Working Paper Series
ISSN 1170-487X

Pace Regression

by Yong Wang and Ian H. Witten

Working Paper 99/12
September 1999

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Pace Regression

Yong Wang

Computer Science Department
University of Waikato, New Zealand
Email: yongwang@cs.waikato.ac.nz

Ian H. Witten

Computer Science Department
University of Waikato, New Zealand
Email: ihw@cs.waikato.ac.nz

Abstract

This paper articulates a new method of linear regression, "pace regression," that addresses many drawbacks of standard regression reported in the literature—particularly the subset selection problem. Pace regression improves on classical ordinary least squares (OLS) regression by evaluating the effect of each variable and using a clustering analysis to improve the statistical basis for estimating their contribution to the overall regression. As well as outperforming OLS, it also outperforms—in a remarkably general sense—other linear modeling techniques in the literature, including subset selection procedures, which seek a reduction in dimensionality that falls out as a natural byproduct of pace regression. The paper defines six procedures that share the fundamental idea of pace regression, all of which are theoretically justified in terms of asymptotic performance. Experiments confirm the performance improvement over other techniques.

Keywords: Linear regression; subset model selection; mixture distribution; orthogonal model; least squares principle

1 Introduction

The basic idea of regression analysis is to fit a linear model to a set of data. The classical ordinary least squares (OLS) estimator is simple, computationally cheap, and has well-established theoretical justification. Nevertheless, the models it produces are often less than satisfactory. For example, OLS does not detect redundancy in the set of dependent variables that are supplied, and when a large number of variables are present, many of which are redundant, the model produced usually has worse predictive performance on future data than simpler models that take fewer variables into account.

Many researchers have investigated methods of *subset selection* in an attempt to neutralize this effect. The most common approach is OLS subset selection: from a set of OLS-fitted subset models, choose the one that optimizes some predetermined modeling criterion. Almost all these procedures are based on the idea of thresholding variation reduction: calculating how much the variation of the model is increased if each variable in turn is taken away, setting a threshold on this amount, and discarding variables that contribute less than the threshold. The rationale is that a noisy variable usually reduces the variation only marginally, whereas the variation accounted for by a meaningful variable is larger and grows with the variable's significance.

Many well-known procedures, including FPE (Akaike, 1970), AIC (Akaike, 1973), C_p (Mallows, 1973) and BIC (Schwarz, 1978) follow this approach. While these certainly work well for some data sets, extensive practical experience and many simulation studies have exposed serious shortcomings in them all. Often, for example, a certain proportion of redundant variables are included in the final model (Derksen and Keselman, 1992). Indeed, there are data sets for which a full regression model outperforms the selected subset model unless most of the variables are redundant (Hoerl et al., 1986; Roecker, 1991).

Shrinkage methods offer an alternative to OLS subset selection. Simulation studies show that the technique of *biased ridge regression* can outperform OLS subset selection, although it generates a more complex model (Frank and Friedman, 1993; Hoerl et al., 1986). The shrinkage idea used in ridge regression was further explored by Breiman (1995) and Tibshirani (1996), who were able to generate models that are less complex than ridge regression models yet still enjoy higher predictive accuracy than OLS subset models. Empirical evidence presented in these papers suggests that shrinkage methods yield greater predictive accuracy than OLS subset selection when a model has many noisy variables, or at most a moderate number of variables with moderate-sized effects—whereas they perform worse when there are variables that have a dramatic effect on the outcome.

These problems are systematic: the performance of modeling procedures can be related to the effects of variables and the extent of these effects. Researchers have sought to understand these phenomena and use them to motivate new approaches. For example, Miller (1990) investigated the selection bias that is introduced when the same data is used both to estimate the coefficients and to choose the subsets. New procedures, including the little bootstrap (Breiman, 1992), RIC (Donoho and Johnstone, 1994; Foster and George, 1994), and CIC (Tibshirani and

Knight, 1997) have been proposed. While these undoubtedly produce good models for many data sets, we will see below that there is no single approach that solves these systematic problems in a general sense.

This paper shows that these problems are the tip of an iceberg. They are manifestations of a much more general phenomenon that can be understood by examining the expected contributions that individual variables make in an orthogonal decomposition of the estimated model. This analysis leads to a new approach called "pace regression," standing for *Projection Adjustment by Contribution Estimation.*

Six procedures of pace regression are developed, denoted PACE$_1$ to PACE$_6$, that share a common fundamental idea—estimating the distribution of the effects of variables from the data and using this to improve modeling. The fact that this distribution can be estimated has apparently not been mentioned in the literature so far. The first four procedures utilize OLS subset selection, and outperform existing OLS methods for subset selection, including OLS itself. By abandoning the idea of selection, PACE$_5$ achieves the highest prediction accuracy of all. It even surpasses the OLS estimate when all variables have large effects on the outcome—in some cases by a substantial margin. Unfortunately, the extensive numerical calculations that PACE$_5$ requires limit its application in practice. However, PACE$_6$ is a very good approximation, and is computationally efficient. Theoretical and experimental comparisons between the new procedures and existing ones are reported in this paper.

We investigate model estimation in a general sense that subsumes subset selection. We do not confine our efforts to finding the best of the subset models; instead we address the whole space of linear models and regard subset models as a special case. But it is an important special case, because simplifying the model structure has wide applications in practice, and we will use it extensively to help sharpen our ideas.

We adopt two separate but complementary criteria for successful modeling. The first is predictive accuracy: the best model is the one with greatest accuracy on future data sampled independently from the same population. The second is parsimony: of models with similar predictive accuracy, prefer the smallest. Because of the inherent uncertainty present in any modeling situation, a small sacrifice in predictive accuracy may allow a substantial reduction in model complexity, significantly improving the comprehensibility of the model. This is particularly important in situations where the only estimates of predictive accuracy are inaccurate ones.

Like most work on model selection, we confine discussion to linear models with normally distributed noise. In fact, most of the ideas presented here generalize to other situations. More importantly, this work challenges many fundamental principles of empirical modeling. We return to these broader issues in Section 9.

The paper proceeds as follows. In Section 2, we discuss issues in linear regression and introduce some notation that will be used throughout. Section 3 discusses an orthogonal decomposition of linear models that forms the basis of all our modeling procedures. In Section 4, the "contribution" of each dimension of the orthogonal decomposition to the model is defined, and related functions are introduced. The notion of *contribution functions* is at the very core of our new scheme, and Section 5 shows how they can be used to understand modeling tasks, giving an intuitive introduction to the procedures that form the principal research contribution of this paper. Section 6 formally defines the six procedures of pace regression and theoretically justifies them, under the condition that a certain distribution function (the distribution of the underlying dimensional absolute distance) is known. Section 7 addresses the question of how to estimate this distribution, followed in Section 8 by some experimental examples that illustrate the power of the new methods. Section 9 discusses some important issues raised by pace regression, and explains how it challenges currently-accepted modeling principles.

2 Issues in linear regression

We begin by introducing some important issues in linear regression to provide a basis for the analysis that follows. We briefly review the major approaches to model selection, focusing on their failure to solve the systematic problems raised above. A common thread emerges: the key to solving the general problem of model selection in linear regression lies in the distribution of the effects of the variables that are involved. Subsection 2.8 pulls this theme together, briefly previewing the ideas underlying this paper.

2.1 Linear models and the distance

Given a set of input variables and an output variable, a linear model is uniquely determined by a parameter vector β of length k, the number of input variables. Suppose we are given the input variables for n independent instances in the form of an $n \times k$ design matrix X, and the corresponding output vector y. Then if β^* is the parameter vector of the true, underlying, model, y can be written as

$$y = X\beta^* + \epsilon, \tag{1}$$

where ϵ is a noise vector whose n components are independently sampled from $N(0, \sigma^2)$. We assume for the most part that the variance σ^2 is known; if not, it can be estimated using the OLS estimator $\hat{\sigma}^2$.

Use $\mathcal{M}$ to denote any model, $\mathcal{M}(\beta)$ the model with parameter vector β, and $\mathcal{M}^*$ as shorthand for the underlying model $\mathcal{M}(\beta^*)$. The entire model space under consideration is $\mathbb{M}_k = \{\mathcal{M}(\beta) : \beta \in \mathbb{R}^k\}$. Given y and X, the modeling task is to find an estimate $\mathcal{M}(\beta) \in \mathbb{M}_k$ of the underlying model $\mathcal{M}^* \in \mathbb{M}_k$ with the greatest predictive accuracy on future data.

Models produced by the OLS method, OLS subset selection methods, and shrinkage methods are all subclasses of

the model space $\mathbb{M}_k$. Any zero entry in β^* corresponds to a redundant variable. In fact, dimensionality reduction is not a problem independent from modeling; it is just a special case in which the discarded dimensions correspond to zero entries in the parameter vector.

We need a way of measuring the *distance* between two models, and choose to characterize this by the difference between the models' prediction vectors, since this relates directly to predictive accuracy. Given a design matrix X, the prediction of the model $\mathcal{M}(\beta)$ is the vector $y_{\mathcal{M}(\beta)} = X\beta$. In particular, the true model $\mathcal{M}^*$ predicts the output vector $y^* = y_{\mathcal{M}^*} = X\beta^*$. We find it convenient to define the distance between two models as

$$\mathcal{D}(\mathcal{M}(\beta_1), \mathcal{M}(\beta_2)) = ||y_{\mathcal{M}(\beta_1)} - y_{\mathcal{M}(\beta_2)}||^2/\sigma^2, \qquad (2)$$

where $||\cdot||$ denotes the L_2 norm. (When the noise variance σ^2 is unknown, the OLS estimator $\hat{\sigma}^2$ is used instead.) The problem of model construction is to find a model $\mathcal{M} \in \mathbb{M}_k$ that is as close as possible to the true model $\mathcal{M}^*$ in the sense that it minimizes the loss function $\mathcal{D}(\mathcal{M}, \mathcal{M}^*)$. Because of the uncertainty involved in the data set from which the model is constructed, it is the *expected* loss—the risk—that is minimized.

2.2 OLS subset models and their ordering

Any model that uses a subset of the k candidate variables and whose parameter vector is an OLS fit is called an "OLS subset model." When determining the best subset to use, it is common practice to generate a sequence of $k+1$ *nested models* $\{\mathcal{M}_j\}$ with increasing numbers j of variables. $\mathcal{M}_0$ is the *null model* with no variables and $\mathcal{M}_k$ is the *full model* with all variables included. The OLS estimate of model $\mathcal{M}_j$'s parameter vector is

$$\hat{\beta}_{\mathcal{M}_j} = (X'_{\mathcal{M}_j} X_{\mathcal{M}_j})^{-1} X'_{\mathcal{M}_j} y, \qquad (3)$$

where $X_{\mathcal{M}_j}$ is the $n \times j$ design matrix for model $\mathcal{M}_j$. Let $P_{\mathcal{M}_j} = X_{\mathcal{M}_j}(X'_{\mathcal{M}_j} X_{\mathcal{M}_j})^{-1} X'_{\mathcal{M}_j}$, which is the orthogonal projection matrix from the original k-dimensional space onto the reduced j-dimensional space. Then $\hat{y}_{\mathcal{M}_j} = P_{\mathcal{M}_j} y$ is the OLS estimate of $y^*_{\mathcal{M}_j} = P_{\mathcal{M}_j} y^*$.

One way of determining subset models is to include the variables in a predefined order using prior knowledge about the modeling situation. For example, in time series analysis it usually makes good sense to give preference to closer points when selecting autoregressive terms, while when fitting polynomials, lower-degree terms are often included before higher-degree ones. When the variable sequence is predefined, a total of $k + 1$ subset models are considered.

In the absence of prior ordering, a data-driven approach must be used to determine appropriate subsets. The final model could involve any subset of the variables. Of course, computing and evaluating all 2^k models rapidly becomes computationally infeasible as k increases. Techniques that are used in practice include forward, backward, and stepwise ranking of variables based on partial-F ratios (Thompson, 1978).

The difference between the prior ordering and the data-driven approach affects the subset selection procedures. If the ordering of variables is predefined, the subsets are determined independently of the data, which implies that the ratio between the residual sum of squares and the estimated variance can be assumed to be F distributed. The subset selection criteria FPE, AIC, and C_p all make this assumption. However, data-driven ordering complicates the situation. Candidate variables compete to enter and leave the model, causing competition bias (Miller, 1990). It is certainly possible to use FPE, AIC and C_p in this situation, but they lack theoretical support, and in practice they perform worse than when the variable order is correctly predefined. For example, suppose underfitting is negligible and the number of redundant variables increases without bound. Then the selected model's predictive accuracy and its expected number of redundant variables both tend to constant values when the variable order is predefined (Shibata, 1976), whereas in the data-driven scenario they both increase without bound.

Predefining the ordering makes use of prior knowledge of the underlying model. As is only to be expected, this will improve modeling if the information is basically correct, and hinder it otherwise. In practice, a combination of predefined and data-driven ordering is often used. For example, when certain variables are known to be relevant, they should definitely be kept in the model; also, it is common practice to always retain the constant term.

2.3 Asymptotics

We will be concerned with two asymptotic situations: *n-asymptotics*, where the number of observations increases without bound, and *k-asymptotics*, where the number of variables increases without bound. In this subsection we review some n-asymptotic results. The remainder of the paper is more concerned with k-asymptotics.

The model selection criteria FPE, AIC and C_p are *n-asymptotically equivalent* (Shibata, 1981) in the sense that they depend on threshold values that become the same—in this case, 2—as n approaches infinity. With reasonably large sample sizes, the performance of different n-asymptotically equivalent criteria are hardly distinguishable—both theoretically and experimentally. When discussing asymptotic situations, we use AIC to represent all three criteria.

Asymptotically speaking, the residual sum of squares of a significant variable is $O(n)$, whereas that of a redundant variable has a weak upper bound $O(1)$ and a strong upper bound $O(\log\log n)$. The model estimator generated by a threshold function bounded between $O(1)$ and $O(n)$ is weakly consistent in terms of model dimensionality, whereas one whose threshold function is bounded between $O(\log\log n)$ and $O(n)$ is strongly consistent.

Some model selection criteria are n-asymptotically strongly consistent. Examples include BIC (Schwarz, 1978), ϕ (Hannan and Quinn, 1979), GIC (Zhao et al., 1986), and Rao & Wu (1989). These all replace AIC's

threshold of 2 by an increasing function of n bounded between $O(\log\log n)$ and $O(n)$. The function value usually exceeds 2 (unless n is very small), giving a threshold that is larger than AIC's. However, employing the rate of convergence in this way is of little help in practice. For any finite data set, a higher threshold runs a greater risk of discarding a nonredundant variable that is only barely contributive. Criteria such as AIC that are n-asymptotically inconsistent do not necessarily perform worse than consistent ones.

Any OLS subset selection criterion minimizes a quantity that becomes, in the sense of n-asymptotic equivalence,

$$||y - y_{\mathcal{M}_j}||^2/\sigma^2 + \tau j \tag{4}$$

with respect to the dimensionality parameter j, where τ is the threshold value. We write this in parameterised form as OLSC(τ), where OLS = OLSC(0), AIC = OLSC(2) and BIC = OLSC($\log n$). The model selected by criterion (4) is denoted by $\mathcal{M}^{\text{OLSC}(\tau)}$; thus we have $\mathcal{M}^{\text{OLS}} = \mathcal{M}^{\text{OLSC}(0)}$, $\mathcal{M}^{\text{AIC}} = \mathcal{M}^{\text{OLSC}(2)}$ and $\mathcal{M}^{\text{BIC}} = \mathcal{M}^{\text{OLSC}(\log n)}$. (When the variance σ^2 is unknown, the OLS estimate $\hat{\sigma}^2$ is used instead.)

2.4 x-fixed *vs* x-random models

Two alternative basic assumptions underly regression modeling. In an "x-fixed" model the design matrix X remains unchanged for future prediction data, while in an "x-random" one each x_j is a random variable with a given distribution, and future data takes on values different from those used for training. Thompson (1978) discusses the implications of these assumptions in the subset selection situation. Some authors (for example, Miller (1990), Breiman (1992) and Breiman and Spector (1992)) treat the two cases differently.

Our work strives to minimize the expected distance between the prediction vector $y_{\mathcal{M}}$ and the underlying y^*, given a data sample X and y. This lies strictly within the classical x-fixed regression scenario. But it differs from most other procedures because they evaluate models according to their expected error on future data—which necessarily requires different handling for the two situations. In our case, we believe the models obtained will work equally well for the x-random situation. Since each model is uniquely determined by its parameter vector β, reducing the distance between $y_{\mathcal{M}}$ and $y_{\mathcal{M}^*}$ given the design matrix X is tantamount to reducing the distance between $\beta_{\mathcal{M}}$ and $\beta_{\mathcal{M}^*}$—in other words, reducing the distance between the prediction vectors of the estimated model and the true model for other samples.

According to the Gauss-Markov Theorem, OLS yields a "best linear unbiased estimator" (or BLUE) in the x-fixed situation (Rao and Toutenborg, 1995). Shaffer (1991) relates this to the x-random situation by establishing that if a best linear unbiased estimator exists for a given x-random situation, the OLS estimator is also a best linear unbiased estimator for the same situation. Shaffer's result makes it plausible to conjecture that that the two situations share (approximately) the same best estimator in the biased model space—although it would be nice to establish this theoretically. Of course, n-asymptotically speaking, there is no difference between the two situations.

2.5 Shrinkage methods

Shrinkage methods provide an alternative to OLS subset selection. Ridge regression gives a biased estimate of the model's parameter vector that depends on a ridge parameter. Increasing this quantity shrinks the OLS parameters toward zero. This may give better predictions by reducing the variance of predicted values, though at the cost of a slight increase in bias. It often improves the performance of the OLS estimate when some of the variables are (approximately) collinear. Experiments show that ridge regression can outperform OLS subset selection if most variables have small to moderate effects (Tibshirani, 1996). Although standard ridge regression does not reduce model dimensionality, its lesser known variants do (Miller, 1990).

The "nn-garrote" (Breiman, 1995) and "lasso" (Tibshirani, 1996) procedures zero some parameters and shrink others by defining linear inequality constraints on the parameters. Experiments show that these methods outperform ridge regression and OLS subset selection when predictors have small to moderate numbers of moderate-sized effects, whereas OLS subset selection based on C_p prevails over others for small numbers of large effects (Tibshirani, 1996).

