On The Limits of Clustering in High Dimensions via Cost Functions

Hoyt A. Koepke* & Bertrand S. Clarke†

September 29, 2010

Abstract

This paper establishes a negative result for clustering: above a certain ratio of random noise to non-random information, it is impossible for a large class of cost functions to distinguish between two partitions of a data set. In particular, it is shown that as the dimension increases, the ability to distinguish an accurate partitioning from an inaccurate one is lost unless the informative components are both sufficiently numerous and sufficiently informative. We examine squared error cost functions in detail. More generally, it is seen that the VC-dimension is an essential hypothesis for the class of cost functions to satisfy for an impossibility proof to be feasible. Separately, we provide bounds on the probabilistic behavior of cost functions that show how rapidly the ability to distinguish two clusterings decays. In two examples, one simulated and one with genomic data, bounds on the ability of squared-error and other cost functions to distinguish between two partitions are computed. Thus, one should not rely on clustering results alone for high dimensional low sample size data and one should do feature selection.

Key Words: Clustering impossibility, High dimensions, cost function, VC-dimension

Running Title: Limits of Clustering in High Dimensions

*Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, p/f 206-685-7431/206-685-7419 hoytak@stat.washington.edu H. Koepke was partially supported by the NSERC operating grant held by the second author.

†Dept. of Medicine, DEPH, and CCS, University of Miami, 1120 NW 14th St. CRB 1055 (C-213) Miami, FL, 33136, p/f 305-243-5457/305-243-9304bclarke2@miami.edu

1 INTRODUCTION

A number of recently emerged data types have dimensions that are beyond the usual scale of conventional statistical techniques. In analyzing such data, it is generally understood that one achieves better results if irrelevant features are discarded, and numerous recent results outline ways to do this. The purpose of this paper is to present a theoretical bound on what can be learned from using *any* technique that evaluates *any* clustering by way of a large class of cost functions that includes squared error. Our findings apply to the results of any clustering technique including partial and two-way clustering.

For our investigation, we assume a standard signal plus noise model

$$\mathbf{Y} = \mathbf{x} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{Y}$, $\mathbf{x}$, and $\boldsymbol{\varepsilon}$ are $D \times n$ dimensional matrices. The D-dimensional data points in the columns of $\mathbf{Y}$ come from n non-random but unknown D-dimensional columns $\mathbf{x}_i$ of $\mathbf{x}$ plus a column from the random noise matrix $\boldsymbol{\varepsilon}$; the entries in $\mathbf{Y}$ are the only values in (1) that are available to the experimenter. Note that the $\mathbf{x}_i$'s are non-stochastic and represent the centroids of clusters or of regions (for which the centroid is the mode). In a clustering context, each column in $\mathbf{x}$ may represent a cluster center associated with each point (so centroids are included with multiplicity). Here we are concerned with the setting that data of this form is high dimensional but low sample size, i.e. large D and small n. Specifically, we assume $D \gg n$ so that standard regression techniques cannot be applied without substantial modification.

It is common to search for useful structure in (1) by using clustering techniques on the columns or rows

of $\mathbf{Y}$. One can cluster over samples, i.e., over n vectors of length D, to find relationships among subjects. Alternatively, one can cluster over variables, i.e., over D vectors of length n, when the goal is to find relationships among the candidate explanatory variables. Here we focus on the first case, clustering over samples, since that is the primary goal in many applications. In this context we show an impossibility result: evaluating different clusterings by a squared error cost function is *only* possible when the sum of squared distances between certain of the explanatory variable values for the i-th subject, the $\mathbf{x}_i$'s, determined by the clusterings has a rate above $\sqrt{D}$ as the dimension D increases, provided the noise terms are well behaved. However, if this rate is smaller than $\sqrt{D}$, meaningful clustering is not possible in the sense that any orderings over clusterings, in terms of squared error loss, is indistinguishable from random. The principle extends to cost functions derived from other L^p norms but has extra complexities we discuss below.

While it is intuitively plausible that clustering of noisy high-dimensional data can lead to spurious clusterings some practitioners seem unaware of the possibility; see [1] and [2], among others. Indeed, our simulations will show clustering starts to have very appreciable probability of spuriousness in relatively benign settings. For instance, one simulation (see Fig. 1(e)) shows that, for $n = 120$ and 2 informative dimensions, by the time there are 20 to 30 variables the probability of distinguishing a good clustering from a bad one can fall to .7 or less in squared error. In a genomic data set from [3], we show that when there are 5 informative variables among approximately $D = 1000$ variables, the probability that a good clustering, i.e. one physically meaningful, is distinguishable from a poor clustering falls to around .6 for n in the low 30's. We suggest that these cases are representative enough of many settings to suggest that clustering results from sparse data, i.e., $D >> n$ and few important variables, should be regarded as unreliable in the absence of further analysis.

To introduce our reasoning, recognize that a clustering is just a partition of the data points into disjoint subsets. So, consider *any* two partitions $\mathcal{P}$ and $\mathcal{Q}$ of the n data points (one of these may be the optimal, though this is not a requirement). That is, suppose $\mathcal{P}$ is of the form $\{P_1, \ldots, P_K\}$ where the P_k's are disjoint, non-empty, exhaustive subsets of $\{1, 2, ..., n\}$ and $\mathcal{Q} = \{Q_1, \ldots, Q_K\}$ is similar. Suppose the squared error cost function is assigned so that the cost of the partition $\mathcal{P}$, i.e., the clustering, is $\mathsf{cost}(\mathbf{Y}, \mathcal{P})$. Then, the costs of the partitions can be compared.

