Sankhyā : The Indian Journal of Statistics
2008, Volume 70-A, Part 1, pp. 1-24

Sparsity and the Possibility of Inference

Peter J. Bickel
University of California, Berkeley, USA
Donghui Yan
University of California, Berkeley, USA

Abstract

We discuss the importance of sparsity in the context of nonparametric regression and covariance matrix estimation. We point to low manifold dimension of the covariate vector as a possible important feature of sparsity, recall an estimate of dimension due to Levina and Bickel (2005) and establish some conjectures made in that paper.

AMS (2000) *subject classification.* Primary 62-02, 62G08, 62G20, 62J10.
Keywords and phrases. Sparsity, statistical inference, nonparametric regression, covariance matrix estimation, dimension estimation.

1 Introduction

In a pathbreaking RSS discussion paper in 1995, Donoho, Johnstone, Hoch and Stern pointed out the importance of sparsity. Subsequent important work by Donoho, Johnstone and collaborators (1995-2006) focused mainly on sparsity in the form of sparse linear representations of regression functions and the like when estimation is carried out using overcomplete dictionaries. My own appreciation of the importance of sparsity came from the following observation that grew out of conversations with my late friend and colleague, Leo Breiman.

Stone (1977) considered the estimation of regression functions

$$\boldsymbol{\eta}(\mathbf{X}) = \mathbb{E}(\mathbf{Y}\big|\mathbf{X})$$

[1]Both authors were supported under NSF Grant DMS0605236.
[2]P. Bickel is responsible for Sections 1-3. Section 4 is based on the work of D. Yan.

on the basis of i.i.d. observations, $(\mathbf{X}_i, \mathbf{Y}_i)$, with $\mathbf{X} \in \mathbb{R}^d$, assuming only that $\boldsymbol{\eta}$ has smoothness of order s, e.g., bounded partial derivatives of order s. He obtained the result that, from a minimax (least favorable) point of view, estimation in the root mean square sense could not be achieved at a rate greater than $n^{-\frac{s}{2s+d}}$.

Qualitatively, what this said to us was that unless we assume an unwarranted degree of smoothness for the function we are estimating, the sample sizes required to get reasonable accuracy are larger that anything possibly available. For example, if $s = 2, d = 12, n = 10^4, n^{-\frac{s}{2s+d}} \cong .33$.

Of course, in principle, constants may favor performance, s may be very large, but the qualitative conclusion is that if Nature is out to get us, we have no chance with today's very high dimensional data with poorly understood models. Yet in practice there routinely are remarkable successes. Here is a table of high dimensional data sets with large to small n where high prediction accuracy has been achieved. R corresponds to misclassification probability.

Table 1. Some empirical evidence

	d	classes	n	R
ZIP code digits	256	10	10,000	<0.025
Microarrays	3-4000	2-3	70-100	0.08
spam	57	2	4,600	<0.07

What are possible explanations? I believe in Einstein's words that "God is subtle but not malicious".

In the next two sections we will discuss the meaning and consequences of sparsity in two contexts, regression and covariance matrix estimation and relate them to some extent.

Roughly speaking we will discuss sparse data, by which we mean points lying on or near a low dimensional sub-manifold of a high dimensional space and sparse models, ones whose probability structures are characterized by low dimensional parameters.

Our treatment is extremely sketchy and covers only a very small part of a huge and growing body of literature. Key topics such as Bayesian methods are only briefly pointed to at the end. But we hope this will only be one of the first of many review papers of aspects of modern statistics which we believe

are both novel and of key importance. In a final more technical section, we take the opportunity to analyze and essentially settle some conjectures on nonparametric dimension estimation raised in Levina and Bickel (2005).

2 Sparsity Issues in Nonparametric Regression

Consider the regression model we introduced in the introduction

$$\mathbf{Y} = \boldsymbol{\eta}(\mathbf{X}) + \epsilon \tag{2.1}$$

where $\mathbb{E}(\epsilon|\mathbf{X}) = 0$ and for simplicity we take $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent of $\mathbf{X}$. We specify the family of possible probability distributions of $(\mathbf{X}, \mathbf{Y})$ by $\mathcal{P}$.

There are two familiar paradigms. One is linear regression where $f_1, ..., f_p : \mathcal{X} \mapsto \mathbb{R}$ is given and

$$\boldsymbol{\eta}(\mathbf{X}) = \boldsymbol{f}^T(\mathbf{X})\boldsymbol{\beta}, \boldsymbol{f} = (f_1, ..., f_p)^T \tag{2.2}$$

for $\boldsymbol{\beta} \in \mathbb{R}^p$ unknown. The other is nonparametric regression where $\boldsymbol{\eta}(\mathbf{X})$ belongs to "a nice function space", e.g., $\mathcal{X} = \mathbb{R}^d, \mathcal{P}_s \leftrightarrow \{|D^k\eta|(.) \leq K < \infty, 1 \leq k \leq s\}$.

A theoretical goal is to obtain a minimax procedure, $\hat{\boldsymbol{\eta}}(\mathbf{X})$, such that,

$$\sup_{\mathcal{P}} \mathbb{E}_{\mathbb{P}}(\hat{\boldsymbol{\eta}}(\mathbf{X}) - \mathbf{Y})^2 = \min_{\boldsymbol{\eta}} \sup_{\mathcal{P}} \mathbb{E}_{\mathbb{P}}(\hat{\boldsymbol{\eta}}(\mathbf{X}) - \mathbf{Y})^2 \tag{2.3}$$

or $\mathbb{E}_{\mathbb{P}}(\hat{\boldsymbol{\eta}}(\mathbf{X}) - \mathbf{Y})^2$ is "small" for $\mathbb{P} \in \mathcal{P}$. The standard solution to prediction for linear regression

$$\boldsymbol{\eta}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta}$$

is least squares,

$$\hat{\boldsymbol{\beta}} = [F^T F]^{-1} F^T \boldsymbol{Y} \tag{2.4}$$

where $F_{N\times p} \equiv ||f_j(\boldsymbol{X}_i)||$, yielding as predictions

$$\hat{\mathbf{Y}} = F[F^T F]^{-1} F^T \mathbf{Y} \tag{2.5}$$

$$\hat{\eta}(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}^T \boldsymbol{f}(\boldsymbol{x}). \tag{2.6}$$

We are faced with the "Overfitting" problem,

$$\mathbb{E}|\hat{\mathbf{Y}} - F^T\boldsymbol{\beta}|^2 = \frac{\sigma^2 p}{n} \tag{2.7}$$

where $F^T\boldsymbol{\beta}$ is the best predictor of $\mathbf{Y}$ if $\boldsymbol{\beta}$ is known. Note that if $p > n$, then $\boldsymbol{\beta}$ is unidentifiable; if $\frac{p}{n} \to \infty$, then $\hat{\mathbf{Y}}$ is worthless.

In the nonparametric case, $p = \infty$. The standard solution is, if $\mathcal{X} = \mathbb{R}^d$ to find a basis $f_1, ..., f_p, ...$, such that for $\boldsymbol{\eta} \in \mathcal{P}_s$,

$$\max_{\mathcal{P}} \mathbb{E}(\boldsymbol{\eta}(\boldsymbol{X}) - \boldsymbol{f}_p^T(\boldsymbol{X})\boldsymbol{\beta}_p)^2 \to 0$$

where $\boldsymbol{f}_p(x) = (f_1(x), ..., f_p(x))^T$ and regularize. For a review of regularization - see Bickel and Li (2007) for instance. Some standard solutions are to choose p so that

$$\hat{\boldsymbol{\beta}}(\lambda) = \arg\min_{\boldsymbol{\beta}_p, p} \left\{ |\mathbf{Y} - \boldsymbol{f}_p^T\boldsymbol{\beta}_p|^2 + \lambda|\boldsymbol{\beta}_p|^r \right\}$$

$$\left.\begin{array}{ll} r = 2 & \text{Ridge Regression} \\ r = 1 & \text{LASSO} \end{array}\right\}, \lambda \to 0.$$

There are many other methods - see, for example, references in Candès and Tao (2007) discussion, and Hastie, Tibshirani, and Friedman (2001).

