat{r}_m = 4\Big[\log \frac{\pi_m}{w_m} - \log \sum_{m'} \pi_{m'} \exp(-\hat{r}_{m'}/4)\Big]$$

$$= \hat{r}_{\hat{m}} + 4\Big[C_{\hat{m}} - \log \frac{w_m}{\pi_m} + \log w_{\hat{m}}\Big]. \tag{24}$$

Thus, the equality follows by averaging over $m \in \mathcal{M}$ with weights w. The inequality results since $D \geq 0$ and $w_{\hat{m}} < 1$ (the logarithm of the latter is strictly negative). □

Therefore, for the first form of weights (19), the average risk estimate (20) is unbiased for the risk of the mixture $\hat{\mu}$, and can be expressed as the minimum of the individual risk estimates plus a price for mixing, a function of the mixing weights w. If the weights w are concentrated on mostly one model $\hat{m}$, then $H(w)$ is close to zero and the combined risk estimate is very close to the minimum $\hat{r}_*$. In any case, since H is less than the log cardinality of $\mathcal{M}$, the average risk estimate cannot exceed $\hat{r}_*$ by a relatively small amount $4 \log M$. (This bound will be improved in the next section.) Moreover, if there are several, say J, models of m with nearly minimal risk estimates $\hat{r}_{\hat{m}}$, then accounting for those J values in the sum on the right side of (24) shows a further reduction of about $4 \log J$ from the bound (21) for the average risk estimate $\hat{r}$, aptly revealing the advantage of the mixing.

For mixing with general weights including π_m, the average risk estimate $\hat{r}$ is the minimum of the complexity-inflated risk estimate plus a reduction due to mixing, a function of w and π. If the data-dependent weights w differ little from the constant weights π, then the quantity $C_m - D(w\|\pi)$ would be close to its upper bound C_m. Moreover, if there are J models of m with nearly minimal $\hat{r}_{\hat{m}} + 4C_{\hat{m}}$, then the bound (23) can be further reduced by about $4 \log J$ again by examining (24).

Remark: The condition $\sum_m \exp(-C_m) \leq 1$ is of course Kraft's inequality [22] in base e and the model complexity is connected to the length of some codeword (in *nats*) that describes the model. However, our theory does not require such an interpretation.

Characterizing the average risk estimate by the minimum is useful as it leads directly to a risk bound.

B. Risk Bound for Mixing Least-Squares Regressions

Corollary 6: The risk $r = \mathbb{E}\,\|\hat{\mu} - \mu\|^2$ of the mixture of least-squares regressions $\hat{\mu} = \sum_m w_m \hat{\mu}^m$ with weights (19) satisfies

$$r \leq \min_{m \in \mathcal{M}} r_m + 4 \log M$$

where $r_m = \mathbb{E}\,\|\hat{\mu}^m - \mu\|^2$, taking value (18), is the risk of $\hat{\mu}^m$. Mixing with weights (22) yields a risk that satisfies

$$r \leq \min_{m \in \mathcal{M}} \{r_m + 4C_m\}$$

where $C_m = \log(1/\pi_m)$. Thus, the risk function $r = r(\mu)$ is upper-bounded by an index of resolvability

$$\mathrm{res}(\mu) = \min_{m \in \mathcal{M}} \{\|\mu^m - \mu\|^2 + d_m + 4C_m\}. \tag{25}$$

Proof: To show the second inequality, we take the expected value of each side of (23). This recovers the risk r by the unbiasedness of $\hat{r}$ on the left. Applying

$$\mathbb{E} \min_{m \in \mathcal{M}} \{\hat{r}_m + 4C_m\} \leq \min_{m \in \mathcal{M}} \mathbb{E}\,[\hat{r}_m + 4C_m]$$

for the right-hand side yields the second statement, from which the resolvability bound follows from (1). The proof for the first statement is the same. □

Note that the mixture $\hat{\mu}$, its risk estimate $\hat{r}$, and risk r all change with the weights w_m, e.g., from (19) to (22). But the risks for the individual models $r_m(\mu)$ (18) and hence, the risk target $\min_m \hat{r}_m$, depend only on μ and not the weights. So, the r_m in the first two displays of Corollary 6 are identical, whereas the two r are different.

The index of resolvability (25) with which we have bounded the risk expresses an idealized tradeoff among error of approximation $\|\mu^m - \mu\|^2$, dimension d_m, and complexity C_m of the models considered. It provides a theoretical calibration of the error the collection $\mathcal{M}$ of models provides as μ varies over $\mathbb{R}^n$. The approximation error term is a sum of squared errors of approximation for the n means, and is typically the dominant term among the three unless the unknown μ is in, or extraordinarily close to, one of the linear spaces considered. Which of the remaining terms, d_m and $4C_m$, is larger depends on the model that yields the best overall tradeoff, which we will discuss at greater length in Section VII.

IV. A Refined Bound

In this section, we bring to the fore the price of mixing estimators with weights (3) (with constant π_m factors) using an arbitrary $\beta > 0$. In short, we shall tighten our risk bounds by replacing the $\log M$ before with a smaller quantity $\psi(M)$.

Definition 7: Let $\psi = \psi(M)$ be a function in $M \geq 2$ defined by the solution to

$$\psi = \log \frac{M-1}{\psi} - 1. \qquad \square$$

Note that $\psi(M)$ is increasing in M. Also, for each $K > 0$

$$\psi \leq \max\left\{K,\ \log\frac{M-1}{K} - 1\right\} \tag{26}$$

by considering separately whether $\psi \leq K$ or not. Then we can also deduce that $\psi(M) < \log M$ by taking $K = \log M$ (treating $M = 2$ as a special case).

Theorem 8: Given the values $\hat{r}_m$ for a finite collection $m \in \mathcal{M}$ and weights

$$w_m = \frac{\exp(-\beta \hat{r}_m/2)}{\sum_{m'} \exp(-\beta \hat{r}_{m'}/2)} \tag{27}$$

with any $\beta > 0$, the weighted average satisfies

$$\sum_{m\in\mathcal{M}} w_m \hat{r}_m \leq \min_{m\in\mathcal{M}} \hat{r}_m + \frac{2\psi(M)}{\beta} \tag{28}$$

where $M = \#\mathcal{M}$ is the cardinality of $\mathcal{M}$.