For fixed n and D, let the partition $\mathcal{P}$ be fixed and regarded as a partition of the whole space. That is, a new point is in P_k if it is closer to the mean of the points in P_k than to the mean of any other $P_{k'}$. Now regard $\mathcal{Q}$ as the analogous partition generated by a new set of n data points. Then, the principle quantity of interest is the probability that one partition has a lower cost than another given partition if the random noise components are redrawn, i.e.,

$$\xi_D = \mathrm{P}(\mathsf{cost}(\mathbf{Y}, \mathcal{P}) \leq \mathsf{cost}(\mathbf{Y}, \mathcal{Q})). \qquad (2)$$

Note that in (2), we are treating cost functions as random variables rather than taking expectations and obtaining a formal decision-theoretic risk as in [4] and [5]. If ξ_D is near 0 or 1, the cost function can reliably determine that $\mathcal{P}$ is much worse or much better, respectively, than $\mathcal{Q}$. However, if $\xi_D \simeq 1/2$, the result is no better than random. This paper gives conditions under which $\xi_D \to 1/2$ as $D \to \infty$, and investigates the rate at which this occurs. We emphasize that (i) our key theorems are asymptotic in D for fixed n for which relatively few results seem to exist, and that (ii) the expression (2) is comparative in the sense that no properties of $\mathcal{P}$ or $\mathcal{Q}$ are assumed; as partitions, either may be good, bad or inbetween.

One way to interpret ξ_D is as follows. Suppose one sample is drawn from a random process described by equation (1). Consider two partitions of the data $\mathcal{P}$ and $\mathcal{Q}$, and rank the goodness of the partitions using the squared error cost. The squared error cost, (see Def. 2.2 below) is essentially the 'within sum of squares' from ANOVA using Euclidean norm and treating the cluster means as the cell means. Then,

if a second sample is drawn

$$\begin{aligned}\zeta_D = &\ \mathrm{P}(\mathcal{P} \text{ is worse in both cases.}) \\ &+ \mathrm{P}(\mathcal{P} \text{ is better in both cases.}) \\ = &\ \xi_D^2 + (1-\xi_D)^2 \end{aligned} \tag{3}$$

gives the probability that the original ranking is preserved. Clearly, ζ_D is highest when $\xi_D \in \{0, 1\}$ and lowest when $\xi_D = 1/2$. So, even though the second draw is hypothetical since we only have one data set, ζ_D is a measure of confidence in the original clustering. Hence, ξ_D is also an assessment of confidence in the clustering and when it goes to $\frac{1}{2}$ our confidence as measured by ζ_D is minimized.

Now suppose that, instead of squared error, some other L^r norm is used. It turns out that expression (2) does not adequately capture the comparison between partitions. The reason is that there is a "bias" term associated with the difference in costs which may deceptively show (2) going to 0 or 1. That is, the asymptotic behavior of $\mathsf{cost}(\mathbf{Y}, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}, \mathcal{Q})$ is of the form "constant plus limiting normal" where the constant depends on the cost function and represents the difference in expected values of the cost function on the partitions. Since the "constant" is the bias, we look at

$$\begin{aligned}\xi_D \ = \ &\mathrm{P}\Big(\mathsf{cost}(\mathbf{Y}, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}, \mathcal{Q}) \\ &\leq \mathrm{E}\left(\mathsf{cost}(\boldsymbol{\varepsilon}, \mathcal{P}) - \mathsf{cost}(\boldsymbol{\varepsilon}, \mathcal{Q})\right)\Big), \end{aligned} \tag{4}$$

the "bias corrected" form of (2). In the case of squared error loss if $\mathcal{P}$ and $\mathcal{Q}$ have the same number of clusters, the bias correction is $\mathrm{E}\, \boldsymbol{\varepsilon}_d^T \mathbf{B} \boldsymbol{\varepsilon}_d = 0$ where $\mathbf{B}$ is the difference matrix for the costs of $\mathcal{P}$ and $\mathcal{Q}$, see Theorem 2.6. More generally, the bias correction is not zero. For these cases, we also give conditions under which $\xi_D \to \frac{1}{2}$. Note that when $\mathcal{P}$ and $\mathcal{Q}$ have different numbers of clusters the bias term can be nonzero even for squared error loss.

It is seen that the bias term in (4) depends on the noise and on the partitions. While the noise is unobservable, there is precedent for comparing a clustering to the dispersion one would expect under another distribution. Indeed, for estimating the number of clusters, the gap statistic, see [6], uses the squared error cost adjusted by the expected cost of a clustering on pure noise. Here, we include the bias term to obtain $\xi_D \to \frac{1}{2}$ in (4) because this is a sense in which the noise terms swamp any informative components. In fact, being unable to determine a difference in costs between two partitions when the difference is costs of clustering noise, even when bias is accounted for, is the essence of impossibility.

Important progress has been made in understanding the limits of high dimensional clustering in senses other than cost functions. For instance, [7] and [8] have explored the fact that for random data satisfying relatively weak distributional assumptions,

$$D_{max} \,/\, D_{min} \xrightarrow{\mathrm{P}} 1 \text{ as } p \to \infty, \tag{5}$$

where D_{max} and D_{min} are the maximum and minimum distances between points in the data set. In a clustering context, this effect, at best, slows down the convergence rates of clustering algorithms, while in the worst case destroys any hope of recovering meaningful structure in the data (see [9]). The same phenomenon is described in [10] with greater attention to the stochastic geometry of high dimensions. [11] examines the limiting geometry of high dimensional clustering in the context of ultrametrics and argues that the histogram of distances between data points is the main source of clustering information. We comment that the geometry of level sets has been used to establish a consistency result for the number of clusters, [12].

