Springer Series in Statistics

Advisors:
P. Bickel, P. Diggle, S. Fienberg, U. Gather,
I. Olkin, S. Zeger

For other titles published in this series go to,
http://www.springer.com/series/692

Bertrand Clarke · Ernest Fokoué · Hao Helen Zhang

Principles and Theory for Data Mining and Machine Learning

Springer

Bertrand Clarke
University of Miami
120 NW 14th Street
CRB 1055 (C-213)
Miami, FL, 33136
bclarke2@med.miami.edu

Ernest Fokoué
Center for Quality and Applied Statistics
Rochester Institute of Technology
98 Lomb Memorial Drive
Rochester, NY 14623
ernest.fokoue@gmail.com

Hao Helen Zhang
Department of Statistics
North Carolina State University
Genetics
P.O.Box 8203
Raleigh, NC 27695-8203
USA
hzhang2@stat.ncsu.edu

ISSN 0172-7397
ISBN 978-0-387-98134-5 e-ISBN 978-0-387-98135-2
DOI 10.1007/978-0-387-98135-2
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009930499

Preface

The idea for this book came from the time the authors spent at the Statistics and Applied Mathematical Sciences Institute (SAMSI) in Research Triangle Park in North Carolina starting in fall 2003. The first author was there for a total of two years, the first year as a Duke/SAMSI Research Fellow. The second author was there for a year as a Post-Doctoral Scholar. The third author has the great fortune to be in RTP permanently. SAMSI was – and remains – an incredibly rich intellectual environment with a general atmosphere of free-wheeling inquiry that cuts across established fields. SAMSI encourages creativity: It is the kind of place where researchers can be found at work in the small hours of the morning – computing, interpreting computations, and developing methodology. Visiting SAMSI is a unique and wonderful experience.

The people most responsible for making SAMSI the great success it is include Jim Berger, Alan Karr, and Steve Marron. We would also like to express our gratitude to Dalene Stangl and all the others from Duke, UNC-Chapel Hill, and NC State, as well as to the visitors (short and long term) who were involved in the SAMSI programs. It was a magical time we remember with ongoing appreciation.

While we were there, we participated most in two groups: Data Mining and Machine Learning, for which Clarke was the group leader, and a General Methods group run by David Banks. We thank David for being a continual source of enthusiasm and inspiration. The first chapter of this book is based on the outline of the first part of his short course on Data Mining and Machine Learning. Moreover, David graciously contributed many of his figures to us. Specifically, we gratefully acknowledge that Figs. 1.1–6, Figs. 2.1,3,4,5,7, Fig. 4.2, Figs. 8.3,6, and Figs. 9.1,2 were either done by him or prepared under his guidance.

On the other side of the pond, the Newton Institute at Cambridge University provided invaluable support and stimulation to Clarke when he visited for three months in 2008. While there, he completed the final versions of Chapters 8 and 9. Like SAMSI, the Newton Institute was an amazing, wonderful, and intense experience.

This work was also partially supported by Clarke's NSERC Operating Grant 2004–2008. In the USA, Zhang's research has been supported over the years by two

grants from the National Science Foundation. Some of the research those grants supported is in Chapter 10.

We hope that this book will be of value as a graduate text for a PhD-level course on data mining and machine learning (DMML). However, we have tried to make it comprehensive enough that it can be used as a reference or for independent reading. Our paradigm reader is someone in statistics, computer science, or electrical or computer engineering who has taken advanced calculus and linear algebra, a strong undergraduate probability course, and basic undergraduate mathematical statistics. Someone whose expertise in is one of the topics covered here will likely find that chapter routine, but hopefully find the other chapters are at a comfortable level.

The book roughly separates into three parts. Part I consists of Chapters 1 through 4: This is mostly a treatment of nonparametric regression, assuming a mastery of linear regression. Part II consists of Chapters 5, 6, and 7: This is a mix of classification, recent nonparametric methods, and computational comparisons. Part III consists of Chapters 8 through 11. These focus on high dimensional problems, including clustering, dimension reduction, variable selection, and multiple comparisons. We suggest that a selection of topics from the first two parts would be a good one semester course and a selection of topics from Part III would be a good follow-up course.

There are many topics left out: proper treatments of information theory, VC dimension, PAC learning, Oracle inequalities, hidden Markov models, graphical models, frames, and wavelets are the main absences. We regret this, but no book can be everything.

The main perspective undergirding this work is that DMML is a fusion of large sectors of statistics, computer science, and electrical and computer engineering. The DMML fusion rests on good prediction and a complete assessment of modeling uncertainty as its main organizing principles. The assessment of modeling uncertainty ideally includes all of the contributing factors, including those commonly neglected, in order to be valid. Given this, other aspects of inference – model identification, parameter estimation, hypothesis testing, and so forth – can largely be regarded as a consequence of good prediction. We suggest that the development and analysis of good predictors is the paradigm problem for DMML.

Overall, for students and practitioners alike, DMML is an exciting context in which whole new worlds of reasoning can be productively explored and applied to important problems.

Bertrand Clarke
University of Miami, Miami, FL

Ernest Fokoué
Kettering University, Flint, MI

Hao Helen Zhang
North Carolina State University, Raleigh, NC

Contents

Chapter 1
Variability, Information, and Prediction

Introductory statistics courses often start with summary statistics, then develop a notion of probability, and finally turn to parametric models – mostly the normal – for inference. By the end of the course, the student has seen estimation and hypothesis testing for means, proportions, ANOVA, and maybe linear regression. This is a good approach for a first encounter with statistical thinking. The student who goes on takes a familiar series of courses: survey sampling, regression, Bayesian inference, multivariate analysis, nonparametrics and so forth, up to the crowning glories of decision theory, measure theory, and asymptotics. In aggregate, these courses develop a view of statistics that continues to provide insights and challenges.