Proof: First, observe that

$$\hat{r}_m = \frac{2}{\beta}\Big[\log\frac{1}{w_m} - \log \textstyle\sum_{m'} \exp(-\beta\hat{r}_{m'}/2)\Big]$$

which, upon averaging with w over m, yields

$$\sum_{m\in\mathcal{M}} w_m \hat{r}_m = \hat{r}_* + \frac{2}{\beta}\big[H(w) + \log w_{\hat{m}}\big] \tag{29}$$

where $\hat{m}$ is a model achieving the minimum risk estimate $\hat{r}_* = \min_m \hat{r}_m$. Let $h(p) = -p\log p - (1-p)\log(1-p)$. Then as in the proof of Fano's inequality [22], we have

$$H(w) = (1 - w_{\hat{m}})H(\tilde{w}) + h(w_{\hat{m}})$$

where $\{\tilde{w}_m : m \neq \hat{m}\}$ are the weights w renormalized on $\mathcal{M}\backslash\{\hat{m}\}$. Thus, (29) becomes

$$\sum_{m\in\mathcal{M}} w_m\hat{r}_m - \hat{r}_* = \frac{2}{\beta}\Big[(1-w_{\hat{m}})H(\tilde{w}) + h(w_{\hat{m}}) + \log w_{\hat{m}}\Big].$$

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

$$(1-w_{\hat{m}})\log(M-1) + h(w_{\hat{m}}) + \log w_{\hat{m}}$$

which is concave in $w_{\hat{m}}$ and equals

$$(1-w_{\hat{m}})\Big[\log(M-1) - \log\frac{1-w_{\hat{m}}}{w_{\hat{m}}}\Big]. \tag{30}$$

Setting to zero the first derivative of (30) with respect to $w_{\hat{m}}$, we see that the maximum of the bound occurs at $w_{\hat{m}} = w_\dagger$ satisfying

$$\log(M-1) - \log\frac{1-w_\dagger}{w_\dagger} = \frac{1}{w_\dagger}.$$

Substituting the result back in (30) yields the bound, taking its optimal value at the odds $(1-w_\dagger)/w_\dagger \overset{\text{def}}{=} O_\dagger$ with

$$O_\dagger = \log\frac{M-1}{O_\dagger} - 1,$$

which is $\psi(M)$. $\square$

Thus, how much the risk estimates averaged with weights (27) exceed the minimum risk estimate $\hat{r}_*$ is related to the odds ratio of m not being the model $\hat{m}$ achieving $\hat{r}_*$, where the odds ratio is optimized over the weight $w_{\hat{m}}$.

The quantity $\psi(M)$ is computed for a range of values of M as shown in Fig. 1 and the table below. It gives a noticeable reduction in the risk bound compared to the use of $\log M$ even for moderate M. For large M, one can approximate $\psi(M)$ by $\log M - \log\log M$.

Now we are ready for the refined risk bound.

Corollary 9: If $\hat{\mu}^m$ are least-squares regressions with risk estimates $\hat{r}_m$ in (2), then the unbiased estimate of risk $\hat{r}$ for the mixture estimator $\hat{\mu} = \sum_m w_m\hat{\mu}^m$ using weights (27) with a fixed $\beta \leq 1/2$ satisfies

$$\hat{r} \leq \min_{m\in\mathcal{M}} \hat{r}_m + \frac{2\psi(\#\mathcal{M})}{\beta}.$$

Hence, with r_m as the risks (18) of the individual estimators

$$\mathbb{E}\,\|\hat{\mu} - \mu\|^2 \leq \min_{m\in\mathcal{M}} r_m + \frac{2\psi(\#\mathcal{M})}{\beta}.$$

Proof: Corollary 2 implies that the unbiased risk estimate for $\hat{\mu}$ is upper-bounded by the average risk estimate for this range of β, which in turn is bounded as in (28). This proves the first claim. The second conclusion follows from taking the expected value of each side of (28) and using $\mathbb{E}\,\min_m \hat{r}_m \leq \min_m \mathbb{E}\,\hat{r}_m$. $\square$

The best of these bounds again occurs at $\beta = 1/2$.

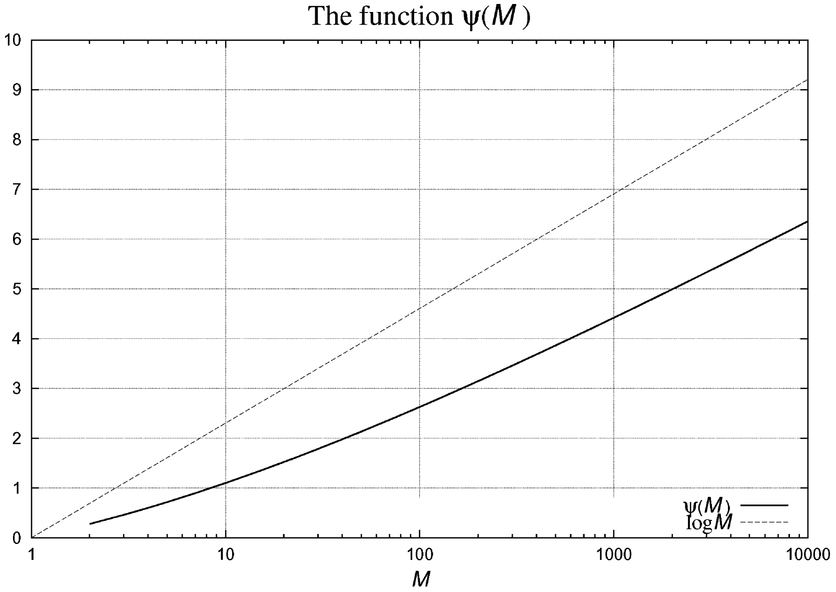

Fig. 1. The term $\psi(M)$ that quantifies the price of mixing M estimators with weights (3) without prior model preferences (π_m are constant).

We compare $\psi(M)$ with $\log M$ (with the coefficient of 4 for the best risk bound) in the following table:

$M = \#\mathcal{M}$	2	5	10	20	40	100	1000
$4 \log M$	2.8	6.4	9.2	12.0	14.8	18.4	27.6
$4\psi(M)$	1.1	2.9	4.4	6.1	7.9	10.5	17.7

We see that the improved bound of order $\psi(M)$ is twice as tight as that of order $\log M$ for $M \leq 20$.

V. Complexity

In this section, we address the choice of the factors π_m in the weights. (These are analogous to prior probabilities of models when $\beta = 1$.) Our bounds assume that $\sum_m \pi_m \leq 1$ and accordingly $C_m = \log(1/\pi_m)$ has an interpretation as a code length, or descriptive complexity, for model m. These factors $\pi_m = \exp(-C_m)$ arise in our risk bounds with $\beta = 1/2$ via the resolvability $\min_m \{r_m + 4C_m\}$.

In general, there may be a very large number of explanatory variables, as may arise from various product basis expansions such as multivariate polynomials. We will say a few words about complexity assignments for such large dictionaries of candidate terms in Section II-B. In what follows, we will focus on the simpler setting of a fixed orthonormal basis of size matching the sample size n for analytical simplifications of the complexities of the approximation errors, and hence of the resolvabilities (as shall be discussed further in Section VII).

