ESTIMATING LOCAL DIMENSIONALITY

David Banks, NIST
Robert Olszewski, Carnegie Mellon University
David Banks, NIST, Gaithersburg MD 20899 (banks@nist.gov)

Key Words: Curse of Dimensionality, Data Mining, Nonparametrics.

Abstract:
The Curse of Dimensionality implies that inference in high dimensions is always hard. Some methods, such as PPR and neural nets, manage to avoid the full force of the curse on narrow technical grounds, but all of the seriously competitive approaches to multivariate nonparametric regression are successful only insofar as they are able to detect and exploit locally simple structure in the data. From this perspective, it is useful for the practicing statistician to estimate the average local dimensionality of a given dataset before investing effort in a difficult analysis. This paper describes two strategies for such estimation; one strategy addresses the case in which there is an identifiable response variable (regression), and the other addresses the case in which no variable is special (principal components). We also present the results from software written to implement these strategies, and discuss how these results inform prudent statistical practice.

1. Introduction

Classical statistical methods were designed to work for low-dimensional data. However, regression analysis and structure discovery (or data mining, in the parlance of computer scientists), quickly become unreliable in high dimensions; this phenomenon is called the “curse of dimensionality” (COD). There are three nearly equivalent formulations of the COD, each offering a usefully different perspective on the problem:

1. The number of possible structural relationships increases faster than exponentially with dimension. To see this, consider just the subclass of polynomial regression models of degree 2 or less with explanatory variables of dimension \( p \). A simple combinatorial argument shows that there are \( 2^{p+p(p+1)/2-1} \) models among which the data must discriminate, and thus one needs large quantities of data.

2. When the dimensionality is high, nearly all datasets are sparse. If \( n \) points are uniformly distributed in the unit cube in \( \mathbb{R}^p \), the proportion of data one expects to find in a subcube of with side length \( d < 1 \) is \( d^p \); which goes to zero as \( p \) increases. Thus the local information about the structure quickly becomes small for a fixed sample size \( n \).

3. In high dimensions, nearly all datasets show multicollinearity (and its nonparametric generalization, concurrence); this makes prediction highly unstable in some regions. Multicollinearity occurs when the values of the explanatory variables lie close to a proper subspace of \( \mathbb{R}^p \), and concurrence occurs when they lie close to a smooth submanifold in \( \mathbb{R}^p \). If \( p \) is large, then the number of possible subspaces and some submanifolds is very large, making it highly likely that such concentration is present simply by chance (Scott and Wand, 1991).

These illustrations have been couched in the context of regression, but the ideas extend to other situations. Detailed discussion of this topic and its consequences may be found in Hastie and Tibshirani (1990).

Historically, methods of multivariate statistical analysis sidestepped the COD by imposing strong model assumptions that restricted the potential complexity of the fitted models, thereby allowing sample information to have non-local influence. But now there is growing demand for data analytical techniques that make weaker model assumptions and use larger datasets. This has led to the rapid development of a number of new methods; in the regression domain, prominent methods are MARS, PPR, LOESS, ACE, AVAS, GAM, CART, neural nets, and wavelets. The structure discovery domain is a bit less developed, but there are now techniques for factor analysis, principal components analysis, and cluster analysis that employ projection pursuit, recursive partitioning, and other ideas that have been central to the success of the regression techniques.

Funded by DARPA Grant No. F30602-96-1-0349
However, none of these methods is truly designed to handle honestly high-dimensional data. The ones that are sometimes successful all assume some kind of locally low-dimensional structure: if they have happened to assume a structure that does an adequate job of describing the situation, then they work very well, but if they have not, then they fail. Hastie and Tibshirani (1990) discuss how the new-wave methods are designed to work well when particular kinds of structure are present. In contrast, methods that don’t assume such local structure, e.g., simple local averaging, are usually unsuccessful because they confront the full force of the COD.

To be a bit more specific, note that MARS, PPR, neural nets, and wavelets look hard to select a small set of variables that are influential in a particular region (MARS and wavelets) or a particular direction (PPR and neural nets). MARS and wavelets do not assume that the same variables are active everywhere, which is a great advantage, but it is unclear whether they are particularly successful in handling regions where the local functional dependence involves more than about four variables. Similarly, PPR and neural nets are most effective when the regression surface in a given direction is dominated by the behavior of a small number of explanatory variables. In an alternate path to a comparable end, GAM, ACE, and AVAS all assume a simple additive structure that precludes the interactions that make high-dimensions so difficult; this assumption posits a different kind of local simplicity that enables an evasion of the COD. Banks, Maxion, and Olzewski (1997) describe a comparison of the performances of new-wave methodologies that highlights these issues.