All of this was very tidy and cosy, but something changed. Maybe it was computing. All of a sudden, quantities that could only be described could be computed readily and explored. Maybe it was new data sets. Rather than facing small to moderate sample sizes with a reasonable number of parameters, there were 100 data points, 20,000 explanatory variables, and an array of related multitype variables in a time-dependent data set. Maybe it was new applications: bioinformatics, E-commerce, Internet text retrieval. Maybe it was new ideas that just didn't quite fit the existing framework. In a world where model uncertainty is often the limiting aspect of our inferential procedures, the focus became prediction more than testing or estimation. Maybe it was new techniques that were intellectually uncomfortable but extremely effective: What sense can be made of a technique like random forests? It uses randomly generated ensembles of trees for classification, performing better and better as more models are used.

All of this was very exciting. The result of these developments is called data mining and machine earning (DMML).

Data mining refers to the search of large, high-dimensional, multitype data sets, especially those with elaborate dependence structures. These data sets are so unstructured and varied, on the surface, that the search for structure in them is statistical. A famous (possibly apocryphal) example is from department store sales data. Apparently a store found there was an unusually high empirical correlation between diaper sales and beer sales. Investigation revealed that when men buy diapers, they often treat themselves to a six-pack. This might not have surprised the wives, but the marketers would have taken note.

B. Clarke et al., *Principles and Theory for Data Mining and Machine Learning*, Springer Series in Statistics, DOI 10.1007/978-0-387-98135-2_1,

Machine learning refers to the use of formal structures (machines) to do inference (learning). This includes what empirical scientists mean by model building – proposing mathematical expressions that encapsulate the mechanism by which a physical process gives rise to observations – but much else besides. In particular, it includes many techniques that do not correspond to physical modeling, provided they process data into information. Here, information usually means anything that helps reduce uncertainty. So, for instance, a posterior distribution represents "information" or is a "learner" because it reduces the uncertainty about a parameter.

The fusion of statistics, computer science, electrical engineering, and database management with new questions led to a new appreciation of sources of errors. In narrow parametric settings, increasing the sample size gives smaller standard errors. However, if the model is wrong (and they all are), there comes a point in data gathering where it is better to use some of your data to choose a new model rather than just to continue refining an existing estimate. That is, once you admit model uncertainty, you can have a smaller and smaller variance but your bias is constant. This is familiar from decomposing a mean squared error into variance and bias components.

Extensions of this animate DMML. Shrinkage methods (not the classical shrinkage, but the shrinking of parameters to zero as in, say, penalized methods) represent a trade-off among variable selection, parameter estimation, and sample size. The ideas become trickier when one must select a basis as well. Just as there are well-known sums of squares in ANOVA for quantifying the variability explained by different aspects of the model, so will there be an extra variability corresponding to basis selection. In addition, if one averages models, as in stacking or Bayes model averaging, extra layers of variability (from the model weights and model list) must be addressed. Clearly, good inference requires trade-offs among the biases and variances from each level of modeling. It may be better, for instance, to "stack" a small collection of shrinkage-derived models than to estimate the parameters in a single huge model.

Among the sources of variability that must be balanced – random error, parameter uncertainty and bias, model uncertainty or misspecification, model class uncertainty, generalization error – there is one that stands out: model uncertainty. In the conventional paradigm with fixed parametric models, there is no model uncertainty; only parameter uncertainty remains. In conventional nonparametrics, there is only model uncertainty; there is no parameter, and the model class is so large it is sure to contain the true model. DMML is between these two extremes: The model class is rich beyond parametrization, and may contain the true model in a limiting sense, but the true model cannot be assumed to have the form the model class defines. Thus, there are many parameters, leading to larger standard errors, but when these standard errors are evaluated within the model, they are invalid: The adequacy of the model cannot be assumed, so the standard error of a parameter is about a value that may not be meaningful. It is in these high-variability settings in the mid-range of uncertainty (between parametric and nonparametric) that dealing with model uncertainty carefully usually becomes the dominant issue which can only be tested by predictive criteria.

There are other perspectives on DMML that exist, such as rule mining, fuzzy learning, observational studies, and computational learning theory. To an extent, these can be regarded as elaborations or variations of aspects of the perspective presented here,

although advocates of those views might regard that as inadequate. However, no book can cover everything and all perspectives. Details on alternative perspectives to the one perspective presented here can be found in many good texts.

Before turning to an intuitive discussion of several major ideas that will recur throughout this monograph, there is an apparent paradox to note: Despite the novelty ascribed to DMML, many of the topics covered here have been studied for decades. Most of the core ideas and techniques have precedents from before 1990. The slight paradox is resolved by noting that what is at issue is the novel, unexpected way so many ideas, new and old, have been recombined to provide a new, general perspective dramatically extending the conventional framework epitomized by, say, Lehmann's books.

1.0.1 The Curse of Dimensionality

Given that model uncertainty is the key issue, how can it be measured? One crude way is through dimension. The problem is that high model uncertainty, especially of the sort central to DMML, rarely corresponds to a model class that permits a finite-dimensional parametrization. On the other hand, some model classes, such as neural nets, can approximate sets of functions that have an interior in a limiting sense and admit natural finite-dimensional subsets giving arbitrarily good approximations. This is the intermediate tranche between finite-dimensional and genuinely nonparametric models: The members of the model class can be represented as limiting forms of an unusually flexible parametrized family, the elements of which give good, natural approximations. Often the class has a nonvoid interior.

In this context, the real dimension of a model is finite but the dimension of the model space is not bounded. The situation is often summarized by the phrase the Curse of Dimensionality. This phrase was first used by Bellman (1961), in the context of approximation theory, to signify the fact that estimation difficulty not only increases with dimension – which is no surprise – but can increase superlinearly. The result is that difficulty outstrips conventional data gathering even for what one would expect were relatively benign dimensions. A heuristic way to look at this is to think of real functions of x, of y, and of the pair (x,y). Real functions f, g of a single variable represent only a vanishingly small fraction of the functions k of (x,y). Indeed, they can be embedded by writing $k(x,y) = f(x) + g(y)$. Estimating an arbitrary function of two variables is more than twice as hard as estimating two arbitrary functions of one variable.