A. Fixed Orthonormal Basis

Here we discuss specific complexity assignments in the case of subsets of a fixed sequence of n explanatory variables, as arises in the context of an orthonormal basis $\{\phi_1, \ldots, \phi_n\}$. The $n+1$ leading term models are those spanned by $\{\phi_1, \ldots, \phi_k\}$ for some $k = 0, 1, \ldots, n$; and the 2^n general subset models are those spanned by arbitrary subsets of the basis, treating all subsets of the same size equally.

Since there are fewer leading-term models, we are content to assign them constant complexity, via $\pi_m = 1/(n+1)$ (or any constant not depending on m). This reduces the weights (3a) to (3), and results in bounds such as (5), (6) derived in Section III or Corollary 9 in Section IV, with terms of order $\log n$. Numerical results with leading-term models are given in the next section, but we note here that flexibility in fitting leading-term models to the observed response can be rather limited.

The situation with general subsets is dramatically different with an exponentially large number of models since mixing these with equal weights (3) would render the bound (6) very loose with a term of order $n \log 2$. Instead, we advocate using weights (3a) with

$$C_m = \log(n+1) + \log \binom{n}{d_m}. \tag{31}$$

This corresponds to a descriptive length of $\log(n+1)$ nats for the subset size $d_m = 0, \ldots, n$ and a descriptive length of $\log \binom{n}{d_m}$ nats to distinguish among the subsets of that size. Alternatively, the probabilistic approach is to directly employ

$$\pi_m = \left[(n+1)\binom{n}{d_m}\right]^{-1}$$

to specify a uniform distribution on the cardinality of the subset and a conditionally uniform distribution on the subsets of that size. When d_m is a small fraction of n (desirably yielding a good tradeoff in $r_m = \|\mu - \mu^m\|^2 + d_m$), this complexity is roughly $d_m \log(n/d_m)$, much smaller than n. The information-theoretic interpretation via Kraft's inequality [22] is that for each subset size d_m, no competing code length can be shorter except for a small fraction of such subsets.

Even though mixing all subset models might at first glance seem computationally prohibitive, the Appendix provides a computation shortcut in the orthonormal basis case.

One may also combine the benefits of both arguments with subsets with different structure. Thus, we may set

$$C_m = \begin{cases} \log(n+1) + \log 2, & \text{if } m \text{ is leading-term} \\ \log\binom{n}{d_m} + \log(n+1) + \log 2, & \text{otherwise} \end{cases}$$

to produce a risk bound that is nearly as good as the best of the two, paying a price of at most $\log 2$ nats.

The dimension term d_m in the resolvability

$$\min_m \{\|\mu - \mu^m\|^2 + d_m + 4C_m\}$$

is negligible compared to the complexity when general subsets are involved. However, when leading-term models have small enough approximate error (that the best resolvability favors them) one sees that the complexity term (of order $\log(n+1)$) can be negligible compared to the dimension d_m, and then the resulting risk tradeoff is not encumbered with multiplicative $\log n$ factors. Implications for this remark will be further discussed next.

B. Large Dictionaries

It can be quite natural for a very large number of candidate basis functions to be available, potentially much larger than n, especially in multivariate settings in which one is modeling nonlinear functions of several variables. For instance, suppose the candidate basis functions of D variables are formed as products based on a countable list of basic one-dimensional functions. Using the first L of such basis functions in each of the variables produces L^D candidate product basis functions. These arise directly in polynomial and trigonometric expansions (and, similarly, in neural net and multivariate wavelet models). So for each $L = 1, 2, \ldots$, models m consist of arbitrary subsets of size $d_m = k$ of these L^D product basis functions, for $k = 0, 1, \ldots, \min\{L^D, n\}$. The associated dictionary of models has a combinatorially large number $\binom{L^D}{k}$ of such subset models for each L and k. We may assign complexity such as

$$C_m = 2\log(k+1) + 2\log L + \log\binom{L^D}{k}$$

for which $w_m = \exp(-C_m)$ is summable over models m indexed by k and L. Because our risk bound depends on the combinatorial term via the logarithm only, a useful risk bound results as long as accurate subset models are available for the target, with $kD\log L$ small compared to n, even though the number of candidate predictors L^D may be much larger than n. However, whether there is a way to compute such provably accurate estimators in subexponential time is doubtful.

VI. An Example With Leading-Term Models

We will show numerical results of the risks of our mixture estimators in the fixed orthonormal basis case in this section.

Consider the $n+1$ nested leading-term models from an orthonormal design. Using $Z = (Z_1, \ldots, Z_n)$ as the coefficients of the basis functions obtained by taking their inner products with Y, we have the canonical setting in which Z is distributed $\mathcal{N}ormal(\theta, I)$ (as in the proof of Lemma 4). Each leading-term model m with dimension d_m posits that $\theta_k = 0$ for $k > d_m$, where d_m ranges from 0 to n. In this case, the least-squares estimators under these models are simply

$$\hat{\theta}_k^m = Z_k \mathbb{1}_{\{k \le d_m\}}. \tag{32}$$

Our discussion will proceed in this suitably transformed space of θ, with emphasis on a moderate problem dimension $n = 20$.

Recalling that the variance σ^2 in each dimension is 1, the naïve maximum-likelihood estimator (for the full model) has a risk of n. The best risk upper bound is obtained with the mixture estimator using $\beta = 1/2$, and is $4\psi(n+1)$ (≈ 6.2 for $n = 20$) beyond the risk target

$$r_*(\theta) = \min_{0 \le k \le n} \Big[k + \sum_{j > k} \theta_j^2\Big]. \tag{33}$$

Simulations with various θ and n show that this margin from the target always seems less than $\log n$ (≈ 3 for $n = 20$), so there is room for improving our risk bounds.

Here we will illustrate a case where the true parameter θ indeed belongs to one of these leading-term models. In particular, only the first 10 elements of θ are nonzero. (If the θ_k's are Fourier-type coefficients where k has a frequency interpretation, then θ, the signal to be estimated, is "ideal low-pass" with a "bandwidth" of 10.) We vary $\|\theta\|^2$ (total signal-to-noise ratio) while restricting the nonzero coefficients θ_k to have constant magnitude. (By symmetry, all risk quantities of interests depend on any coefficient θ_k via θ_k^2 only.) Hence, the true parameter can be described by

$$\theta_k^2 \propto \mathbb{1}_{\{k \le 10\}}$$

or $\theta_k^2 = 1/10\, \|\theta\|^2 \mathbb{1}_{\{k \le 10\}}$ to be more precise. The risk target (33) reduces to

$$r_*(\theta) = \min\{\|\theta\|^2, 10\}.$$

In confirming this target, note that if $\|\theta\|^2 < 10$, we are better off leaving out all the terms (i.e., $k = 0$), since the bias so incurred is less than the variance of 10 if we included them; whereas if $\|\theta\|^2 > 10$, then the best k is seen to be 10.