Because equation (1) has both random and non-random terms, it leads to conditions on the rate of increase in the distance between the non-random components. So, our conditions are more general than those for which (5) has been shown to hold. Furthermore, our result applies to the evaluation of any output that can be expressed in terms of the squared error cost function. This includes principal component analysis, where finding the first principal component is equivalent to optimally performing K-means with two centroids [13].

Our theoretical results complement several other important theoretical results on optimality conditions for squared error loss. In particular, [5] uses bounds on the excess risk to show that in high dimensions, squared error loss can be nearly optimized.

They do not directly consider sparsity of informative components or noise on the observed data points, however, so the only implication is that the squared distance between informative components grows as $\mathcal{O}(D)$. We show that if this rate is $o\left(\sqrt{D}\right)$, and the variance of the noise doesn't go to zero, the results will be meaningless. Between rates $\mathcal{O}(D)$ and $o\left(\sqrt{D}\right)$ examples where squared error loss is meaningful and where it is not both occur.

There has also been important work on bounds for classification in high dimensions. Similar to us, [14] describes an impossibility region in terms of the number and dimension of the data points and the strength and frequency of the meaningful dimension components. Likewise, [15] prove an impossibility theorem for classification in high dimensions. They also show that under plausible large D, small n conditions, the accuracy of any classifier using linear discriminants becomes essentially random. (A deeper problem with linear discriminants is even seen with finite dimensions, see [4], Sec. 4.6 and Problem 4.9.) [16] establish the surprising result that linear discriminant classifiers that neglect the covariance structure of the explanatory variables can actually lead to better classification than linear classifiers that take the covariances into consideration. This is not an impossibility result; however, it does suggest that part of the problem with high dimensional classification is that the dimensions of the explanatory variables make the problem over-complex for many existing methods. Recall that linear classifiers arise from norms defined by a variance matrix, in essence a squared error criterion.

One of our points here is that including all the available features in clustering can be detrimental. Much work has been done on this for unsupervised learning; we refer the reader to [17] or [18] for more thorough overviews of recent results. Our main result implies that variable selection will be necessary if the clusters generated from most cost-function based clustering procedures are to be believed.

In the next section, we formally define our model for the squared error setting and discuss the relevant properties of partitions and the noise component $\boldsymbol{\varepsilon}$. We then state and prove our main asymptotic impossibility theorem in section 3, giving conditions under which $\xi_D \to \frac{1}{2}$ as $D \to \infty$ for squared error. In section 4, we give theoretical bounds on the rate of this convergence along with some simple Monte Carlo algorithms for computing it under different assumptions of the noise model. In section 5, we provide a contrast to our work. By examining the link between clustering and classification we are able to argue informally that in some large n, small D cases (the reverse of the setting here) the consistency of classification can be used to show consistency of clustering results. This introduces VC-dimension as an important criteria in considering consistency of clustering results. In Section 6, we extend our impossibility results for squared error cost to other cost functions, obtaining qualitatively similar, but more complex, results involving bias corrections. Finally, in 7.1, we present empirical evidence for our results in the form of simulations and a re-analysis of real genomic data to reveal the limits of clustering and how these can be bounded by rates such as those in Section 4. Proofs of select results are given in the Appendices.

2 FORMAL SETTING

Here we treat a clustering as a partition of a data set into disjoint, exhaustive subsets. The disjointness is essential for the proofs below. We use K to denote the number of subsets and restrict ourselves to nontrivial partitions containing at least two partition elements and no void partition elements. As a matter of notation, the points in our partitions are generically represented as integers in $1, 2, ..., n$. We formalize this in the following.

Definition 2.1. *Given n points and a number of clusters $K \leq n$, a partitioning $\mathcal{P} = \{P_1, P_2, ..., P_K\}$ is a set of K non-empty, disjoint subsets of $\{1, 2, ..., n\}$ such that $\cup_{k=1}^{K} P_k = \{1, 2, ..., n\}$. We use $\mathscr{P}_{nK}$ to denote the set of all partitionings of n points into K partitions.*

Next, we formally define the squared error cost function.

Definition 2.2. *Given a partitioning $\mathcal{P} = \{P_1, P_2, ..., P_K\}$ on a set of data points $\mathbf{Y} \in \mathbb{R}^{p \times n}$,*

the squared error cost function is

$$\mathsf{cost}(\mathbf{Y}, \mathcal{P}) = \sum_k \sum_{i \in P_k} \|\mathbf{Y}_{:i} - \overline{\mathbf{Y}}_k\|_2{}^2$$

where $\mathbf{Y}_{:i} = (Y_{1i}, Y_{2i}, ..., Y_{Di})$, $\overline{\mathbf{Y}}_\mathbf{k} = \text{mean}\{\mathbf{Y}_{:i} : i \in P_k\}$ *is the k-th cluster mean, and* $\|\cdot\|_2$ *indicates the Euclidean norm.*

First note that the cost function can be separated into dimension components, i.e.