An extreme case of the Curse of Dimensionality occurs in the "large p, small n" problem in general regression contexts. Here, p customarily denotes the dimension of the space of variables, and n denotes the sample size. A collection of such data is $(\boldsymbol{y}_i, \boldsymbol{x}_{1,i}, ..., \boldsymbol{x}_{p,i})$ for $i = 1, ...n$. Gathering the explanatory variables, the $\boldsymbol{x}_{i,j}$s, into an $n \times p$ matrix X in which the ith row is $(\boldsymbol{x}_{1,i}, ..., \boldsymbol{x}_{p,i})$ means that X is short and fat when $p >> n$. Conventionally, design matrices are tall and skinny, $n >> p$, so there is a relatively high ratio n/p of data to the number of inferences. The short, fat data problem occurs when $n/p << 1$, so that the parameters cannot be estimated directly at all, much

less well. These problems need some kind of auxiliary principle, such as shrinkage or other constraints, just to make solutions exist.

The finite-dimensional parametric case and the truly nonparametric case for regression are settings in which it is convenient to discuss some of the recurrent issues in the treatments here. It will be seen that the Curse applies in regression, but the Curse itself is more general, applying to classification, and to nearly all other aspects of multivariate inference. As noted, traditional analysis avoids the issue by making strong model assumptions, such as linearity and normality, to get finite-dimensional behavior or by using distribution-free procedures, and being fully nonparametric. However, the set of practical problems for which these circumventions are appropriate is small, and modern applied statisticians frequently use computer-intensive techniques on the intermediate tranche that are designed to minimize the impact of the Curse.

1.0.2 The Two Extremes

Multiple linear regression starts with n observations of the form $(Y_i, \boldsymbol{X}_i)$ and then makes the strong modeling assumption that the response Y_i is related to the vector of explanatory variables $\boldsymbol{X}_i = (X_{1,i}, ..., X_{p,i})$ by

$$Y_i = \boldsymbol{\beta}^T \boldsymbol{X}_i + \varepsilon_i = \beta_0 + \beta_1 X_{1,i} + \ldots \beta_p X_{p,i} + \varepsilon_i,$$

where each random error ε_i is (usually) an independent draw from a normal distribution with mean zero and fixed but unknown variance. More generally, the ε_is are taken as symmetric, unimodal, and independent. The $\boldsymbol{X}_i$s can be random, or, more commonly, chosen by the experimenter and hence deterministic. In the chapters to follow, instances of this setting will recur several times under various extra conditions.

In contrast, nonparametric regression assumes that the response variable is related to the vector of explanatory variables by

$$Y_i = f(\boldsymbol{X}_i) + \varepsilon_i,$$

where f is some smooth function. The assumptions about the error may be the same as for linear regression, but people tend to put less emphasis on the error structure than on the uncertainty in estimates $\hat{f}$ of f. This is reasonable because, outside of large departures from independent, symmetric, unimodal ε_is, the dominant source of uncertainty comes from estimating f. This setting will recur several times as well; Chapter 2, for instance, is devoted to it.

Smoothness of f is central: For several nonparametric methods, it is the smoothness assumptions that make theorems ensuring good behavior (consistency, for instance) of regression estimators $\hat{f}$ of f possible. For instance, kernel methods often assume f is in a Sobolev space, meaning f and a fixed number, say s, of its derivatives lie in a Hilbert space, say $L_q(\Omega)$, where the open set $\Omega \subset R^p$ is the domain of f.

Other methods, like splines for instance, weaken these conditions by allowing f to be piecewise continuous, so that it is differentiable between prespecified pairs of points, called knots. A third approach penalizes the roughness of the fitted function, so that the data help determine how wiggly the estimate of f should be. Most of these methods include a "bandwidth" parameter, often estimated by cross-validation (to be discussed shortly). The bandwidth parameter is like a resolution defining the scale on which solutions should be valid. A finer-scale, smaller bandwidth suggests high concern with very local behavior of f; a large-scale, higher bandwidth suggests one will have to be satisfied, usually grudgingly, with less information on the detailed behavior of f.

Between these two extremes lies the intermediate tranche, where most of the action in DMML is. The intermediate tranche is where the finite-dimensional methods confront the Curse of Dimensionality on their way to achieving good approximations to the nonparametric setting.

1.1 Perspectives on the Curse

Since almost all finite-dimensional methods break down as the dimension p of $\boldsymbol{X}_i$ increases, it's worth looking at several senses in which the breakdown occurs. This will reveal impediments that methods must overcome. In the context of regression analysis under the squared error loss, the formal statement of the Curse is:

- The mean integrated squared error of fits increases faster than linearly in p.

The central reason is that, as the dimension increases, the amount of extra room in the higher-dimensional space and the flexibility of large function classes is dramatically more than experience with linear models suggests.

For intuition, however, note that there are three nearly equivalent informal descriptions of the Curse of Dimensionality:

- In high dimensions, all data sets are too sparse.
- In high dimensions, the number of possible models to consider increases superexponentially in p.
- In high dimensions, all data sets show multicollinearity (or concurvity , which is the generalization that arises in nonparametric regression).

In addition to these near equivalences, as p increases, the effect of error terms tends to increase and the potential for spurious correlations among the explanatory variables increases. This section discusses these issues in turn.

These issues may not sound very serious, but they are. In fact, scaling up most procedures highlights unforeseen weaknesses in them. To dramatize the effect of scaling from two to three dimensions, recall the high school physics question: What's the first thing that would happen if a spider kept all its proportions the same but was suddenly 10 feet tall? Answer: Its legs would break. The increase in volume in its body

(and hence weight) is much greater than the increase in cross-sectional area (and hence strength) of its legs. That's the Curse.