Any mixture of these leading-term estimators (32) with weights w_m will have the form $\hat{\theta}_k = c_k Z_k$ where the data-driven coefficients

$$c_k = \sum_{m: d_m \ge k} w_m$$

are between 0 and 1 and monotonically (strictly) decreasing in k (for w_m strictly positive). We have examined both choices of $\beta = 1/2, 1$ in our mixture estimator, as they correspond to the estimator with the tightest risk upper bound and a Bayes procedure. In addition, we also examine the AIC model selection estimator (mixture with $\beta \to \infty$) for comparison. The performance of the mixture estimator is not very sensitive to the choice of β between $1/2$ and 2.

Fig. 2 says that all three estimators have risks just over 2 worse than the target at small and large $\|\theta\|$, but the mixtures ($\beta = 1/2, 1$) even beat the target around $\|\theta\|^2 = 10$. The AIC model selection estimator is often worse than the mixtures. In fact, the advantage of mixing over selection seems uniform

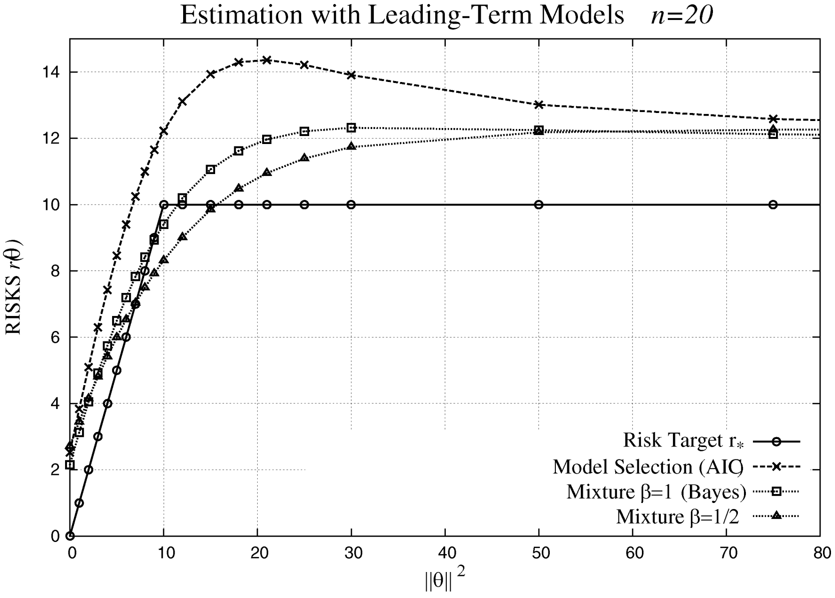

Fig. 2. Risks and target with blockwise constant $\theta_k^2 \propto \mathbb{1}_{\{k \leq 10\}}$.

(over the entire parameter space) in the Bayes case $\beta = 1$, and almost uniform for the $\beta = {}^1\!/_2$ mixture (AIC is slightly better at the origin $\theta = 0$). The risks of all three estimators are similar for large $\|\theta\|^2$. This is expected since the true θ is in one of the models considered (the one with $d_m = 10$). Indeed, when the signal-to-noise ratio $\|\theta\|^2$ is large, AIC picks the correct model with high probability, while the adaptive weights in our mixture give strong emphasis on the right model.

Note that the mixture with $\beta = {}^1\!/_2$ outperforms the Bayes mixture for a large range of $\|\theta\|^2$ between 2.5 and 55. Thus, besides analytical convenience, using $\beta = {}^1\!/_2$ indeed provides nontrivial risk advantage over Bayes $\beta = 1$ in some cases.

VII. Approximation and Resolvability

This section exhibits classes of behavior for the true coefficients θ that permit control of the approximation error arising in our resolvability bound on the risk of the mixture estimator. The point is to observe how the mixture simultaneously adapts to multiple such classes, and to differentiate when certain types of mixtures are suitable. For example, leading-term mixtures are appropriate for cases with ellipsoidal controls on θ (in which the axis widths decay), and general subset mixtures are appropriate when measures of the sparsity of the coefficients θ are controlled (regardless of their order).

To facilitate discussion of approximation and risk on a standardized scale, we shall use the average squared error $\|\hat{\mu} - \mu\|_n^2$ as the loss function, where $\|\cdot\|_n^2 = \sum_{i=1}^n (\cdot)_i^2/n$ (with division by n). The risk $r_m = \mathbb{E}\,\|\hat{\mu}^m - \mu\|_n^2$ of the least-squares estimator $\hat{\mu}^m$ for model m is $\|\mu^m - \mu\|_n^2 + d_m/n$, and likewise, the risk $r = \mathbb{E}\,\|\hat{\mu} - \mu\|_n^2$ of the combined estimator $\hat{\mu}$ is bounded by the index of resolvability

$$\begin{aligned} r &\leq \min_{m \in \mathcal{M}} \{r_m + 4C_m/n\} \\ &= \min_{m \in \mathcal{M}} \left\{ \|\mu^m - \mu\|_n^2 + \frac{d_m}{n} + \frac{4C_m}{n} \right\} \qquad (34) \end{aligned}$$

trading off among approximation error, dimension relative to sample size, and complexity relative to sample size.

Recall that our models are the linear subspaces m spanned by a subset of the orthonormal basis vectors $\{\phi_1, \ldots, \phi_n\}$ with $\|\phi_k\|_n^2 = 1$ for each $k \leq n$ (e.g., these may arise from evaluation of a function at given input values $x_1, \ldots, x_n$). For convenience, we abuse notation by identifying m with the set of all indices k such that ϕ_k is a basis vector for m, i.e., $\{k : \phi_k \in m\}$. Thus, the best approximation to $\mu = \sum_k \theta_k \phi_k$ in m is $\mu^m = \sum_{k \in m} \theta_k \phi_k$, keeping only the terms in m. The resulting approximation error is

$$\|\mu^m - \mu\|_n^2 = \sum_{k \notin m} \theta_k^2.$$

A. General Subsets and Adaptation to Sparsity

Let $\mathcal{M}$ consist of all subsets of $\{\phi_1, \ldots, \phi_n\}$. Here we assign for general subset models (31) a complexity C_m which depends on the subset m only through its dimension d_m.