All such methods can be interpreted as estimating "sparse" - low dimensional approximations to $\boldsymbol{\eta}$ in the belief that these are adequate. It is in fact methods such as these which yield Stone's rates. But they do not therefore a priori eliminate our difficulty, since we identify approximate sparsity with smoothness of $\boldsymbol{\eta}$. There are however often classes of models, we shall refer to as sparse models (SM) which do - and these are, in some form, familiar from classical statistics. By SM we think of model families, such as

a) $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \sum_{j=1}^d g_j(\boldsymbol{X}_j)$ (Additive models)

   Type a) are the natural generalizations of no interaction models in ANOVA. They were introduced by Breiman and Friedman (1985) and as one might expect correspond to difficulty $d = 1$.

b) There exists $S \subset \{1, ..., d\}$ such that $|S| \ll \min(d, n)$ and

   (i) $\{\boldsymbol{X}_j : \ j \in S\} \perp \{\boldsymbol{X}_k : \ k \in S^c\}$

   (ii) $\boldsymbol{Y} \perp \{\boldsymbol{X}_k : \ k \in S^c\}$

   (iii) $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}) = \mathbb{E}(\boldsymbol{Y}|\boldsymbol{X}_j : \ j \in S)$

Note that $\mathbf{X}_j$ are used generically and could be (known) functions of original $\mathbf{X}_1, ..., \mathbf{X}_d$.

Type b) corresponds to the belief that most variables X_j are irrelevant in that they are independent of both Y and the set of relevant variables in S.

The other main feature of regression models which makes Table 1 possible is, we believe, the presence of what we call "Sparse data" (SD). By SD we mean, $\boldsymbol{X}$ lives on (or close to) $\mathbb{M}$, a m dimensional Riemannian manifold $\boldsymbol{X} = \boldsymbol{T}_{d\times 1}(\boldsymbol{U})$ for $\boldsymbol{U} \in \mathcal{O} \subset \mathbb{R}^m, m \ll d$, $\boldsymbol{T}$ 1-1 and smooth or $|\boldsymbol{X} - \boldsymbol{T}(\boldsymbol{U})| \leq \epsilon$ for some $\boldsymbol{U}$. Note a special case is that $\mathbb{M}$ is a hyperplane, and this leads to principal components regression.

Of course, one can have both, irrelevant variables corresponding to SM and predictors highly correlated with each other as well as $\mathbf{Y}$. It is, however, clear that the irrelevant variables have to be removed before we can take advantage of the correlated good predictors - since independent variables result in high dimension.

It is worth pointing out that the goal we are focusing on here is prediction and sparsity really means being well approximable by low dimensional linear submodels. This is different from the goal of model selection where there is implicitly a belief that the data has been generated by an $\boldsymbol{\eta}(\boldsymbol{x})$ for which most of the $\boldsymbol{\beta}_j$ are 0.

For prediction, for instance, we choose λ in the lasso so as to ensure optimal prediction for a smoothness class $\mathcal{P}$ with $f_1(.), ..., f_k(.), ...$, a complete basis for representing $\eta(.)$ in $\mathcal{P}$. Then, we expect most estimated coefficients of the $f_j(\boldsymbol{x})$ to be small but not necessarily 0. This choice of λ does not lead to what is known as consistency, which is defined by: If $\boldsymbol{\eta}(\boldsymbol{x})$ has (sparsest) representation

$$\boldsymbol{\eta}(\boldsymbol{x}) = \sum \{\boldsymbol{\beta}_j f_j(x) : \ j \in S\},$$

then the $\hat{\boldsymbol{\beta}}_j$ of the estimated $\hat{\boldsymbol{\eta}}$ are $\neq 0$ iff $j \in S$. On the other hand, the choice of λ which does lead to consistency does not yield the best minimax prediction risks - see for instance Atkinson (1980) and Yang (2005) for a discussion. However, nothing prevents us from first optimizing for prediction and then doing model selection or vice versa, see Fan and Lv (2008) for instance. These questions are still being explored. We believe the practical question is different. With Box (1979) we agree that,

> "All models are false but some models are useful".

Our real interest is in determining which "factors" among $\boldsymbol{X}_1, ..., \boldsymbol{X}_d$ are "important". A possibly more fruitful but yet unexplored point of view is

to isolate all models of dimension at most m whose predictive power is close to optimal and then study the factors which appear in these.

Bickel and Li (2007) studied the theoretical effect on nonparametric regression if the high dimensional vector of covariates satisfies our notion of SD. They noted that, if the manifold is unknown, employing local linear or higher order regression methods using the d dimensional covariates but choosing the bandwidth by cross validation or some other data determined way yields the same minimax risks as if $\mathbf{X} \in \mathbb{R}^m$ rather than $\mathbf{X} \in \mathbb{R}^d$. But see also Niyogi (2007). This result can be thought of as the nonlinear analogue of the observation that for prediction collinearity of predictive variables is immaterial since the p in (2.7) is the dimension of the linear space of predictors.

To what extent does SD appear? One way of checking is to construct a nonparametric estimate of dimension suitably defined. There are many notions of dimension and a number of estimates have been proposed in the physics and dynamical systems literature. Levina and Bickel (2005) developed a simple estimate of dimension at a point which can be extended to estimating manifold dimension when it is an integer. We will present this estimate in Section 4 and study some conjectures they made. Although the estimate was applied successfully to some examples in their paper, its practical applicability requires a lot of further exploration. In line with our theme we have applied it to the famous handwritten ZIP code digits (of apparent dimension 256) example (see table 2). Here is a table of estimated dimension

Table 2.Estimated dimension of the ZIP code digits of apparent dimension 256.

Digit	0	1	2	3	4	5	6	7	8	9
Dimension	9	8	11	12	9	11	9	8	11	9

3 Covariance Matrix Estimation

In the regression context sparsity and approximate sparsity are relatively easy to define. This is not so evident when we are interested in covariance matrices. Our model here is

$$\boldsymbol{X}_1, ..., \boldsymbol{X}_n \text{ i.i.d., } \mathbb{E}(\boldsymbol{X}_1) = \mu, \ Var(\boldsymbol{X}_1) \equiv \mathbb{E}(\boldsymbol{X}_1 - \mu)(\boldsymbol{X}_1 - \mu)^T \equiv \Sigma.$$

Estimating Σ is important for a number of purposes, for example, principal component analysis (PCA), linear or quadratic discriminant analysis

(LDA/QDA), inferring independence and conditional independence (graphical models), implicit estimation of linear regression, $\Sigma_{\boldsymbol{XX}}^{-1}\Sigma_{\boldsymbol{XY}}$ with $\boldsymbol{X}$ regressors and $\boldsymbol{Y}$ the response.

Through recent results in random matrix theory a major pathology of the empirical covariance matrix has been pointed out. The eigenstructure is inconsistent for i.i.d. component models as soon as $\frac{p}{n} \to c > 0$ - see Johnstone and Lu (2008) for a review.

We have implicitly noted in the previous section that when used in regression, i.e., least squares, using the empirical covariance matrix is a part of least squares which breaks down even for prediction. Bickel and Levina (2004) point out the breakdown of LDA when $\frac{p}{n} \to \infty$.

These problems do not simply emerge because we use the empirical covariance matrix, as the minimax results of the previous section suggest. The issue is again that, without sparsity assumptions, estimation of large covariance matrices is hopeless. A number of notions of sparsity are discussed in Bickel and Levina (2008)(i)(ii). The simplest is permutation invariant sparsity where,

A) Each row of Σ is sparse or sparsely approximable in operator norm, e.g., $S_i = \{j : \sigma_{ij} \neq 0\}$ and $|S_i| \leq s$ for all i

or

B) Each row of Σ^{-1} is sparsely approximable.