All shrinkage methods rely on a parameter: the ridge parameter for ridge regression, the garrote parameter for the nn-garrote, and the tuning parameter for the lasso. In each case the parameter value significantly influences the result. However, there is no consensus on how to determine suitable values, which may explain the unstable performance of these methods. In Section 5, we offer a new explanation of shrinkage methods.

2.6 Data resampling

Standard techniques of data resampling, such as cross-validation and the bootstrap, can be applied to the subset selection problem. Theoretical work has shown that, despite their computational expense, these methods perform no better than the OLS subset selection procedures. For example, under weak conditions, Shao (1993) shows that the model selected by leave-d-out cross-validation or CV(d) is n-asymptotically consistent only if $d/n \to 1$ and $n - d \to \infty$ as $n \to \infty$. This suggests that the training set in each fold should be chosen to be as small as possible. Zhang (1993) further establishes that under similar conditions, CV(d) = OLSC($(2n - d)/(n - d)$) n-asymptotically. This means that AIC = CV(1) and BIC = CV($n(\log n - 2)/(\log n - 1)$) n-asymptotically.

The behavior of the bootstrap for subset selection is examined by Shao (1996), who proves that if the bootstrap using sample size m, BS(m), satisfies $m/n \to 0$, it

is n-asymptotically equivalent to CV($n-m$); in particular, BS(n) = CV(1) n-asymptotically. Therefore, BS(m) = OLSC($(n+m)/m$) n-asymptotically.

One problem with these data resampling methods is the difficulty of choosing an appropriate number of folds d for cross-validation, or an appropriate sample size m for the bootstrap. More fundamentally, data resampling methods do not address problems caused by chance features of competing variables whose biases are rooted in the training set, and so cannot solve the problems raised in Section 1.

2.7 The RIC and CIC

When there is no predefined ordering of variables, it is necessary to take account of the process by which a suitable ordering is determined. The expected value of the ith largest squared t-statistic of k noisy variables approaches $2\log(k/i)$ as k increases indefinitely. This property can help with variable selection.

The soft thresholding procedure developed in the context of wavelets (Donoho and Johnstone, 1994) and the RIC for subset selection (Foster and George, 1994) both aim to eliminate all non-contributory variables, up to the largest, by replacing the threshold 2 in AIC with $2\log k$; that is, RIC = OLSC($2\log k$). The more variables, the higher the threshold. When the true hypothesis is the null hypothesis (that is, there are no contributive variables), or the contributive variables all have large effects, RIC finds the correct model by eliminating all noisy variables up to the largest. However, when there are significant variables with small to moderate effects, these can be erroneously eliminated by the higher threshold value.

The CIC procedure adjusts the training error by taking into account the average covariance of the predictions and responses, based on the permutation distribution of the dataset (Tibshirani and Knight, 1997). In an orthogonally decomposed model space, the criterion simplifies to

$$\text{CIC}(j) = ||y - y_{\mathcal{M}_j}||^2 + 2\text{E}^0\left[\sum_{i=1}^{j} t^2_{(i:k)}\right]\hat{\sigma}^2 \quad (5)$$

$$\approx ||y - y_{\mathcal{M}_j}||^2 + 4\sum_{i=1}^{j}\log(k/i)\hat{\sigma}^2. \quad (6)$$

where $t^2_{(i:k)}$ is the ith largest squared t-statistic out of k, and E^0 is the expectation over the permutation distribution. As this equation shows, CIC uses a threshold value that is twice the expected sum of the squared t statistics of the j largest noisy variables out of k. Because $\lim_{k\to\infty} P[t^2_{(1:k)} \geq 2\text{E}t^2_{(1:k)}] = 0$, this means that, for the null hypothesis, even the largest noisy variable is almost always eliminated from the model. Furthermore, this has the advantage over RIC that smaller contributive variables are more likely to be recognized and retained.

Nevertheless, shortcomings exist. For example, if most variables have strong effects and will certainly not be discarded, the remaining noisy variables are treated by CIC as though they were the smallest out of k noisy variables—whereas in reality, the number should be reduced to reflect the smaller number of noisy variables. An overfitted model will likely result. Analogously, underfitting will occur when there are just a few contributive variables (Section 8 gives an experimental illustration of this effect.) CIC is based on an expected ordering of the squared t-statistics for noisy variables, and does not deal properly with situations where variables have mixed effects.

2.8 Remarks

We have now completed our brief review of the major procedures for linear regression. They all fail to solve the systematic problems raised in Section 1 in any general sense. From our discussion, it has emerged that each procedure's performance is closely related to the proportions of the different effects of the individual variables. It seems that the essential feature of any particular regression problem is the distribution of the effects of the variables it involves. This raises three questions: how to define this distribution; how to estimate it from the data, if indeed this is possible; and how to formulate satisfactory general regression procedures if the distribution is known.

In Section 3, we introduce the orthogonal decomposition of models. In the resulting model space, the effects of variables, which correspond to dimensions in the space, are mutually independent. Once the effects of individual variables have been teased out in this way, the distribution of these effects is easily defined.

The second question asks whether we can estimate this distribution from the data. Surprisingly, the answer is "yes." Moreover, estimators exist which are k-asymptotically strongly consistent. In fact, estimation is simply a clustering problem—to be more precise, it is the estimation of the mixing distribution of a semiparametric mixture. Section 7 shows how to perform this.

We answer the third question by demonstrating three successively more powerful techniques. First, following conventional ideas of model selection, the distribution of the effects of the variables can be used to derive an optimal threshold for OLS subset model selection. The resulting estimator is provably superior to all existing OLS subset selection techniques that are based on the idea of thresholding. Second, by showing that there are limitations to the idea of thresholding variation reduction, we develop an improved selection procedure that does not involve thresholding. Third, abandoning the idea of selection entirely results in a new adjustment technique that substantially outperforms all other procedures—outperforming OLS even when all variables have large effects. Section 6 introduces these procedures, which are all based on analyzing the dimensional contributions of the estimated models introduced in Section 4, and discusses their properties. Section 5 illustrates the ideas of these procedures through examples.

3 Orthogonal decomposition of models

In this section, we discuss issues concerning orthogonal decomposition of linear models. Advantages of an orthogonal model space include additivity over individual dimensions of the distance measure between models, dimensional independence, and the ability to define the distribution of the effects of variables. We assume in the following that no variables are collinear—that a model with k variables has k degrees of freedom. We return to the problem of collinear variables in Section 9.

Given a model $\mathcal{M}(\beta)$ with parameter vector β, its prediction vector is $y_{\mathcal{M}} = X\beta$ where X is the $n \times k$ design matrix X. This vector is located in the space spanned by the k separate n-vectors that represent the values of the individual variables. For any orthogonal basis of this space $b_1, \ldots, b_k$, let $P_1, \ldots, P_k$ be the corresponding projection matrices onto the axes. $y_{\mathcal{M}}$ decomposes into k components $P_1 y_{\mathcal{M}}, \ldots, P_k y_{\mathcal{M}}$, each being a projection on to a different axis. Clearly the whole is the sum of the parts: $y_{\mathcal{M}} = \sum_{j=1}^{k} P_j y_{\mathcal{M}}$.

3.1 Decomposing distances

The distance $\mathcal{D}(\mathcal{M}(\beta_1), \mathcal{M}(\beta_2))$ between models $\mathcal{M}(\beta_1)$ and $\mathcal{M}(\beta_2)$ has been defined in (2) above. Although this measure involves the noise variance σ^2 for convenience of both analysis and computation, it is $||y_{\mathcal{M}(\beta_1)} - y_{\mathcal{M}(\beta_2)}||^2$ that is the center of interest.

Given an orthogonal basis, the distance between two models can be decomposed as follows.

$$\mathcal{D}(\mathcal{M}(\beta_1), \mathcal{M}(\beta_2)) = \sum_{j=1}^{k} \mathcal{D}_j(\mathcal{M}(\beta_1), \mathcal{M}(\beta_2)), \quad (7)$$

where

$$\mathcal{D}_j(\mathcal{M}(\beta_1), \mathcal{M}(\beta_2)) = ||P_j y_{\mathcal{M}(\beta_1)} - P_j y_{\mathcal{M}(\beta_2)}||^2/\sigma^2 \quad (8)$$

is the jth *dimensional distance* between the models. The property of *additivity* of distance in this orthogonal space will turn out to be crucial for our purposes: the distance between the models is equal to the sum of the distances between the models' projections.

Denote by $\mathcal{M}_0$ the *null model* $\mathcal{M}(0)$, whose every parameter is zero. The distance between $\mathcal{M}$ and the null model is the *absolute distance* of $\mathcal{M}$, denoted by $\mathcal{A}(\mathcal{M})$; that is, $\mathcal{A}(\mathcal{M}) = \mathcal{D}(\mathcal{M}, \mathcal{M}_0)$. Decomposing the absolute distance yields

$$\mathcal{A}(\mathcal{M}) = \sum_{j=1}^{k} \mathcal{A}_j(\mathcal{M}), \quad (9)$$

where

$$\mathcal{A}_j(\mathcal{M}) = ||P_j y_{\mathcal{M}}||^2/\sigma^2 \quad (10)$$

is the jth *dimensional absolute distance* of $\mathcal{M}$.

3.2 Decomposing the estimation task

Two models are of central interest in the process of estimation: the true model $\mathcal{M}^*$ and the estimated model $\mathcal{M}$. The distance between them is defined as the *loss* of the estimated model, denoted by $\mathcal{L}(\mathcal{M})$; that is, $\mathcal{L}(\mathcal{M}) = \mathcal{D}(\mathcal{M}, \mathcal{M}^*)$. The aim of estimation is to minimize the loss.

Being a distance, the loss can be decomposed into dimensional components

$$\mathcal{L}(\mathcal{M}) = \sum_{j=1}^{k} \mathcal{L}_j(\mathcal{M}), \quad (11)$$

where

$$\mathcal{L}_j(\mathcal{M}) = \mathcal{D}_j(\mathcal{M}, \mathcal{M}^*) = ||P_j y_{\mathcal{M}} - P_j y_{\mathcal{M}^*}||^2/\sigma^2 \quad (12)$$

is the jth *dimensional loss* of model $\mathcal{M}$.

Orthogonal decomposition breaks the estimation task down into individual estimation tasks for each of the k dimensions. $\mathcal{A}_j(\mathcal{M}^*)$ is the underlying absolute distance in the jth dimension, and $\mathcal{A}_j(\mathcal{M})$ is an estimate of it. The loss incurred by this estimate is $\mathcal{L}_j(\mathcal{M})$. The sum of the losses in each dimension is the total loss $\mathcal{L}(\mathcal{M})$ of the model $\mathcal{M}$. This reduces the modeling task to estimating $\mathcal{A}_j(\mathcal{M}^*)$ for all j. Once these estimated distances have been found for each dimension $j = 1, \ldots, k$, the estimated model can be reconstructed from them; we tackle the details in the next subsection.

Our estimation process has two steps: an initial estimate of $\mathcal{A}_j(\mathcal{M}^*)$ followed by a further refinement stage. This implies that there is some information not used in the initial estimate that can be exploited to improve it; we use the loss function to guide refinement. The first step is to find a relationship between the initial estimate $\mathcal{A}_j(\mathcal{M})$ and $\mathcal{A}_j(\mathcal{M}^*)$. The classic OLS estimate $\widehat{\mathcal{M}}$, which has parameter vector $\widehat{\beta} = (X'X)^{-1}X'y$, provides a basis for such a relationship, because it is well known that for all j, $\mathcal{A}_j(\widehat{\mathcal{M}})$ are independently, noncentrally χ^2 distributed with one degree of freedom and noncentrality parameter $\mathcal{A}_j(\mathcal{M}^*)/2$ (Schott, 1997, p390). We write this as

$$\mathcal{A}_j(\widehat{\mathcal{M}}) \sim \chi^2_1(\mathcal{A}_j(\mathcal{M}^*)/2) \text{ independently for all } j. \quad (13)$$

When σ^2 is unknown and therefore replaced by the unbiased OLS estimate $\hat{\sigma}^2$, the χ^2 distribution in (13) becomes an F distribution: $\mathcal{A}_j(\widehat{\mathcal{M}}) \sim F(1, n-k, \mathcal{A}_j(\mathcal{M}^*)/2)$ independently for all j. The F-distribution can be accurately approximated by (13) when $n/k \gg 1$.

This relationship forms a cornerstone of this paper.

3.3 Reconstructing the model from estimated absolute distances

Once final estimates for the absolute distances in each dimension have been found, the model needs to be reconstructed from them. Consider how to build a model from a set of absolute distances, denoted by $A_1, \ldots, A_k$. Let

$\alpha = (s_1\sqrt{A_1}, \ldots, s_k\sqrt{A_k})'$, where s_j is either $+1$ or -1 depending on whether or not the jth projection of the prediction vector $y_{\widehat{\mathcal{M}}}$ has the same direction as the orthogonal base b_j. This choice of s_j's value is based on the fact that the projections of $y_{\widehat{\mathcal{M}}}$ and $y_{\mathcal{M}^*}$ are most likely in the same direction. (It will become clear later that alternative choices would degrade the estimate.)

Our estimate of the parameter vector is

$$\beta = (X'X)^{-1}X'B\alpha\sigma, \tag{14}$$

where B is a column matrix formed from the bases $b_1, \ldots, b_k$. For example, if $\{A_1, \ldots, A_k\}$ are the OLS estimates of the absolute distances in the corresponding dimensions, (14) gives β the value of the OLS estimate $\widehat{\beta}$.

It may be that not all the A_j's are available, but a prediction vector is known that corresponds to all missing A_j's. This situation will occur if some dimensions are forced to be in the final model—for example, the constant term, or dimensions that give very great reductions in variation (very large A_j's). Suppose the number of dimensions with known A_j's is k', and call the overall prediction vector for the remaining $k - k'$ dimensions y_{rest}. Then the estimated parameter vector is

$$\beta = (X'X)^{-1}X'(y_{\text{rest}} + B\alpha\sigma), \tag{15}$$

where B is an $n \times k'$ matrix and α a k'-vector.

The estimation of β from A_j's is fully described by (14) and (15). However, in practice the computation takes a different, more efficient, route. Once the $n \times k$ approximation equation of the original least-squares problem has been orthogonally transformed, finding the least squares solution reduces to solving a matrix equation

$$U\beta = d, \tag{16}$$

where U is a $k \times k$ upper-triangular matrix and d is a k-vector (Lawson and Hanson, 1974). As a matter of fact, the square of the jth element in d is exactly the OLS estimate $\widehat{A}_j\sigma^2$. When a new set of estimates, say $\widetilde{A}_j$ $(j = 1, \ldots, k)$, is obtained, the corresponding estimate of β^* is the solution of (16) with the jth element in d replaced by $\sqrt{\widetilde{A}_j}\sigma$ without changing sign. If not all $\widehat{A}_j$'s are known, so that (15) is used instead of (14), only dimensions with known $\widehat{A}_j$'s are replaced.

3.4 A special case: OLS subset models

OLS subset models are just a special case of orthogonally decomposed models in which each variable is associated with an orthogonal dimension. This makes it easy to simplify the model structure: discarding variables in a certain order is the same as deleting dimensions in the same order. If the discarded variables are actually redundant, deleting them makes the model more accurate.

Denote the subset models of $\mathcal{M}$, from the null model up to the full one, by $\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_k$ respectively. Denote by $y_{\mathcal{M}_j}$ the prediction vector of the j-dimensional model $\mathcal{M}_j$, and by $P_{\mathcal{M}_j} = X_{\mathcal{M}_j}(X'_{\mathcal{M}_j}X_{\mathcal{M}_j})^{-1}X'_{\mathcal{M}_j}y$ the orthogonal projection matrix from the space of k dimensions to the j-dimensional subspace corresponding to $\mathcal{M}_j$. Then $y_{\mathcal{M}_j} = P_{\mathcal{M}_j}y_{\mathcal{M}}$ and $P_j = P_{\mathcal{M}_j} - P_{\mathcal{M}_{j-1}}$. Furthermore, the jth base can be written as $b_j = P_jy/||P_jy|| = (y_{\mathcal{M}_j} - y_{\mathcal{M}_{j-1}})/||(y_{\mathcal{M}_j} - y_{\mathcal{M}_{j-1}})||$.

This demonstrates that OLS subset models are indeed orthogonally decomposed models.

3.5 Summary of orthogonal decomposition

There are two advantages of using orthogonally decomposed models for estimation. The first is additivity: the distance between two models is the sum of their distances in each dimension. This convenient property is inherited by two special distances: the absolute distance of the model itself and the loss function of the estimated model. Of course, any quantity that involves addition and subtraction of distances between models is additive too.

The second advantage is the mutual independence of the model's components in the different dimensions. This makes the dimensional distances between models independent too. In particular, the absolute distances in each dimension, and the losses incurred by an estimated model in each dimension—as well as any measure derived from these by additive operators—are independent between one dimension and another.

These two features allow the process of estimating the overall underlying model to be broken down into estimating its components in each dimension separately.

4 Contributions and contribution functions

We next explore a new measure for a model: its *contribution.* An estimated model's contribution is zero for the null model and reaches a maximum when the model is the same as the underlying one. It can be decomposed into k independent, additive components in k-dimensional orthogonal space—the "dimensional contributions." In practice, these quantities are random variables, and we can define both a *cumulative expected contribution function* and an *expected contribution density function* of a given dimensional contribution of a given estimated model. These two functions can be estimated for any particular regression problem, and will turn out to play key roles in understanding the modeling process and in building actual models in practice.

Definition 4.1 *The contribution of an estimate $\mathcal{M}$ of the underlying model $\mathcal{M}^*$ is defined to be*

$$\mathcal{C}(\mathcal{M}) = \mathcal{A}(\mathcal{M}^*) - \mathcal{L}(\mathcal{M}). \tag{17}$$

The contribution is calculated as the difference between two distances: the maximum gain that the estimated model can possibly offer over the null model, and the loss

that the actual estimated model incurs from this ideal situation. The maximum gain is the gain achieved by the true underlying model $\mathcal{M}^*$, namely $\mathcal{A}(\mathcal{M}^*)$, the distance of $\mathcal{M}^*$ from the null model. The loss incurred by $\mathcal{M}$ is its distance from the true model: $\mathcal{D}(\mathcal{M}, \mathcal{M}^*)$. For a given underlying model, the contribution reaches a maximum when the estimated model is the same as the underlying one, while the contribution is zero if the estimated model is the null model. A positive contribution means that the estimated model is better than the null one in terms of predictive accuracy; a negative value means it is worse.