When performing the minimization for each dimension d, the smallest approximation error occurs when m is the model consisting of the d largest magnitude coefficients. Thus, we denote by $\{\theta_{(j)}\}$ the coefficients $\{\theta_k\}$ sorted descendingly as such, $|\theta_{(1)}| \geq |\theta_{(2)}| \geq \cdots \geq |\theta_{(n)}|$. Consequently, the index of resolvability takes the form

$$\mathrm{res}_n(\theta) = \min_{d \geq 0} \left\{ \sum_{j > d} \theta_{(j)}^2 + \frac{d}{n} + \frac{4C_n(d)}{n} \right\}$$

where we have rewritten (31), the complexity C_m for a model m with dimension $d = d_m$ as

$$C_n(d) = \log(n+1) + \log \binom{n}{d}. \qquad (35)$$

The mixture estimator is constructed without knowledge of the subsets for which the true coefficients are largest. Nevertheless, it achieves risk

$$r_n(\theta) \le \mathrm{res}_n(\theta), \qquad \text{for all } \theta \in \mathbb{R}^n.$$

To enunciate the relationship between sparsity and approximation of θ, we define a sparsity index

$$\|\theta\|_s^s = \sum_k |\theta_k|^s, \qquad 0 < s \le 2$$

and (by taking the limit $s \searrow 0$) denote

$$\|\theta\|_0^0 = \#\{\theta_k : \theta_k \ne 0\}$$

as the number of the nonzero elements in θ. If $\|\theta\|_s^s$ is not large for some $s \in [0,2)$, then general subset models permit control of the approximation error as a function of the number of coefficients d. For example, if θ is such that $|\theta_{(j)}| \le K/j^2$ for each j and some $K > 0$, then we can control its sparsity index for all $s > 1/2$, whereas if there are few nonzero elements in θ, then we can control its sparsity index all the way down to $s = 0$.

Lemma 10: With $\{\theta_{(j)}\}$ being the elements of θ reordered in descending magnitude, we have

$$\sum_{j>d} \theta_{(j)}^2 \le \frac{\|\theta\|_s^2}{(d+1)^{(2-s)/s}}, \qquad 0 < s \le 2.$$

Proof: Since the $|\theta_{(j)}|^s$ sum to $\|\theta\|_s^s$ and are nonincreasing, we have $|\theta_{(j)}|^s \le \|\theta\|_s^s/j$. Write

$$\sum_{j>d} \theta_{(j)}^2 = \sum_{j>d} |\theta_{(j)}|^{2-s} |\theta_{(j)}|^s$$

and use the inequality $|\theta_{(j)}| \le \|\theta\|_s/(d+1)^{1/s}$ in the first factor inside the sum to yield the bound. □

Consider $s = 1$, for instance. This is the case that the unknown μ, suitably scaled, is in the convex hull of $\{\pm\phi_k\}$. Then the approximation error bound is $\|\theta\|_1^2/(d+1)$. If s is smaller, e.g., $1/2$, then one has a faster decay in d for the error bound, $\|\theta\|_{1/2}^2/(d+1)^3$. For $s = 0$, the approximation error vanishes when the model dimension exceeds the number of nonzero elements such that we may put $d = \|\theta\|_0^0$, which with $\binom{n}{d} \le d \log n$ yields the risk bound

$$r_n(\theta) \le \|\theta\|_0^0 \frac{1 + 4\log n}{n} + \frac{4\log(n+1)}{n}.$$

Putting the ingredients together we have a result which says that the mixture estimator, formulated without specification of the sparsity index $\|\theta\|_s^s$, estimates as well as if one knew in advance which index s produces the best tradeoff in approximation error and dimension plus complexity.

Theorem 11: The risk of the mixture of all subset models with weights (3a) and complexity (35) satisfies

$$r_n(\theta) \le \min_{d \ge 0} \left\{ \frac{\|\theta\|_s^2}{(d+1)^{(2-s)/s}} + \frac{d}{n} + \frac{4C_n(d)}{n} \right\}$$

for each $s < 0 \le 2$. Moreover

$$r_n(\theta) \le \min_{s \in [0,2]} \left\{ 2\|\theta\|_s^s \left(\frac{1 + 4\log n}{n} \right)^{1-s/2} + \frac{4\log(n+1)}{n} \right\}.$$

Proof: The first line is by Theorem 5(b) together with the approximation bound from the previous lemma. For the second line, use $C_n(d) \le d\log n + \log(n+1)$ to show that the bracketed bound holds for each $s \in [0,2]$, which for $s = 0$ is immediate from the comment above (with an inflation by a factor of 2). For $s > 0$, we optimize the right hand side of

$$r_n(\theta) \le \min_{d \ge 0} \left\{ \frac{\|\theta\|_s^2}{(d+1)^{(2-s)/s}} + \frac{d+1}{n} + \frac{4C_n(d)}{n} \right\}$$

over d. In particular, putting

$$d + 1 = \left[\left(\frac{2-s}{s} \right) \frac{\|\theta\|_s^2 \, n}{1 + 4\log n} \right]^{s/2}$$

(rounded down to an integer) yields the bracketed bound. Now the stated conclusion follows after minimization over s. □

Remark: For $s = 1$, the appropriate bound $\|\mu^m - \mu\|_n^2 \le \|\theta\|_1^2/d$ is known to hold for when $\mu = \sum_k \theta_k \phi_k$ not only when the ϕ_k are orthonormal but in fact for any basis functions with $\|\phi_k\|_n^2 \le 1$, as shown in [23], [24]. It follows in this case that the risk of the mixture over all subsets satisfies

$$r_n(\theta) \le \mathrm{res}_n(\theta) \le 2\|\theta\|_1 \left(\frac{1 + 4\log n}{n} \right)^{1/2} + O\left(\frac{\log n}{n} \right).$$

This extends the results available for the risk of selection criteria in this convex hull setting from [15], [25]–[28], [2] to the mixture estimator. The results in these references are primarily cast in multivariate settings where there is an exponentially large dictionary of candidate basis functions and where training data tends to be sparse so that risk bounds are perhaps better cast for random designs (new inputs are independent from but identically distributed as training data). The recent work [29] takes a step to develop analogous conclusions for the more challenging case with sparsity indices $1 < s < 2$ for nonorthogonal candidate basis functions.

In the case of wavelet models with wavelet coefficients $\theta_{j,l}$ on each translate j and level l, natural conditions on the coefficients, expressed via bounds on $\sum_j |\theta_{j,l}|^s$ on each level l, correspond to certain Besov spaces. Similar risk bounds for model selection procedures are given in [30]. Analogous conclusions are possible for mixture estimators by our techniques here. For certain problems with piecewise-constant models, the logarithmic factor in the risk is necessary [30].

If instead of having the nonzero elements of θ scattered throughout the indices, it happens that the $|\theta_k|$ are bounded by a decreasing function in k, then mixtures of leading-term models can avoid the logarithmic factor, as this is a generic phenomenon of certain ellipsoidal classes of functions (discussed next).

To summarize the story for general subsets of basis vectors from a dictionary, we have, in this case, that the complexity, essential to the risk bounds, is larger than the dimension of the models. Small approximation error by models of moderate dimension requires adaptation of subsets, and one achieves these good approximations in optimal balance with complexity by mixing estimators over these models.