Each allows estimation of the inverse but the conditions are different. See El Karoui (2008) for a more general notion. Note the interpretations of A), B) in the Gaussian case where,

A) $\boldsymbol{X}_i \perp \boldsymbol{X}_j$ for all $j \in S_i^c$

or

B) $\boldsymbol{X}_i \perp \boldsymbol{X}_j \mid \boldsymbol{X}_k,\ k \neq j, j \in S_i^c$ where $S_i \equiv \{j : \sigma_{ij} \neq 0\}$.

If sparsity is present as in the regression case, regularization can lead to excellent performance. A forthcoming issue of the Annals of Statistics will contain a number of papers including some of those cited as well as others relating to the area.

One salient feature of all approaches is that in the presence of sparsity, $\frac{\log p}{n} \to 0$ is enough for interesting and very compelling conclusions for general problems. This is an area of continuing research which may shed light on the sparsity issues in regression as well. For instance, it is appealing to estimate the inverse covariance matrix of $\boldsymbol{X}$ taking advantaging of potential sparsity and estimate the covariances of $\boldsymbol{X}$ and $\boldsymbol{Y}$ sparsely and put them together in a sparse version of least squares.

Even more than in regression, the order in which things are done may also matter. Is it good to first estimate the covariance matrix sparsely and then look at its eigenstructure and inverses or aim at each feature of the matrix separately? Approaches of this type may be found in d'Aspremont et al (2007) and Johnstone and Lu (2008) and El Karoui (2007) .

There are two topics we have not touched on which are of key importance and need much additional work

1. The choice, in practice, of regularization parameters for methods that take advantage of sparsity. Theoretical order of magnitude is well understood, but is of little value in this choice. We have great faith in V fold cross validation - see Bickel and Levina (2008)(ii) for one analysis.

2. Bayesian methods. There is clearly an intimate link between regularization in both regression and covariance estimation and Bayesian methods. But in high dimensional situations which Bayesian methods are trustworthy from a frequentist point of view remains to be explored.

4 Some Results on Dimension Estimation

Levina and Bickel (2005) proposed the following estimator for the "true" dimension of a high dimensional dataset

$$\hat{m_k} = \frac{1}{n}\sum_{i=1}^{n} \hat{m_k}(\boldsymbol{X}_i)$$

where $\hat{m_k}(\boldsymbol{x})$ is defined as

$$\hat{m_k}(\boldsymbol{x}) = \Big[\frac{1}{k-1}\sum_{j=1}^{k-1} \log \frac{R_k}{R_j}(\boldsymbol{x})\Big]^{-1} \triangleq (W(\boldsymbol{x}))^{-1}$$

and $R_j(\boldsymbol{x})$ is the $j-$th nearest neighbor distance to $\boldsymbol{x} \in \mathbb{R}^m$.

Here $\hat{m}_k(\boldsymbol{x})$ is a local estimate of dimension and may be more useful than the global estimate $\hat{m}_k$. We will state and prove three theorems about estimation of the "true" dimension m where $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are a sample of d dimensional observations, which however take their values with probability 1 in a flat smooth manifold of dimension m, and have a smooth density with respect to Lebesgue measure on $\mathbb{R}^m$. What we mean by this is explained in the statement of Theorem 4.1 on local dimension estimation below. All theorems can be viewed as proving corrected versions of conjectures in Levina and Bickel (2005).

Our first theorem deals with the local estimate. The idea of the proof is to establish the result for the case $\mathbf{X}$ is uniformly distributed locally on the manifold and the manifold is locally described as an affine transformation of the m-cube. For the uniform case, we can first apply the delta method and then employ some known distributional results. The next theorem deals with the more difficult global behavior which we carefully establish only for the uniform. We only make a second moment calculation and use a theorem of Chatterjee for distributional results. The main observation is that the covariance contribution occurs only if k-th nearest neighbor spheres intersect which means that the corresponding centers are no more than $O((k/n)^{\frac{1}{m}})$ apart. But that event occurs with probability $O(\frac{k}{n})$. The rest of the argument rests upon scaling two such spheres, one centered at $\mathbf{0}$, and their intersection by a common factor. A stronger distributional result due to Yukich (2008) was brought to our attention. His proof is both more elegant and more general. Nevertheless, we believe our more special method which yields both order bounds as $k, n \to \infty$ and local results is worth considering.

THEOREM 1. *Let $\boldsymbol{U}_{m\times 1}$ be a r.v. in $\mathbb{R}^m$. If $\boldsymbol{U}_{m\times 1} \sim f$ with f twice continuously differentiable. $\boldsymbol{X}_{d\times 1} = \boldsymbol{T}\boldsymbol{U}^{\tau}_{m\times 1}$ where $\mathbf{T} = (\mathbf{T}_1, ..., \mathbf{T}_d)^{\tau}$. $\dot{\mathbf{T}}(\boldsymbol{u}) \triangleq ||\frac{\partial \mathbf{T}_i}{\partial u_j}||_{d\times m}$. Assume $\dot{\mathbf{T}}$ is of rank m for all $\boldsymbol{u} \in \mathbb{R}^m$, and $\mathbf{T}$ is twice continuously differentiable with*

$$\left|\frac{\partial \mathbf{T}_i(\boldsymbol{u})}{\partial u_a \partial u_b}\right| \leq M$$

for all $\boldsymbol{u}, i, a, b$. Then if $k \to \infty, k^{(1+\frac{m}{4})}/n \to 0, n \to \infty$ for all $\boldsymbol{x}_0 \in \mathbf{T}\mathbb{R}^m$,

$$\sqrt{k}(\hat{m}_k(\boldsymbol{x}_0) - m) \Rightarrow \mathcal{N}(0, m^2).$$

Here are some preliminaries. Let $\boldsymbol{x}_0 = \mathbf{T}(\boldsymbol{u}_0)$. Then, note

$$\mathbf{T}(\boldsymbol{u}) - \boldsymbol{x}_0 = \dot{\mathbf{T}}(\boldsymbol{u}_0)(\boldsymbol{u} - \boldsymbol{u}_0) + O(|\boldsymbol{u} - \boldsymbol{u}_0|^2). \tag{4.1}$$

We claim that w.l.o.g. we may take

$$\dot{\mathbf{T}} = (\boldsymbol{e}_1, ..., \boldsymbol{e}_m) \triangleq \mathbf{E}$$

where $\boldsymbol{e}_j$ are the coordinate vectors in $\mathbb{R}^d$. Consider the mapping, $S(\boldsymbol{u}) = \mathbf{T}(\mathbf{A}^{-1}(\boldsymbol{u}))$ where $\mathbf{A}_{m\times m}$ is such that,

$$\dot{\mathbf{T}}(\boldsymbol{u}_0) = \mathbf{E}\mathbf{A}.$$

By assumption $\mathbf{A}$ is nonsingular.

If we now redefine $\boldsymbol{X} = S(\mathbf{A}\boldsymbol{U})$, then we have

$$\dot{\boldsymbol{S}}(\mathbf{A}(\boldsymbol{u}_0)) = \mathbf{E}$$

and $\mathbf{A}\boldsymbol{U}$ has density $g(\boldsymbol{\nu}) = |det(\mathbf{A})|^{-1} f(\mathbf{A}^{-1}\boldsymbol{\nu})$ which satisfies the same conditions as f.

Proof. Our proof is of a fairly standard type. We note that

$$S_k \triangleq \hat{m}_k^{-1}(\boldsymbol{x}_0) = \frac{1}{k-1}\sum_{j=1}^{k-1} \log \frac{R_k}{R_j}(\boldsymbol{x}_0)$$

given $R_k(\boldsymbol{x}_0)$ is the mean of $k-1$ i.i.d. variables whose distribution is that of $|\mathbf{T}(\boldsymbol{U}) - \boldsymbol{x}_0|$ given $|\mathbf{T}(\boldsymbol{U}) - \boldsymbol{x}_0| < R_k(\boldsymbol{x}_0)$.

Let $Z(r)$ have the distribution of $-\log \frac{|\mathbf{T}(\boldsymbol{U})-\boldsymbol{x}_0|}{r}$ given $|\mathbf{T}(\boldsymbol{U}) - \boldsymbol{x}_0| < r$. To establish the theorem, we need to establish

$$\mathbb{E}\left(Var(Z(R_k)|R_k)\right) \xrightarrow{p} \sigma^2 \tag{4.2}$$

$$\sqrt{k}\left(\mathbb{E}(Z(R_k)|R_k) - m^{-1}\right) \xrightarrow{p} 0 \tag{4.3}$$

and $\sigma^2 + \tau^2 = \frac{1}{m^2}$.