$$\mathsf{cost}(\mathbf{Y}, \mathcal{P}) = \sum_{d=1}^{D} \sum_k \sum_{i \in P_k} \left(Y_{di} - \overline{Y}_{di}\right)^2,$$

where the subscript d's indicates dimension. Second, some straightforward algebra allows us to write

$$\mathsf{cost}(\mathbf{Y}, \mathcal{P}) = \sum_{d=1}^{D} \frac{1}{2} \sum_k \frac{1}{|P_k|} \sum_{i,j \in P_k} (Y_{di} - Y_{dj})^2, \quad (6)$$

where $|P_k|$ is the cardinality of the partition element P_k. Hence, (6) can be restated as

$$\mathsf{cost}(\mathbf{Y}, \mathcal{P}) = \sum_{d=1}^{D} \mathbf{Y}_d^T \mathbf{A}^\mathcal{P} \mathbf{Y}_d = \text{trace}\left(\mathbf{Y}^T \mathbf{A}^\mathcal{P} \mathbf{Y}\right), \quad (7)$$

where $\mathbf{A}^\mathcal{P}$ is

$$\mathbf{A}^\mathcal{P} = \left[\mathbf{a}_{ij}^\mathcal{P}\right]_{i,j=1,2,...,n} = \frac{1}{2}\left(\mathbf{1}_{\{i=j\}} - \frac{1}{|P_k|}\mathbf{1}_{\{i,j \in P_k\}}\right).$$

Similarly, if $\mathcal{Q}$ is another partition, there is a matrix $\mathbf{A}^\mathcal{Q}$. Having established this notation, we can examine the the difference between two cost functions.

2.1 Differences of Cost Functions

Many of our results depend on the difference in cost between two partitions, say $\mathcal{P}$ and $\mathcal{Q}$ of the same data. It is convenient to define a cost difference matrix as follows.

Definition 2.3. *Let* $\mathbf{B} = [b_{ij}] = \mathbf{A}^\mathcal{P} - \mathbf{A}^\mathcal{Q}$ *be the cost difference matrix between partitionings* $\mathcal{P}$ *and* $\mathcal{Q}$ *assumed to have only non-void partition elements.*

Using the definition for $\mathbf{A}^\mathcal{P}$ and $\mathbf{A}^\mathcal{Q}$ given in (8), we have

$$b_{ij} = \frac{1}{2\,|Q_\ell|}\mathbf{1}_{\{i,j \in Q_\ell\}} - \frac{1}{2\,|P_k|}\mathbf{1}_{\{i,j \in P_k\}}.$$

Given this definition of $\mathbf{B}$, it's easy to see that

$$\mathsf{cost}(\mathbf{Y}, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}, \mathcal{Q}) = \text{trace}\left(\mathbf{Y}^T \mathbf{B} \mathbf{Y}\right).$$

The matrix $\mathbf{B}$ has several desirable properties listed in the following.

Theorem 2.4. *Let* $\mathbf{B}$ *be the cost difference matrix between two partitions* $\mathcal{P}, \mathcal{Q} \in \mathscr{P}_{nK}$. *Let* $\lambda_1, ..., \lambda_n$ *be the eigenvalues of* $\mathbf{B}$, *ordered so nonzero values are indexed first. Then*

A. $\mathbf{B} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$ *for some orthogonal matrix* $\mathbf{U}$ *and* $\mathbf{\Lambda} = \text{diag}\,(\lambda_1, \lambda_2, ..., \lambda_n)$.

B. $\lambda_i \in \left[-\frac{1}{2}, \frac{1}{2}\right]$ *for* $i = 1, ..., 2(K - M)$ *and* $\lambda_i = 0$ *for* $i = 2(K - M) + 1, ..., n$, *where* $M = |\mathcal{P} \cap \mathcal{Q}|$ *is the number of common partitions.*

C. $\text{rank}(\mathbf{B}) \leq 2(K - M)$.

E. $\text{trace}\,(\mathbf{B}) = \sum_i \lambda_i = 0$.

Proof. See appendix. □

2.2 Noise

We now formalize what we mean by the noise variable $\boldsymbol{\varepsilon}$ in equation (1). Note, our asymptotic results hold for any noise generating distribution with mean zero and finite, non-zero fourth moment. The existence of the fourth moment is necessary to control the second moment of the squared error cost function. Our class of noise terms is the following.

Definition 2.5. *Suppose a sequence of distribution functions* $F_1, F_2, ...$ *with* $\varepsilon_{di} \sim F_d$, $d = 1, 2..., i = 1, 2, ..., n$ *satisfies the following properties:*

1. $\text{E}\,\varepsilon_{d_i} = 0$.

2. $\text{E}\,\varepsilon_{di}^4 = \rho_d$, *and,* $\forall\ d, \exists\ L, U$ *s. t.* $0 < L \leq \rho_d \leq U < \infty$.

We refer to any such sequence as a noise generating sequence, and any distribution function in a such a sequence as a noise generating distribution.

The following theorem lists some useful connections between the cost difference matrix and noisy data.

Theorem 2.6. *Let $\mathbf{Y}_d = \mathbf{x}_d + \boldsymbol{\varepsilon}_d$ be as in equation (1), and let $\mathbf{B}$ be a cost difference matrix for two partitions $\mathcal{P}, \mathcal{Q} \in \mathscr{P}_{nK}$, $\mathcal{P} \neq \mathcal{Q}$. Then, for $Z_d = \mathbf{Y}_d^T \mathbf{B} \mathbf{Y}_d$, we have the following:*

A. $\mathrm{E}\, \varepsilon_d^T B \varepsilon_d = 0$.

B. $\mathrm{E}\, Z_d = \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d$.

C. $\mathrm{Var}\left(\boldsymbol{\varepsilon}_d^T \mathbf{B} \boldsymbol{\varepsilon}_d\right) \leq K \rho_d \,/\, 2$.

Proof. See appendix. □

It is seen that the variance has a bound independent of n.