Given a k-dimensional orthogonal basis, the contribution function decomposes into k components that retain the properties of additivity and dimensional independence:

$$\mathcal{C}(\mathcal{M}) = \sum_{j=1}^{k} \mathcal{C}_j(\mathcal{M}) \tag{18}$$

where

$$\mathcal{C}_j(\mathcal{M}) = \mathcal{A}_j(\mathcal{M}^*) - \mathcal{L}_j(\mathcal{M}) \tag{19}$$

is the jth *dimensional contribution* of the model $\mathcal{M}$.

In the subset selection task, each dimension is either retained or discarded. It is clear that this decision should be based on the sign of the corresponding dimensional contribution. If a dimension's contribution is positive, retaining it will give better predictive accuracy than discarding it, and conversely, if the contribution is negative then discarding it will improve accuracy. If the dimensional contribution is zero, it makes no difference to predictive accuracy whether that dimension is retained or not.

In following we focus on a single dimension, the jth. The results of individual dimensions can easily be combined because dimensions are independent and the contribution measure is additive. Focusing on the contribution in this dimension, we write ${a_j}^2 = \mathcal{A}_j(\mathcal{M})$ and ${a_j^*}^2 = \mathcal{A}_j(\mathcal{M}^*)$. Without loss of generality, assume that $a_j^* \geq 0$. If the projection of $y_{\mathcal{M}}$ in the jth dimension is in the same direction as that of $y_{\mathcal{M}^*}$, then a_j is the positive square root of $\mathcal{A}_j(\mathcal{M})$; otherwise it is the negative square root. In either case, the contribution can be written

$$\mathcal{C}_j(\mathcal{M}) = {a_j^*}^2 - (a_j - a_j^*)^2. \tag{20}$$

Clearly, $\mathcal{C}_j(\mathcal{M})$ is zero when a_j is 0 or $2a_j^*$. When a_j lies between these values, the contribution is positive. For any other values of a_j, it is negative. The maximum contribution is achieved when $a_j = a_j^*$, and has value ${a_j^*}^2$, which occurs when—in this dimension—the estimated model is the true model. This is obviously the best that the estimated model can do in this dimension.

In practice, however, only ${a_j}^2$ is available. Neither the value of ${a_j^*}^2$ nor the directional relationship between the two projections are known. Denote $\mathcal{C}_j(\mathcal{M})$ by $C({a_j}^2)$, altering the notion of the contribution of $\mathcal{M}$ in this dimension to the contribution of ${a_j}^2$. Notice, however, that $C({a_j}^2)$ is used below as shorthand for $C({a_j}^2; {a_j^*}^2, s_j) = \mathcal{C}_j(\mathcal{M})$, where s_j is the sign of a_j. In the following we drop the subscript j when only one dimension is under consideration, giving a^2 and ${a^*}^2$ for ${a_j}^2$ and ${a_j^*}^2$ respectively. We also use A for a^2 since it is this, rather than a, that is available; likewise we use A^* for ${a^*}^2$.

We have argued that the performance of an estimated model can be analyzed in terms of the value of its contribution in each dimension. Unfortunately, this value is unavailable in practice. What can be computed is the expected value, denoted by $E[C(A)]$, where the expectation is taken over all other factors that affect the value of $C(A)$.

In practice, the estimate A is a random variable. For example, according to (13), the OLS estimate is $\widehat{A} \sim \chi_1^2(A^*/2)$. Analogously to the definitions of CDF and pdf of a random variable, we define the *cumulative expected contribution function* and the *expected contribution density function* of A, denoted by $H(A)$ and $h(A)$ respectively.

Definition 4.2 *The cumulative expected contribution function of A is defined as*

$$H(A) = \int_0^A E[C(t)] f(t)\, dt, \tag{21}$$

where $f(A)$ is the pdf of A.

Definition 4.3 *The expected contribution density function of A is defined as*

$$h(A) = \frac{dH(A)}{dA}. \tag{22}$$

For reason of convenience, we call $H(A)$ and $h(A)$ the *H-* and *h-functions* of A. Immediately from the above definitions, we have

$$E[C(A)] = \frac{h(A)}{f(A)}. \tag{23}$$

Now we derive an expression for $h(\widehat{A})$ given A^*, where $\widehat{A}$ is the OLS estimate of the dimensional absolute distance. We utilize the property (13). The pdf of the noncentral chi-squared distribution has an infinite series form, and thus is inconvenient for both analysis and computation. Instead, setting $a^* = +\sqrt{A^*}$, we use the fact that $\widehat{A} = \hat{a}^2 \sim \chi_1^2({a^*}^2/2)$ where $\hat{a} \sim N(a^*, 1)$. The pdf of $\hat{a}$ is given by

$$p(\hat{a}; a^*) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(\hat{a}-a^*)^2}{2}}. \tag{24}$$

Then the CDF of $\widehat{A}$ given A^* is

$$F(\widehat{A}; A^*) = \int_{-\sqrt{\widehat{A}}}^{\sqrt{\widehat{A}}} p(t; \sqrt{A^*})\, dt, \tag{25}$$

hence

$$\begin{aligned} f(\widehat{A}; A^*) &= \frac{dF(\widehat{A}; A^*)}{d\widehat{A}} \\ &= \frac{p(\sqrt{\widehat{A}}; \sqrt{A^*}) + p(-\sqrt{\widehat{A}}; \sqrt{A^*})}{2\sqrt{\widehat{A}}}. \end{aligned} \tag{26}$$

Using (20), rewrite the contribution of $\widehat{A}$ given A^* by a two-argument function $c(\hat{a}; a^*)$

$$C(\widehat{A}) = c(\hat{a}; a^*) = a^{*2} - (\hat{a} - a^*)^2. \tag{27}$$

Here only the sign of $\hat{a}$ can affect the value of the contribution, and so the expected contribution of $\widehat{A}$ given A^* is

$$E[C(\widehat{A})|A^*] = \frac{c(\sqrt{\widehat{A}}; \sqrt{A^*})p(\sqrt{\widehat{A}}; \sqrt{A^*}) + c(-\sqrt{\widehat{A}}; \sqrt{A^*})p(-\sqrt{\widehat{A}}; \sqrt{A^*})}{p(\sqrt{\widehat{A}}; \sqrt{A^*}) + p(-\sqrt{\widehat{A}}; \sqrt{A^*})}. \tag{28}$$

Using (23),

$$h(\widehat{A}; A^*) = \frac{1}{2\sqrt{\widehat{A}}}[c(\sqrt{\widehat{A}}; \sqrt{A^*})p(\sqrt{\widehat{A}}; \sqrt{A^*}) + c(-\sqrt{\widehat{A}}; \sqrt{A^*})p(-\sqrt{\widehat{A}}; \sqrt{A^*})]. \tag{29}$$

In particular, $h(0; A^*) = 0$ for every A^* (see Appendix A). This gives the following theorem.

Theorem 4.1 *The expected contribution density function $h(\widehat{A}; A^*)$ of the* OLS *estimate $\widehat{A}$ given A^* is determined by (29), while the pdf $f(\widehat{A}; A^*)$ is determined by (26).*

Because $E[C(A)] = h(A)/f(A)$ by (23), and $f(A)$ is always positive, the value of $h(A)$ has the same sign as $E[C(A)]$. Therefore the sign of h can be used as a criterion to determine whether a dimension should be discarded or not. Within a positive interval, where $h(A) > 0$, A is expected to contribute positively to the predictive accuracy, whereas within a negative one, where $h(A) < 0$, it will do the opposite. At a zero of $h(A)$ the expected contribution is zero.

Figure 1(a) shows examples of $h(\widehat{A}; A^*)$ where A^* is $0, 0.5, 1, 2$ and 5. All the curves start from the origin. When $A^* = 0$, the curve first decreases as $\widehat{A}$ increases, and then gradually increases, approaching the horizontal axis asymptotically and never rising above it. As the value of A^* grows, the $h(\widehat{A}; A^*)$ curve generally rises. However, it always lies below the horizontal axis until A^* becomes 0.5. When $A^* > 0.5$, there is one positive interval. A maximum is reached not far from A^* (the maximum approaches A^* as the latter increases), and thereafter each curve slowly descends to meet the axis at around $4A^*$ (the ordinate approaches this value as A^* increases). Thereafter the interval remains negative; within it, each h-function reaches a minimum and then ascends to approach the axis asymptotically from below.

Most observations in the last paragraph are secured in the following theorem. For convenience, denote the zeros of h by $\mathcal{Z}_1$, $\mathcal{Z}_2$, $\mathcal{Z}_3$ in increasing order along the horizontal axis, and assume that h is properly defined at ∞. Since A^* in practice is always finite, we only consider the situation $A^* < \infty$.

Theorem 4.2 *Properties of $h(\widehat{A}; A^*)$.*

1. *Every $h(\widehat{A}; A^*)$ has three zeros (two of which may coincide).*
2. *When $A^* \leq 0.5$, $\mathcal{Z}_1 = \mathcal{Z}_2 = 0$, $\mathcal{Z}_3 = \infty$; when $A^* > 0.5$, $\mathcal{Z}_1 = 0$, $0 < \mathcal{Z}_2 < \infty$, $\mathcal{Z}_3 = \infty$.*
3. *$\lim_{A^* \to \infty}(\mathcal{Z}_2 - 4A^*) = 0$.*
4. *$h(\widehat{A}; A^*) > 0$ for $\widehat{A} \in (0, \mathcal{Z}_2)$ and $h(\widehat{A}; A^*) < 0$ for $\widehat{A} \in (\mathcal{Z}_2, \infty)$.*
5. *When $A^* > 0.5$, $h(\widehat{A}; A^*)$ has a unique maximum, at $A_{\max}$, say. Then $\lim_{A^* \to \infty}(A_{\max} - A^*) = 0$.*
6. *$h(\widehat{A}; A^*)$ is continuous for every $\widehat{A}$ and A^*.*

The proof can be easily established using formulae (24)–(29). Almost all these properties are evident in Figure 1(a). The critical value 0.5—the largest value of A^* for which $h(\widehat{A}; A^*)$ has no positive interval—can be obtained by setting to zero the first derivative of $h(\widehat{A}; A^*)$ with respect to $\widehat{A}$ at point $\widehat{A} = 0$. The derivatives around $\widehat{A} = 0$ can be obtained using the Taylor expansion of $h(\widehat{A}; A^*)$ with respect to $\sqrt{\widehat{A}}$ (see Appendix A).

As noted earlier, the sign of h can be used as a criterion for subset selection. In Section 5, Figure 1(a) is interpreted from a subset selection perspective.

Figure 1(b) shows the expected contribution curves $E[C(\widehat{A})|A^*] = h(\widehat{A}; A^*)/f(\widehat{A}; A^*)$. The location of the maximum converges to A^* as $A^* \to \infty$, and it converges very quickly—when $A^* = 2$, the difference is almost unobservable.

When deriving the H- and h-functions, we have assumed that the underlying dimensional absolute distance A^* is given. However, A^*, is, in practice, unknown—only the value of A is known. To allow the functions to be calculated, we consider A^* to be a random variable with a distribution $G(A^*)$, a distribution which is estimable from a sample of A. This is exactly the problem of estimating a mixing distribution. In our situation, the pdf of the corresponding mixture distribution has the general form

$$f(A; G) = \int_{\mathbb{R}^+} f(A; A^*)\, dG(A^*), \tag{30}$$

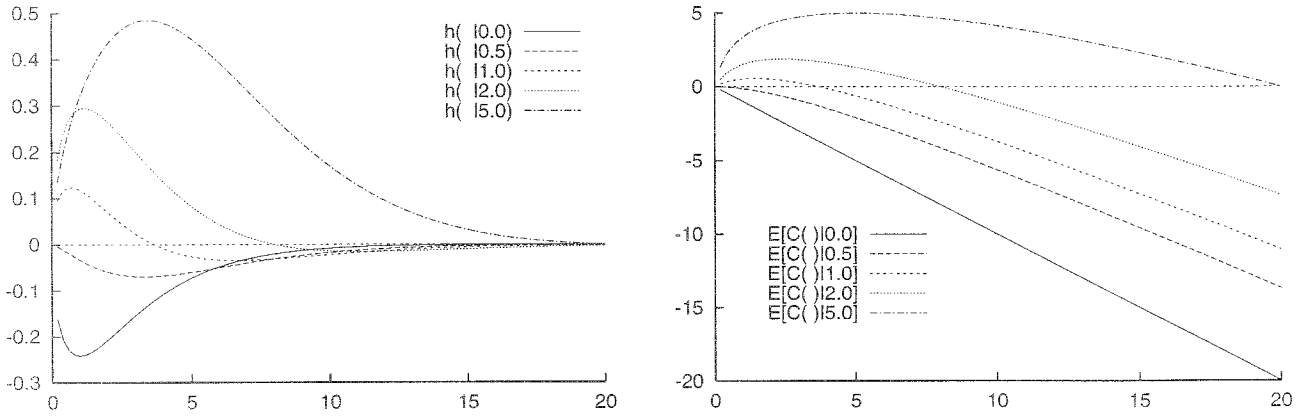
where $A^* \in \mathbb{R}^+$, the nonnegative half real line. $G(A^*)$ is the mixing distribution function, $f(A; A^*)$ the pdf of the component distribution, and $f(A; G)$ the pdf of the mixture distribution. Section 7 shows how to estimate the mixing distribution from a sample of A.

If the mixing distribution $G(A^*)$ is given, the h-function of A can be obtained from the following theorem.

Theorem 4.3 *Let A^* be distributed according to $G(A^*)$ and A be a random variable sampled from the mixture distribution defined in (30). Then*

$$h(A; G) = \int_{\mathbb{R}^+} h(A; A^*)\, dG(A^*), \tag{31}$$

where $h(A; A^*)$ is the h-function determined by $f(A; A^*)$.



(a) (b)

Figure 1: The expected contribution density function $h(\widehat{A}; A^*)$ and the expected contribution $E[C(\widehat{A})|A^*]$, for $A^* = 0, .5, 1, 2, 5$.

The proof follows easily from the definition of h.

In the modeling situation, each A corresponds to an A^*. Therefore $G(A^*)$ is discrete and the mixture is a countable one. Suppose the discrete random variable A^* take values from the set $\{\alpha_i^*; i = 1, \ldots, m\}$—$m$ could be as large as k—and define the pdf $g(A^*)$ as

$$g(\alpha_i^*) = w_i, \text{ where } \sum_{i=1}^{m} w_i = 1. \tag{32}$$

Then (30) can be re-written as

$$f(A; G) = \sum_{i=1}^{m} w_i f(A; \alpha_i^*), \tag{33}$$

and (31) as

$$h(A; G) = \sum_{i=1}^{m} w_i h(A; \alpha_i^*). \tag{34}$$

Although the general forms (30) and (31) are adopted in the following analysis, it is (33) and (34) that are used in practical computations. Note that if $h(A; G)$ and $f(A; G)$ are known, the expected contribution of A given $G(A^*)$ is given by (23) as $h(A; G)/f(A; G)$.

Since all $\widehat{A}_j$'s of an OLS estimated model which is decomposed in an orthogonal space are mutually independent (see (13)), they form a sample from the mixture distribution with a discrete mixing distribution $G(A^*)$. The pdf of the mixture distribution takes the form (33), while the pdf of the component distribution is provided by (26). Likewise the h-function of the mixture has the form (34), while the component h-function is given by (29).

From now on, the mixing distribution $G(A^*)$ becomes our major concern. If $f(A; A^*)$ and $h(A; A^*)$ are well defined (as they are in OLS estimation), $G(A^*)$ uniquely determines $f(A; G)$ and $h(A; G)$ by (30) and (31) respectively. The following sections analyze the modeling process with known $G(A^*)$, show how to build the best model with known $G(A^*)$, and finally tackle the question of estimating $G(A^*)$.

5 The role of H- and h-functions in modeling

The H- and h- functions illuminate our understanding of the modeling process, and help with building models too. Here we use them to illustrate issues associated with the OLS subset selection procedures described in Section 1, and to elucidate new, hitherto unreported, phenomena. While we concentrate on OLS subset selection criteria, we touch on shrinkage methods too.

We also illustrate with examples the basis for the new procedures that are formally defined in the next section. Not only does subset selection by thresholding variation reduction severely restrict the modeling space, but the very idea of subset selection is a limited one—when a wider modeling space is considered, better estimators emerge. Consequently we expand our horizon from subset selection to the general modeling problem, producing a final model that is not a least-squares fit at all. This improves on OLS modeling even when all dimensions contribute significantly.

The methodology we adopt is suggested by contrasting the model-based selection problem that we have studied so far with the "dimension-based selection" that is used in principal component analysis. Dimension-based selection tests each orthogonal dimension independently for elimination, whereas model-based selection analyses a set of orthogonal nested models in sequence (as discussed in subsection 2.2, the sequence may be defined a priori or computed from the data). In dimension-based selection, deleting a dimension removes the transformed variable associated with it, and although this reduces the number

of dimensions, it does not generally reduce the number of original variables.

Section 2 showed that all OLS subset selection criteria share the idea of thresholding the amount by which the variation is reduced in each dimension. While straightforward for dimension-based selection, this needs some adjustment in model-based selection because the variation reductions of the nested models may not be in the desired decreasing order. The necessary adjustment, if p variables are tested, is to compare the reductions of the variation by these variables with p times the threshold value. The central idea remains the same.

The key issue in OLS subset selection is the choice of threshold. CIC's threshold depends on the number of dimensions in the model, whereas other methods use fixed thresholds. We denote these schemes by OLSC(τ), where τ is the threshold (see (4)). The optimum value of τ, in the sense of minimum expected risk, is denoted by τ^*. We now consider how τ^* can be determined using the H-function, assuming that all dimensions have the same known H. We begin with dimension-based selection and tackle the nested model situation later.