1.1.1 Sparsity

Nonparametric regression uses the data to fit local features of the function f in a flexible way. If there are not enough observations in a neighborhood of some point $\boldsymbol{x}$, then it is hard to decide what $f(\boldsymbol{x})$ should be. It is possible that f has a bump at $\boldsymbol{x}$, or a dip, some kind of saddlepoint feature, or that f is just smoothly increasing or decreasing at $\boldsymbol{x}$. The difficulty is that, as p increases, the amount of local data goes to zero.

This is seen heuristically by noting that the volume of a p-dimensional ball of radius r goes to zero as p increases. This means that the volume of the set centered at $\boldsymbol{x}$ in which a data point $\boldsymbol{x}_i$ must lie in order to provide information about $f(\boldsymbol{x})$ has fewer and fewer points per unit volume as p increases.

This slightly surprising fact follows from a Stirling's approximation argument. Recall the formula for the volume of a ball of radius r in p dimensions:

$$V_r(p) = \frac{\pi^{p/2} r^p}{\Gamma(p/2+1)}. \tag{1.1.1}$$

When p is even, $p = 2k$ for some k. So,

$$\ln V_r(p) = k \ln(\pi r^2) - \ln(k!)$$

since $\Gamma(k+1) = k!$. Stirling's formula gives $k! \approx \sqrt{2\pi} k^{k+1/2} e^{-k}$. So, (1.1.1) becomes

$$\ln V_r(p) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln k + k[1 + \ln(\pi r^2)] - k \ln k.$$

The last term dominates and goes to $-\infty$ for fixed r. If $p = 2k+1$, one again gets $V_r(p) \to 0$. The argument can be extended by writing $\Gamma(p/2+1) = \Gamma((k+1)+1/2)$ and using bounds to control the extra "$1/2$". As p increases, the volume goes to zero for any r. By contrast, the volume of a cuboid of side length r is r^p, which goes to 0, 1, or ∞ depending on $r < 1$, $r = 1$, or $r > 1$. In addition, the ratio of the volume of the p-dimensional ball of radius r to the volume of the cuboid of side length r typically goes to zero as p gets large.

Therefore, if the $\boldsymbol{x}$ values are uniformly distributed on the unit hypercube, the expected number of observations in any small ball goes to zero. If the data are not uniformly distributed, then the typical density will be even more sparse in most of the domain, if a little less sparse on a specific region. Without extreme concentration in that specific region – concentration on a finite-dimensional hypersurface for instance – the increase in dimension will continue to overwhelm the data that accumulate there, too. Essentially, outside of degenerate cases, for any fixed sample size n, there will be too few data points in regions to allow accurate estimation of f.

To illustrate the speed at which sparsity becomes a problem, consider the best-case scenario for nonparametric regression, in which the $\boldsymbol{x}$ data are uniformly distributed in the p-dimensional unit ball. Figure 1.1 plots r^p on $[0,1]$, the expected proportion of the data contained in a centered ball of radius r for $p = 1, 2, 8$. As p increases, r must grow large rapidly to include a reasonable fraction of the data.

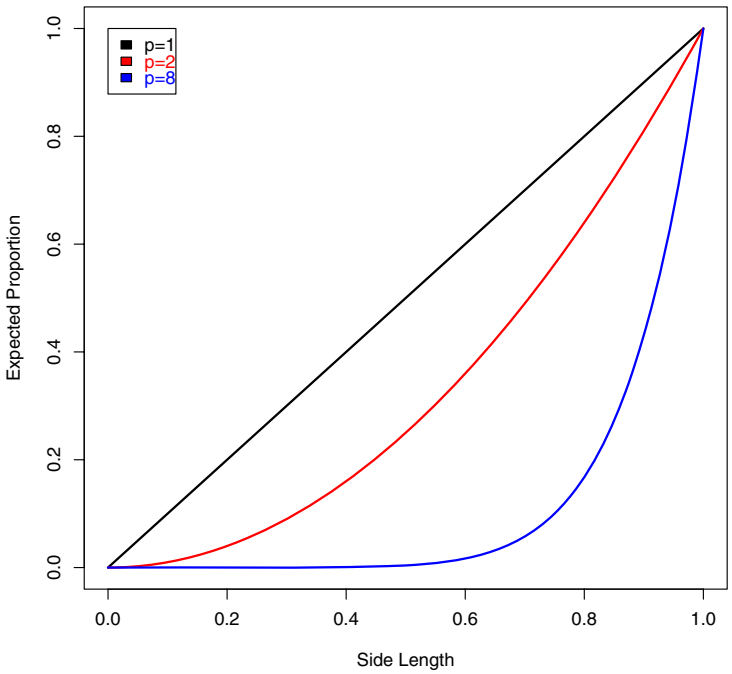

Fig. 1.1 This plots r^p, the expected proportion of the data contained in a centered ball of radius r in the unit ball for $p = 1, 2, 8$. Note that, for large p, the radius needed to capture a reasonable fraction of the data is also large.

To relate this to local estimation of f, suppose one thousand values of are uniformly distributed in the unit ball in $\mathbb{R}^p$. To ensure that at least 10 observations are near $\boldsymbol{x}$ for estimating f near $\boldsymbol{x}$, (1.1.1) implies the expected radius of the requisite ball is $r = \sqrt[p]{.01}$. For $p = 10$, $r = 0.63$ and the value of r grows rapidly to 1 with increasing p. This determines the size of the neighborhood on which the analyst can hope to estimate local features of f. Clearly, the neighborhood size increases with dimension, implying that estimation necessarily gets coarser and coarser. The smoothness assumptions mentioned before – choice of bandwidth, number and size of derivatives – govern how big the class of functions is and so help control how big the neighborhood must be to ensure enough data points are near an $\boldsymbol{x}$ value to permit decent estimation.

Classical linear regression avoids the sparsity issue in the Curse by using the linearity assumption. Linearity ensures that all the points contribute to fitting the estimated surface (i.e., the hyperplane) everywhere on the $\boldsymbol{X}$-space. In other words, linearity permits the estimation of f at any $\boldsymbol{x}$ to borrow strength from all of the $\boldsymbol{x}_i$s, not just the $\boldsymbol{x}_i$s in a small neighborhood of $\boldsymbol{x}$.