We need

Lemma 4.1. *Under the conditions of the theorem,*

$$\mathcal{L}\left(Z(R_k)|R_k\right) \Rightarrow \mathcal{E}(\tfrac{1}{m}) \tag{4.4}$$

in probability where $\Rightarrow$ denotes weak convergence.

PROOF. Let $S(\boldsymbol{u}_0, t)$ be the t sphere around $\boldsymbol{u}_0$ in $\mathbb{R}^m$. Then, for $0 \leq t \leq 1$, let $A_t = \{\boldsymbol{u} : \ |\mathbf{T}(\boldsymbol{u}) - \boldsymbol{x}_0| \leq tr\}$. Then

$$S\left(\boldsymbol{u}_0, tr(1 - O(r^2))\right) \subset A_t \subset S\left(\boldsymbol{u}_0, tr(1 + O(r^2))\right) \tag{4.5}$$

since

$$|\boldsymbol{u} - \boldsymbol{u}_0|\left(1 - O(|\boldsymbol{u} - \boldsymbol{u}_0|^2)\right) \leq |\mathbf{T}(\boldsymbol{u}) - \boldsymbol{x}_0| \leq |\boldsymbol{u} - \boldsymbol{u}_0|\left(1 + O(|\boldsymbol{u} - \boldsymbol{u}_0|^2)\right)$$

by (4.1). But, then

$$\int_{A_t} f(\boldsymbol{u})d\boldsymbol{u} = \int_{A_t} \left[f(\boldsymbol{u}_0) + \dot{f}(\boldsymbol{u}_0)(\boldsymbol{u} - \boldsymbol{u}_0) + (\boldsymbol{u} - \boldsymbol{u}_0)^T \mathbf{T} \ddot{f}(\boldsymbol{u}^*)(\boldsymbol{u} - \boldsymbol{u}_0)\right] d\boldsymbol{u} \tag{4.6}$$

where $\dot{f}$ and $\ddot{f}$ are the differentials of f.

By (4.1) we can argue from (4.5) and (4.6) that, for $r \to 0$, if $V(S)$ is the volume of S,

$$\int_{A_t} f(\boldsymbol{u})d\boldsymbol{u} = f(\boldsymbol{u}_0)V\left(S(\boldsymbol{u}_0, tr)\right) + O(r^2)V\left(S(\boldsymbol{u}_0, tr)\right). \tag{4.7}$$

Arguing similarly for $\int_{S(\boldsymbol{u}_0, tr)} f(\boldsymbol{u})d\boldsymbol{u}$, we conclude that

$$\frac{\int_{A_t} f(\boldsymbol{u})d\boldsymbol{u}}{\int_{A_1} f(\boldsymbol{u})d\boldsymbol{u}} = \frac{V(S(\boldsymbol{u}_0, tr))}{V(S(\boldsymbol{u}_0, r))}(1 + O(r^2)). \tag{4.8}$$

But the first term in (4.8) is just the uniform distribution on $S(\mathbf{0}, r)$ in $\mathbb{R}^m$ and thus (4.8) implies that

$$\mathbb{P}[Z(r) > z] = e^{-mz}(1 + O(r^2)), z > 0.$$

Since, for Q uniform on $S(\mathbf{0}, r)$, $-\log \frac{V(S(\mathbf{0},Q))}{V(S(\mathbf{0},r))} = -m \log \frac{Q}{r}$ is well known to have a standard exponential distribution, the lemma follows. □

To complete the proof of (4.2) we need only show that, say

$$\mathbb{E}Z^4(R_k) = O_p(1).$$

But if $k/n \to 0$ evidently $R_k \xrightarrow{p} 0$ and (4.8) implies that all moments of $Z(R_k)$ are uniformly bounded. Finally for (4.3) we need to show not only that

$$\mathbb{E}(Z(R_k)|R_k) \xrightarrow{p} \frac{1}{m}$$

but that the difference is $o(k^{-\frac{1}{2}})$. But again (4.8) yields this since our argument shows that

$$\mathbb{E}(Z(R_k)|R_k) = \tfrac{1}{m} + O_p(R_k^2). \tag{4.9}$$

Again by (4.8) $\mathbb{E}R_k^2 = (\frac{k}{n})^{\frac{2}{m}}(1 + o(1))$. Therefore

$$\sqrt{k}\left[\mathbb{E}(Z(R_k)|R_k) - \tfrac{1}{m}\right] = o_p(1)$$

if $(\frac{k}{n})^{\frac{2}{m}} k^{\frac{1}{2}} \to 0$ which is our assumption. We have established (4.3) and the theorem is proved. □

DISCUSSION. Theorem 4.1 is unsatisfactory in two aspects

1. If m is an integer as is used in the proof then convergence at this rate is not too informative but rather exponential rate convergence theorems are suitable. However, this is mitigated if we note that we can define *local dimension* at $\boldsymbol{x} \in \mathbb{R}^d$ in a set S by dimension $= \gamma$ iff

   $$\lim_{t\to 0} \frac{\lambda(B_t(\boldsymbol{x}) \cap S)}{t^\gamma} \to c > 0$$

   where $B_t(\boldsymbol{x})$ is a ball of radius t centered at $\boldsymbol{x}$ and λ is Lebesgue measure (Volume). It would seem that if we obtain a set of local dimension γ at $\mathbf{T}_{m\times d}(\boldsymbol{u}_0)$ by mapping a set of local dimension γ at $\boldsymbol{u}_0$ in $\mathbb{R}^m, m \geq \gamma$ to $\mathbb{R}^d$ with $\mathbf{T}$ as before then our results should hold. Unfortunately it is not clear that local dimension in this sense is related to any of the global notions of dimension, see e.g. Falconer (1990).

2. We have no guidance on the choice of k from the result. A possible approach which has given plausible answers in some examples is to choose k so as to minimize the difference between $\hat{m}_k(\boldsymbol{x})$ and the empirical standard deviation of the $\log \frac{R_k}{R_j}(\boldsymbol{x})$ which by our result provides another consistent estimate of m^{-1}.

We will establish

THEOREM 4.2. *Under the conditions of Theorem 4.1, if* $\hat{m}_k \triangleq \frac{1}{n}\sum_{i=1}^n \hat{m}_k(\boldsymbol{X}_i)$, *then*

(i)

$$\hat{m}_k - m = \Omega_p\left[\left(\tfrac{k}{n}\right)^{\frac{1}{2}}\right] \quad \textit{if } k^{(1+\frac{m}{4})}/n \to 0, k^3/n \to \infty, n \to \infty.$$

(ii) There exists a polynomial in m of order $\frac{1}{k}$, $P(m, L, k)$, such that

$$\hat{m}_k - m - P(m, L, k) = \Omega_p\Big[\big(\tfrac{k}{n}\big)^{\frac{1}{2}}\Big]$$

if $k^{(1+\frac{m}{4})}/n \to 0, k^{(2L+1)}/n \to \infty, n \to \infty$.

Here $A = \Omega_p(B)$ means $A = O_p(B)$ and $B = O_p(A)$.

DISCUSSION.

1. This result has the same unsatisfactory aspects as Theorem 4.1 in that it is not too instructive for integer dimensions. Although there are well-established notions of fractal dimensions for sets, we do not know how our estimate will behave.
2. Levina and Bickel (2005) had conjectured that $\hat{m}_k - m$ was asymptotically normal with variance of order n^{-1} for suitable k_n. This conjecture appears to be true only for k bounded. We sketch a proof after that of Theorem 4.2. Asymptotic normality may hold at scale $(\frac{k}{n})^{\frac{1}{2}}$ in general but we have not shown this.
3. The same lack of guidance on k holds although we may obviously adopt our local variance based estimate to the global case and apply the same principle as that of Theorem 4.1.