Theorem 5.1 *Given an orthogonal decomposition of a model space $\mathbb{M}_k$, let $\mathcal{M}^* \in \mathbb{M}_k$ be any underlying model and $\mathcal{M} \in \mathbb{M}_k$ be its estimate. Assume that all dimensional absolute distances of $\mathcal{M}$ have the same distribution function $F(A)$ and thus the same H-function. Rank the $\mathcal{A}_j(\mathcal{M})$'s in decreasing order as j increases. Then the estimator $\mathcal{M}^{\text{OLSC}(\tau^*)}$ of $\mathcal{M}^*$ has the minimum expected risk with respect to $F(A)$ for every $\mathcal{M}^{\text{OLSC}(\tau)}$ if and only if*

$$\tau^* = \arg\min_{A \geq 0} \; H(A). \tag{35}$$

Proof. For dimension j, we know from (19) that

$$\mathcal{L}_j(\mathcal{M}) = \mathcal{A}_j(\mathcal{M}^*) - \mathcal{C}_j(\mathcal{M}).$$

OLSC(τ) discards dimension j if $\mathcal{A}_j(\mathcal{M}) \leq \tau$, so

$$\mathcal{C}_j(\mathcal{M}^{\text{OLSC}(\tau)}) = \begin{cases} 0 & \text{if } \mathcal{A}_j(\mathcal{M}) \leq \tau \\ \mathcal{C}_j(\mathcal{M}) & \text{if } \mathcal{A}_j(\mathcal{M}) > \tau, \end{cases}$$

where $\mathcal{A}_j(\mathcal{M})$ is a random variable which is distributed according to the CDF $F(\cdot)$.

$$\begin{aligned} &\int_{\mathbb{R}^+} E[\mathcal{L}_j(\mathcal{M}^{\text{OLSC}(\tau)})]\, dF(A) \\ &= \int_{\mathbb{R}^+} E[\mathcal{A}_j(\mathcal{M}^*)]\, dF(A) - \int_{\tau}^{\infty} E[\mathcal{C}_j(\mathcal{M})]\, dF(A) \\ &= \mathcal{A}_j(\mathcal{M}^*) - H(\infty) + H(\tau). \end{aligned}$$

Taking advantage of additivity and dimensional independence, sum the above equation over all k dimensions:

$$\int_{\mathbb{R}^+} E[\mathcal{L}(\mathcal{M}^{\text{OLSC}(\tau)})]\, dF(A) = \mathcal{A}(\mathcal{M}^*) - kH(\infty) + kH(\tau). \tag{36}$$

$H(\tau)$ is the only term that varies with τ on the right-hand side of (36). Thus minimizing the expected risk is equivalent to minimizing $H(\tau)$, and *vice versa*. This completes the proof. □

Theorem 5.1 requires the dimensional absolute distances of the initially estimated model to be sorted into decreasing order. This is easily accomplished in the dimension-based situation, but not in the model-based situation. However, the nested models are invariably generated in a way that attempts to establish such an order, and so this condition is approximately satisfied in practice. Thus $\mathcal{M}^{\text{OLSC}(\tau^*)}$ is a good approximation to the minimum expected risk estimator even in the model-based situation.

From Theorem 5.1, we have

Corollary 5.1.1 *Properties of τ^*.*

1. $\tau^* = \arg\min_{\mathcal{Z} \in \{\mathcal{Z}_i\}} H(\mathcal{Z})$, *where $\{\mathcal{Z}_i\}$ is the set of zeros of h.*
2. *If $\tau^* > 0$, $h(\tau^*-) > 0$; if $\tau^* < \infty$, $h(\tau^*+) < 0$.*
3. *$H(\tau^*) \leq 0$ and $H(\infty) - H(\tau^*) \geq 0$.*
4. *If there exists A such that $H(\infty) - H(A) > 0$, then $\tau^* < \infty$.*

Properties 1 and 2 show that the optimum τ^* must be a zero of h—moreover, one that separates a negative interval to the left from a positive interval to the right (unless $\tau^* = 0$ or ∞). Properties 3 and 4 narrow the set of zeros that includes the optimal value τ^*, and thus help to establish which one is the optimum.

Four examples follow. The first two illustrate Theorem 5.1 in a dimension-based situation in which each dimension is processed individually. In the first example, each dimension's underlying A^* is known—equivalently, its $h(A; A^*)$ is known. In the second, the underlying value of each dimensional absolute distance is chosen from two possibilities, and only the mixing distribution of these two values and the corresponding h-functions are known.

The last two examples introduce the ideas that we will explore in the next section for building models.

Throughout these examples, notice that the dimensional contributions are only ever used in expected-value form, and the component h-function is the OLS $h(\widehat{A}; A^*)$.

Example 1 *Subset selection from a single mixture.* Consider the function $h(\widehat{A}; A^*)$ illustrated in Figure 1(a). We suppose that all dimensions have the same $h(\widehat{A}; A^*)$.

Noisy dimensions, and ones whose effects are undetectable. A noisy dimension, for which $h(\widehat{A}; 0)$ is always negative, will be eliminated from the model no matter how large its absolute distance $\widehat{A}$. Since $\lim_{A^* \to 0} h(\widehat{A}; A^*) = h(\widehat{A}; 0)$, nonredundant dimensions behave more like noisy ones as their underlying effect decreases—in other words, their contribution eventually becomes undetectable. When $A^* \leq 0.5$, any contribution is completely overwhelmed by the noise, and no subset selection procedure can detect it.

Dimensions with small vs. large effects. When the estimate resides in a negative interval of h, its contribution is negative. All hs, no matter how large their A^*, have at least one negative interval $(4A^*, \infty)$. This invalidates all subset selection schemes that eliminate dimensions based on thresholding their variation reductions with fixed threshold, because a large estimate $\widehat{A}$ does not necessarily mean that the corresponding variable is contributive—its contribution also depends on A^*. The reason that threshold-type selection works at all is that the estimate $\widehat{A}$ in a dimension whose effect is large is less likely to fall into a negative interval than one whose effect is small.

The OLSC(τ) *criterion.* The OLS subset selection criterion OLSC(τ) eliminates dimensions whose OLS estimate falls below the threshold τ, where $\tau = 2$ for AIC, $\log n$ for BIC, $2 \log k$ for RIC, and the optimal value is τ^* as defined in (35). Since dimensions should be discarded based on the sign of their expected contribution, we consider three cases: dimensions with zero and small effects, those with moderate effects, and those with large effects.

When a dimension is redundant, i.e. $A^* = 0$, it should always be discarded no matter how large the estimate $\widehat{A}$. This can only be done by OLSC(τ^*), with $\tau^* = \infty$ in this case. Whenever $\tau < \infty$, dimensions whose $\widehat{A}$ exceeds τ are kept inside the model: thus a certain proportion of redundant variables are included in the final model. Dimensions with small effects behave similarly to noisy ones, and the threshold value $\tau^* = \infty$ is still best—which results in the null model.

Suppose that dimensions have moderate effects. As the value of A^* increases from zero, the value of the cumulative expected contribution $H(\tau)$ will at some point change sign. At this point, the model found by OLSC(τ), which heretofore has been better than the full model, becomes worse than it. Hence there is a value of A^* for which the predictive ability of $\mathcal{M}^{\text{OLSC}(\tau)}$ is the same as that of the full model. Furthermore, there exists a value of A^* at which the predictive ability of the null model is the same as that of the full model. In these cases, model $\mathcal{M}^{\text{OLSC}(\tau^*)}$ is either the null model or the full one, since τ^* is either 0 or ∞ depending on the value of A^*.

When each dimension has a large effect—large enough that the position of the second zero of $h(\widehat{A}; A^*)$ is at least τ—any OLSC(τ) with fixed τ will inevitably eliminate contributive dimensions. This means that the full model is a better one than $\mathcal{M}^{\text{OLSC}(\tau)}$. Furthermore, OLSC($\tau^*$) with $\tau^* = 0$ will always choose the full model, which is the optimal model for every $\mathcal{M}^{\text{OLSC}(\tau)}$.

Shrinkage methods in orthogonal space. In orthogonal regression, when $X'X$ is a diagonal matrix, contribution functions help explain why shrinkage methods work. These methods shrink the parameter values of OLS models and use smaller values than the OLS estimates. This may or may not change the signs of the OLS estimated parameters; however, for orthogonal regressions, the signs of the parameters are left unchanged. In this situation, therefore, shrinking parameters is tantamount to shrinking the $\widehat{A}$'s. Ridge regression shrinks all the parameters while the nn-garrote and lasso shrink the larger parameters and zero the smaller ones.

When A^* is small, it is possible to choose a shrinkage parameter that will shrink $\widehat{A}$'s that lie between A^* and $4A^*$ to around A^*, and shrink the negative contributions outside $4A^*$ to become positively contributive—despite the fact that $\widehat{A}$ around the maximum point A^* are shrunk to smaller values. This may give the resulting model lower predictive error than any model selected by OLSC(τ), including $\tau = \tau^*$. Zeroing the smaller $\widehat{A}$'s by the nn-garrote and lasso does not guarantee better predictive accuracy than ridge regression, for these dimensions might be contributive. When A^* is large, shrinkage methods perform badly because the distribution of $\widehat{A}$ tends to be sharper around A^*. This is why the OLS subset selection often does better in this situation.

Example 2 *Subset selection from a double mixture.* Suppose

$$h(\widehat{A}) = \frac{k_1}{k} h(\widehat{A}; \alpha_1^*) + \frac{k_2}{k} h(\widehat{A}; \alpha_2^*), \tag{37}$$

where $k = k_1 + k_2$. For k_1 dimensions the underlying A^* is α_1^*, while for the remaining k_2 dimensions it is α_2^*. $\widehat{A}$ is an observation sampled from the mixture distribution $f(\widehat{A}) = \frac{k_1}{k} f(\widehat{A}; \alpha_1^*) + \frac{k_2}{k} f(\widehat{A}; \alpha_2^*)$. Altering the values of k_1, α_1^*, k_2 and α_2^* yields the different hs illustrated in Figure 2.

Let us consider the optimal threshold τ^* of OLSC(τ). In Figure 2(a), where $\alpha_1^* = 0$, $\alpha_2^* = 3$ and $k_1 : k_2 = 80 : 20$, no positive interval exists despite the fact that there are 20 nonredundant dimensions. This is because the effect of all the nonredundant dimensions is overwhelmed by the noisy ones. No matter how large its effect, any dimension can be overwhelmed by a sufficient number of noisy ones. In this case $\tau^* = \infty$ and OLSC(τ^*) selects the null model.

In Figure 2(b), which is obtained from the previous situation by altering $k_1 : k_2$ to $75 : 25$, there is a positive interval. But $H(\infty)$ is the minimum of all zeros, so that τ^* remains ∞ and the model chosen by OLSC(τ^*) is still the null one. The contributive dimensions are still submerged by the noisy ones.

If a finite threshold value τ^* exists, it must satisfy $H(\infty) - H(\tau^*) > 0$ (Property 4 of Corollary 5.1.1). Then it can take on any nonnegative value by adjusting the four parameters k_1, α_1^*, k_2 and α_2^*. Figure 2(c) shows three functions h, obtained by setting α_1^* and α_2^* to 0 and 20 respectively and making the ratio between k_1 and k_2 $5 : 95$, $50 : 50$, and $95 : 5$. As these curves show, the corresponding values of τ^* are about 2, 4 and 8.

In Figure 2(d), $k_1 : k_2 = 95 : 5$ and α_1^* and α_2^* are set to make $H(\mathcal{Z}_3) = 0$, where $\mathcal{Z}_3$ is the third zero of h. In this case, there are two possibilities for τ^*: the origin $\mathcal{Z}_1$, and $\mathcal{Z}_3$. $\mathcal{Z}_3$ gives a simpler model. Notice that the number of parameters of the two models is in the approximate ratio $5 : 100$. However, the balance is easily broken—for example, if α_1^* increases slightly, then there is a single value 0 for

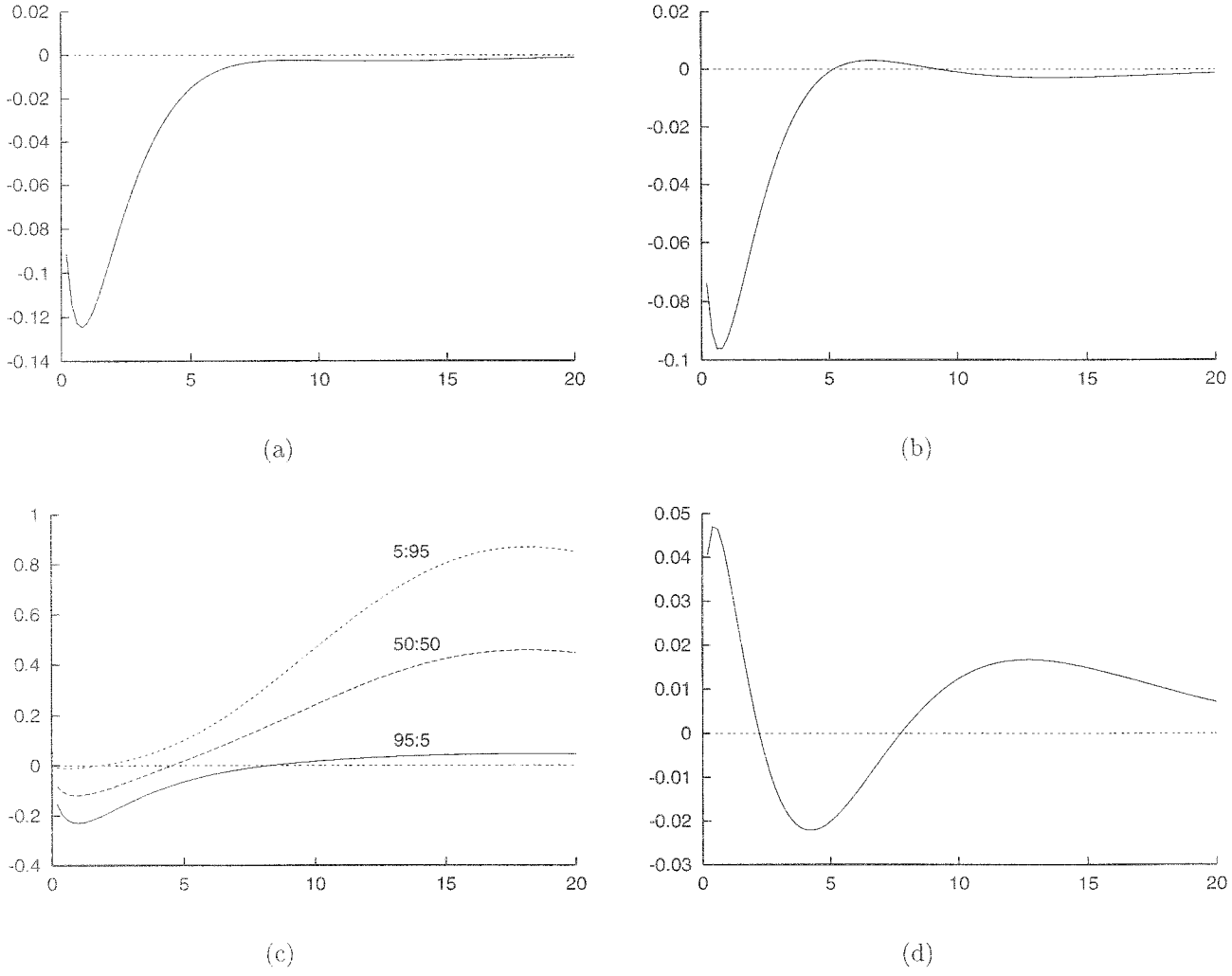


Figure 2: $h(\widehat{A}) = \frac{k_1}{k} h(\widehat{A}; \alpha_1^*) + \frac{k_2}{k} h(\widehat{A}; \alpha_2^*)$. (a). $\alpha_1^* = 0, \alpha_2^* = 3, k_1 : k_2 = 80 : 20$. (b). $\alpha_1^* = 0, \alpha_2^* = 3, k_1 : k_2 = 75 : 25$. (c). $\alpha_1^* = 0, \alpha_2^* = 20$, $k_1 : k_2$ are respectively $5 : 95, 50 : 50, 95 : 5$. (d). α_1^* and α_2^* are carefully set with fixed $k_1 : k_2 = 95 : 5$ such that $H(Z_3) = 0$.

τ^*. Although the larger model has slightly smaller predictive error than the smaller one, it is much more complex. Here is a situation where a far more succinct model can be obtained with a small sacrifice in predictive accuracy.

Example 3 *Subset selection based on the sign of h.* Although OLSC(τ^*) is optimal among every OLSC(τ), it has limitations. It always deletes dimensions whose $\widehat{A}$'s fall below the threshold, and retains the remaining ones. Thus it may delete dimensions that lie in the positive intervals of h, and retain ones in the negative intervals. We know that the sign of h is the sign of the expected contribution, and the selection can be improved by using this fact: we simply retain dimensions whose $h(\widehat{A})$ is positive and discard the remainder.

In the next section we formalize this idea and prove that predictive accuracy always increases. Because the positive and negative intervals of h can lie anywhere along the half real line, this procedure may retain dimensions with smaller $\widehat{A}$ and discard ones with larger $\widehat{A}$. Or it may delete a dimension whose $\widehat{A}$ lies between those of the other dimensions. For example, in Figure 2(b), the new procedure will keep all the dimensions whose $\widehat{A}$'s lie within the small positive interval, despite the fact that OLSC(τ^*) chooses the null model.

Example 4 *General modeling.* So far we have only discussed subset selection, where the estimate $\widehat{A}$ is either altered to zero or remains the same. A natural question is whether better results might be obtained by relaxing this constraint—and indeed they are. For example, in Example 1, where all dimensions are noisy, OLSC(τ^*) deletes them all and chooses the null model. This in effect replaces the estimate $\widehat{A}$ by the underlying A^*, which is 0 in this case. Similarly, when all estimates $\widehat{A}$ have the same underlying A^* (which is non-zero), and all $\widehat{A}$'s are updated to A^*, the estimated model improves significantly—even though no dimension is redundant.

A similar thing happens when the underlying A^* is distributed according to a mixture distribution. In Figure 2(a), the h-function of the mixture has no positive interval, although 20 out of 100 dimensions have $A^* = 3$. The best that subset selection can do is to discard all dimensions. However, a dimension with $\widehat{A} = 20$—despite $h(\widehat{A})$ being less than 0—is unlikely to be a redundant one; it is more likely to belong to the group for which $A^* = 3$. Altering its OLS estimate $\widehat{A}$ from 20 to 3 is likely to convert the dimension into a contributive one.

The next section formulates a formal estimation procedure based on this idea.

6 Modeling with known $G(A^*)$

We now assume that the underlying $G(A^*)$ is known and define a group of six procedures, collectively called *pace regression*, which build models by adjusting the orthogonal projections based on estimations of the expected dimensional contributions. They are denoted $\text{PACE}_1, \text{PACE}_2, \ldots, \text{PACE}_6$, and the model produced by PACE_i is written $\mathcal{M}^{\text{PACE}_i}$. "Pace" stands for "projection adjustment by contribution estimation." We will learn in the next section how to estimate $G(A^*)$.