3 AN IMPOSSIBILITY THEOREM FOR SQUARED ERROR LOSS

Our technique of proof is a variant on the classical Lindeberg-Feller central limit theorem and rests on the structure of the squared error cost function, as previously discussed. The basic strategy is to show that the limiting distribution of $(\mathsf{cost}(\mathbf{Y}, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}, \mathcal{Q}))/\sqrt{D}$ tends to a mean-zero normal. If this happens, then, asymptotically, the cost function of any fixed clustering has probability one half of being larger than the cost function of any other fixed clustering, independently of how representative of the modal structure of the underlying distribution either clustering is.

Because we focus almost exclusively on the difference of cost functions, we define the variable Z_d to refer to the difference in cost of one of the D components of $\mathbf{Y}$ over all n samples. Specifically, define

$$\begin{aligned} Z_d &= \mathsf{cost}(\mathbf{Y}_d, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}_d, \mathcal{Q}) \\ &= \mathbf{Y}_d^T \mathbf{B} \mathbf{Y}_d = (\mathbf{x}_d + \boldsymbol{\varepsilon}_d)^T \mathbf{B} (\mathbf{x}_d + \boldsymbol{\varepsilon}_d). \end{aligned} \quad (8)$$

where $\mathbf{Y}_d = (Y_{d1}, \ldots, Y_{dn})$, $\mathbf{x}_d = (x_{d1}, \ldots, x_{dn})$, and $\boldsymbol{\varepsilon}_d = (\boldsymbol{\varepsilon}_{d1}, \ldots, \boldsymbol{\varepsilon}_{dn})$ for each $d = 1, \ldots, D$, and $\varepsilon_{di} \sim F_d$, $i = 1, 2, ..., n$, for some noise generating sequence $F_1, F_2, ...$ We examine the properties of the cost for fixed D and n and then let D increase to investigate $D >> n$.

3.1 Asymptotic Behavior

One of the key steps in the proof of the main result, theorem 3.2 below, is a central limit theorem showing that a sequence of controlled random variables, defined as follows, converges to a normal distribution centered at 0. We have the following.

Theorem 3.1. *Let $T_d, d = 1, 2, ...$ be a sequence of independent random variables with $\mathrm{E}\, T_d = r_d$, where r_d is a non-random sequence such that*

$$\frac{1}{\sqrt{D}} \sum_{d=1}^{D} r_d \to 0 \text{ as } D \to \infty.$$

Suppose $\mathrm{E}\left(T_d^2\right) = \sigma_d^2$, $0 < L \leq \sigma_d \leq U < \infty$. Let $S_d = T_1 + T_2 + \cdots + T_d$ and $c_d^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2$. Then

$$S_D \,/\, c_D \xrightarrow{D} \varphi \text{ as } D \to \infty \quad (9)$$

where φ is a random variable having a standard normal distribution with mean 0 and variance 1.

Proof. See appendix. □

To present our clustering impossibility theorem, we now focus on the asymptotic behavior of Z_d as defined in equation (8). Since the event $[\sum_d Z_d \geq 0]$ is identical to the event $[\mathsf{cost}(\mathbf{Y}, \mathcal{P}) \geq \mathsf{cost}(\mathbf{Y}, \mathcal{Q})]$ for finite dimensions, it is enough to show that $\mathrm{P}(\sum_d Z_d \geq 0) \to \frac{1}{2}$. However, $\mathrm{Var}\,(\sum_d Z_d) \to \infty$, so we use the central limit theorem 3.1 above. To state our result, let

$$R(\boldsymbol{\varepsilon}_d) = ((x_{d1} + \varepsilon_{d1})(x_{d1} + \varepsilon_{d1}), ..., (x_{dn} + \varepsilon_{dn})(x_{dn} + \varepsilon_{dn}))^T$$

be the vector of length n^2 consisting of the coefficients of the b_{ij} in Z_d. Next, let $\mathbf{e}_d$ be a fixed value in the range of the n dimensional error term $\boldsymbol{\varepsilon}_d$, fix $\delta \geq 0$ and write

$$N_\delta(\mathbf{e}_d) = \{\mathbf{e}'_d : \ \|\mathbf{e}'_d - \mathbf{e}_d\| \leq \delta\}$$

for the δ neighborhood around $\mathbf{e}_d$. Now, we impose a full rank condition that partitions the range of the error term and so can be used for both continuous and discrete error terms $\boldsymbol{\varepsilon}_d$.

Condition F: For $j = 1, \ldots, n^2$, there is a $\delta \geq 0$ and n^2 neighborhoods of the form $N_{\delta,j}(\mathbf{e}_d^{(j)})$ with centers $\mathbf{e}_d^{(j)}$ so that for all j, $\mathrm{P}\Big(\boldsymbol{\varepsilon}_d \in N_{\delta,j}(\mathbf{e}_d^{(j)})\Big) > 0$ and so that for some $\eta > 0$ and any selection of $\mathbf{e}_d^{(j),\prime} \in N_{\delta,j}(\mathbf{e}_d^{(j)})$,

$$\left| \det \begin{bmatrix} R(\mathbf{e}^{(1),\prime})^T \\ \vdots \\ R(\mathbf{e}^{(n^2)^T,\prime}) \end{bmatrix} \right| > \eta.$$

Our result is the following.