Our perspective is that some problems are intrinsically too hard for statisticians to solve with the data available. It then becomes valuable to find a screening tool to determine when a dataset has simple structure that makes it potentially amenable to a modern computer-intensive analysis, as opposed to when a dataset is so complex that no method has realistic hopes of discovering useful information. To this end, we describe a strategy for screening datasets that applies to either regression or structure discovery, and show the results of a designed experiment that implements our strategy for the structure discovery form of the problem.

As a final point, we note that recent work has shown that in a narrow, technical sense, neural nets (Barron, 1994) and PPR (Zhao and Atkeson, 1992) evade the COD. For a rich class of functions, the mean squared error in their regression fits increases linearly in the dimension of the data. However, these results are asymptotic, and do not appear to translate into any useful comparative dominance, as demonstrated either by practice or simulation studies. In fact, Banks, Maxion, and Olzewski (1997) found that neural net methods were notably poor in their comparative performance across a range of situations.

2. Screening Strategy

The simplest and first case to consider is that of structure discovery, which is more akin to principal components analysis than to regression. Here the data consist of vectors \( \mathbf{X}_1, \ldots, \mathbf{X}_n \) in \( \mathbb{R}^p \), and one wants to discover whether there is some hidden regularity.

We call this the “crumpled napkin problem” because of the following illustration. Suppose during dinner with Persi Diaconis (a mathematician who is also a magician), he draws a large number of dots on his napkin, crumples it loosely, and causes the napkin to become invisible, so that only the dots may be seen. From cursory inspection, it appears that the dots are a featureless blob in space; but if one looks very closely, and if there are enough dots, one could in principle discover that the points lie on the hidden two-dimensional manifold that is the surface of the napkin. The clue that enables that insight is the fact that in sufficiently small subregions of the volume containing the dots, the dots tend to lie on nearly flat two-dimensional surfaces. Thus the local dimensionality is in fact two except near the folds in the napkin, despite the apparent three-dimensionality.

Following this direction, our analytical strategy is to place the center of a small hypersphere in \( \mathbb{R}^p \) at random in the smallest right hyperparallellepiped containing the data. The radius of the sphere is then adjusted so that the sphere contains a fixed number of points. (In our work, that fixed number is \( 2p + 1 \), but this should probably be adjusted upward; the danger with larger values is that the sphere is so large that it typically includes “folds” of the napkin, distorting the results, and pointing up the delicate balance between sample size requirements and the amount of folding that can be accommodated.) At each placement, we record the points that lie inside the sphere, and perform a principal components analysis on them. The number of eigenvalues that contribute appreciably to the trace of the covariance matrix indicates the approximate local dimensionality of the data. (In our work, we estimate the local dimensionality as the number of eigenvalues needed to sum to 80% of the trace, but this is more a choice of convenience than principle.) Finally, we average
the local estimates of dimension; if the result is much smaller than \( p \), we have good reason to hope that there is simple structure hidden in the data, and thus the dataset might repay further study. But if the result is not much smaller than \( p \), this suggests that no currently available analysis can succeed, and our time would be better spent on other projects.

The previous strategy is still a bit too simple. Suppose that one's data fill out the locus between two \( p \)-dimensional spheres of different radii, each centered at \( 0 \). Then a randomly-placed fixed-radius sphere will typically find either no points in its ambit (when it lies in empty region of the smaller sphere), or it will estimate a local dimensionality of \( p \) when it is located inside the outer sphere but outside the inner sphere. Thus the proposed strategy does not notice structure caused by regions of data sparsity. To repair this deficiency, we make a second analysis that is a slight variation on the strategy delineated previously. As before, we place random hyperspheres in the data cloud, but these have fixed radius (as is described in the following section). We then count the number of observations within the sphere at each random position; if the number falls below a threshold (we use \( 2p \)), then the region is declared sparse. At the end of the run, we record the proportion of positions that were sparse.

When both analyses are complete, we examine the proportion of sparse sites and the average local dimensionality. If the former number is large or the latter number is small, this implies the presence of interesting structure.