$G(A^*)$ is the distribution function of a random variable A^* that represents the dimensional absolute distance of the underlying model $\mathcal{M}^*$. These distances, denoted by $A_1^*, \ldots, A_k^*$, form a sample of size k from $G(A^*)$. When the mixing distribution $G(A^*)$ is known, (30) can be used to obtain the mixture pdf $f(A; G)$ from the component distribution pdf $f(A; A^*)$. From this, along with the component h-function $h(A; A^*)$, the functions $h(A; G)$ and $H(A; G)$ can be found from (31) and (23) respectively. As Theorem 4.1 shows, OLS estimation provides both $f(A; A^*)$ and $h(A; A^*)$.

The pace procedures consist of two steps. The first, which is the same for all procedures, generates an initial model $\mathcal{M}$. We always use the OLS estimate $\widehat{\mathcal{M}}$ for this, although any initial model can be used so long as the component distribution and the component h-function are available. Decomposing $\widehat{\mathcal{M}}$ in a given orthogonal space yields the model's dimensional absolute distances, say $\widehat{A}_1, \ldots, \widehat{A}_k$. These are in fact a sample from a mixture distribution $F(\widehat{A}; G)$ with known component distribution $F(\widehat{A}; A^*)$ and mixing distribution $G(A^*)$. In the second step, the final model is generated from either (14) or (15), where $\widetilde{A}_1, \ldots, \widetilde{A}_k$ are obtained by updating $\widehat{A}_1, \ldots, \widehat{A}_k$. The new procedures differ in how the updating is done.

To characterize the resulting performance, we define a class of estimators and show that pace estimators are optimal within this class, or a subclass of it. The class of estimators $\mathfrak{M}_k$ (where k is the number of orthogonal dimensions) is as follows: given the initial model $\widehat{\mathcal{M}}$ and its absolute distances in an orthogonal decomposed model space, every member of $\mathfrak{M}_k$ is an updating of $\widehat{\mathcal{M}}$ by (14) and *vice versa*, where the updating is entirely dependent on the set of absolute distances of $\widehat{\mathcal{M}}$. Clearly, $\mathcal{M}^{\text{OLSC}(\tau)} \in \mathfrak{M}_k$ for any τ.

Various corollaries below establish that the pace estimators are better than others, although the proofs are omitted. Each estimator's optimality is established by a theorem that applies to a specific subclass, and proving each corollary reduces to exhibiting an example where the performance is actually better, which is easily done. When we say that one estimator is "better than" another, we mean in the usual sense of risk. With one exception, better estimators have lower expected risk (the exception, as noted below, is Corollary 6.2.1, which compares PACE_1 and PACE_3).

Of the six procedures, PACE_1 and PACE_2 perform model-based selection, that is, they select a subset model from a sequence. Procedures PACE_3 and PACE_4 address the dimension-based situation, where each orthogonal dimension is tested and selected (or not) individually. If these procedures are used for a sequence of orthogonal nested models, the resulting model may not belong to the sequence. The last two procedures, PACE_5 and PACE_6, are not selection procedures. Instead, they update the absolute distances of the estimated model to values chosen appropriately from the nonnegative half real line.

Theorem 5.1 shows that there is an optimal threshold for threshold-type OLS subset selection that minimizes the expected risk. We use this idea for nested models.

Procedure 1 (PACE_1). *Given a sequence of orthogonal nested models, let* $\tau = \arg\min_{\mathcal{Z} \in \{\mathcal{Z}_i\}} H(\mathcal{Z})$, *where* $\{\mathcal{Z}_i\}$ *is the set of zeros of* h. *Output the model in the sequence selected by* $\text{OLSC}(\tau)$.

According to Corollary 5.1.1, Procedure 1 finds the optimal threshold τ^*. Therefore $\text{PACE}_1 = \text{OLSC}(\tau^*)$ and $\mathcal{M}^{\text{PACE}_1} = \mathcal{M}^{\text{OLSC}(\tau^*)}$. Since the sequence of dimensional absolute distances of the model $\widehat{\mathcal{M}}$ is not necessarily always decreasing, $\mathcal{M}^{\text{OLSC}(\tau^*)}$ is not guaranteed to be the minimum expected risk estimator. However, in practice these distances do generally decrease, and so in the nested model situation $\mathcal{M}^{\text{OLSC}(\tau^*)}$ is an excellent approximation to the minimum expected risk estimator. In this sense, PACE_1 is superior to other OLS subset selection procedures—OLS, AIC, BIC, RIC, and CIC—for these do not use the optimal threshold τ^*. In particular, the procedure CIC uses a threshold that depends on the number of variables in the subset model as well as the total number of variables. The relative performance of these procedures depends on the particular experiments used for comparison, since datasets exist for which any selection criterion's threshold coincides with the optimal value τ^*.

Instead of approximating the optimum as PACE_1 does, PACE_2 always selects the optimal model from a sequence of nested models, where optimality is in the usual sense of risk.

Procedure 2 (PACE_2). *Among a sequence of orthogonal nested models, output the one which has the largest value of* $\sum_{i=1}^{j} h(\widehat{A}_i)/f(\widehat{A}_i)$.

Theorem 6.1 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_2}$ has the smallest risk in a subclass of $\mathfrak{M}_k$ in which each estimator can only select from the sequence of orthogonal nested models that is provided.*

Proof. Let $\mathcal{M}_j$ be the selected model, then $\widetilde{A}_i = 0$ if $i > j$ and $\widetilde{A}_i = \widehat{A}_i$ if $i \leq j$. From the definition of dimensional contribution (19), we have

$$E[\mathcal{L}(\mathcal{M}_j)] = \mathcal{A}(\mathcal{M}^*) - \sum_{i=1}^{j} E[C(\widehat{A}_i)]. \tag{38}$$

Because $E[C(\widehat{A}_i)] = h(\widehat{A}_i)/f(\widehat{A}_i)$ by (23), minimizing $E[\mathcal{L}(\mathcal{M}_j)]$ is equivalent to maximizing $\sum_{i=1}^{j} h(\widehat{A}_i)/f(\widehat{A}_i)$ with respect to j. This completes the proof. □

Corollary 6.1.1 *Given $G(A^*)$ and a sequence of orthogonal nested models, $\mathcal{M}^{\text{PACE}_2}$ is a better estimator than $\mathcal{M}^{\text{OLSC}(\tau)}$ for any τ. This includes $\mathcal{M}^{\text{PACE}_1}$, $\mathcal{M}^{\text{OLS}}$, $\mathcal{M}^{\text{AIC}}$, $\mathcal{M}^{\text{BIC}}$, $\mathcal{M}^{\text{RIC}}$ and $\mathcal{M}^{\text{CIC}}$.*

Since n-asymptotically the cross-validation subset selection procedure $\text{CV}(d) = \text{OLSC}((2n-d)/(n-d))$ and the bootstrap subset selection procedure $\text{BS}(m) = \text{OLSC}((n+m)/m)$ (see subsection 2.6), we have

Corollary 6.1.2 *Given $G(A^*)$ and a sequence of orthogonal nested models, n-asymptotically $\mathcal{M}^{\text{PACE}_2}$ is a better estimator than $\mathcal{M}^{\text{CV}(d)}$ for any d and $\mathcal{M}^{\text{BS}(m)}$ for any m.*

In fact, the difference between the models generated by PACE_1 and PACE_2 is small, because we have

Corollary 6.1.3 *Given $G(A^*)$, if the elements of $\{\widehat{A}_j; j = 1, \ldots, k\}$ are in decreasing order as j increases, almost surely, the difference between the values of the risk of $\mathcal{M}^{\text{PACE}_2}$ and the expected risk of $\mathcal{M}^{\text{PACE}_1}$ tends to zero k-asymptotically. In particular, if $\mathcal{M}^{\text{PACE}_1}$ is unique, $\mathcal{M}^{\text{PACE}_1} = \mathcal{M}^{\text{PACE}_2}$ a.s. k-asymptotically.*

Proof. Since $\{\widehat{A}_j; j = 1, \ldots, k\}$ are in decreasing order as j increases, $\mathcal{M}^{\text{PACE}_2}$ is in effect model that minimizes $\int_{\mathbb{R}^+} E[\mathcal{L}(\mathcal{M}_j)]\, dF_k(\widehat{A})$ with respect to j, where $F_k(\widehat{A})$ is Kolmogrov's empirical CDF of $\widehat{A}$, and $\mathcal{M}^{\text{PACE}_1}$ is the model that minimizes $\int_{\mathbb{R}^+} E[\mathcal{L}(\mathcal{M}_j)]\, dF(\widehat{A})$ with respect to j. Clearly, $\lim_{k\to\infty} \int_{\mathbb{R}^+} E[\mathcal{L}(\mathcal{M}_j)]\, dF_k(\widehat{A}) = \int_{\mathbb{R}^+} E[\mathcal{L}(\mathcal{M}_j)]\, dF(\widehat{A})$ almost surely, because of the Dominated Convergence theorem(Galambos, 1995). Apparently, if $\mathcal{M}^{\text{PACE}_2} \neq \mathcal{M}^{\text{PACE}_1}$ as $k \to \infty$, there would exist more than one $\mathcal{M}^{\text{PACE}_1}$, contradicting the uniqueness of $\mathcal{M}^{\text{PACE}_1}$. □

These two procedures show how to select the best model in a sequence of $k+1$ nested models. However, no model sequence can guarantee that the optimal model is one of the nested models. Thus we now consider dimension-based modeling, where the final model can be a combination of any dimensions. With selection, the number of potential models given the projections on k dimensions is as large as 2^k. When the orthogonal basis is provided by a sequence of orthogonal nested models, this kind of selection means that the final model may not be one of the nested models, and its parameter vector may not be the OLS fit in terms of the original variables.

Procedure 3 (PACE_3). *Let $\tau = \arg\min_{\mathcal{Z}\in\{\mathcal{Z}_i\}} H(\mathcal{Z})$, where $\{\mathcal{Z}_i\}$ is the set of zeros of h. Set $\widetilde{A}_j = 0$ if $\widehat{A}_j \leq \tau$; otherwise $\widetilde{A}_j = \widehat{A}_j$. Output the model determined by $\{\widetilde{A}_1, \ldots, \widetilde{A}_k\}$.*

Theorem 6.2 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_3}$ is the minimum expected risk estimator of $\mathcal{M}^*$ with respect to $F(A)$ in an estimator class which is the subclass of $\mathfrak{M}_k$ in which every estimator is determined by $\{\widetilde{A}_1, \ldots, \widetilde{A}_k\}$ where $\widetilde{A}_j \in \{0, \widehat{A}_j\}$ for all j.*

The proof is omitted; it is similar to that of Theorem 5.1.

Since $\mathcal{M}^{\text{PACE}_1}$ is the model determined by $\{\widetilde{A}_1, \ldots, \widetilde{A}_k\}$ where $\widetilde{A}_j$ is either $\widehat{A}_j$ or 0 depending on whether the associated variable is included or not, this estimator belongs to the class described in Theorem 6.2. This gives the following corollary.

Corollary 6.2.1 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_3}$ is a better estimator (in the sense of expected risk) than $\mathcal{M}^{\text{PACE}_1}$.*

The difference between the models generated by PACE_1 and PACE_3 is also small, because

Corollary 6.2.2 *Given $G(A^*)$, if the elements of $\{\widehat{A}_j; j = 1, \ldots, k\}$ are in decreasing order as j increases, $\mathcal{M}^{\text{PACE}_1} = \mathcal{M}^{\text{PACE}_3}$.*

As we have seen, whether or not a dimension is contributive is indicated by the sign of the corresponding h-function. This leads to the next procedure.

Procedure 4 (PACE_4). *Set $\widetilde{A}_j = 0$ if $h(\widehat{A}_j) \leq 0$; otherwise $\widetilde{A}_j = \widehat{A}_j$. Output the model determined by $\{\widetilde{A}_1, \ldots, \widetilde{A}_k\}$.*

PACE_4 does not rank dimensions in order of absolute distance and eliminate those with smaller distances, as do conventional subset selection procedures and the preceding pace procedures. Instead, it eliminates dimensions that are not contributive in the estimated model irrespective of the magnitude of their dimensional absolute distance. It may eliminate a dimension with a larger absolute distance than another dimension that is retained. (In fact the other procedures may end up doing this occasionally, but they do so only because of incorrect ranking of variables.)

Theorem 6.3 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_4}$ has the smallest risk in the subclass of $\mathfrak{M}_k$ in which every estimator is determined by $\{\widetilde{A}_1, \ldots, \widetilde{A}_k\}$ where $\widetilde{A}_j \in \{0, \widehat{A}_j\}$ for all j.*

The proof is omitted; it is similar to that of Theorem 6.1.

Because the estimator class defined in Theorem 6.3 covers the classes defined in Theorems 6.1 and 6.2,

Corollary 6.3.1 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_4}$ is a better estimator than $\mathcal{M}^{\text{PACE}_2}$ and $\mathcal{M}^{\text{PACE}_3}$.*

PACE_1, PACE_2, PACE_3 and PACE_4 are all selection procedures: each updated dimensional absolute distance of the estimated model must be either 0 or $\widehat{A}_j$. The optimal value of $\widetilde{A}_j$ is often neither of these. If the possible values are chosen from $\mathbb{R}^+$ instead, the best updated estimate $\widetilde{A}_j$ is the one that maximizes the expected contribution of the jth dimension given $\widehat{A}_j$ and $G(A^*)$. The optimality is achieved over an uncountably infinite set of potential models. This relaxation can improve performance dramatically even when there are no noisy dimensions.

Procedure 5 (PACE_5). *Output the model determined by $\{\widetilde{A}_1, \ldots, \widetilde{A}_k\}$, where*

$$\widetilde{A}_j = \arg\max_{A \in \mathbb{R}^+} \int_{\mathbb{R}^+} \frac{h(A; A^*)}{f(A; A^*)} f(\widehat{A}_j; A^*)\, dG(A^*). \tag{39}$$

Theorem 6.4 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_5}$ has the smallest risk of all estimators in $\mathfrak{M}_k$.*

Proof. Each OLS estimate $\widehat{A}_j$ is an observation sampled from the mixture pdf $f(\widehat{A}; G)$ determined by the component pdf $f(\widehat{A}; G)$ and the mixing distribution $G(A^*)$. If $\widehat{A}_j$ is replaced by any $A \in \mathbb{R}^+$, the expected contribution of A given $\widehat{A}_j$ and $G(A^*)$ is

$$E[C(A)|\widehat{A}_j, G] = \frac{\int_{\mathbb{R}^+} E[C(A)|A^*] f(\widehat{A}_j; A^*)\, dG(A^*)}{\int_{\mathbb{R}^+} f(\widehat{A}_j; A^*)\, dG(A^*)}. \tag{40}$$

Since $\int_{\mathbb{R}^+} f(\widehat{A}_j; A^*)\, dG(A^*)$ is constant for every A, and $E[C(A)|A^*] = h(A; A^*)/f(A; A^*)$ from (23), (39) actually maximizes over the expected contribution of A. From (19), this is equivalent to minimizing the expect loss in the jth dimension. Because all dimensions are mutually independent, the risk of the updated model given the set $\{\widehat{A}_j\}$ and $G(A^*)$ is minimized. □

Corollary 6.4.1 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_5}$ is a better estimator than $\mathcal{M}^{\text{PACE}_4}$.*

This motivates a new shrinkage method: shrink the magnitude of the orthogonal projections of the model $\widehat{\mathcal{M}}$. This is equivalent to updating the OLS estimate $\widehat{\mathcal{M}}$ to a model that satisfies $\widetilde{A}_j \leq \widehat{A}_j$ for every j. Since all shrinkage estimators of this type obviously yield a member of $\mathfrak{M}_k$, $\mathcal{M}^{\text{PACE}_5}$ is a better estimator than any of them. Models produced by shrinkage methods in the literature—ridge regression (including ridge regression for subset selection), the nn-garrote, and the lasso—do not necessarily belong to this subclass, and so we cannot show that the new estimator is superior to them in general. However, in the important special case of orthogonal regression, when the vectors corresponding to each variable are taken as the orthogonal axes, these shrinkage methods do belong to the above subclass. Therefore,

Corollary 6.4.2 *Given $G(A^*)$, $\mathcal{M}^{\text{PACE}_5}$ is a better estimator for orthogonal regression than $\mathcal{M}^{\text{ridge}}$, $\mathcal{M}^{\text{nn-garrote}}$ and $\mathcal{M}^{\text{lasso}}$.*

A general explicit solution to (39) does not seem to exist. Rather than resorting to numerical techniques, however, a good approximate solution is available. Considering that

$$\frac{h(A; A^*)}{f(A; A^*)} = \frac{c(\sqrt{A}; \sqrt{A^*}) p(\sqrt{A}; \sqrt{A^*}) + c(-\sqrt{A}; \sqrt{A^*}) p(-\sqrt{A}; \sqrt{A^*})}{p(\sqrt{A}; \sqrt{A^*}) + p(-\sqrt{A}; \sqrt{A^*})}, \tag{41}$$

the dominant part on the right-hand side is $c(\sqrt{A}; \sqrt{A^*}) p(\sqrt{A}; \sqrt{A^*})$—and its dominance increases dramatically as A^* increases. Replacing $h(A; A^*)/f(A; A^*)$ in (39) by $c(\sqrt{A}; \sqrt{A^*})$, we obtain the following approximation to PACE_5.

Procedure 6 (PACE_6). *Output the model determined by $\{\widetilde{A}_1, \ldots, \widetilde{A}_k\}$ where*

$$\widetilde{A}_j = \arg\max_{A \in \mathbb{R}^+} \int_{\mathbb{R}^+} c(\sqrt{A}; \sqrt{A^*}) f(\widehat{A}_j; A^*)\, dG(A^*). \tag{42}$$

Equation (42) can be solved by setting the first derivative of the right-hand side to zero, resulting in

$$\widetilde{A}_j = \left[\frac{\int_{\mathbb{R}^+} \sqrt{A^*} f(\widehat{A}_j; A^*)\, dG(A^*)}{\int_{\mathbb{R}^+} f(\widehat{A}_j; A^*)\, dG(A^*)} \right]^2. \tag{43}$$

In (39), (42) and (43) the true distribution $G(A^*)$ is discrete (as in (32)), so they become respectively

$$\widetilde{A}_j = \arg\max_{A \in \mathbb{R}^+} \sum_{i=1}^{m} \frac{h(A; \alpha_i^*)}{f(A; \alpha_i^*)} f(\widehat{A}_j; \alpha_i^*) w_i, \tag{44}$$

$$\widetilde{A}_j = \arg\max_{A \in \mathbb{R}^+} \sum_{i=1}^{m} c(\sqrt{A}; \sqrt{\alpha_i^*}) f(\widehat{A}_j; \alpha_i^*) w_i, \tag{45}$$

and

$$\widetilde{A}_j = \left[\frac{\sum_{i=1}^{m} \sqrt{\alpha_i^*} f(\widehat{A}_j; \alpha_i^*) w_i}{\sum_{i=1}^{m} f(\widehat{A}_j; \alpha_i^*) w_i} \right]^2. \tag{46}$$

The following loose bound can be obtained for the increased risk suffered by the PACE_6 approximation.