Theorem 3.2. *For $\mathbf{Y}_d$, $\mathbf{x}_d$, and $\boldsymbol{\varepsilon}_d$ as in (8) and suppose $\mathcal{P}$ and $\mathcal{Q}$ are any partitions in $\mathscr{P}_{nK}$, $\mathcal{P} \neq \mathcal{Q}$, are two partitions of the n data points into K clusters, with corresponding cost difference matrix $\mathbf{B}$. If Condition F holds and if*

$$\frac{1}{\sqrt{D}} \sum_{d=1}^{D} \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d \;\to\; 0 \tag{10}$$

then

$$\mathrm{P}(\mathsf{cost}(\mathbf{Y}, \mathcal{P}) \leq \mathsf{cost}(\mathbf{Y}, \mathcal{Q})) \;\to\; \frac{1}{2} \tag{11}$$

as $D \to \infty$.

Note that $\sum_d \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d = o\left(\sqrt{D}\right)$ is trivially satisfied if $\sum_d \|\mathbf{x}_d\|_2{}^2 = o\left(\sqrt{D}\right)$. Also, note that the condition in theorem 3.2 on the growth rate of the cost of the informative components is tight; if they grow at rate $\mathcal{O}(\sqrt{D})$, then $\sum_d Z_d/\sqrt{D}$ may converge to a normal distribution shifted by a non-zero constant, which would therefore have a non-zero mean. A higher rate of growth would mean that the informative components would eventually win out decisively over the noise.

Proof. With the previous theorems and lemmas in place, all that remains is to show that Z_d satisfies the assumptions of theorem 3.1. From theorem 2.6, we know that

$$\mathrm{E}\, Z_d = \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d$$

which is a non-random, deterministic quantity. Furthermore, theorem 2.6 tells us that $\mathrm{Var}\left(Z_d^2\right)$ is finite, so $\exists U'$ such that $\mathrm{E}\, Z_d^2 \leq U' < \infty$. More, when $\mathcal{P} \neq \mathcal{Q}$ we have $B \neq 0$ and this implies $\mathrm{E}\, Z_d^2 > 0$ because $\mathrm{E}\, Z_d^2 = 0$ only if $\mathbf{B} = 0$.

Indeed, we know that $Z_d^2 \geq 0$ so if $\mathrm{E}\, Z_d^2 = 0$ then $Z_d^2 = 0$ a.e. and so $Z_d = (\mathbf{x}_d + \boldsymbol{\varepsilon}_d)^T \mathbf{B} (\mathbf{x}_d + \boldsymbol{\varepsilon}_d) = 0$, a.e. This means that

$$\mathbf{b}^T R(\boldsymbol{\varepsilon}_d) = \sum_{i,j=1}^{n} b_{ij}(x_{di} + \varepsilon_{di})(x_{dj} + \varepsilon_{dj}) = 0 \quad a.e. \tag{12}$$

where $\mathbf{b}^T = (b_{11}, \ldots, b_{nn})$ are the entries of $\mathbf{B}$ written out as a vector (rather than a matrix) in the same order as in $R(\boldsymbol{\varepsilon}_d)$. Under Condition F, we can find n^2 linearly independent equations of the form of (12) constraining $\mathbf{b}$, each holding on a set of strictly positive measure. So, their only solution is $\mathbf{b} = 0$, i.e., $\mathbf{B} = 0$.

Now, the conditions of theorem 3.1 are satisfied, and we have that

$$\frac{1}{\sqrt{D}} \sum_{d=1}^{D} \frac{Z_d}{s_D} \xrightarrow{D} \varphi \text{ as } D \to \infty$$

where $s_D^2 = \sum_{d=1}^{D} \sigma_d^2$ and $\varphi \sim \mathcal{N}(0,1)$. However, $\mathrm{P}(\varphi \geq 0) = 1/2$, so

$$\mathrm{P}\left(\sum\nolimits_{d=1}^{D} Z_d \geq 0\right) \;\to\; \frac{1}{2}.$$

But

$$\sum_{d=1}^{D} Z_d = \mathsf{cost}(\mathbf{Y}, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}, \mathcal{Q}),$$

so the theorem is proved. □

To verify that Condition F is nontrivial and readily satisfied for a large class of continuous error distributions, we show the following.

Corollary 3.3. *Suppose $\boldsymbol{\varepsilon}_d$ has a density with respect to Lebesgue measure that is continuous on an open*

set that has positive probability. Suppose also that the components ε_{di} are independent and identical with support containing an open set. Then, Condition F is satisfied.

Proof. Recall that $R(\boldsymbol{\varepsilon}_d)$ has length n^2 and $\mathbf{b}^T R(\boldsymbol{\varepsilon}_d) = 0$, a.e. Partition the range of each ε_{di} into n^2 disjoint subsets W_k for $k = 1, \ldots, n^2$ each having strictly positive probability and containing an open set. Then, on each subset (12) holds. That is, $\mathbf{b}^T R(\boldsymbol{\varepsilon}_d) = 0$ a.e. can be regarded as a second order polynomial in the ε_{di}'s. The only way a polynomial can be zero a.e. is if all its coefficients, the b_{ij}'s, are zero. So, choose a single equation from each W_k for $k = 1, \ldots, n^2$ and use its coefficients as a row in the matrix in Condition F. These rows can be chosen to be linearly independent because for each subset the value of $\mathbf{e}_d$ at which R is evaluated ranges over a set of strictly positive measure. □

We suggest that Condition F is valid even for large classes of discrete error terms provided D is large enough. This can be readily conjectured from (12) but is not formally proved here. Note that our real data example in Sec. 7 satisfies Corollary 3.3 but other biomedical examples do not.