More generally, nonlinear models may avoid the Curse when the parametrization does not "pick off" local features. To see the issue, consider the nonlinear model:

$$f(\boldsymbol{x}) = \begin{cases} 17 & \text{if } \boldsymbol{x} \in B_r = \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}_0\| \le r\} \\ \beta_0 + \sum_{j=1}^{p} \beta_j x_j & \text{if } \boldsymbol{x} \in B_r^c. \end{cases}$$

The ball B_r is a local feature. This nonlinear model borrows strength from the data over most of the space, but even with a large sample it is unlikely that an analyst can estimate f near $\boldsymbol{x}_0$ and the radius r that defines the nonlinear feature. Such cases are not pathological – most nonlinear models have difficulty in some regions; e.g., logistic regression can perform poorly unless observations are concentrated where the sigmoidal function is steep.

1.1.2 Exploding Numbers of Models

The second description of the Curse is that the number of possible models increases superexponentially in dimension. To illustrate the problem, consider a very simple case: polynomial regression with terms of degree 2 or less. Now, count the number of models for different values of p.

For $p = 1$, the seven possible models are:

$$\begin{array}{lll} \mathbb{E}(Y) = \beta_0, & \mathbb{E}(Y) = \beta_1 x_1, & \mathbb{E}(Y) = \beta_2 x_1^2, \\ \mathbb{E}(Y) = \beta_0 + \beta_1 x_1, & \mathbb{E}(Y) = \beta_0 + \beta_2 x_1^2, & \mathbb{E}(Y) = \beta_1 x_1 + \beta_2 x_1^2, \\ \mathbb{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2. & & \end{array}$$

For $p = 2$, the set of models expands to include terms in x_2 having the form x_2, x_2^2 and $x_1 x_2$. There are 63 such models. In general, the number of polynomial models of order at most 2 in p variables is $2^a - 1$, where $a = 1 + 2p + p(p-1)/2$. (The constant term, which may be included or not, gives 2^1 cases. There are p possible first order terms, and the cardinality of all subsets of p terms is 2^p. There are p second-order terms of the form $\boldsymbol{x}_i^2$, and the cardinality of all subsets is again 2^p. There are $C(p,2) = p(p-1)/2$ distinct subsets of size 2 among p objects. This counts the number of terms of the form $\boldsymbol{x}_i \boldsymbol{x}_j$ for $i \neq j$ and gives $2^{p(p-1)/2}$ terms. Multiplying and subtracting 1 for the disallowed model with no terms gives the result.)

Clearly, the problem worsens if one includes models with more terms, for instance higher powers. The problem remains if polynomial expansions are replaced by more general basis expansions. It may worsen if more basis elements are needed for good approximation or, in the fortunate case, the rate of explosion may decrease somewhat

if the basis can express the functions of interest parsimoniously. However, the point remains that an astronomical number of observations are needed to select the best model among so many candidates, even for low-degree polynomial regression.

In addition to fit, consider testing in classical linear regression. Once p is moderately large, one must make a very large number of significance tests, and the family-wise error rate for the collection of inferences will be large or the tests themselves will be conservative to the point of near uselessness. These issues will be examined in detail in Chapter 10, where some resolutions will be presented. However, the practical impossibility of correctly identifying the best model, or even a good one, is a key motivation behind ensemble methods, discussed later.

In DMML, the sheer volume of data and concomitant necessity for flexible regression models forces much harder problems of model selection than arise with low-degree polynomials. As a consequence, the accuracy and precision of inferences for conventional methods in DMML contexts decreases dramatically, which is the Curse.

1.1.3 Multicollinearity and Concurvity

The third description of the Curse relates to instability of fit and was pointed out by Scott and Wand (1991). This complements the two previous descriptions, which focus on sample size and model list complexity. However, all three are different facets of the same issue.

Recall that, in linear regression, multicollinearity occurs when two or more of the explanatory variables are highly correlated. Geometrically, this means that all of the observations lie close to an affine subspace. (An affine subspace is obtained from a linear subspace by adding a constant; it need not contain $\mathbf{0}$.)

Suppose one has response values Y_i associated with observed vectors $\boldsymbol{X}_i$ and does a standard multiple regression analysis. The fitted hyperplane will be very stable in the region where the observations lie, and predictions for similar vectors of explanatory variables will have small variances. But as one moves away from the observed data, the hyperplane fit is unstable and the prediction variance is large. For instance, if the data cluster about a straight line in three dimensions and a plane is fit, then the plane can be rotated about the line without affecting the fit very much. More formally, if the data concentrate close to an affine subspace of the fitted hyperplane, then, essentially, any rotation of the fitted hyperplane around the projection of the affine subspace onto the hyperplane will fit about as well. Informally, one can spin the fitted plane around the affine projection without harming the fit much.

In p-dimensions, there will be p elements in a basis. So, the number of proper subspaces generated by the basis is $2^p - 2$ if $\mathbb{R}^p$ and $\mathbf{0}$ are excluded. So, as p grows, there is an exponential increase in the number of possible affine subspaces. Traditional multicollinearity can occur when, for a finite sample, the explanatory variables concentrate on one of them. This is usually expressed in terms of the design matrix $\boldsymbol{X}$ as $\det \boldsymbol{X}'\boldsymbol{X}$ near zero; i.e., nearly singular. Note that $\boldsymbol{X}$ denotes either a matrix or a vector-valued

outcome, the meaning being clear from the context. If needed, a subscript i, as in $\boldsymbol{X}_i$, will indicate the vector case. The chance of multicollinearity happening purely by chance increases with p. That is, as p increases, it is ever more likely that the variables included will be correlated, or seem to be, just by chance. So, reductions to affine subspaces will occur more frequently, decreasing $|\det \boldsymbol{X}'\boldsymbol{X}|$, inflating variances, and giving worse mean squared errors and predictions.