Theorem 6.5

$$0 \leq E[\mathcal{L}_j(\mathcal{M}^{\text{PACE}_6})|\widehat{A}_j, G(A^*)] - E[\mathcal{L}_j(\mathcal{M}^{\text{PACE}_5})|\widehat{A}_j, G(A^*)] < 2e^{-1}. \tag{47}$$

Proof. See Appendix B.

Appendix B actually obtains the tighter bound $4\sqrt{\widetilde{A}_j^{\text{PACE}_6}\alpha^*}e^{-2\sqrt{\widetilde{A}_j^{\text{PACE}_6}\alpha^*}}$. This bound rises from zero at the origin, achieves the maximum $2e^{-1}$ at the point $\widetilde{A}_j^{\text{PACE}_6}\alpha^* = 0.5$, and thereafter drops exponentially to zero as $\widetilde{A}_j^{\text{PACE}_6}\alpha^*$ increases. α^* is one of $G(A^*)$'s discrete points and has a maximum value of $c(\sqrt{\widetilde{A}_j^{\text{PACE}_6}}; \sqrt{\alpha_i^*}) - E[C(\widetilde{A}_j^{\text{PACE}_6})|\alpha_i^*]$ for $i = 1, \ldots, m$. It follows that the increased risk caused by approximating PACE$_6$ is usually close to zero.

All these pace procedures adjust the magnitude of the orthogonal projections of the OLS estimate $\widehat{\mathcal{M}}$, based on an estimate of the expected dimensional contributions. Among them, PACE$_5$ and PACE$_6$ go the furthest: each projection of $\widehat{\mathcal{M}}$ onto the orthogonal axis can be adjusted to any nonnegative value and the adjusted value achieves (or approximately achieves) the greatest expected contribution, corresponding to the minimum risk. These two procedures can shrink, retain or even expand the values of the absolute dimensional distances. Surprising though it may sound, increasing a zero distance to a much higher value can improve predictive accuracy.

Of the six pace procedures, PACE$_2$, PACE$_4$ and PACE$_6$ are most appropriate for practical applications. PACE$_6$ generates a very good approximation to the model from PACE$_5$, which is the best of the six procedures. Procedure PACE$_2$ chooses the best member of a sequence of subset models that is provided to it, which is useful if prior information dictates the sequence of subset models. PACE$_1$ and PACE$_3$ involve numerical integration and have higher risk than other procedures. PACE$_4$, which is a lower risk procedure than PACE$_2$, is useful for dimension-based subset selection.

7 The estimation of $G(A^*)$

Now it is time to consider how to estimate $G(A^*)$ from $\widehat{A}_1, \ldots, \widehat{A}_k$, which are the dimensional absolute distances of the OLS estimate $\widehat{\mathcal{M}}$. Once this is accomplished, the procedures described in the last section become fully defined by replacing the true $G(A^*)$, which we assumed in the last section was known, with the estimate. The estimation of $G(A^*)$ is an independent step in these modeling procedures, and can be investigated independently. It critically influences the quality of the final model—better estimates of $G(A^*)$ give better estimators for the underlying model.

$\widehat{A}_1, \ldots, \widehat{A}_k$ are actually a sample from a mixture distribution whose component distribution $F(\widehat{A}; A^*)$ is known and whose mixing distribution is $G(A^*)$. Estimating $G(A^*)$ from data points $\widehat{A}_1, \ldots, \widehat{A}_k$ is tantamount to estimating the mixing distribution. Note that the mixture here is a countable one—the underlying $G(A^*)$ has support at $\widehat{A}_1, \ldots, \widehat{A}_k$, and the number of support points is unlimited as $k \to \infty$.

There are many ways of estimating mixing distributions: the method of moments, maximum likelihood, Bayesian methods, and minimum distance methods (Titterington et al., 1985; McLachlan and Basford, 1988; Lindsay, 1995). However, only the minimum distance approach is suitable for estimating unlimited countable mixtures. This approach has some advantages over others: it can estimate arbitrary mixing distribution (others cannot); it is computationally cheap (being a mathematical programming problem which has an efficient solution); it always yields a globally optimal solution (other methods may converge to a local optimum and need careful selection of initial values).

The following theorem guarantees that if the mixing distribution is estimated sufficiently well, the pace regression procedures continue to enjoy the various properties proved above in the limit of large k.

Theorem 7.1 *Let $\{G_k(A^*)\}$ be a sequence of CDF estimators. If $G_k(A^*) \to_w G(A^*)$ a.s. as $k \to \infty$, and the known $G(A^*)$ is replaced by the estimator $G_k(A^*)$, Theorems 5.1–6.5 (and all their corollaries) hold k-asymptotically a.s.*

Proof. According to the Helly-Bray theorem(Galambos, 1995), $G_k(A^*) \to_w G(A^*)$ a.s. as $k \to \infty$ implies the almost sure pointwise convergence of all the objective functions used in these theorems (and their corollaries) to the underlying corresponding functions, because these functions are continuous. This further implies the almost sure convergence of the optimal values, and of the locations where these optima are achieved, as $k \to \infty$. This completes the proof. □

All of the above results utilize the loss function $||y_{\mathcal{M}} - y^*||^2/\sigma^2$. However, our real interest is $||y_{\mathcal{M}} - y^*||^2$. Therefore we need the following corollary.

Corollary 7.1.1 *If the loss function $||y_{\mathcal{M}} - y^*||^2/\sigma^2$ is replaced by $||y_{\mathcal{M}} - y^*||^2$ in Theorems 5.1–6.5 (and all their corollaries),*

1. *Theorem 7.1 continues to hold, if σ^2 is known;*
2. *Theorem 7.1 holds almost surely as $n \to \infty$ if σ^2 is replaced with an n-asymptotically strongly consistent estimator.*

It is well known that both the unbiased OLS estimator $\hat{\sigma}^2$ and the biased maximum likelihood estimator are n-asymptotically strongly consistent, if $\lim_{n\to\infty} k/n = 0$.

In view of Theorem 7.1, any estimator of the mixing distribution is able to provide the desired theoretic results in the limit if it is *strongly consistent* in the sense that, almost surely, it converges weakly to the underlying mixing distribution as $k \to \infty$. A few minimum distance estimators are known to be strongly consistent. Before introducing them, we need some notation.

Let $x_1, \ldots, x_k$ be a sample chosen according to a mixture distribution, and suppose (without loss of generality) that the data is ordered so that $x_1 \leq x_2 \leq \cdots \leq$

x_k. Let $G_k(\theta)$ be a discrete estimator of the underlying mixing distribution with a set of (potential) support points at $\{\theta_{kj}; j = 1, \ldots, m_k\}$. Each θ_{kj} provides a component of the final mixture with weight $w_{kj} \geq 0$, where $\sum_{i=1}^{m_k} w_{ki} = 1$. Given the support points, obtaining $G_k(\theta)$ is equivalent to computing the weight vector $w_k = (w_{k1}, w_{k2}, \ldots, w_{km_k})'$. Denote by $F_{G_k}(x)$ the estimated mixture CDF with respect to $G_k(\theta)$.

Choi and Bulgren (1968) investigated a minimum distance estimator with

$$\frac{1}{k}\sum_{j=1}^{k}[F_{G_k}(x_j) - j/k]^2 \tag{48}$$

as the distance measure. Minimizing this quantity with respect to G_k yields a strongly consistent estimator. A slight improvement is obtained by using the Cramér-von Mises statistic

$$\frac{1}{k}\sum_{j=1}^{k}[F_{G_k}(x_j) - (j - 1/2)/k]^2 + 1/(12k^2), \tag{49}$$

which essentially replaces j/k in (48) by $(j - \frac{1}{2})/k$ without affecting the asymptotic result. As might be expected, this reduces the bias for small-sample cases, as was demonstrated empirically by Macdonald (1971) in a note on Choi and Bulgren's paper.

Deely and Kruse (1968) investigated a similar estimator that uses the sup-norm associated with the Kolmogorov-Smirnov test. The minimization is over

$$\sup_{1\leq j\leq k}\{|F_{G_k}(x_j) - (j-1)/k|, |F_{G_k}(x_j) - j/k|\}, \tag{50}$$

and this leads to a linear programming problem. Deely and Kruse also established the strong consistency of their estimator. This approach was extended by Blum and Susarla (1977) by using any sequence $\{f_k\}$ of functions that satisfies $\sup|f_k - f_G| \to 0$ a.s. as $k \to \infty$. Each f_k can be obtained by a kernel-based density estimator. Blum and Susarla approximated f_k by the estimated mixture pdf f_{G_k}, and established the strong consistency of the estimator under weak conditions.

If any of these estimators are used to obtain $G_k(A^*)$, Theorem 7.1 is secured. This, finally, closes our circle of analysis.

However, it has been pointed out that all these minimum distance methods suffer from a serious defect in a finite-sample situation: they may completely ignore small numbers of data points in the estimated mixture, no matter how distant they are from the dominant data points (Wang and Witten, 1999). This severely impacts their use in our modeling procedures, because the value of one dimensional absolute distance is frequently quite different to all the others—and this implies that the underlying absolute distance has a very high probability of being different too. Unfortunately, other approaches such as maximum likelihood or Bayesian methods, which could be employed if computational cost was not an issue, suffer from the same problem.

A new minimum distance method has been developed recently that overcomes this problem (Wang and Witten, 1999). For the existing minimum distance methods, the problem stems from the use of CDFs for approximation and the normalization constraint $\sum_{j=1}^{m_k} w_{kj} = 1$. The new method is based on the idea of local rather than global fitting, and uses the estimated mixture measure to approximate the empirical measure over selected intervals. Let G'_k be a discrete function with masses $\{w_{kj}\}$ at $\{\theta_{kj}\}$; note that we do not require the w_{kj}'s to sum to one. The new method seeks to approximate the empirical measure P_k (not necessarily a probability measure) by an estimated one $P_{G'_k}$, which we denote

$$P_{G'_k} \cong P_k. \tag{51}$$

Here, the symbol $\cong$ implies with respect to G'_k the minimization of the distance between the measures on either side. The intervals over which the approximation takes place are called "fitting intervals." It is important to note that (51) is not subject to the normalization constraint, and so G'_k is not a CDF; neither is $P_{G'_k}$ a probability measure. However, G'_k can be converted into a CDF estimator by normalizing it after (51) has been solved.

To define the estimation method fully, we need to determine (a) the set of support points, (b) the set of fitting intervals, (c) the empirical measure, and (d) the distance measure. Wang and Witten (1999) show how to determine these in a suitable way. They report experiments on mixtures of normal distributions whose results illustrate the superiority of their method to other minimum-distance estimators when small clusters are present in finite data sets, and also suggest that it is more accurate and stable than other methods even when there are no small clusters.

For all the above minimum distance estimators, the following three conditions must be satisfied in order to ensure strong consistency (Robbins, 1964).

(C1) $F(x;\theta)$ is continuous on $\mathcal{X} \times \Theta$.

(C2) Define $\mathcal{G}$ to be the class of CDFs on Θ. If $F_{G_1} = F_{G_2}$ for $G_1, G_2 \in \mathcal{G}$, then $G_1 = G_2$.

(C3) Either Θ is a compact subset of $\mathbb{R}$, or $\lim_{\theta\to\pm\infty,\theta\in\Theta} F(x;\theta)$ exists for each $x \in \mathcal{X}$ and is not a distribution function on $\mathcal{X}$.

Pace regression involves mixtures of $\chi^2_1(r^2/2)$, where r^2 is the mixing parameter. Conditions C1 and C3 are clearly satisfied. C2, the identifiability condition, is verified in Appendix C.

There might well exist better estimators of mixing distributions in finite sample situations, particularly for mixtures of $\chi^2_1(r^2/2)$ distributions. If so, their use might improve the performance of the modeling procedures in Section 6.

8 Experimental examples

It is time for some experimental results to illustrate the idea of pace regression and give some indication how it performs in practice. The results are very much in accordance

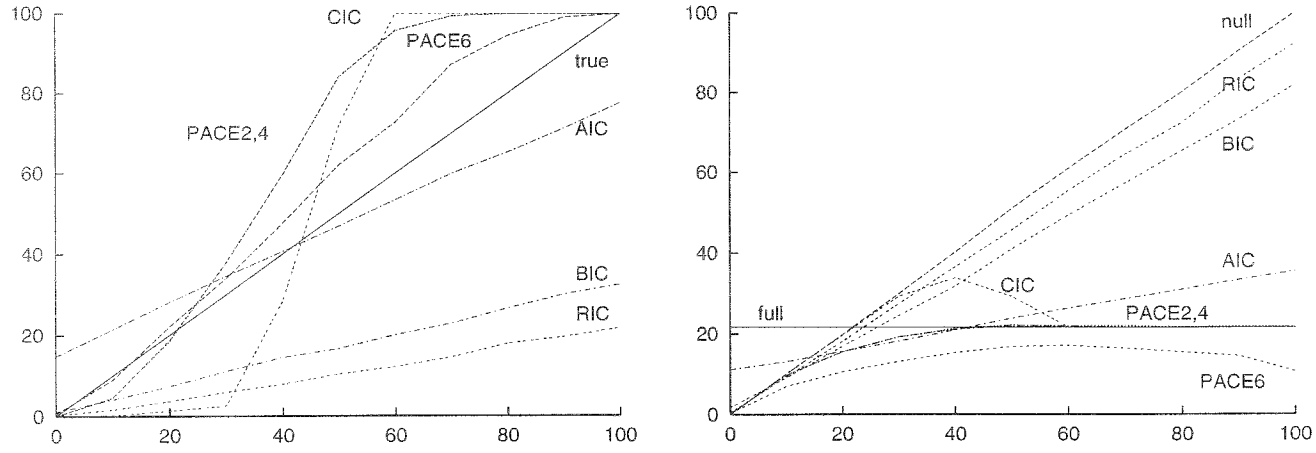


Figure 3: Example 5, $k = 100, \sigma^2 = 200$. (a) Model complexity. (b) Prediction errors.

with the theoretical analysis above. We give three examples that compare pace regression with other procedures in terms of both predictive accuracy and model complexity. In the first, contributive variables have small effects; in the second they have large effects. The third example tests the influence of the number of candidate variables on pace regression. Results are given in the form of graphs; corresponding tabular data appears in Appendix D.

We use the partial-F test and backward elimination of variables to determine the orthogonalization, and employ the OLS unbiased $\hat{\sigma}^2$ estimator. The non-central χ^2 distribution is used instead of the F-distribution for computational reasons. For clustering, we use the measure-based minimum distance estimation procedure of (51), along with the quadratic programming algorithm NNLS provided by Lawson and Hanson (1974). Support points are the set $\{\widehat{A}_j\}$ plus zero (if there is any $\widehat{A}_j$ near zero), except that points lying between zero and three are discarded in order to simplify the model. For PACE_6, $\widetilde{A}_j$'s smaller than 0.5 are discarded to simplify the model without sacrificing much loss of accuracy. These are rough and ready decisions, taken for practical expediency.

We tested AIC, BIC, RIC, CIC (in the form (6)), PACE_2, PACE_4, and PACE_6. The OLS full model and the null model—which are actually generated by procedures OLS and $\text{OLSC}(\infty)$—are included for comparison. Shrinkage methods are not included because the choice of shrinkage parameter is in each case somewhat controversial. Procedures PACE_1, PACE_3, and PACE_5 are not included because they involve numerical solution.

Example 5 *Variables with small effects.* The underlying model is $y = b_0 + \sum_{j=1}^{k} b_j x_j + N(0, \sigma^2)$, where $b_0 = 0, x_j \sim N(0,1)$, the covariance between x_i and x_j is zero if $i \neq j$, and $\sigma^2 = 200$. Each parameter b_j is either 1 or 0, depending on whether it has an effect or is redundant. For each sample, $n = 1000$ and $k = 100$ (excluding b_0 which is always left inside the model). The number of nonredundant variables k^* is set to $0, 10, 20, \ldots, 100$. For each value, the result is the average of twenty runs over an independent test set. The test set is generated from the same model structure with 10,000 observations and zero noise variance. The results are shown in Figure 3 and recorded in Table 1.

Figure 3(a) shows the dimensionality of the estimated models, while Figure 3(b) shows their predictive errors; the horizontal axis is k^* in both cases. In Figure 3(a), the solid line gives the size of the underlying models. Since predictive accuracy rather than model complexity is used as the standard for modeling, the best estimated model does not necessarily have the same complexity as the underlying one—dimensionality reduction is merely a byproduct of eliminating redundant variables. Models generated by PACE_2, PACE_4 and PACE_6 find the underlying null hypothesis H_0 and the underlying full hypothesis H_f correctly; their seeming inconsistency between these extremes is in fact necessary for the estimated models to produce optimal predictions. AIC overfits H_0 and underfits H_f. BIC and RIC both fit H_0 well, but dramatically underfit all the other hypotheses—including H_f. CIC successfully selects both H_0 and H_f but either underfits or overfits models in between.