We finally state

THEOREM 4.3. *If $k \le K < \infty$ for all K and the condition on $\boldsymbol{T}$ and f of Theorem 4.1 hold, then*

$$\sqrt{n}\Big(\hat{m}_k - \frac{k-1}{k-2}m\Big) \Rightarrow \mathcal{N}(0, \sigma^2(m)).$$

DISCUSSION

1. As we indicated earlier, we really would like large deviation theorems. Although asymptotic normality does not establish this it at least suggests that global integer dimensions can be established with probability going to 0 exponentially in n. We conjecture that it is not too hard to prove exponential in k convergence for local dimension.

2. In those cases, as in dynamical systems, where it seems plausible that observations have global fractal dimension, we conjecture that our methods should be fruitful if our definition of local definition is replaced by local ball covering dimension - e.g., partition a k-th nearest neighbor d-cube of side L into ρ^{-d} cubes of side ρL. For each little cube B_j, let $\hat{\rho}_j$ be the fraction of the k nearest neighbors contained in this cube. Now estimate $\hat{m} \triangleq \sum \log \hat{\rho}_j / \log(\rho L)$ and let $k \to \infty, k/n \to 0, \rho \to 0$ sufficiently slowly that $\rho k \to \infty$. Unfortunately we expect that the speed of convergence of such an estimate to be $(\rho k)^{-\frac{1}{2}}$ so that large number of observations would be needed.

PROOF OF THEOREM 4.2. Consider the expansion,

$$\hat{m}_k(\boldsymbol{X}) - m - \frac{W(\boldsymbol{X}) - \mu}{\mu^2} = \sum_{j=2}^{2L-1} \frac{(W(\boldsymbol{X}) - \mu)^j (-1)^j}{\mu^{j+1}} + \frac{(W(\boldsymbol{X}) - \mu)^{2L}}{\mu^{2L+1}} \tag{4.10}$$

To prove Theorem 4.2 for the uniform case, we will argue for (i) that the expectation of the first term on the right of (4.10) for $L = 1$ is $O(\frac{1}{k})$ and then compute the variance of $\hat{m}_{k_0}^{-1} \triangleq \frac{1}{n} \sum_{i=1}^{n} W(\boldsymbol{X}_i)$, and show that

$$\tfrac{n}{k} Var(\hat{m}_{k_0}^{-1}) \to \sigma^2(m) > 0.$$

We can then conclude that (i) holds for all m and k such that $\frac{k^3}{n} \to \infty$.

The other requirement on k comes when we do not have a uniform distribution.

More generally, for part (ii), we will argue that the right centering is

$$P(m, L, k) = \sum_{j=2}^{2L-1} \frac{\mathbb{E}(W(\boldsymbol{X}_1) - \mu)^j}{\mu^{j+1}}.$$

Note that $P(m, L, k) = O(k^{-1})$ as claimed by using the central moments of the Gamma distribution.

We begin our argument for part (i) with the fundamental

PROPOSITION 4.1. *Suppose Θ is bounded and f is uniform on Θ and $\boldsymbol{T}$ is the identity. Then*

$$kVar(W(\boldsymbol{X})) = \tfrac{1}{m^2}(1 + o(1)) \tag{4.11}$$

$$\tfrac{n}{k} Cov(W(\boldsymbol{X}_1), W(\boldsymbol{X}_2)) = \sigma^2(m)(1 + o(1)). \tag{4.12}$$

We first prove (4.11). (4.12) requires a series of lemmas. For (4.11), as we noted before, $n(\frac{R_j}{R_k})^m, 1 \leq j \leq k-1$ are distributed as the order statistics of a sample of size $k-1$ from $\mathcal{U}(0,1)$. Thus $(k-1)W(\boldsymbol{X})$ has a Gamma distribution with parameters $k-1$ and m and (4.11) is immediate. For (4.12), we will use the following standard formula. Let $\boldsymbol{X}_1, \boldsymbol{X}_2$ be random variables and $\mathbf{Y}$ be a random vector. Then

$$Cov(\boldsymbol{X}_1, \boldsymbol{X}_2) = \mathbb{E}[Cov(\boldsymbol{X}_1, \boldsymbol{X}_2)|\mathbf{Y}] + Cov((\mathbb{E}\boldsymbol{X}_1|\mathbf{Y}), (\mathbb{E}\boldsymbol{X}_2|\mathbf{Y})).$$

Let $\mathbb{A}$ denote the event $(\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, R_k(\boldsymbol{X}_1) = r_1, R_k(\boldsymbol{X}_2) = r_2)$ and let

$$\begin{aligned} A_1 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2) : \ S(\boldsymbol{x}_2, r_2) \cap S(\boldsymbol{x}_1, r_1) = \emptyset\} \\ A_2 &= \{(\boldsymbol{x}_1, \boldsymbol{x}_2) : \ S(\boldsymbol{x}_2, r_2) \cap S(\boldsymbol{x}_1, r_1) \neq \emptyset\}. \end{aligned}$$

Write

$$\begin{aligned} Cov(W(\boldsymbol{X}_1), W(\boldsymbol{X}_2)) &= \mathbb{E}[Cov(W(\boldsymbol{X}_1), W(\boldsymbol{X}_2))|\mathbb{A}] \\ &+ Cov(\mathbb{E}W(\boldsymbol{X}_1)|\mathbb{A}, \mathbb{E}W(\boldsymbol{X}_2)|\mathbb{A}). \end{aligned}$$

Write the first term as

$$\begin{aligned} &\mathbb{E}[Cov(W(\boldsymbol{X}_1), W(\boldsymbol{X}_2)).\mathbf{1}_{A_1}|\mathbb{A}] + \mathbb{E}[Cov(W(\boldsymbol{X}_1), W(\boldsymbol{X}_2)).\mathbf{1}_{A_2}|\mathbb{A}] \\ = \ & I_1 + I_2. \end{aligned}$$

Denote $c = \frac{\pi^{m/2}}{\Gamma[m/2+1]}$, the volume of a unit sphere in $\mathbb{R}^m$.

LEMMA 4.2. $I_1 = 0$.

PROOF. Given $\mathbb{A}$, we can write

$$W(\boldsymbol{X}_1) =_d \frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{r_1}{\| \boldsymbol{V}_i^1 - \boldsymbol{x}_1 \|} \tag{4.13}$$

$$W(\boldsymbol{X}_2) =_d \frac{1}{k-1} \sum_{i=1}^{k-1} \log \frac{r_2}{\| \boldsymbol{V}_i^2 - \boldsymbol{x}_2 \|} \tag{4.14}$$

where $=_d$ denotes equal in distribution, for $\boldsymbol{V}_1^1, ..., \boldsymbol{V}_{k-1}^1$ and $\boldsymbol{V}_1^2, ..., \boldsymbol{V}_{k-1}^2$ i.i.d. uniform on spheres $S(\boldsymbol{x}_1, r_1)$ and $S(\boldsymbol{x}_2, r_2)$ respectively. Since the two

sequences of random variables $\boldsymbol{V}_1^1, ..., \boldsymbol{V}_{k-1}^1$ and $\boldsymbol{V}_1^2, ..., \boldsymbol{V}_{k-1}^2$ are conditionally independent on set A_1,

$$\begin{aligned} & Cov(W(\boldsymbol{X}_1), W(\boldsymbol{X}_2)).\mathbf{1}_{A_1}|\mathbb{A} \\ = \; & \mathbb{E}\big[\big(W(\boldsymbol{X}_1) - \mathbb{E}W(\boldsymbol{X}_1)|\mathbb{A}\big)\big(W(\boldsymbol{X}_2) - \mathbb{E}W(\boldsymbol{X}_2)|\mathbb{A}\big).\mathbf{1}_{A_1} \mid \mathbb{A}\big] \\ = \; & \mathbb{E}\Big[\Big(\frac{1}{k-1}\sum_{i=1}^{k-1}\log\frac{r_1}{\parallel \boldsymbol{V}_i^1 - \boldsymbol{x}_1 \parallel} - \mu\Big).\mathbf{1}_{A_1}|\mathbb{A}\Big]. \\ & \mathbb{E}\Big[\Big(\frac{1}{k-1}\sum_{i=1}^{k-1}\log\frac{r_2}{\parallel \boldsymbol{V}_i^2 - \boldsymbol{x}_2 \parallel} - \mu\Big).\mathbf{1}_{A_1}|\mathbb{A}\Big] \\ = \; & 0. \end{aligned}$$

Therefore $I_1 = \mathbb{E}[Cov(W(\boldsymbol{X}_1), W(\boldsymbol{X}_2)).\mathbf{1}_{A_1}|\mathbb{A})] = 0$. □

Consider the case A_2, the two spheres $S(\boldsymbol{x}_1, r_1)$ and $S(\boldsymbol{x}_2, r_2)$ intersect (see Figure 1 below).