B. Leading-Term Models and Adaptation to Ellipsoids

Next, we consider models in which the subsets of terms arise in prescribed forms. Those models have complexity smaller than dimension and are also important in theory and applications. Among the simplest such models are those of leading-term type such as polynomials (of adjustable degree) and truncated Fourier series (of adjustable maximal frequency). These linear models are indexed by $m = \{1, 2, \ldots, k\}$ with dimension $d_m = k \leq n$.

The model complexity can be set to either $C_m = \log^* d_m$ with $\log^* d = \log(d+1) + 2\log\log(d+1)$ to slightly favor small models, or $C_m = \log(n+1)$ which gives uniform weights. We need not restrict the models to be nested. For instance, polynomial splines on equal spaced knots provide a sequence of models indexed by (k, r), where k is the number of knots and r is the degree of the local polynomials, and we may set $C_m = \log^* k + \log^* r$.

In these cases, the complexity is seen to be of smaller order than the dimension (which we allow to be large to improve approximation error). Now when the complexity is negligibly small compared to the dimension, the interpretation of the resolvability simplifies to just the optimal tradeoff between squared bias and variance among the linear models. This is preferred for cases in which a good approximation is achieved without taking all subsets of terms.

For example, suppose $(\underline{\phi}_k)_{k \leq n}$ are orthonormal basis functions and that the mean $\underline{\mu} = \sum_k \theta_k \underline{\phi}_k$ is in an ellipsoidal (also called Sobolev) class $\mathcal{E}_{a,b}$ which is the collection of points θ in $\mathbb{R}^n$ such that $\sum_k \theta_k^2 a_k^2 \leq b^2$, where $(a_k^2)_{k \leq n}$ is an increasing sequence. Now the leading-term model which stops at dimension m provides an approximation $\mu^m = \sum_{k \leq m} \theta_k \underline{\phi}_k$ for which the approximation error $\|\mu^m - \mu\|_n^2 = \sum_{k > m} \theta_k^2$ is bounded by b^2/a_{m+1}^2 uniformly for points in $\mathcal{E}_{a,b}$. Adding the variance term m/n and minimizing over m yields a risk minimum $r_* = \min_m \{b^2/a_{m+1}^2 + m/n\}$, which is known to be the minimax rate over all possible estimators for each such ellipsoid $\mathcal{E}_{a,b}$ (see, e.g., [30]).

For example, when $a_k^2 = k^{2s}$ (as arise in characterizing Sobolev classes using Fourier series), we recover the rate $C_s b^{2s/(2s+1)} n^{-2s/(2s+1)}$ optimal with respect to b and n, as laid out in Pinsker [31] (though our bound based on adaptive mixing of least-squares projections reflects a possibly larger constant than that with optimal Pinsker filtering). Note that in the construction of the mixture there is no presumption of any particular regularity sequence (a_k), smoothness index s, or size of ball b. The mixture across model dimensions is adaptive in that, in providing risk bounded by the risk of the best linear model, for each μ, it will be simultaneously minimax rate-optimal for all ellipsoids $\mathcal{E}_{a,b}$ (all a and b). Beran and Dümbgen [32] has another approach (see also the discussion in [33]).

C. Asymptotic Optimality and Improved Oracle Inequalities

The adaptation ability of our mixture estimator is quite general: a sequence of linear models m which are not necessarily nested and not necessarily built from orthonormal terms. The cleanness of the resolvability bound, with constant multiplier of 1 for the squared bias and dimension terms, provides an oracle inequality that exhibits already in finite samples the type of optimality previously studied in asymptotic settings. For example, Shibata [34], Li [35], and others have shown that estimators based on certain model selection criteria are risk ratio-optimal. In particular, the ratio of risk relative to the minimum of risks over all size models converges to 1 as $n \to \infty$ for fixed sequences of means μ, provided the sequence is such that $nr_*(\mu) \to \infty$ as $n \to \infty$, and provided that the log-cardinality of models of each dimension d is of a lower order than d. However, that convergence is not uniform in μ.

We provide a similar result here for out mixture of estimators. For sample size n (on which all risk quantities implicitly depend), let the risk of the least-squares estimator for model m be $r_m(\mu) = \|\mu^m - \mu\|_n^2 + d_m/n$. Then our combined estimator achieves risk $r(\mu) \leq \min_m \{r_m(\mu) + 4C_m/n\}$, which is in turn less than $\min_m \{r_m(\mu) + 4C_m/n : \ d_m \geq \gamma C_m\}$, where in the latter, we have restricted our attention to the models with dimensions greater than a multiple $\gamma > 0$ of their complexities. Thus, relative to the risk target

$$r_*^\gamma(\mu) \overset{\text{def}}{=} \min_m \{r_m(\mu) : \ d_m \geq \gamma C_m\}$$

our mixture achieves a risk ratio

$$\frac{r(\mu)}{r_*^\gamma(\mu)} \leq 1 + \frac{4}{\gamma}$$

uniformly in μ for each γ, such that the ratio can be arbitrarily close to one. To see this result in a setting similar to that of [34], [35], suppose for a fixed sequence of μ, the models achieving the target $r_* = \min_m r_m$ have dimensions that grow unboundedly, yet the complexities of these models are of a smaller order than their dimensions. Then $r_*(\mu)/r_*^\gamma(\mu)$ converges to 1 for each γ, and hence $r(\mu)/r_*(\mu)$ converges to 1 as well.

In any case, the risk of the combined estimator is never worse than the best risk among the models for which the complexity is negligible compared to the dimension. More precisely, this can be quantified using a multiplicative constant of 1 for the risk target plus a term for the complexity relative to n as in the resolvability bound (34). For a similar spirit of oracle inequalities but with larger multiplicative constants, see works by Birgé, Massart, and Barron [30], [36], [37] for model selection in least-square regressions; Donoho and Johnstone [38]–[41] for shrinkage estimation in orthonormal basis; and Devroye and Lugosi [42] for density estimation; Yang [1] for prediction; Wegkamp [43] for L_1 risk in regression; and Juditsky *et al.* [28] and Tsybakov [2] for function aggregation in regression.

D. Summary of Approximation Tradeoffs

Whether it is better to use all subset models or complete models of various orders in regression depends on the nature of the unknown target μ. If coefficients (θ_k) in a suitably transformed representation are scattered throughout the indices, then the target requires all subsets associated with sparse approximations, achieving good risk properties when mixed with weights that account for appropriate model complexity. On the other hand, if the magnitude of (θ_k) decays with k, typical of those in ellipsoid classes, then mixtures of nested leading-term projections can achieve the best tradeoff in approximation and dimension, with a small model complexity penalty. If one does not know in advance which of the two settings is more appropriate for a case at hand, then they may be combined, adding only a $(\log 2)/n$ price to the complexity terms. The resulting estimator achieves risk corresponding to the best tradeoff in approximation, dimension, and complexity.