In Figure 3(b), the vertical axis represents the average mean squared error in predicting independent test sets. The models built by BIC and RIC have errors nearly as great as the null model. AIC is slightly better for H_f than BIC and RIC, but fails on H_0. CIC eventually coincides with the full model as k^* increases, and produces large errors for some model structures between H_0 and H_f. PACE_2 always performs as well as the best of $\text{OLSC}(\infty)$ (the null model), RIC, BIC, AIC, CIC, and OLS (the full model): no OLS subset selection procedures produce models that are sensibly better. Recall that PACE_2 selects the optimal subset model

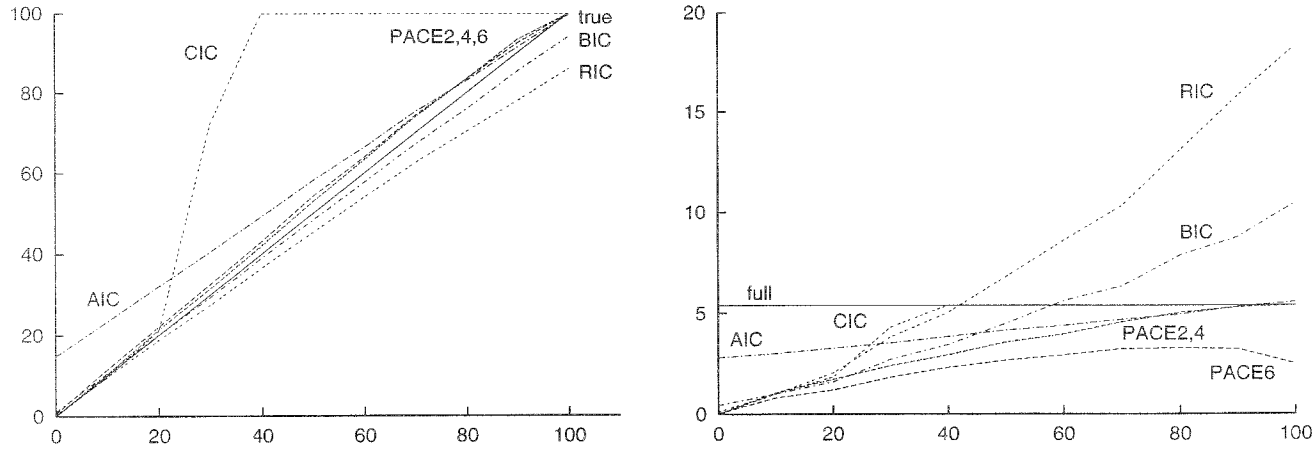


Figure 4: Example 6. $k = 100, \sigma^2 = 50$. (a) Selected dimensions. (b) Prediction errors.

from the sequence, and in this respect resembles PACE_1, which is the optimal threshold-based OLS subset selection procedure $\text{OLSC}(\tau^*)$. PACE_4 performs similarly to PACE_2. Remarkably, PACE_6 outperforms PACE_2 and PACE_4 by a large margin, even when there are no redundant variables.

Example 6 *Variables with large effects.* The underlying model is the same as before except that $\sigma^2 = 50$; results are shown in Figure 4 and Table 2.

The results are similar to those in the previous example. As Figure 4(a) shows, models generated by PACE_2, PACE_4 and PACE_6 lie closer to the line of underlying models than in the last example. AIC generally overfits by an amount that decreases as k^* increases. RIC and BIC generally underfit by an amount that increases as k^* increases. CIC settles on the full model earlier than in the last example.

In terms of prediction error (Figure 4(b)), models built by PACE_2 and PACE_4 are still the best of all OLS subset selection procedures, while PACE_6 is significantly superior. When there are no redundant variables, PACE_6 chooses a full-sized model but uses different coefficients from OLS, yielding a model with much smaller prediction error than the OLS full model. This defies conventional wisdom, which views the OLS full model as the best possible choice when all variables have large effects.

Example 7 *Rate of convergence.* Our third example explores the influence of the number of candidate variables. The value of k is chosen to be $10, 20, \ldots, 100$ respectively, and for each value k^* is chosen as 0, $k/2$ and k; otherwise the experimental conditions are as in Example 5. Note that variables with small effects are harder to distinguish in the presence of noisy variables. Results are shown in Figure 5 and recorded in Tables 3–5.

The pace regression procedures are always among the best in terms of prediction error, a property enjoyed by none of the conventional procedures. None of the pace procedures suffer noticeably as the number of candidate variables k decreases. Apparently, they are stable when k is as small as ten.

9 Discussion

Several issues related to pace regression deserve further discussion, issues that are either necessary to complete the definition of the pace regression procedures in special situations, or to expand their implications into a broader arena.

9.1 Finite k *vs.* k-asymptotics

We have seen that the pace regression procedures are optimal in a k-asymptotic sense. Larger numbers of variables tend to produce estimators that are closer to optimal. If there are only a few candidate variables, pace regression will not necessarily outperform other methods. Since k is inevitably finite in practice, it is worth expanding on this.

The central idea of pace regression is to decompose the prediction vector of an estimated model into k orthogonal components, and then adjust each according to aggregated magnitude information from all components. The more diverse the magnitudes of the different components, the less they can inform the adjustment of any particular one. If one component's magnitude differs greatly from that of all others, there is little basis on which to alter its value.

Pace regression shines when many variables have similar effects—a common special situation is when many variables have zero effect. As the effects of the variables disperse, pace regression's superiority over other procedures gradually fades. When the effect of each variable is isolated from that of all others, the pace estimator is exactly the OLS one. In principle, the worst case is when no improvement over OLS is possible.

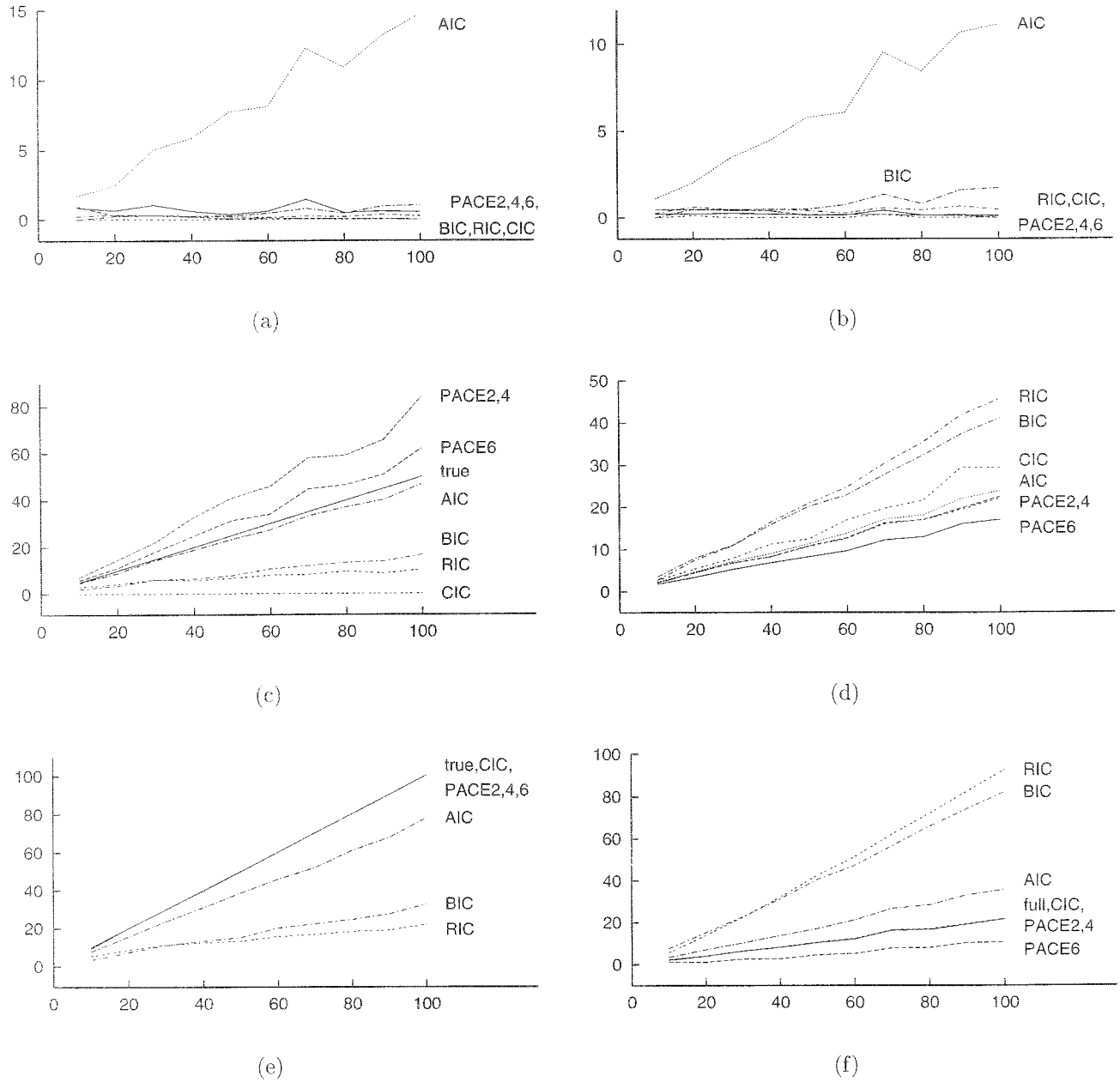


Figure 5: Example 7. (a), (c) and (e) show the dimensionalities of the estimated models when the underlying models are the null, half, and full hypotheses. (b), (d) and (f) show corresponding prediction errors over large independent test sets.

Although pace regression k-asymptotically optimal, this does not mean that increasing the number of candidate variables necessarily improves prediction. New contributive variables should increase predictive accuracy, but new redundant variables will decrease it. Pre-selection of variables based on background knowledge will always help modeling, if suitable variables are selected.

9.2 Collinearity

Pace regression, like almost any other linear regression procedure, fails when presented with collinear variables. Such variables should be detected and eliminated before applying the procedure. We suggest eliminating collinearity by discarding variables. The number of candidate variables k should be reduced accordingly, because collinearity does not provide new independent dimensions, a prerequisite of pace regression. In other words, the model has the same degrees of freedom without collinearity as with it. Appropriate variables can be identified by examining the matrix $(X'X)^{-1}$ (Lawson and Hanson, 1974).

Note that OLS subset selection procedures are sometimes described as a protection against collinearity. However, the fact is that none of these automatic procedures

can reliably eliminate collinearity.

9.3 Regression for partial models

The full OLS model forms the basis for pace regression. In some situations, however, the full model may be unavailable. For example, there may be too many candidate variables—perhaps more than the number of observations. Although the largest available partial model can be substituted for the full model, this causes some practical difficulties.

The clustering step in pace regression must take *all* candidate variables into account. This is possible so long as the statistical test used to determine the initial partial model can supply an approximate distribution for the dimensional absolute distances of the variables that do not participate in it. For example, the partial-F test may be used to discard variables based on forward selection; typically, dimensional absolute distances are smaller for the discarded variables than for those used in the model. It seems likely that a sufficiently accurate approximate distribution for the dimensional absolute distances of the discarded variables can be derived for this test, though further investigation is necessary to confirm this.

In many practical applications, some kind of feature selection is performed before the modeling procedure is invoked. However, it is generally not acknowledged that bias is introduced by discarding variables without passing relevant information on to the modeling procedure—though admittedly most modeling procedures cannot make use of this kind of information.

Estimating the noise variance, should it be unknown a priori, is another issue that is affected when only a partial initial model is available. The noise component will contain the effects of all variables that are not included in the partial model. Moreover, because of competition between variables, the OLS estimate of σ^2 from the partial model is biased downwards. How to compensate for this is an interesting topic worth further investigation.

9.4 Pace regression *vs.* major modeling principles

As noted in Section 1, pace regression measures the success of modeling using two separate criteria. The primary one is accuracy, or minimization of the expected loss. The secondary one is parsimony, or a preference for the smallest model, and is only used when it does not conflict with the first—that is, to decide between several models that have about the same accuracy. The secondary criterion is essential if dimensionality reduction is of interest. For example, if the support points found by PACE_5 and PACE_6 are not all zero, none of the upgraded $\widetilde{A}_j$'s will be zero. But many may be tiny, and eliminating tiny $\widetilde{A}_j$ has negligible influence on predictive ability.

When the goal is to minimize the expected loss, pace regression casts doubt on four widely-accepted general principles of modeling.

First, pace regression challenges the Least Squares Principle in a general sense. All six pace procedures outperform OLS estimation: PACE_1 and PACE_3 in the sense of expected risk and the remainder in the sense of expected loss. According to the Gauss-Markov Theorem, OLS yields a "best linear unbiased estimator" (or BLUE). OLS's inferiority, however, is due precisely to the unbiasedness constraint, which fails to utilize all the information implicit in the data. It is well known that biased estimators such as subset selection and shrinkage can outperform the OLS estimator in particular situations. We have shown that pace regression estimators, which are also biased, outperform OLS in all situations.

Second, pace regression challenges the Maximum Likelihood Principle. In the linear regression situation, the maximum likelihood estimator is exactly the same as the OLS one.

Third, Bayes' Rule itself comes under attack. The Bayesian estimation is threshold-based OLS subset selection, where the a priori density is used to determine the threshold. Yet the six procedures of pace regression equal or outperform the best OLS subset selection estimator. This improvement is not based on prior information—the validity of which has long been questioned—but on hitherto unexploited information that is implicit in the very same data. Furthermore, when the noninformative prior is used—and the use of the noninformative prior is widely accepted, even by many non-Bayesians—the Bayes estimator is the same as the maximum likelihood estimator, i.e., the OLS estimator, and therefore inferior to the pace regression estimators.

Fourth, questions arise concerning complexity-based modeling. According to the Minimum Description Length (MDL) Principle (Rissanen, 1978), the best model is the one that minimizes the sum of the model complexity and the data complexity given the model. In practice the first part is an increasing function of the number of parameters required to define the model, while the second is the resubstitution error. Our analysis and experiments do not support this principle. We have found that pace regression often chooses models of the same or even larger size, and with larger resubstitution errors, than those of other procedures, yet gives much smaller prediction errors on independent test sets. In addition, the MDL estimator derived by Rissanen (1978) is the same as the BIC estimator, which has already been shown inferior to the pace regression estimators.

10 Conclusions

This paper explores a new approach to linear regression. Not only does this approach yield accurate prediction models; it also reduces model dimensionality. It outperforms all other modeling procedures in the literature, both in a theoretical sense and experimentally.

The consequences of this work reach well beyond linear regression. We have seen that the least squares principle is outperformed in a very general sense by pace regression.

This raises significant questions about the validity of this principle--and of other principles widely used in empirical modeling. The analysis we have presented disputes many conventional ideas about modeling from data.

We have limited our investigation to linear models with normally distributed noise, but the ideas are so fundamental that we believe they will soon find application in other realms of empirical modeling.

References

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math., 22*, 203–217.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory* (pp. 267–281). New York: Akadémiai Kiadó.

Blum, J. R. and Susarla, V. (1977). Estimation of a mixing distribution function. *Ann. Probab, 5*, 200–209.

Breiman, L. (1992). The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistics Association, 87*, 738–754.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Techometrics, 37(4)*, 373–384.

Breiman, L. and Spector, P. (1992). Submodel selection and evaluation in regression. the x-random case. *International Statistical Review, 60*, 291–319.

Choi, K. and Bulgren, W. B. (1968). An estimation procedure for mixtures of distributions. *J. R. Statist. Soc. B, 30*, 444–460.

Deely, J. J. and Kruse, R. L. (1968). Construction of sequences estimating the mixing distribution. *Ann. Math. Statist., 39*, 286–288.

Derksen, S. and Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45*, 265–282.

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

Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics, 22*, 1947–1975.

Frank, E. and Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussions). *Technometrics, 35*, 109–148.

Galambos, J. (1995). *Advanced Probability Theory.* A series of Textbooks and Reference books/10. Marcel Dekker, Inc.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of autoregression. *J. R. Statist. Soc. B, 41*, 190–195.

Hoerl, R. W., Schuenemeyer, J. H. and Hoerl, A. E. (1986). A simulation of biased estimation and subset selection regression techniques. *Technometrics, 9*, 269–380.

Lawson, C. L. and Hanson, R. J. (1974). *Solving Least Squares Problems.* Prentice-Hall, Inc.

Lindsay, B. G. (1995). *Mixture models: theory, geometry, and applications*, Volume 5 of *NSF-CBMS Regional Conference Series in Probability and Statistics.* Institute for Mathematical Statistics: Hayward, CA.

Macdonald, P. D. M. (1971). Comment on a paper by Choi and Bulgren. *J. R. Statist. Soc. B, 33*, 326–329.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics, 15*, 661–675.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering.* Marcel Dekker, New York.

Miller, A. J. (1990). *Subset Selection in Regression*, Volume 40 of *Monographs on Statistics and Applied Probability.* Chapmen & Hall.

Rao, C. R. and Toutenborg, H. (1995). *Linear Models—Least Squares and Alternatives.* New York: Springer.

Rao, C. R. and Wu, Y. (1989). A strongly consistent procedure for model selection in a regression problem. *Technometrics, 72(2)*, 369–374.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*, 465–471.

Robbins, H. (1964). The empirical bayes approach to statistical decision problems. *Ann. Math. Statist., 35*, 1–20.

Roecker, E. B. (1991). Prediction error and its estimation of subset selected models. *Technometrics, 33*, 459–468.

Schott, J. R. (1997). *Matrix analysis for statistics.* New York: Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6(2)*, 461–464.

Shaffer, J. P. (1991). The Gauss-Markov theorem and random regressors. *The American Statistician, 45*, 269–273.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc., 88*, 486–494.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc., 91*, 655–665.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika, 63*, 117–126.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika, 68*, 45–54.

Teicher, H. (1960). On the mixture of distributions. *Ann. Math. Statist., 31*, 55–57.

Thompson, M. L. (1978). Selection of variables in multiple regression, part 1 and part 2. *International Statistical Review, 46*, 1–21 and 129–146.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B, 58*, 267–288.

Tibshirani, R. and Knight, K. (1997). The covariance inflation criterion for model selection. Technical report, Department of Statistics, University of Stanford.

Titterington, D. M., Smith, A. F. M. and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. John Wiley & Sons.

Wang, Y. and Witten, I. H. (1999). Clustering for finite samples from semi-parametric mixtures. In *Proceedings of Interface'99*. The Interface Foundation of North America.

Zhang, P. (1993). Model selection via multifold cross-validation. *Ann. Statist., 21*, 299–313.

Zhao, L. C., Krishnaiah, P. R. and Bai, Z. D. (1986). On the detection of the number of signals in the presence of white noise. *Journal of Multivariate Analysis, 20*, 1–25.