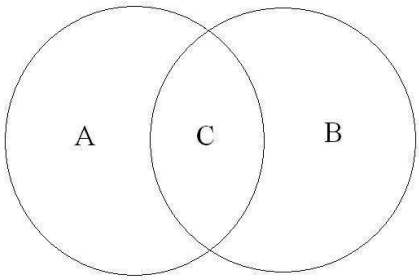

Figure 1: *Given* $\boldsymbol{X}_1 = \boldsymbol{x}_1, \boldsymbol{X}_2 = \boldsymbol{x}_2, R_k(\boldsymbol{X}_1) = r_1, R_k(\boldsymbol{X}_2) = r_2$, *the two spheres* $S(\boldsymbol{X}_1, R_k(\boldsymbol{X}_1))$ *and* $S(\boldsymbol{X}_2, R_k(\boldsymbol{X}_2))$ *intersect.*

LEMMA 4.3. *Let* $g_m(r_1, r_2, \Delta)$ *be the volume of the intersection of two spheres* $S(\mathbf{0}, r_1)$ *and* $S(\boldsymbol{x}, r_2)$ *where* $r_1 \geq r_2$. *Let*

$$\Delta = |\boldsymbol{x}|, \;\; \rho = \frac{r_1}{r_1 + r_2} \geq \frac{1}{2}, \;\; \gamma = \frac{\Delta}{r_1 + r_2}.$$

Then
(i) $g_m(r_1, r_2, \Delta) = d(m)(r_1 + r_2)^m h_m(\rho, \gamma)$, *for* $d(m)$ *a universal constant, on the set*

$$\mathcal{C} \triangleq \{(\rho, \gamma) : \; 2\rho - 1 \leq \gamma \leq 1\}$$

and
(ii)

$$\begin{cases} h_m = 0, & \text{for } \gamma \geq 1 \\ h_m = (1-\rho)^m c_m, & \text{for } \gamma \leq 2\rho - 1 \end{cases}$$

since $\gamma > 1$ is equivalent to $\Delta \geq r_1 + r_2$ which means that the spheres are tangent or disjoint, while $\gamma \leq 2\rho - 1$ is equivalent to $\Delta + r_2 \leq r_1$ which means that the smaller sphere is contained in the larger.

PROOF. By definition

$$g_m(r_1, r_2, \Delta) = \int ... \int_{S(0,r_1) \cap S(\boldsymbol{x}, r2)} d\boldsymbol{x}.$$

Without loss of generality assume $\boldsymbol{x} = (\Delta, 0, ..., 0)$ and change to polar coordinate $(r, \varphi_1, ..., \varphi_{m-1})$

$$\begin{cases} x_1 = r \sin\varphi_1 ... \sin\varphi_{m-2} \cos\varphi_{m-1} \\ x_2 = r \sin\varphi_1 ... \sin\varphi_{m-2} \sin\varphi_{m-1} \\ \quad ..... \\ x_{m-1} = r \sin\varphi_1 \cos\varphi_2 \\ x_m = r \cos\varphi_1 \end{cases}$$

where $(r, \varphi_1, ..., \varphi_{m-1}) \in [0, \infty) \times [0, \pi) \times [0, \pi) \times [0, 2\pi)$. Then

$$g_m(r_1, r_2, \Delta) \quad = \quad \int_{\mathcal{T}} r^{m-1} \prod_{i=1}^{m-2} (\sin\varphi_i)^{m-i-1} dr d\boldsymbol{\varphi}$$

where

$$\mathcal{T} = \left\{ r^2 \leq r_1^2,\ r^2 - 2\Delta r \prod_{i=1}^{m-2} (\sin\varphi_i)^{m-i-1} + \Delta^2 \leq r_2^2 \right\}.$$

Change variables to

$$v = \frac{r}{r_1 + r_2}.$$

Then

$$g_m(r_1, r_2, \Delta) = (r_1 + r_2)^m \int_{\mathcal{U}} v^{m-1} \prod_{i=1}^{m-2} (\sin\varphi_i)^{m-i-1} dv d\boldsymbol{\varphi}$$

where

$$\mathcal{U} = \left\{ v^2 \leq \rho^2,\ v^2 - 2\gamma v \prod_{i=1}^{m-2} (\sin\varphi_i)^{m-i-1} + \gamma^2 \leq (1-\rho)^2 \right\}.$$

The result follows. □

We continue with the proof. Let

$$Z_j = cnR_j^m/k,\ j = 1, 2$$
$$\Delta = |\boldsymbol{X}_1 - \boldsymbol{X}_2|.$$

Let $P_{v,t,d}$ be the conditional distribution of the data given, $Z_1 = t, Z_2 = v, \Delta = (\frac{k}{nc}d)^{\frac{1}{m}} \triangleq \Delta^0$. Let

$$N_{12}(r_1, r_2) = \sum_{i \neq 1,2} \mathbf{1}\left(\boldsymbol{X}_i \in S(\boldsymbol{X}_1, r_1) \cap S(\boldsymbol{X}_2, r_2)\right)$$
$$\bar{N}_{12}(r_1, r_2) = \sum_{i \neq 1,2} \mathbf{1}\left(\boldsymbol{X}_i \in S(\boldsymbol{X}_2, r_2) \cap \bar{S}(\boldsymbol{X}_1, r_1)\right)$$
$$\bar{N}_{21}(r_1, r_2) = \sum_{i \neq 1,2} \mathbf{1}\left(\boldsymbol{X}_i \in S(\boldsymbol{X}_1, r_1) \cap \bar{S}(\boldsymbol{X}_2, r_2)\right).$$

where $\bar{S}$ denotes complement. Note that the distribution of $N_{12}, \bar{N}_{12}, \bar{N}_{21}$ depend on Δ as well as R_1, R_2.

LEMMA 4.4. *Under $P_{v,t,d}$,*

$$\frac{N_{12}(R_1, R_2) - kp(v,d)}{\sqrt{kp(1-p)}} \text{ and } \frac{\bar{N}_{12}(R_1, R_2) - kvt}{\sqrt{kvt}} \text{ and } \frac{\bar{N}_{21}(R_1, R_2) - kt}{\sqrt{kt}}$$

are asymptotically independent $\mathcal{N}(0,1)$. Here

$$p \triangleq p(v,d) = Fg_m(r_1^0, r_2^0, \Delta) = Fd(m)h_m\left(\tfrac{1}{2}, d^{\frac{1}{m}}\right)(1 + o(1))$$

when $r_1^0 = [kt/(cn)]^{\frac{1}{m}}, r_2^0 = [kvt/(cn)]^{\frac{1}{m}}$.

PROOF. Without loss of generality we can assume $\mathbf{X}_1 = \mathbf{0}$.

Now given $\boldsymbol{X} = \mathbf{0}$ and r_1^0 as above, the $k-1$ points in the interior of $S(\mathbf{0}, r_1^0)$ other than $\boldsymbol{X}_2$ are distributed uniformly and independently. Therefore conditionally $N_{12}(r_1^0, r_2^0)$ is Binomial$(k-2, \pi_{12})$ where

$$\pi_{12} = \frac{\text{Volume of } \left[S\left(\boldsymbol{X}_2, \left(vt\frac{k}{n}c^{-1}\right)^{\frac{1}{m}}\right) \cap S(\mathbf{0}, r_1^0)\right]}{\frac{k}{n}t}$$

where the denominator is the volume of $S(\mathbf{0}, r_1^0)$. By Lemma 4.3, $\pi_{12} = Fg_m(r_1^0, r_2^0, \Delta^0)$.