Appendix

Given a fixed orthonormal basis of size n with all 2^n subset models, here we examine computation for the all-subset mixture. At first, it might seem impractical to combine so many components as the mixture involves calculating all 2^n associated least-square fits and their respective weights. But we provide an alternative route in obtaining this mixture with simplified computation due to a Bayesian interpretation.

Let $Z_1, \dots, Z_n$ be the inner products of the data vector Y with the orthonormal basis vectors $\phi_1, \dots, \phi_n$, which provide the coefficients for the representation of Y in this full basis. The least-squares estimator for any subset model simply zeros out the coefficients for the variables outside the given subset. Consequently, the coefficients in the representation of the combined estimator are given by $\hat{\theta}_k = \hat{c}_k Z_k$ where $\hat{c}_k = \sum_{m \ni k} w_m$, between 0 and 1, are weights aggregated from the models which include term k. As in (3a), w_m is proportional to $\pi_m \exp(-\beta \hat{r}_m/2)$. Here we provide, for certain natural π_m, more direct means to compute the filter coefficients $\hat{c}_k$ that does not require summing w_m over the exponentially many models including term k.

Toward this end, we first note that the factor $\exp(-\beta \hat{r}_m/2)$ equals a constant times the product $\prod_{k \notin m} e^{-\beta(Z_k^2/2-1)}$, which we may also write as

$$p_\beta(Z \,|\, U) = \prod_{k=1}^{n} \exp[-\beta(Z_k^2/2 - 1)(1 - U_k)]$$

where $U = U(m) = (U_1, \dots, U_n)$ with $U_k = \mathbb{1}_{\{k \in m\}}$ as either 1 or 0 depending on whether m includes k (and, hence, $\sum_k U_k = d_m$). Here, U in $\{0,1\}^n$ provides a standard alternative way to refer to subsets m of $\{1, \dots, n\}$. The notation $p_\beta(Z \,|\, U)$ arises from a probabilistic interpretation we shall come to shortly. Denoting $\pi_m = \pi(U)$ we may write

$$\hat{c}_k = \sum_{m \ni k} w_m = \frac{\sum\limits_{U: U_k = 1} p_\beta(Z \,|\, U) \pi(U)}{\sum\limits_{U'} p_\beta(Z \,|\, U') \pi(U')}.$$

The point we want to make here is that if $\pi_m = \pi(U)$ is expressible as a mixture distribution for U over some hidden parameter q, as in $\pi(U) = \int p(U \,|\, q)\, p(q)\, dq$, then in calculating the numerator and the denominator of the preceding expression, we may exchange the order of the sums and the integrals. For instance, the denominator above becomes

$$\int \Big[\sum_U p_\beta(Z \,|\, U)\, p(U \,|\, q)\Big]\, p(q)\, dq = \int p_\beta(Z \,|\, q)\, p(q)\, dq.$$

This is the case with the π_m we recommended for all-subset mixtures using complexity (31)

$$\pi_m = \pi(U) = \frac{1}{n+1} \frac{1}{\binom{n}{d_m}} = \int_0^1 q^{d_m} (1-q)^{n-d_m}\, dq.$$

The product form $p(U \,|\, q) = \prod_{k=1}^n q^{U_k}(1-q)^{1-U_k}$ and the binary nature of U_k allows us to express the shrinkage coefficients as $\hat{c}_k = I_k / I$, where

$$I_k = \int_0^1 c_k(q) \ell_n(q)\, dq \quad \text{and} \quad I = \int_0^1 \ell_n(q)\, dq$$

$$\ell_n(q) = \prod_{k=1}^{n} [q + (1-q)\exp(-\beta(Z_k^2/2 - 1))] \tag{36}$$

$$c_k(q) = \frac{q}{q + (1-q)\exp[-\beta(Z_k^2/2 - 1)]}.$$

Note that $c_k \in [0,1]$ is greater or less than q according to whether $|Z_k|^2$ exceeds 2 (an evidence that the true parameter contains term k), and c_k is near 1 for large $|Z_k|$.

We evaluate the $n+1$ integrals I and I_k numerically. This can be done by summing over a fine uniform grid on $q \in [0,1]$, with care taken to note that $\ell_n(q)$ peaks around its maximizer $\hat{q}$. In accordance with standard Laplace approximation of integrals, the grid width should be narrower than order $1/\sqrt{n}$ (order $1/n$ based on n uniformly spaced grid points suffices) so as to ensure that we capture the peak. Also, for large n, such Laplace approximation shows that the shrinkage factors $\hat{c}_k$ are numerically close to $c_k(\hat{q})$. In essence, this is an adaptive shrinkage factor in which the magnitudes of all elements of Z are used to adapt to levels of q that appear to give rise to the individual Z_k.

A probabilistic interpretation emerges when $\beta = 1$, giving rise to a hierarchical model in which each variable, when conditioned on its sole dependent variable, is independent of all other

$$\begin{aligned} q &\sim \text{Uniform } [0,1] \\ U_k \,|\, q &\sim \text{i.i.d. Bernoulli } (q) \\ \theta_k \,|\, U_k &\sim \text{i.i.d.} \begin{cases} \text{point mass at } 0, & \text{if } U_k = 0 \\ \text{Uniform } (\mathbb{R}) \text{ density } h, & \text{if } U_k = 1 \end{cases} \\ Z_k \,|\, \theta_k &\sim \text{i.i.d. Normal } (\theta_k, 1). \end{aligned}$$

Thus, $p(Z_k \,|\, U_k = 0) = \varphi(Z_k)$ where φ is the standard normal density, and $p(Z_k \,|\, U_k = 1) = h$ and

$$p(U_k = 1 \,|\, Z_k, q) = \frac{qh}{qh + (1-q)\varphi(Z_k)}$$

leading to $p(Z \,|\, q) = \prod_{k=1}^n [qh + (1-q)\varphi(Z_k)]$. And the Bayes shrinkage factor $\hat{c}_k = \mathbb{E}[U_k \,|\, Z]$ agrees with expression (36) with the choice $h = 1/(\sqrt{2\pi e})$. See Hartigan [10] and the references therein for Bayesian considerations of this model.

Even for $\beta \neq 1$, one can still interpret all the above quantities probabilistically, with the distribution $\theta_k \,|\, U_k$ scaled by $1/\beta$ and the normal $Z_k \,|\, \theta_k$ having variance $1/\beta$ instead of 1. For example, our best bound occurs at $\beta = 1/2$, meaning that by being twice as conservative about the error variance, we end up mixing across models more indiscriminately. Occasionally, the risk obtained this way is lower than that when $\beta = 1$ (Section VI).

In summary, it is equivalent to consider our estimator either as a mixture across all subsets specified by U (with q integrated out) or as a mixture across q (with U summed out). We have found the former to be more conducive to our risk analysis and the latter more conducive to computation.