There has been previous research that bears on this problem. Shepard and Carroll (1966) made an early attempt at “uncrumpling the napkin” using two approaches; one involves a proximity measure (which is a bit like minimum spanning tree methods in cluster analysis) and the other involves examination of local smoothness. The latter strategy is more closely related to our perspective, but their technique is parametric in spirit and entails the explicit determination of the napkin's folds. They call this second approach “parametric mapping” and seek a numerical solution that identifies the smoothest function that expresses the observed data (with noise) as a function of values in a space of lower dimension \( q \), where \( q \) is determined by the user. From their examples, it seems most useful when the apparent \( p \) is very small (about \( p = 4 \)) and the dimensionality of \( q \) is even smaller. By construction, their solution does not allow different values of \( q \) for different regions of \( \mathbb{R}^p \), whereas our method explicitly recognizes the potential for such variation.

More recently, Kruskal (1975) developed a multidimensional scaling approach to the parametric mapping problem. This method is based upon the Gateaux derivative and poses a number of difficult numerical subproblems, for which Kruskal provides solution strategies. Also, Hastie and Stuetzle (1989) have developed a method of principal curves that offers (in two dimensions) an attractive approach to the parametric mapping problem. However, the goal of uncrumpling seems too ambitious, at least at the first stage. Our interest is the humbler task of simply determining whether an estimate of local dimensionality might warrant a more athletic analysis of the kind these other authors consider.

3. The Designed Experiment

This experiment calibrates the screening strategy described previously for the situation in which data on a \( q \)-dimensional manifold is presented in \( \mathbb{R}^p \), for \( p \geq q \). The particular structure of the manifold that we consider is the \( q \)-dimensional boundary of the unit \( p \)-dimensional hypercube. For example, uniform data on the 1-boundary of a 3-dimensional hypercube would have points that lay (apart from noise) on the 12 edges of the cube; in contrast, uniform data on the 2-boundary of the cube would have points that lay (apart from noise) on the 6 faces of the cube. And clearly, when \( q = p = 3 \), there is no structure present that anyone would be interested to discover. We wish to examine how well average local dimensionality and sparsity do in finding the cases in which \( q < p \).

Before describing the simulation experiment in more detail, we note that the choice of the \( q \)-dimensional boundary of a \( p \)-dimensional cube as the example manifold is driven by several considerations:

1. For a fixed \( p \) and all values of \( q \) (\( q \leq p \)), the covariance matrix of the entire dataset is (up to noise) of the form \( \sigma^2 I \); thus a typical naive analysis based on correlations would fail to find the hidden structure.

2. If one uses X-gobi or some other data projection and rotation scheme (cf. Swayne, Cook, and Buja, 1991), then from nearly every angle, the data appear uninteresting (i.e., spherically symmetric noise). This is an instantiation of a theorem by Diaconis and Freedman (1986), which says that in high dimensions, nearly every projection of a dataset appears normally distributed.
3. The simulation task is easy. As shown by an induction argument given in Sommerville (1958), a \( p \)-dimensional cube has \( 2^{p-q} \binom{p}{q} \) different \( q \)-dimensional boundaries, each of which is itself a \( q \)-dimensional cube. Thus one simulates uniformly in the unit cube in \( \mathbb{R}^q \), then pads the vectors with \( p - q \) additional components, at each of the \( \binom{p}{q} \) possible choices for insertions, with zeroes and/or ones. That is, to produce a particular \( q \)-dimensional face, one selects the \( p - q \) components at which insertions will be made, and then assigns one of the \( 2^{p-q} \) possible patterns of zeroes and ones.

This geometric structure makes generalization to higher dimensions straightforward. Of course, a different kind of manifold might lead to different simulation results. But to preclude nonidentifiability, one needs to ensure that the underlying manifold cannot assume arbitrarily large numbers of folds or creases, as that situation cannot be distinguished from a truly \( p \)-dimensional manifold.

The experiment consists of 20 replications at each combination of the following factor levels:

1. Dimension. We take all values of \((p, q)\) such that \( 7 \geq p \geq q \geq 1 \). (A preliminary experiment took even larger values of \( p \), but that proved time-intensive as data sparsity became a problem.)

2. Noise. All sample observations (uniform on the \( q \)-surface) are corrupted by independent \( p \)-variate Gaussian noise with mean 0 and covariance matrix \( \sigma^2 I \) where \( \sigma^2 \) takes two levels: 0.02 and 0.1. (This amount of noise is relatively small, but since the effect of noise becomes confounded with the amount of crumpling, as reflected in the number of joinings of \( q \)-dimensional faces, we chose to focus on the simplest version of the problem.)