But the problem gets worse. Nonparametric regression fits smooth curves to the data. In analogy with multicollinearity, if the explanatory variables tend to concentrate along a smooth curve that is in the family used for fitting, then the prediction and fit will be good near the projected curve but poor in other regions. This situation is called concurvity . Roughly, it arises when the true curve is not uniquely identifiable, or nearly so. Concurvity is the nonparametric analog of multicollinearity and leads to inflated variances. A more technical discussion will be given in Chapter 4.

1.1.4 The Effect of Noise

The three versions of the Curse so far have been in terms of the model. However, as the number of explanatory variables increases, the error component typically has an ever-larger effect as well.

Suppose one is doing multiple linear regression with $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$; i.e., all convenient assumptions hold. Then, from standard linear model theory, the variance in the prediction at a point $\boldsymbol{x}$ given a sample of size n is

$$\mathrm{Var}[\hat{Y}|\boldsymbol{x}] = \sigma^2(1 + \boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}), \tag{1.1.2}$$

assuming $(\boldsymbol{X}^T\boldsymbol{X})$ is nonsingular so its inverse exists. As $(\boldsymbol{X}^T\boldsymbol{X})$ gets closer to singularity, typically one or more eigenvalues go to 0, so the inverse (roughly speaking) has eigenvalues that go to ∞, inflating the variance. When $p \gg n$, $(\boldsymbol{X}^T\boldsymbol{X})$ is singular, indicating there are directions along which $(\boldsymbol{X}^T\boldsymbol{X})$ cannot be inverted because of zero eigenvalues. If a generalized inverse, such as the Moore-Penrose matrix, is used when $(\boldsymbol{X}^T\boldsymbol{X})$ is singular, a similar formula can be derived (with a limited domain of applicability).

However, consider the case in which the eigenvalues decrease to zero as more and more explanatory variables are included, i.e., as p increases. Then, $(\boldsymbol{X}^T\boldsymbol{X})$ gets ever closer to singularity and so its inverse becomes unbounded in the sense that one or more (usually many) of its eigenvalues go to infinity. Since $\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}$ is the norm of $\boldsymbol{x}$ with respect to the inner product defined by $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, it will usually tend to infinity (as long as the sequence of $\boldsymbol{x}$s used doesn't go to zero). That is, typically, $\mathrm{Var}[\hat{Y}|\boldsymbol{x}]$ tends to infinity as more and more explanatory variables are included. This means the Curse also implies that, for typically occurring values of p and n, the instability of estimates is enormous.

1.2 Coping with the Curse

Data mining, in part, seeks to assess and minimize the effects of model uncertainty to help find useful models and good prediction schemes. Part of this necessitates dealing with the Curse.

In Chapter 4, it will be seen that there is a technical sense in which neural networks can provably avoid the Curse in some cases. There is also evidence (not as clear) that projection pursuit regression can avoid the Curse in some cases. Despite being remarkable intellectual achievements, it is unclear how generally applicable these results are. More typically, other methods rest on other flexible parametric families, nonparametric techniques, or model averaging and so must confront the Curse and other model uncertainty issues directly. In these cases, analysts reduce the impact of the Curse by designing experiments well, extracting low-dimensional features, imposing parsimony, or aggressive variable search and selection.

1.2.1 Selecting Design Points

In some cases (e.g., computer experiments), it is possible to use experimental design principles to minimize the Curse. One selects the $\boldsymbol{x}$s at which responses are to be measured in a smart way. Either one chooses them to be spread as uniformly as possible, to minimize sparsity problems, or one selects them sequentially, to gather information where it is most needed for model selection or to prevent multicollinearity.

There are numerous design criteria that have been extensively studied in a variety of contexts. Mostly, they are criteria on $\boldsymbol{X}^T\boldsymbol{X}$ from (1.1.2). *D*-optimality, for instance, tries to maximize $\det \boldsymbol{X}^T\boldsymbol{X}$. This is an effort to minimize the variance of the parameter estimates, $\hat{\beta}_i$. *A*-optimality tries to minimize $\text{trace}(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. This is an effort to minimize the average variance of the parameter estimates. *G*-optimality tries to minimize the maximum prediction variance; i.e., minimize the maximum of $\boldsymbol{x}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}$ from (1.1.2) over a fixed range of $\boldsymbol{x}$. In these and many other criteria, the major downside is that the optimality criterion depends on the model chosen. So, the optimum is only optimal for the model and sample size the experimenter specifies. In other words, the uncertainty remaining is conditional on n and the given model. In a fundamental sense, uncertainty in the model and sampling procedure is assumed not to exist.

A fundamental result in this area is the Kiefer and Wolfowitz (1960) equivalence theorem. It states conditions under which *D*-optimality and *G*-optimality are the same; see Chernoff (1999) for an easy, more recent introduction. Over the last 50 years, the literature in this general area has become vast. The reader is advised to consult the classic texts of Box et al. (1978), Dodge et al. (1988), or Pukelsheim (1993).

Selection of design points can also be done sequentially; this is very difficult but potentially avoids the model and sample-size dependence of fixed design-point criteria. The full solution uses dynamic programming and a cost function to select the explanatory

values for the next response measurement, given all the measurements previously obtained. The cost function penalizes uncertainty in the model fit, especially in regions of particular interest, and perhaps also includes information about different prices for observations at different locations. In general, the solution is intractable, although some approximations (e.g., greedy selection) may be feasible. Unfortunately, many large data sets cannot be collected sequentially.

A separate but related class of design problems is to select points in the domain of integration so that integrals can be evaluated by deterministic algorithms. Traditional Monte Carlo evaluation is based on a Riemann sum approximation,

$$\int_S f(\boldsymbol{x})d\boldsymbol{x} \approx \sum_{i=1}^{n} f(\boldsymbol{X}_i)\Delta(S_i),$$

where the S_i form a partition of $S \subset \mathbb{R}^p$, $\Delta(S_i)$ is the volume of S_i, and the evaluation point $\boldsymbol{X}_i$ is uniformly distributed in S_i. The procedure is often easy to implement, and randomness allows one to make uncertainty statements about the value of the integral. But the procedure suffers from the Curse; error grows faster than linearly in p.