3.2 Connection to Subspaces

In practice, it is often assumed that the true data is sparse in the sense that a small number of features contain almost all the information but we do not know which those are. The following theorem considers this case explicitly, partly to emphasize the point that considering all the components of the dataset can make matters worse, and partly because this model is used in section 4 to investigate convergence rates.

Corollary 3.4. *Suppose $\mathbf{Y} = \mathbf{x} + \boldsymbol{\varepsilon}$, and suppose the columns of $\mathbf{x}$ vary over a fixed finite-dimensional subspace $S \subset \mathbb{R}^D$ as D increases. If the components of $\boldsymbol{\varepsilon}$ are independent and distributed according to a noise generating sequence, then $\xi_D \to \frac{1}{2}$ as $D \to \infty$.*

Proof. Let S be a fixed c-dimensional subspace of $\mathbb{R}^D$ that contains the span of the columns of $\mathbf{x}$. Then,

$$\mathbf{x}_{d_1}^T \mathbf{B} \mathbf{x}_{d_2} = o\left(\sqrt{D}\right),$$

for any d_1 and d_2 between 1 and D since any $\mathbf{x}_d$ is $n \times 1$, $\mathbf{B}$ is $n \times n$, and n is fixed. Since there are only finitely many such terms that are non-zero,

$$\sum_{d=1}^{D} \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d = o\left(\sqrt{D}\right),$$

which is the main hypothesis of Theorem 3.2. □

The model used in Corollary 3.3 is, essentially, the case that we have a finite number of informative components and the result shows that the effect of the differences among the non-random parts disappears as D increases. That is, clustering is impossible under a squared error evaluation criterion.

4 CONVERGENCE RATES

Since the impossibility results in Sec. 3 hold in the limit, the important question becomes how quickly their conclusions are observed as D increases. In this section, we explore this question by approximating and bounding the sequence ξ_D, given in equation (2), as D increases and $\xi_D \to \frac{1}{2}$.

4.1 Monte Carlo Approximations to ξ_D

Assuming random variables can be drawn from noise distributions $F_1, F_2, ...$, it is easy to obtain a simple Monte Carlo approximation to ξ_D. Since $\mathbf{B}$ is $n \times n$, we can write, as in Theorem 2.5, $\mathbf{B} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^T$, where $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_n]$ and the u_i's are unit vectors and $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_n)$. Theorem 2.5 also provides the bound $\text{rank}(\mathbf{B}) \leq 2(K - M)$, where M is the number of partition elements in common between $\mathcal{P}$ and $\mathcal{Q}$. So, letting $\ell = \min(n, 2(K - M))$ we can reduce $\mathbf{U}$ to $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_\ell]$ and $\boldsymbol{\Lambda}$ to $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_\ell)$, by eliminating eigenvectors with eigenvalue zero. The noise term has distribution $F = (F_1, \ldots, F_D)$ as in

Definition 2.6 so that for each $d = 1, \ldots, D$, $\varepsilon_{di} \sim F_d$ for $i = 1, \ldots, n$ so that over $j = 1, \ldots, N$ i.i.d. replications $\tilde{\varepsilon}_{dij} \sim F_d$. Now, a Monte Carlo approach is given in the following.

Theorem 4.1. *The probability that the cost of one partition $\mathcal{P}$ of size K is less than the cost of another partition $\mathcal{Q}$, also of size K is*

$$\begin{aligned}\xi_D &= \mathrm{P}(\mathsf{cost}(\mathbf{Y}, \mathcal{P}) \leq \mathsf{cost}(\mathbf{Y}, \mathcal{Q})) \\ &= \mathrm{P}\left(\sum_{i=1}^{\ell} \lambda_i \sum_{d=1}^{D} \left(w_{id} + \boldsymbol{\varepsilon}_d^T \mathbf{u}_i\right)^2 \geq 0\right) \\ &\overset{a.s.}{=} \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left\{\sum_{i=1}^{\ell} \lambda_i \sum_{d=1}^{D} \left(w_{id} + \tilde{\boldsymbol{\varepsilon}}_{dj}^T \mathbf{u}_i\right)^2 \geq 0\right\},\end{aligned} \tag{13}$$

where $\mathbb{I}\{\}$ is the indicator function and $\mathbf{w} = \mathbf{x}^T\mathbf{u}$.

The disadvantage of this Monte Carlo algorithm is its computational complexity, $\mathcal{O}(N\ell nD)$. The task is easier if we assume i.i.d. Gaussian noise.

Corollary 4.2. *Suppose that all ε_{di}, $d = 1, \ldots, D$ and $i = 1, \ldots, n$ are i.i.d. $\mathcal{N}(0, \sigma^2)$. Furthermore, let $\chi^2_{c,i}(\delta)$, $i = 1, ..., \ell$, denote i.i.d. Chi-square random variables with c degrees of freedom and non-centrality parameter δ. Then ξ_D can be expressed as a weighted sum of Chi-square variables, i.e.,*

$$\xi_D = \mathrm{P}\left(\frac{1}{\sqrt{D}} \sum_{i=1}^{\ell} \lambda_i \chi^2_D(\delta_i) \geq 0\right) \tag{14}$$

$$\overset{a.s.}{=} \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left\{\frac{1}{\sqrt{D}} \sum_{i=1}^{\ell} \lambda_i \chi^2_{D,j}(\delta_i) \geq 0\right\}, \tag{15}$$

where $\delta(i) = \sum_{d=1}^{D} (w_{id}/\sigma)^2$.