3. Sample Size. We consider two levels; the general formula is \( n = 2^q k \), where \( k = 10, 15 \). (This allows the sample sizes at each level of this factor to vary according to the dimension, reflecting the unequal informational demands imposed by the COD.)

This design is not yet intended as a stringent test of the strategy, but rather establishes the general sensitivity of our implementational code to choices of sphere size and the threshold values for the eigenvalue sum.

For completeness, we note that the algorithm used in the simulation experiment differs slightly from the heuristic sketched in the previous section. In order to avoid the technical difficulties that result from trying to place hyperspheres at random in an arbitrary dataset volume, we make it a practice to first “sphere” the data. This is done by subtracting off the sample mean and premultiplying the result by the square root of the inverse of the sample covariance matrix \( S \); i.e., the transformed data vector \( Y_i \) is found as

\[
Y_i = S^{-1/2}(X_i - \bar{X}).
\]

The result is that the sphered dataset has mean 0 and covariance matrix \( I \). This affine transformation does not affect the local dimensionality of the problem.

The first result of the experiment are shown in Table 1. The entries show the average of the average local dimensionality over the 20 replications of the two levels of the treatment combinations, and the the corresponding standard errors. The factor levels determine the small sample size, low-noise regime (or \( k = 10, \sigma^2 = .02 \)).

<table>
<thead>
<tr>
<th>( q )</th>
<th>5.03</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>2</td>
<td>4.25</td>
</tr>
<tr>
<td>3</td>
<td>3.49</td>
</tr>
<tr>
<td>4</td>
<td>2.75</td>
</tr>
<tr>
<td>5</td>
<td>3.55</td>
</tr>
<tr>
<td>6</td>
<td>3.69</td>
</tr>
<tr>
<td>7</td>
<td>0.01</td>
</tr>
<tr>
<td>1</td>
<td>0.00</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
</tr>
<tr>
<td>3</td>
<td>0.01</td>
</tr>
<tr>
<td>4</td>
<td>0.01</td>
</tr>
<tr>
<td>5</td>
<td>0.01</td>
</tr>
<tr>
<td>6</td>
<td>0.01</td>
</tr>
<tr>
<td>7</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Table 1: The value of \( p \) indicates the apparent dimension, while \( q \) is the true dimension of the data. In each double row, the top number is the estimate of \( q \), while the bottom number is the standard error of that estimate.

From Table 1, it is clear that the estimates of \( q \) are extremely stable across different values of \( p \). The estimates tend to increase slightly with \( p \), but appear to reach a bounding asymptote, as one would expect. Also, note that the estimates are all somewhat lower than the true value of \( q \). This is also to be expected, since within each hypersphere, the principal components analysis declares only as many dimensions as
are needed to explain 80% of the total variation in the enclosed data.

For other levels of the design factors, a similar stability is found. Space limitations preclude a complete listing of the tables obtained in the experiment, but interested readers may obtain them over the web at http://ftp.cs.cmu.edu/user/bobski/napkin.

The second portion of the simulation experiment concerns sparsity, and the results are shown in Table 2. This table displays the average sparsity proportion over the 20 replications of the design. The factor levels shown in the table correspond to large sample size and the high-noise situation (or \( k = 15 \) and \( \sigma^2 = 0.1 \)). The full set of tables, which are too extensive to include, are also available on the web at http://ftp.cs.cmu.edu/user/bobski/napkin.

<table>
<thead>
<tr>
<th>( q )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>41.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>39.1</td>
<td>52.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>34.4</td>
<td>45.3</td>
<td>53.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>32.3</td>
<td>36.2</td>
<td>46.1</td>
<td>55.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>29.1</td>
<td>27.0</td>
<td>34.7</td>
<td>48.6</td>
<td>57.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>28.4</td>
<td>26.5</td>
<td>32.0</td>
<td>41.6</td>
<td>46.6</td>
<td>55.2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>28.5</td>
<td>40.1</td>
<td>45.8</td>
<td>51.0</td>
<td>51.3</td>
<td>51.0</td>
<td>52.5</td>
</tr>
</tbody>
</table>

\( p=1 \) \( 2 \) \( 3 \) \( 4 \) \( 5 \) \( 6 \) \( 7 \)

Table 2: The value of \( p \) indicates the apparent dimension, while \( q \) is the true dimension of the data. The number in a row indicates the proportion of spheres that did not contain at least \( 2p \) observations.