One can sometimes improve the accuracy of the approximation by using nonrandom evaluation points $\boldsymbol{x}_i$. Such sets of points are called quasi-random sequences or low-discrepancy sequences. They are chosen to fill out the region S as evenly as possible and do not depend on f. There are many approaches to choosing quasi-random sequences. The Hammersley points discussed in Note 1.1 were first, but the Halton sequences are also popular (see Niederreiter (1992a)). In general, the grid of points must be fine enough that f looks locally smooth, so a procedure must be capable of generating points at any scale, however fine, and must, in the limit of ever finer scales, reproduce the value of the integral exactly.

1.2.2 Local Dimension

Nearly all DMML methods try to fit the local structure of a function. The problem is that when behavior is local it can change from neighborhood to neighborhood. In particular, an unknown function on a domain may have different low-dimensional functional forms on different regions within its domain. Thus, even though the local low-dimensional expression of a function is easier to uncover, the region on which that form is valid may be difficult to identify.

For the sake of exactitude, define $f : \mathbb{R}^p \to \mathbb{R}$ to have locally low dimension if there exist regions $R_1, R_2, \ldots$ and a set of functions $g_1, g_2, \ldots$ such that $\bigcup R_i \approx \mathbb{R}^p$ and for $\boldsymbol{x} \in R_i$, $f(\boldsymbol{x}) \approx g_i(\boldsymbol{x})$, where g_i depends only on q components of $\boldsymbol{x}$ for $q \ll p$. The sense of approximation and meaning of $\ll$ is vague, but the point is not to make it precise (which can be done easily) so much as to examine the local behavior of functions from a dimensional standpoint.

As examples,

$$f(\boldsymbol{x}) = \begin{cases} 3x_1 & \text{if } x_1 + x_2 < 7 \\ x_2^2 & \text{if } x_1 + x_2 > 7 \\ x_1 + x_2 & \text{if } x_1 = x_2, \end{cases} \quad \text{and} \quad f(\boldsymbol{x}) = \sum_{k=1}^{m} \alpha_k I_{R_k}(\boldsymbol{x})$$

are locally low-dimensional because they reduce to functions of relatively few variables on regions. By contrast,

$$f(\boldsymbol{x}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j \text{ for } \beta_j \neq 0 \quad \text{and} \quad f(\boldsymbol{x}) = \prod_{j=1}^{p} x_j$$

have high local dimension because they do not reduce anywhere on their domain to functions of fewer than p variables.

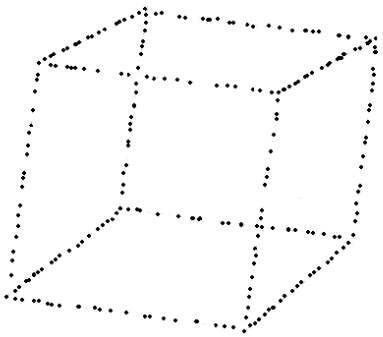

Fig. 1.2 A plot of 200 points uniformly distributed on the 1-cube in $\mathbb{R}^3$, where the plot is tilted 10 degrees from each of the natural axes (otherwise, the image would look like points on the perimeter of a square).

As a pragmatic point, outside of a handful of particularly well-behaved settings, success in multivariate nonparametric regression requires either nonlocal model assumptions or that the regression function have locally low dimension on regions that are not too hard to identify.

Since most DMML methods use local fits (otherwise, they must make global model assumptions), and local fitting succeeds best when the data have locally low dimension, the difficulty is knowing in advance whether the data have simple, low-dimensional structure. There is no standard estimator of average local dimension, and visualization methods are often difficult, especially for large p.

To see how hidden structure, for instance a low-dimensional form, can lurk unsuspected in a scatterplot, consider q-cubes in $\mathbb{R}^p$. These are the q-dimensional boundaries of a p-dimensional cube: A 1-cube in $\mathbb{R}^2$ is the perimeter of a square; a 2-cube in $\mathbb{R}^3$ consists of the faces of a cube; a 3-cube in $\mathbb{R}^3$ is the entire cube. These have simple structure, but it is hard to discern for large p.

Figure 1.2 shows a 1-cube in $\mathbb{R}^3$, tilted 10 degrees from the natural axes in each coordinate. Since $p = 3$ is small, the structure is clear.

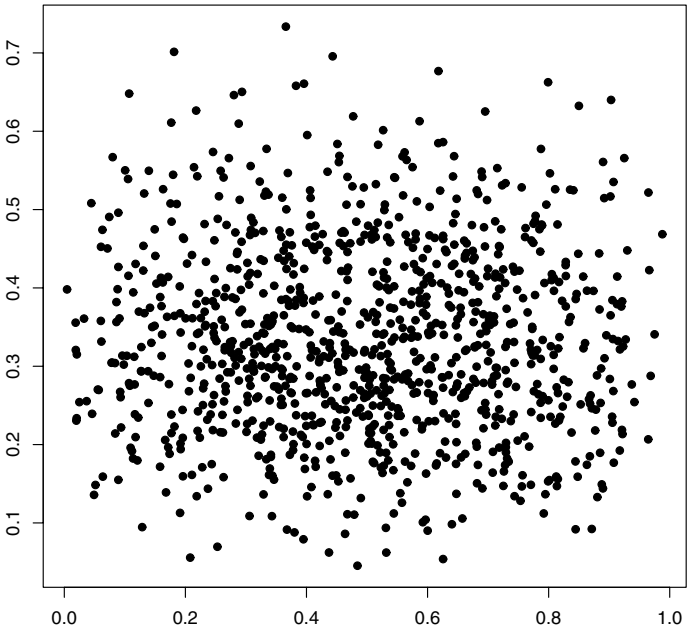

Fig. 1.3 A plot of 200 points uniformly distributed on the 1-cube in $\mathbb{R}^{10}$, where the plot is tilted 10 degrees from each of the natural axes (otherwise, the image would look like points on the perimeter of a square).