Acknowledgment

G. Leung would like to thank J. Hartigan, H. Leeb, and D. Pollard for their useful comments on his Ph.D. dissertation.

References

[1] Y. Yang, "Combining forecasting procedures: Some theoretical results," *Econometric Theory*, vol. 20, pp. 176–222, 2004.

[2] A. B. Tsybakov, "Optimal rates of aggregation," in *Computational Learning Theory and Kernel Machines (Lecture Notes in Artificial Intelligence)*, B. Scholkopf and M. Warmuth, Eds.. Heidelberg, Germany: Springer-Verlag, 2003, vol. 2777, pp. 303–313.

[3] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial (with discussions)," *Statist. Sci.*, vol. 14, pp. 382–417, 1999.

[4] C. Stein, "Estimation of the mean of a multivariate normal distribution," in *Proc. Prague Symp. Asymptotic Statistics*, Prague, Czechoslovakia, 1973, pp. 345–381.

[5] ——, "Estimation of the mean of a multivariate normal distribution," *Ann. Statist.*, vol. 9, pp. 1135–1151, 1981.

[6] G Leung, "Improving regression through model mixing," Ph.D. dissertation, Yale Univ., New Haven, CT, 2004.

[7] H. Akaike, "Statistical predictor identification," *Ann. Inst. Statist. Math.*, vol. 22, pp. 203–217, 1970.

[8] ——, "Information theory and an extension of the maximum likelihood principle," in *Proc. 2nd Int. Symp. Information Theory,* B. N. Petrov and F. Csáki, Eds. Budapest, Hungary: Akadémia Kiado, 1973, pp. 267–281.

[9] C. Mallows, "Some comments on C_p," *Technometrics*, vol. 15, pp. 661–675, 1973.

[10] J. A. Hartigan, Bayesian Regression Using Akaike Priors. New Haven, CT, Yale Univ., 2002, Preprint.

[11] S. T. Buckland, K. P. Burnham, and N. H. Augustin, "Model selection: An integral part of inference," *Biometrics*, vol. 53, pp. 603–618, 1997.

[12] A. R. Barron, "Are Bayes rules consistent in information?," in *Open Problems in Communication and Computation*, T. Cover and B. Gopinath, Eds. New York: Springer-Verlag, 1987.

[13] ——, "Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems," in *Bayesian Statistics,* J. Bernardo, J. Berger, A. Dawid, and A. Smith, Eds. Oxford, U.K.: Oxford Univ. Press, 1998.

[14] O. Catoni, Mixture Approach to Universal Model Selection Laboratoire de Mathématiques de l'Ecole Normale Supérieure, Paris, France, 1997, Preprint 30, to be published.

[15] Y. Yang, "Combining different regression procedures for adaptive regression," *J. Multivariate Anal.*, vol. 74, pp. 135–161, 2000.

[16] ——, "Regression with multiple candidate models: Selecting or mixing?," *Statistica Sinica*, vol. 13, pp. 783–809, 2003.

[17] O. Catoni, 'Universal' Aggregation Rules with Exact Bias Bounds Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris, 1999, Preprint 510, to be published.

[18] E. I. George, "Minimax multiple shrinkage estimation," *Ann. Statist.*, vol. 14, pp. 188–205, 1986.

[19] ——, "Combining minimax shrinkage estimators," *J. Amer. Statist. Assoc.*, vol. 81, pp. 431–445, 1986.

[20] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer-Verlag, 1998.

[21] E. L. Lehmann, *Testing Statistical Hypotheses*, 2nd ed. New York: Springer-Verlag, 1986.

[22] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[23] L. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, pp. 608–613, 1992.

[24] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[25] ——, "Approximation and estimation bounds for artificial neural networks," *Machine Learning*, vol. 14, pp. 115–133, 1994.

[26] W. Lee, P. Bartlett, and R. Williamson, "Efficient agnostic learning of neural networks with bounded fan-in," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2118–2132, Nov. 1996.

[27] Y. Yang and A. R. Barron, "Information-theoretic determination of minimum rates of convergence," *Ann. Statist.*, vol. 27, pp. 1564–1599, 1999.

[28] A. Juditsky and A. Nemirovski, "Functional aggregation for nonparametric estimation," *Ann. Statist.*, vol. 28, pp. 681–712, 2000.

[29] A. Barron, A. Cohen, W. Dahmen, and R. DeVore, "Approximation and learning by greedy algorithms," *Ann. Statist.*, Feb. 2006, submitted for publication.

[30] A. R. Barron, L. Birgé, and P. Massart, "Risk bounds for model selection via penalization," *Probab. Theory Related Fields*, vol. 113, pp. 301–413, 1999.

[31] M. S. Pinsker, "Optimal filtering of square integrable signals in Gaussian white noise," (in Russian) *Probl. Inf. Transm. (Probl. Pered. Inform.)*, vol. 16, pp. 120–133, 1980.

[32] R. Beran and L. Dümbgen, "Modulation of estimators and confidence sets," *Ann. Statist.*, vol. 26, no. 5, pp. 1826–1856, 1998.

[33] G. Leung and A. R. Barron, "Information theory, model selection and model mixing for regression," in *Proc. Conf. Information Sciences and Systems*, Princeton, NJ, Mar. 2004, pp. 579–584.

[34] R. Shibata, "An optimal selection of regression variables," *Biometrika*, vol. 68, no. 1, pp. 45–54, 1981.

[35] K. Li, "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," *Ann. Statist.*, vol. 15, no. 1, pp. 958–975, 1987.

[36] L. Birgé and P. Massart, "From model selection to adaptive estimation," in *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, D. Pollard, E. Torgersen, and G. Yang, Eds. New York: Springer-Verlag, 1997, pp. 55–87.

[37] ——, "Gaussian model selection," *J. Eur. Math. Soc.*, vol. 3, pp. 203–268, 2001.

[38] D. L. Donoho and I. M. Johnstone, "Ideal denoising in an orthonormal basis chosen from a library of bases," *C. R. Acad. Sci. Paris: Sér. I Math.*, vol. 319, pp. 1317–1322, 1994.

[39] ——, "Adapting to unknown smoothness via wavelet shrinkage," *J. Amer. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[40] ——, "Minimax estimation via wavelet shrinkage," *Ann. Statist.*, vol. 26, pp. 1879–921, 1998.

[41] I. M. Johnstone, *Function Estimation in Gaussian Noise: Sequence Models* 1998 [Online]. Available: www-stat.stanford.edu

[42] L. Devroye and G. Lugosi, *Combinatorial Methods in Density Estimation*. New York: Springer-Verlag, 2001.

[43] M. Wegkamp, "Model selection in nonparametric regression," *Ann. Statist.*, vol. 31, pp. 252–273, 2003.