The results in Table 2 show that as \( p \) increases, the proportion of hyperspheres that are sparse increases. This is a natural consequence of the second formulation of the COD; to a degree, it also follows from the geometry of the particular crumpled surface that we explore. In general, \( p \)-cubes built from \( q \)-dimensional boundaries are hollow when \( q < p \), and thus by construction the data should contain a volume that is sparse. This points up the delicate interplay between sample size, noise, and the amount of crumpledness in the underlying structure.

In the simulations used to produce Table 2, our program took the fixed radius to be

\[
 r = \left( \frac{3p}{2^p k} \right)^{\frac{p}{q}},
\]

which aims at ensuring that the expected number of points inside the hypersphere is \( 3p/n \), with an inflation factor to scale for the COD (this assumes uniformity in the \( p \)-dimensional unit sphere, which seems a sensible approximate description of the null distribution after sphering the data). But other choices are worth consideration; in particular, we now recommend that one try to ensure that (under uniformity) the probability of a random hypersphere being sparse be fixed across comparisons at some common value \( v \) for all values of \( p \) and \( n \). To achieve this, one would use the relationship between the binomial sum and the beta function by setting

\[
 v = P[\text{less than } 2p + 1 \text{ values}]
 = \sum_{j=0}^{2p} (r^p)^j (1 - r^p)^{n-j}
 = \frac{n!}{(2p)![(n - 2p - 1)!]} \int_0^1 x^{2p}(1 - x)^{n-2p-1} \, dx
\]

where \( r^p \) is the probability of inclusion in the sampling hypersphere as derived from the formula for the volume of a \( p \)-dimensional sphere. Solving this incomplete beta function for \( r \) is computationally intensive, but standardizes comparisons across the rows and columns in Table 2 and the related tables not shown in this paper.

4. Regression

We want to extend the structure discovery approach to regression analysis. The way to do this is to replace the principal components analysis at each random placement of the small hypersphere with a principal components stepwise regression. The number of components that are selected for inclusion in the model represents the local dimensionality of the functional relationship between the explanatory variables and the response. When this number is low, there is hope that MARS, ACE, GAM, PPR, or one of the other new wave methods might be tuned to discover the kind of relationship that exists.

Of course, this implementation requires a bit more judgment. One must determine the levels of significance in the repeated tests for the stepwise regression’s variable selection, and one should also set a minimum value for the coefficient of determination, below which one does not feel that the strength of the local relationship is sufficiently strong to warrant further study. In this spirit, one can also calculate the local average coefficient of determination which, if small, might persuade one that an analysis is bootless, despite a low average local dimensionality. This would occur, for example, if the explanatory variables had locally low-dimensional structure, but were independent or nearly independent of the response variables.
Of course, even if the proposed technique finds that there is a locally low-dimensional structure relating to the explanatory variables, there is no guarantee that any of the new-wave regression methods will be successful in sorting this out. Our sense is that in many cases one will need to use the results of the method we employ to segment the space into disjoint regions within which a particular regression method is useful. This segmentation might be carried out with respect to either the explanatory variables, the response variable, or both. Related discussion occurs in Li (1991).

Space limitations preclude further development of the regression application here, but the issues and strategy are clear. We are currently involved in work that will illustrate how the method can be used to coordinate a full analysis of a complex regression problem.

5. Conclusions

Our preliminary findings indicate that the estimated local dimensionality is a key piece of information in forecasting the success of a multivariate analysis. The algorithm we’ve developed provides stable estimates of this quantity, and the simulation experiment allows users to calibrate a particular result against an easily understood, but usefully general, example on hypercube boundaries.

For further work, it would be worthwhile to extend the simulation experiment to show that our results are insensitive to the kind of distribution chosen, or the convenient geometry of the hypercube. Conceptually, it would be very surprising if either of these factors would substantially affect the outcome.

There is a range of additional work to undertake. One major problem involves the proper determination of the radius of the hypersphere used to examine local behavior, and how this might be chosen to reflect the amount of noise in the data, the available sample size, and the degree of crumpling that is present. A second issue concerns the definition of sparsity; in the work performed so far, we believe that we have been too latitudinarian, and that we should require non-sparse hyperspheres to contain more than 2p observations. However, this choice is also bound up with sample size considerations.

References