A The Taylor expansion of $h(A;A^*)$ with respect to $\sqrt{A}$

Let $a=\sqrt{A}\geq 0$ and $a^*=\sqrt{A^*}\geq 0$. Denote $\psi=1/\sqrt{2\pi}$. Because

$$c(a;a^*)=a^{*2}-(a-a^*)^2=2aa^*-a^2$$

and

$$p(a;a^*)=\psi e^{-\frac{(a-a^*)^2}{2}}=\psi e^{\frac{2aa^*-a^2}{2}}e^{-\frac{a^{*2}}{2}}$$

hence

$$\begin{aligned}&\frac{c(a;a^*)p(a;a^*)}{\psi e^{-\frac{a^{*2}}{2}}}\\ =\;&(2a^*-a)ae^{\frac{2a^*-a}{2}a}\\ =\;&(2a^*-a)a+\frac{(2a^*-a)^2}{2}a^2+\frac{(2a^*-a)^3}{8}a^3+O(a^4)\\ =\;&2a^*a+(-1+2a^{*2})a^2+(-2a^*+a^{*3})a^3+O(a^4)\end{aligned}$$

Similarly

$$c(-a;a^*)=a^{*2}-(a+a^*)^2=-2aa^*-a^2$$

and

$$p(-a;a^*)=\psi e^{-\frac{(a+a^*)^2}{2}}=\psi e^{\frac{-a^2-2aa^*}{2}}e^{-\frac{a^{*2}}{2}}$$

hence

$$\begin{aligned}&\frac{c(-a;a^*)p(-a;a^*)}{\psi e^{-\frac{a^{*2}}{2}}}\\ =\;&(-2a^*-a)ae^{\frac{-2a^*-a}{2}a}\\ =\;&-(2a^*+a)a+\frac{(2a^*+a)^2}{2}a^2-\frac{(2a^*+a)^3}{8}a^3+O(a^4)\\ =\;&-2a^*a+(-1+2a^{*2})a^2+(2a^*-a^{*3})a^3+O(a^4)\end{aligned}$$

Therefore

$$\begin{aligned}\frac{h(A;A^*)}{\psi e^{-\frac{a^{*2}}{2}}}&=\frac{c(a;a^*)p(a;a^*)+c(-a;a^*)p(-a;a^*)}{2a\psi e^{-\frac{a^{*2}}{2}}}\\ &=(-1+2a^{*2})a+O(a^3)\end{aligned}$$

that is,

$$h(A;A^*)=\frac{1}{\sqrt{2\pi}}e^{-\frac{A^*}{2}}[(-1+2A^*)\sqrt{A}+O(A^{\frac{3}{2}})]. \tag{52}$$

B Proof of Theorem 6.5

To prove Theorem 6.5, we need a simple lemma.

Lemma 1 *For any $A\geq 0$ and $A^*\geq 0$,*

$$0\leq c(\sqrt{A};\sqrt{A^*})-E[C(A)|A^*]<4\sqrt{AA^*}e^{-4\sqrt{AA^*}}\leq 2e^{-1}. \tag{53}$$

Proof. For every $A\geq 0$ and $A^*\geq 0$, $c(\sqrt{A};\sqrt{A^*})\geq 0$ and $c(-\sqrt{A};\sqrt{A^*})\leq 0$, hence

$$\begin{aligned}&c(\sqrt{A};\sqrt{A^*})\\ =\;&\frac{c(\sqrt{A};\sqrt{A^*})p(\sqrt{A};\sqrt{A^*})+c(\sqrt{A};\sqrt{A^*})p(-\sqrt{A};\sqrt{A^*})}{p(\sqrt{A};\sqrt{A^*})+p(-\sqrt{A};\sqrt{A^*})}\\ \geq\;&\frac{c(\sqrt{A};\sqrt{A^*})p(\sqrt{A};\sqrt{A^*})+c(-\sqrt{A};\sqrt{A^*})p(-\sqrt{A};\sqrt{A^*})}{p(\sqrt{A};\sqrt{A^*})+p(-\sqrt{A};\sqrt{A^*})}\\ =\;&E[C(A)|A^*],\end{aligned}$$

and

$$\begin{aligned}&c(\sqrt{A};\sqrt{A^*})-E[C(A)|A^*]\\ =\;&\frac{[c(\sqrt{A};\sqrt{A^*})-c(-\sqrt{A};\sqrt{A^*})]p(-\sqrt{A};\sqrt{A^*})}{p(\sqrt{A};\sqrt{A^*})+p(-\sqrt{A};\sqrt{A^*})}\\ =\;&\frac{4\sqrt{AA^*}p(-\sqrt{A};\sqrt{A^*})}{p(\sqrt{A};\sqrt{A^*})+p(-\sqrt{A};\sqrt{A^*})}\\ <\;&\frac{4\sqrt{AA^*}p(-\sqrt{A};\sqrt{A^*})}{p(\sqrt{A};\sqrt{A^*})}\\ =\;&4\sqrt{AA^*}e^{-2\sqrt{AA^*}}\\ \leq\;&2e^{-1},\end{aligned}$$

thus completing the proof of the lemma. □

Proof of Theorem 6.5. According to the definition of dimensional contribution, to prove Theorem 6.5 is the same as to prove

$$0\leq E[C(\widehat{A}_j^{\text{PACE}_5})|\widehat{A}_j,G(A^*)]-E[C(\widehat{A}_j^{\text{PACE}_6})|\widehat{A}_j,G(A^*)]<2e^{-1}. \tag{54}$$

The first inequality in (54) is obvious because $\widetilde{A}_j^{\text{PACE5}}$ is the optimal solution. For the second inequality, due to the above lemma and the fact that $\widetilde{A}_j^{\text{PACE6}}$ is the optimal solution of (45), we have

$$\begin{aligned}
& E[C(\widetilde{A}_j^{\text{PACE5}})|\widehat{A}_j, G(A^*)] \\
= & \frac{\sum_{i=1}^m E[C(\widetilde{A}_j^{\text{PACE5}})|\alpha_i^*] f(\widehat{A}_j;\alpha_i^*) w_i}{\sum_{i=1}^m f(\widehat{A}_j;\alpha_i^*) w_i} \\
\leq & \frac{\sum_{i=1}^m c(\sqrt{\widetilde{A}_j^{\text{PACE5}}};\sqrt{\alpha_i^*}) f(\widehat{A}_j;\alpha_i^*) w_i}{\sum_{i=1}^m f(\widehat{A}_j;\alpha_i^*) w_i} \\
\leq & \frac{\sum_{i=1}^m c(\sqrt{\widetilde{A}_j^{\text{PACE6}}};\sqrt{\alpha_i^*}) f(\widehat{A}_j;\alpha_i^*) w_i}{\sum_{i=1}^m f(\widehat{A}_j;\alpha_i^*) w_i}.
\end{aligned}$$

Hence

$$\begin{aligned}
& E[C(\widetilde{A}_j^{\text{PACE5}})|\widehat{A}_j, G(A^*)] - E[C(\widetilde{A}_j^{\text{PACE6}})|\widehat{A}_j, G(A^*)] \\
\leq & \frac{\sum_{i=1}^m \{c(\sqrt{\widetilde{A}_j^{\text{PACE6}}};\sqrt{\alpha_i^*}) - E[C(\widetilde{A}_j^{\text{PACE6}})|\alpha_i^*]\} f(\widehat{A}_j;\alpha_i^*) w_i}{\sum_{i=1}^m f(\widehat{A}_j;\alpha_i^*) w_i}.
\end{aligned}$$

According to the lemma, for every i,

$$c(\sqrt{\widetilde{A}_j^{\text{PACE6}}};\sqrt{\alpha_i^*}) - E[C(\widetilde{A}_j^{\text{PACE6}})|\alpha_i^*] \geq 0,$$

and let

$$\alpha^* = \arg\max_{\{\alpha_i^*\}} \; c(\sqrt{\widetilde{A}_j^{\text{PACE6}}};\sqrt{\alpha_i^*}) - E[C(\widetilde{A}_j^{\text{PACE6}})|\alpha_i^*].$$

Therefore, according to the lemma,

$$\begin{aligned}
& E[C(\widetilde{A}_j^{\text{PACE5}})|\widehat{A}_j, G(A^*)] - E[C(\widetilde{A}_j^{\text{PACE6}})|\widehat{A}_j, G(A^*)] \\
\leq & \; c(\sqrt{\widetilde{A}_j^{\text{PACE6}}};\sqrt{\alpha^*}) - E[C(\widetilde{A}_j^{\text{PACE6}})|\alpha^*] \\
< & \; 4\sqrt{\widetilde{A}_j^{\text{PACE6}}\alpha^*} e^{-2\sqrt{\widetilde{A}_j^{\text{PACE6}}\alpha^*}} \\
\leq & \; 2e^{-1}
\end{aligned}$$

which finishes the proof of Theorem 6.5. □

C Identifiability for mixtures of $\chi_1^2(r^2/2)$ distributions

Although we have been unable to locate the following theorem in the literature, it seems unlikely to be original. Note that here, identifiability applies only to the situation where the mixing function is limited to being a CDF. However, it is possible to show that identifiability of mixtures using CDFs as mixing functions implies the identifiability of mixtures using any finite nonnegative functions as mixing functions, as required by (51).

Theorem C.1 *The mixture of $\chi_1^2(A^*/2)$ distributions, where A^* is the mixing parameter, is identifiable.*

Proof. Let $P(x;\mu)$ be the CDF of the distribution $N(\mu,1)$, and $F(A;A^*)$ be the CDF of the distribution $\chi_1^2(A^*/2)$. From the definition of $\chi_1^2(A^*/2)$ distribution and the symmetry of the normal distribution,

$$\begin{aligned}
F(A;A^*) &= P(\sqrt{A};\sqrt{A^*}) - P(-\sqrt{A};\sqrt{A^*}) \\
&= P(\sqrt{A};\sqrt{A^*}) + P(\sqrt{A};-\sqrt{A^*}) - 1, \quad (55)
\end{aligned}$$

where $P(\sqrt{A};\sqrt{A^*})$ is the CDF of $N(\sqrt{A^*},1)$ and $P(\sqrt{A};-\sqrt{A^*})$ is the CDF of $N(-\sqrt{A^*},1)$. Therefore, if the mixture of $F(A;A^*)$ is unidentifiable, the mixture of normal distributions where the mean μ is the mixing parameter would be unidentifiable. Clearly this contradicts the well-known identifiability result for mixtures of normal distributions (Teicher, 1960), thus completing the proof. □

D Results for Examples 5, 6 and 7

Table 1: Data for the experimental results in Example 5.

	k^*	0	10	20	30	40	50	60	70	80	90	100
Model complexities	AIC	14.8	21.4	28.1	34.5	40.6	46.9	53.4	59.9	65.2	71.2	77.6
	BIC	1.1	4.2	7.4	11.1	14.5	16.7	20.1	22.9	26.6	30.1	32.5
	RIC	0.2	1.6	3.7	6.0	7.9	10.4	12.2	14.6	17.9	19.6	21.8
	CIC	0.0	0.2	1.3	2.6	28.4	72.0	100	100	100	100	100
	$PACE_2$	0.0	4.6	18.6	37.8	59.6	84.0	95.7	99.4	100	100	100
	$PACE_4$	0.0	4.6	18.8	37.9	59.6	84.2	95.8	99.4	100	100	100
	$PACE_6$	0.6	8.9	22.1	34.0	47.6	62.0	72.8	87.2	94.5	99.1	100
Prediction errors	AIC	11.2	13.1	15.6	18.2	20.9	23.8	26.4	28.8	31.1	33.5	35.7
	BIC	1.7	9.1	16.9	24.5	31.8	41.0	49.4	57.3	65.6	73.3	82.1
	RIC	0.5	9.3	18.2	27.5	36.6	45.6	55.4	64.6	72.6	83.5	92.5
	CIC	0.0	10.1	19.7	29.3	33.9	29.3	21.6	21.6	21.6	21.6	21.6
	$PACE_2$	0.0	9.8	15.6	19.3	21.1	22.1	21.8	21.6	21.6	21.6	21.6
	$PACE_4$	0.0	9.8	15.6	19.3	21.3	22.2	22.0	22.0	21.8	21.8	21.7
	$PACE_6$	0.1	7.0	10.6	13.2	15.4	16.9	17.1	16.5	15.5	14.5	10.7
	full	21.6	21.6	21.6	21.6	21.6	21.6	21.6	21.6	21.6	21.6	21.6

Table 2: Data for the experimental results in Example 6.

	k^*	0	10	20	30	40	50	60	70	80	90	100
Model complexities	AIC	14.8	23.4	32.0	40.5	49.2	58.2	66.4	75.3	83.0	91.4	99.7
	BIC	1.1	10.6	20.1	29.4	39.0	48.4	57.8	67.2	75.9	85.4	94.0
	RIC	0.2	9.6	18.8	27.4	36.6	45.3	54.0	62.5	70.3	78.0	86.0
	CIC	0.0	9.6	21.1	72.4	100	100	100	100	100	100	100
	$PACE_2$	0.0	10.3	21.4	31.5	42.0	53.0	63.2	74.0	83.4	93.3	100
	$PACE_4$	0.0	10.3	21.4	31.5	42.1	53.0	63.4	74.0	83.5	93.3	100
	$PACE_6$	0.6	11.5	22.1	32.6	43.0	54.2	63.9	74.3	83.4	92.5	100
Prediction errors	AIC	2.80	3.00	3.27	3.53	3.85	4.16	4.40	4.70	4.95	5.32	5.56
	BIC	0.43	1.00	1.59	2.71	3.44	4.50	5.62	6.35	7.88	8.76	10.5
	RIC	0.12	1.00	2.03	3.82	5.04	6.81	8.59	10.3	13.1	15.8	18.3
	CIC	0.00	0.99	1.83	4.31	5.40	5.40	5.40	5.40	5.40	5.40	5.40
	$PACE_2$	0.00	1.06	1.71	2.40	2.95	3.56	3.96	4.57	5.03	5.29	5.40
	$PACE_4$	0.00	1.06	1.71	2.40	2.92	3.58	3.98	4.57	5.04	5.29	5.40
	$PACE_6$	0.03	0.80	1.19	1.82	2.31	2.66	2.92	3.24	3.28	3.22	2.48
	full	5.40	5.40	5.40	5.40	5.40	5.40	5.40	5.40	5.40	5.40	5.40

Table 3: Example 7—null hypothesis ($k^* = 0$).

	k	10	20	30	40	50	60	70	80	90	100
Model complexities	AIC	1.7	2.5	5.0	5.8	7.8	8.2	12.3	10.9	13.2	14.8
	BIC	0.0	0.2	0.3	0.2	0.2	0.5	0.8	0.5	0.9	1.1
	RIC	0.2	0.3	0.3	0.2	0.2	0.1	0.2	0.2	0.3	0.2
	CIC	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
	$PACE_2$	0.9	0.3	0.3	0.2	0.1	0.1	0.1	0.1	0.1	0.0
	$PACE_4$	0.9	0.3	0.3	0.2	0.1	0.1	0.1	0.1	0.1	0.0
	$PACE_6$	0.8	0.7	1.1	0.6	0.3	0.6	1.4	0.5	0.6	0.6
Prediction errors	AIC	1.1	2.0	3.5	4.4	5.8	6.1	9.5	8.5	10.7	11.2
	BIC	0.0	0.5	0.5	0.5	0.5	0.7	1.4	0.8	1.6	1.7
	RIC	0.3	0.6	0.5	0.5	0.4	0.3	0.6	0.4	0.7	0.5
	CIC	0.0	0.1	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0
	$PACE_2$	0.5	0.5	0.4	0.4	0.1	0.2	0.2	0.1	0.1	0.0
	$PACE_4$	0.5	0.5	0.4	0.4	0.1	0.2	0.2	0.1	0.1	0.0
	$PACE_6$	0.2	0.2	0.3	0.2	0.2	0.1	0.4	0.1	0.2	0.1

Table 4: Example 7—half hypothesis ($k^* = k/2$).

	k	10	20	30	40	50	60	70	80	90	100
Model complexities	AIC	4.8	8.9	14.6	18.8	23.4	27.2	33.2	37.0	40.2	46.9
	BIC	2.0	3.5	6.2	6.5	7.9	10.6	12.0	13.3	13.9	16.7
	RIC	3.0	4.2	6.2	5.8	6.7	8.0	8.3	9.7	8.7	10.4
	CIC	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
	PACE_2	7.2	14.6	22.3	32.8	41.0	45.9	57.9	59.1	65.6	84.0
	PACE_4	7.2	14.6	22.3	32.8	41.0	46.0	58.0	59.1	65.8	84.2
	PACE_6	5.9	11.1	18.2	24.8	31.4	34.0	44.8	46.6	51.0	62.0
Prediction errors	AIC	2.3	4.7	7.2	9.0	11.4	13.8	17.1	18.1	22.0	23.8
	BIC	3.5	8.0	11.0	15.8	20.2	22.9	27.8	32.2	37.3	41.0
	RIC	2.8	7.5	10.9	16.5	20.9	24.7	30.4	35.3	41.8	45.6
	CIC	2.8	5.4	8.0	11.3	12.5	16.9	19.7	21.6	29.4	29.3
	PACE_2	2.1	4.5	6.8	8.3	10.7	12.6	16.0	16.9	19.4	22.1
	PACE_4	2.1	4.5	6.8	8.3	10.8	12.7	16.2	16.9	19.5	22.2
	PACE_6	1.8	3.4	5.3	6.8	8.3	9.6	12.2	12.9	15.9	16.9

Table 5: Example 7—full hypothesis ($k^* = k$).

	k	10	20	30	40	50	60	70	80	90	100
Model complexities	AIC	7.7	15.6	23.6	31.0	38.6	45.9	52.0	60.9	67.5	77.6
	BIC	3.3	7.2	11.1	13.2	15.2	19.9	22.1	24.4	27.4	32.5
	RIC	5.3	8.5	11.3	12.3	13.1	15.6	16.6	18.1	18.8	21.8
	CIC	10.0	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100
	PACE_2	9.4	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100
	PACE_4	9.4	20.0	30.0	40.0	50.0	60.0	70.0	80.0	90.0	100
	PACE_6	9.2	20.0	29.9	39.6	49.9	60.0	70.0	80.0	90.0	100
Prediction errors	AIC	3.5	7.0	10.3	13.8	17.1	21.2	26.7	28.3	33.1	35.7
	BIC	7.7	15.3	22.8	31.3	40.5	47.2	56.4	65.7	74.0	82.1
	RIC	5.6	13.9	22.6	32.1	42.5	51.3	61.9	71.8	82.4	92.5
	CIC	2.0	4.0	6.4	8.2	10.4	12.1	16.2	16.6	19.1	21.6
	PACE_2	2.3	4.0	6.4	8.2	10.4	12.1	16.2	16.6	19.1	21.6
	PACE_4	2.3	4.0	6.4	8.3	10.6	12.3	16.4	16.9	19.4	21.7
	PACE_6	1.2	1.0	2.6	2.7	4.5	5.2	7.8	8.0	10.4	10.7
	full	2.0	4.0	6.4	8.2	10.4	12.1	16.2	16.6	19.1	21.6