Further, $\bar{N}_{12}$ is independent of N_{12} and follows a Binomial $\left(n-k-2, \frac{k}{n}vt + o(1)\right)$ distribution. Thus

$$\frac{N_{12}(R_1, R_2) - kp(v,d)}{\sqrt{kp(1-p)}} \text{ and } \frac{\bar{N}_{12}(R_1, R_2) - kv}{\sqrt{kv}}$$

are asymptotically independent $\mathcal{N}(0,1)$. The same argument applies to $\bar{N}_{21}$. □

Let $v = v_0(d)$ be the unique solution of

$$v + p(v,d) = 1.$$

The solution exists, since $p(v,d)$ increases strictly from 0 to 1 as v increases, by the definition of π_{12}. Call the solution $v_0(d)$.

LEMMA 4.5. *Let $W(\mathbf{X}_1)$ and $W(\mathbf{X}_2)$ be as defined in (4.13) and (4.14). Let $A(d)$ be the event $|\mathbf{X}_1 - \mathbf{X}_2| = \Delta^0$. Let $D = c\frac{n}{k}|\mathbf{X}_1 - \mathbf{X}_2|^m$. Then,*

$$\begin{aligned} & n.\mathbb{E}[Cov(W(\mathbf{X}_1), W(\mathbf{X}_2)) \mid A(d)] \\ = \;& F.\mathbb{E}[h_m(\tfrac{1}{2}, D)\kappa_0(v_0(D))](1 + o(1)) \end{aligned} \tag{4.15}$$

where D is uniformly distributed on $(0,1)$ and $\kappa(v_0(d))$ is the covariance of $(\log \boldsymbol{A}_1, \log \boldsymbol{A}_2)$ where $\boldsymbol{A}_1 = |\boldsymbol{X}|$ and $\boldsymbol{A}_2 = |\boldsymbol{X} - \boldsymbol{\Delta}|$ and $\boldsymbol{X}$ is uniform on the intersection of $S(0,1)$ and $S(\boldsymbol{\Delta}, v_0(d))$ with $|\boldsymbol{\Delta}| = d^{\frac{1}{m}}$.

PROOF. Note that,

$$\begin{aligned} & Cov\left((W(\boldsymbol{X}_1), W(\boldsymbol{X}_2)) \mid \Delta = \Delta^0, Z_1 = t, Z_2 = v, N_{12}\right) \\ = \;& k^{-2} N_{12} Cov(\log A_1, \log A_2) \end{aligned} \tag{4.16}$$

Since given $A(s)$, if V_i^1 is in $S(\mathbf{X}_1, r_1^0) \cap \bar{S}(\mathbf{X}_2, r_2^0)$ and V_i^2 in $S(\mathbf{X}_2, r_2^0)$ then the points are independent and the same holds if $(\mathbf{X}_1, r_1^0)$ and $(\mathbf{X}_2, r_2^0)$ are interchanged, and given membership in $S(\mathbf{X}_1, r_1) \cap S(\mathbf{X}_2, r_2)$ and (Δ, R_1, R_2). The variables clearly have the identical distribution.

Note that

$$\mathbb{E}(N_{12} \mid A(d)) = kp(v,d) = kFd(m)h_m\left(\tfrac{1}{2}, d^{\frac{1}{m}}\right)(1 + o(1)) \tag{4.17}$$

by Lemma 4.4.

Finally note that

$$P\left[|\boldsymbol{X}_2 - \boldsymbol{X}_1| \leq \left(dc^{-1}\frac{k}{n}\right)^{\frac{1}{m}}\right] = d\frac{k}{n}, \quad 0 \leq d^m < 2. \tag{4.18}$$

Since for $d^m > 2, \pi_{12} = 0$, we conclude that (4.15) holds provided that the variables defined by (4.15) are uniformly integrable. To see this note that $\mathbb{E}[|W(\mathbf{X}_1)|^4 \mid D]$ is clearly bounded since D only provides information about one point in $S(\mathbf{X}_1, R_1)$. □

LEMMA 4.6. *Under the assumptions of Lemma 4.5,*

$$Cov[\mathbb{E}(W(\mathbf{X}_1) \mid A(D_m)), \mathbb{E}(W(\mathbf{X}_2) \mid A(D_m))] = \lambda(m)\frac{k}{n}(1 + o(1))$$

for a suitable $\lambda(m)$.

PROOF. Let $A(d, t, v)$ be the event

$$Z_1 = t, \quad Z_2 = vt, \quad |\mathbf{X}_1 - \mathbf{X}_2| = \left(\frac{k}{nc}d\right)^{\frac{1}{m}}.$$

Then,

$$\begin{aligned} & \mathbb{E}[W(\mathbf{X}_1) \mid A(d, t, v), N_{12}, \bar{N}_{12}, \bar{N}_{21}] \\ = \quad & \frac{\bar{N}_{12}}{k}\mathbb{E}[-\log Q_1(d)] + \frac{N_{12}}{k}\mathbb{E}[-\log Q_2(d)] + \frac{\bar{N}_{21}}{k}\mathbb{E}[-\log Q_3(d)] \end{aligned}$$

where Q_1 is uniform over $S(\mathbf{0}, 1) \cap \bar{S}(\mathbf{\Delta}, v)$ with $|\mathbf{\Delta}| = d^{\frac{1}{m}}$, Q_2 is uniform over $S(\mathbf{0}, 1) \cap S(\mathbf{\Delta}, v)$, and Q_3 is uniform over $\bar{S}(\mathbf{0}, 1) \cap S(\mathbf{\Delta}, v)$.

$\mathbb{E}(W(\mathbf{X}_2) \mid A(d, t, v))$ has the same expression as $\mathbb{E}(W(\mathbf{X}_1) \mid A(d, t, v))$ but with $\bar{N}_{12}$ and $\bar{N}_{21}$ switched. By Lemma 4.4

$$\frac{N_{12}}{k} \to_p p(d, v(d)), \quad \frac{\bar{N}_{12}}{k} \to_p v(d), \quad \frac{\bar{N}_{12}}{k} \to_p v(d).$$

Hence, using uniform integrability as before,

$$\mathbb{E}[W(\mathbf{X}_1) \mid A(D_m)] = f(D_m)(1 + o_p(1)) = \mathbb{E}[W(\mathbf{X}_2) \mid A(D_m)].$$

Thus

$$\begin{aligned} & \mathbb{E}Cov[\mathbb{E}(W(\mathbf{X}_1) \mid A(D_m)), \mathbb{E}(W(\mathbf{X}_2) \mid A(D_m))].\mathbf{1}_{A(D_m)} \\ = \quad & \mathbb{E}[Var(f(D_m)(1 + op(1))).\mathbf{1}_{A(D_m)}] \\ = \quad & \tfrac{k}{n}\lambda(m)(1 + o(1)) \end{aligned}$$

again using (4.18) and the fact that f is continuous and $f(d) = 0, d \geq 2$. □

A similar calculation applies to the variances of the individual terms in (4.10). This completes the proof of (ii) for the case $\boldsymbol{T}$ is identity on the first m coordinate vectors. To show that the results remain valid if $\boldsymbol{T}$ is as specified and $k^{(1+\frac{m}{4})}/n \to 0$, we simply apply the bounds of (4.5), (4.6) to the joint density of $(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X})$ when computing covariances. For instance, if $B(s) = \big\{|\boldsymbol{T}(\boldsymbol{U}_1) - \boldsymbol{T}(\boldsymbol{U}_2)| \le (c^{-1}s\frac{k}{n})^{\frac{1}{m}}\big\}$

$$P\Big[|\boldsymbol{T}(\boldsymbol{U}) - \boldsymbol{T}(\boldsymbol{U}_1)| \le (c^{-1}\tfrac{k}{n})^{\frac{1}{m}}, |\boldsymbol{T}(\boldsymbol{U}) - \boldsymbol{T}(\boldsymbol{U}_2)| \le (c^{-1}v_0(s)\tfrac{k}{n})^{\frac{1}{m}} \;\Big|\; B(s)$$

$$= \quad \pi(s, v_0(s))\Big(1 + O\big[(\tfrac{k}{n})^{\frac{2}{m}}\big]\Big)$$

since we can replace $|\boldsymbol{T}(\boldsymbol{U}) - \boldsymbol{T}(\boldsymbol{U}_j)|$ by $|(\boldsymbol{T}(\boldsymbol{U}) - \boldsymbol{x}_0) - (\boldsymbol{T}(\boldsymbol{U}_j) - \boldsymbol{x}_0)|$. The theorem follows. □

Finally we sketch the proof of Theorem 4.3.