In contrast, Fig. 1.3 is a projection of a 1-cube in $\mathbb{R}^{10}$, tilted 10 degrees from the natural axes in each coordinate. This is a visual demonstration that in high dimensions, nearly all projections look Gaussian, see Diaconis and Freedman (1984). This shows that even simple structure can be hard to see in high dimensions.

Although there is no routine estimator for average local dimension and no standard technique for uncovering hidden low-dimensional structures, some template methods are available. A template method is one that links together a sequence of steps but many of the steps could be accomplished by any of a variety of broadly equivalent

techniques. For instance, one step in a regression method may involve variable selection and one may use standard testing on the parameters. However, normal-based testing is only one way to do variable selection and one could, in principle, use any other technique that accomplished the same task.

One way to proceed in the search for low local dimension structures is to start by checking if the average local dimension is less than the putative dimension p and, if it is, "grow" sets of data that can be described by low-dimensional models.

To check if the local dimension is lower than the putative dimension, one needs to have a way to decide if data can locally be fit by a lower-dimensional surface. In a perfect mathematical sense, the answer is almost always no, but the dispersal of a portion of a data set in a region may be tight enough about a lower-dimensional surface to justify the approximation. In principle, therefore, one wants to choose a number of points at least as great as p and find that the convex hull it forms really only has $q < p$ dimensions; i.e., in the leftover $p - q$ dimensions, the convex hull is so thin it can be approximated to thickness zero. This means that the solid the data forms can be described by q directions. The question is how to choose q.

Banks and Olszewski (2004) proposed estimating average local dimension in structure discovery problems by obtaining M estimates of the number of vectors required to describe a solid formed by subsets of the data and then averaging the estimates. The subsets are formed by enlarging a randomly chosen sphere to include a certain number of data points, describing them by some dimension reduction technique. We specify principal components, PCs, even though PCs will only be described in detail in Chapter 8, because it is popular. The central idea of PCs needed here is that it is a method that produces vectors from explanatory variable inputs in order of decreasing ability to explain observed variability. Thus, the earlier PCs are more important than later PCs. The parallel is to a factor in an ANOVA: One keeps the factors that explain the biggest portions of the sum of squared errors, and may want to ignore other factors.

The template is as follows.

Let $\{\boldsymbol{X}_i\}$ denote n data points in $\mathbb{R}^p$.

- ☐ Select a random point $\boldsymbol{x}_m^*$ in or near the convex hull of $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ for $m = 1, \ldots, M$.
- ☐ Find a ball centered at $\boldsymbol{x}_m^*$ that contains exactly k points. One must choose $k > p$; $k = 4p$ is one recommended choice.
- ☐ Perform a principal components regression on the k points within the ball.
- ☐ Let c_m be the number of principal components needed to explain a fixed percentage of the variance in the Y_i values; 80% is one recommended choice.

The average $\hat{c} = (1/M)\sum_{m=1}^{M} c_m$ estimates the average local dimension of f. (This assumes a locally linear functional relationship for points within the ball.) If $\hat{c}$ is large relative to p, then the regression relationship is highly multivariate in most of the space; no method has much chance of good prediction. However, if $\hat{c}$ is small, one infers there

are substantial regions where the data can be described by lower-dimensional surfaces. It's just a matter of finding them.

Note that this really is a template because one can use any variable reduction technique in place of principal components. In Chapter 4, sliced inverse regression will be introduced and in Chapter 9 partial least squares will be explained, for instance. However, one needn't be so fancy. Throwing out variables with coefficients too close to zero from goodness-of-fit testing is an easily implemented alternative. It is unclear, a priori, which dimension reduction technique is best in a particular setting.

To test the PC-based procedure, Banks and Olszewski (2004) generated $10 * 2^q$ points at random on each of the $2^{p-q}\binom{p}{q}$ sides of a q-cube in $\mathbb{R}^p$. Then independent $N(\mathbf{0}, .25\boldsymbol{I})$ noise was added to each observation. Table 1.1 shows the resulting estimates of the local dimension for given putative dimension p and true lower-dimensional structure dimension q. The estimates are biased down because the principal components regression only uses the number of directions, or linear combinations, required to explain only 80% of the variance. Had 90% been used, the degree of underestimation would have been less.

q							
7							5.03
6						4.25	4.23
5					3.49	3.55	3.69
4				2.75	2.90	3.05	3.18
3			2.04	2.24	2.37	2.50	2.58
2		1.43	1.58	1.71	1.80	1.83	1.87
1	.80	.88	.92	.96	.95	.95	.98
	p=1	2	3	4	5	6	7

Table 1.1 Estimates of the local dimension of q-cubes in $\mathbb{R}^p$ based on the average of 20 replications per entry. The estimates tend to increase up to the true q as p increases.

Given that one is satisfied that there is a locally low-dimensional structure in the data, one wants to find the regions in terms of the data. However, a locally valid lower-dimensional structure in one region will typically not extend to another. So, the points in a region where a low-dimensional form is valid will fit well (i.e., be good relative to the model), but data outside that region will typically appear to be outliers (i.e., bad relative to the model).

One approach to finding subsamples is as follows. Prespecify the proportion of a sample to be described by a linear model, say 80%. The task is to search for subsets of size $.8n$ of the n data points to find one that fits a prechosen linear model. To begin, select k, the number of subsamples to be constructed, hoping at least one of them matches 80% of the data. (This k can be found as in House and Banks (2004) where this method is described.) So, start with k sets of data, each with $q+2$ data points randomly assigned to them with replacement. This is just enough to permit estimation of q coefficients and assessment of goodness of fit for a model. The q can be chosen near $\hat{c}$ and then nearby values of q tested in refinements. Each of the initial samples can be augmented