PROOF OF THEOREM 3. We note that

$$\hat{m}_k(\boldsymbol{X}_i) = h(\boldsymbol{X}_i : \; N_k(\boldsymbol{X}_i))$$

where $N_k(\boldsymbol{X}_i)$ is the set of neighbors up to the k-th of $\boldsymbol{X}_i$. Therefore we can apply Theorem 3.4 of Chatterjee (2006) to establish Theorem 4.3 provided we can show that

$$\mathbb{E}(\hat{m}_k) \quad = \quad m\left(\frac{k-1}{k-2}\right) \quad \text{and} \tag{4.19}$$

$$Var(\hat{m}_k) \quad \to \quad \sigma^2(m). \tag{4.20}$$

Since $\Big[\frac{1}{(k-1)m}\hat{m}_k(\boldsymbol{X}_i)\Big]^{-1}$ has a Gamma$(k-1, 1)$ distribution, (4.19) is immediate.

We can similarly compute

$$\begin{aligned} Var(\hat{m}_k(\boldsymbol{X}_i)) &= m^2\Big[\frac{(k-1)^2}{(k-2)(k-3)} - \frac{(k-1)^2}{(k-2)^2}\Big] \\ &= m^2\frac{(k-1)^2}{(k-2)^2}\frac{1}{k-3}. \end{aligned}$$

We are left to check that,

$$nCov(\hat{m}_k(\boldsymbol{X}_1), \hat{m}_k(\boldsymbol{X}_2)) \sim \gamma(m, k)$$

for suitable γ. Decomposing as before it is clear we need only be concerned with

$$\mathbb{E}\big[U(\boldsymbol{X}_1)U(\boldsymbol{X}_2)\ \big|A(s)\big] \quad \text{and} \quad \mathbb{E}\big[U(\boldsymbol{X}_1)\ \big|A(s)\big] \tag{4.21}$$

Given A_2 and $A(s)$, $\boldsymbol{X}_2 = \boldsymbol{X}_1 + (\frac{s}{c}\frac{k}{n})^{\frac{1}{m}}\boldsymbol{V}$ where $\boldsymbol{V}$ has a uniform distribution on $S(\mathbf{0}, 2)$. By Lemma 4.3, $\pi(s, v(s)) = Fh_m(v, s)$. If we now define

$$\boldsymbol{W} = \boldsymbol{Z}\left(\frac{s}{c}\frac{k}{n}\right)^{-\frac{1}{m}}$$

for $\boldsymbol{Z}$ uniform on $S(\mathbf{0}, r_1) \cap S(\mathbf{1}, r_2)$, then $\boldsymbol{W}$ has a distribution which depends on s only.

Using this rescaling, the quantities in (4.21) depend only on the distribution of $\boldsymbol{V}$ and $N_k(r_2)$. Therefore for k bounded, the quantities in (4.21) depend on s and not on n to first order. But

$$P\big[\cup\{A(t):\ t \le s\}\big] = \frac{s}{2}\frac{k}{n}$$

for $0 \le s \le 2$. From

$$nVar(\hat{m}_k) = Var(\hat{m}_k(\boldsymbol{X}_1)) + 2(n-1)Cov\big[\hat{m}_k(\boldsymbol{X}_1), \hat{m}_k(\boldsymbol{X}_2)\big],$$

the theorem follows in the uniform case. The general case is argued as in Theorem 4.2 again taking advantage of the boundedness of k. □

References

Abramovich, F., Benjamini, Y., Donoho, D.L., Johnstone, I.M.(2000). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, **34**, 584-653.

Atkinson, A.C. (1980). A note on the generalized information criterion for choice of a model. *Biometrika*, **67**, 413-418.

Bickel, P.J. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, **10**, 989-1010.

Bickel, P.J. and Levina, E. (2008)(i). Covariance Regularization by Thresholding. *Ann. Statist.* To appear.

Bickel, P.J. and Levina, E. (2008)(ii). Regularized Estimation of Large Covariance Matrices. *Ann. Statist.*, **36**, 199-227.

Bickel, P.J. and Li, B. (2006). Regularization in Statistics. *Test*, **15**, 271-344.

Bickel, P.J. and Li, B. (2007). Local polynomial regression on unknown manifolds. *IMS Lecture Notes*, **54.**

BOX, G.E.P. (1979). *Robustness in the strategy of scientific model building, in Robustness in Statistics*, R.L. Launer and G.N. Wilkinson, Editors. Academic Press, New York.

BREIMAN, L. and FRIEDMAN, J. H. (1985). Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion). *J. Amer. Statist. Assoc.*, **80**, 580-619.

CANDES, E. J., DONOHO, D.L. (2004). New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Comm. Pure Appl. Math.*, **57**, 219-266.

CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n (with discussion). *Ann. Statist.*, **35**, 2313-2351.

CHATTERJEE, S. (2006). A new method of normal approximation. *Ann. Probab.* To appear.

CHEN, S.S., DONOHO, D.L., SAUNDERS, M.A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20**, 33-61.

D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M.I., LANCKRIET, G.R.G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.*. **49**, 434-448.

DONOHO, D.L., ELAD, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci.*, USA, **100**, 2197-2202.

DONOHO, D.L., JOHNSTONE, I.M. (1994). Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sci. Paris Sr. I Math.*, **319**, 1317-1322.

DONOHO, D.L., JOHNSTONE, I.M. (1994). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Related Fields*, **99**, 277-303.

DONOHO, D.L., JOHNSTONE, I.M., PICARD, D. and KERKYACHARIAN, G. (1995). Wavelet shrinkage: asymptotia? (with discussion). *Journal of Royal Statistical Society (B)*, 301-369.

EL KAROUI, N. (2007). Spectrum estimation for large dimensional covariance matrices using random matrix theory. *Ann. Statist.* To appear.

EL KAROUI, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* To appear.

FALCONER, K. (1990). *Fractal Geometry, Mathematical Foundations and Applications*. Wiley.

FAN, J. and LV, J. (2008). Sure independence screening for ultra-high dimensional feature space (with discussion). *J. Roy Statist. Soc. (B)*. To appear.

HASTIE, T., TIBSHIRANI, R and FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

JOHNSTONE, I.M. and LU, A.Y. (2008). Sparse principal components analysis. *J. Amer. Statist. Assoc.* To appear.

LEVINA, E. and BICKEL, P.J. (2005). Maximum likelihood estimation of the intrinsic dimension. *Advances in NIPS 17*.

NIYOGI, P. (2007). Manifold Regularization and Semi-supervised Learning: Some Theoretical Analyses. *Technical Report, Computer Science Dept., University of Chicago*.

STONE, C.J. (1977). Consistent nonparametric regression (with discussion). *Ann. Statist.*, **5**, 595-645.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? *Biometrika*, **92**, 937-950.

Yukich, J. (2008). Point process stabilization methods and dimension estimation. *Discrete Mathematics and Theoretical Computer Science*, Nancy, France.

Peter J. Bickel
Department of Statistics
University of California, Berkeley
CA 94720, USA
E-mail: bickel@stat.berkeley.edu

Donghui Yan
Department of Statistics
University of California, Berkeley
CA 94720, USA
E-mail: dhyan@stat.berkeley.edu

Paper received December 2008; revised January 2009.