The setting of the last corollary can be simplified further if the dimension of the span of the columns of $\mathbf{x}$ is finite; this assumption was used in Corollary 3.3. Without loss of generality, we suppose that these informative components i.e., having nonzero signal, are in the first c entries of $\mathbf{x}$ and that all later entries are only noise i.e., they have zero signal.

Theorem 4.3. *Assume all the conditions of Cor. 4.2, in particular, $\varepsilon_{id} \overset{iid}{\sim} \mathcal{N}(0,1)\ \forall\ i, d$ and λ_i, for $i = 1, ..., \ell$ are the ℓ nonzero eigenvalues of $\mathbf{B}$. Then*

$$\xi_D = \mathrm{P}\left(\frac{1}{\sqrt{D-c}} \sum_{i=1}^{\ell} \lambda_i \chi^2_{D-c,i} \geq 0\right) + o_P(1) \quad as \quad D \to \infty \tag{16}$$

$$= \lim_{N\to\infty} \left[\frac{1}{N} \sum_{j=1}^{N} \mathbb{I}\left\{\frac{1}{\sqrt{D-c}} \sum_{i} \lambda_i \chi^2_{D-c,i,j} \geq 0\right\} + o_P(1)\right] \tag{17}$$

where $\chi^2_{D-c,i}$ and $\chi^2_{D-c,i,j}$ for $i = 1, \ldots, \ell$ and $j = 1, \ldots, n$ are i.i.d. Chi-squared random variables with $D - c$ degrees of freedom.

4.2 General Bounds on the Convergence of ξ_D

While the above approximations are useful for calculating ξ_D in practice, other theoretical bounds on these rates are also of interest. In this section, we present bounds on the convergence of $\xi_D \to \frac{1}{2}$ as $D \to \infty$ in terms of the higher moments of the cost function. Often these may be easier to obtain as they can be calculated empirically by using estimates of the parameters that appear in the expressions below.

One standard way to bound convergence rates in central limit theory is the well-known Berry-Esseen theorem. One version of this is the following. Let $V_1, \ldots, V_D$ be a sequence of i.i.d. random variables such that $\mathrm{E}(V_d) = 0$, $\mathrm{E}\left(V_d^2\right) = \sigma^2$, and $\mathrm{E}\left(|V_d|^3\right) = \rho < \infty$. Let $\overline{V_D} = \frac{1}{D}\sum_{d=1}^{D} V_d$, and let F_D be the cumulative distribution function of $\sqrt{D}\overline{V_D}/\sigma$. Then there exists a constant δ such that

$$|F_n(t) - \Phi(t)| \leq \frac{\delta\rho}{\sigma^3\sqrt{D}}$$

where $\Phi(t)$ is the cumulative distribution function of the standard normal distribution. The constant δ, while not known exactly, has been bounded above by 0.7655 [19]. Using this version of the Berry-Esseen theorem for the Z_ds will require that the noise distribution from which the ε_{id}'s are drawn to have finite sixth moment and be i.i.d. along the dimension component d.

To give rates on the convergence, some notation is needed. Let

$$C = \sum_{d=1}^{c} \mathbf{x}_d^T \mathbf{B} \mathbf{x}_d,$$

so that

$$\begin{aligned}\sum_{d=1}^{c} Z_d &= C + \sum_{d=1}^{c} (\boldsymbol{\varepsilon}_d)^T \mathbf{B} (\boldsymbol{\varepsilon}_d) + \sum_{d=1}^{c} (\boldsymbol{\varepsilon}_d)^T \mathbf{B} (\mathbf{x}_d) \\ &\quad + \sum_{d=1}^{c} (\mathbf{x}_d)^T \mathbf{B} (\boldsymbol{\varepsilon}_d) \\ &= C + V_c \end{aligned} \tag{18}$$

where (18) defines V_c as a sum of normal and Chi-square random variables. We have the following.

Theorem 4.4. *Suppose the first c dimension components are the only ones with non-zero signals and the later $D-c$ components are drawn from an i.i.d. noise distribution with finite sixth moment. Then for $\alpha = \alpha(D)$ satisfying*

$$\frac{e^{-\alpha(D)/8}}{\sqrt{D}} \to 0 \tag{19}$$

we have that

$$\xi_D \in [\Phi^*(-a_D) - b_D, \Phi^*(-a_D) + b_D] \tag{20}$$

where

$$\begin{aligned} a_D &= \frac{C + \alpha'}{\sigma\sqrt{D-c}} \\ b_D &= \frac{\delta\rho}{\sigma^3\sqrt{D-c}} \\ \rho &= \mathrm{E}\left|\mathsf{cost}(\mathbf{Y}_d, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}_d, \mathcal{Q})\right|^3 = \mathrm{E}\left|\boldsymbol{\varepsilon}_d^T \mathbf{B} \boldsymbol{\varepsilon}_d\right|^3 \\ \sigma^2 &= \mathrm{E}\,\mathsf{cost}(\mathbf{Y}_d, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}_d, \mathcal{Q})^2 = \mathrm{E}\left(\boldsymbol{\varepsilon}_d^T \mathbf{B} \boldsymbol{\varepsilon}_d\right)^2, \end{aligned}$$

and Φ^* indicates the normal distribution integrated over $V_c = \alpha'$ for $\alpha' < \alpha$ multiplied by $1/\mathrm{P}(\{V_c \le \alpha\})$ and $-a_D$ indicates the argument over which the integration is done.

Note that the confidence intervals are distorted by the integration, however, the rate is preserved for each $\alpha' > \alpha$ giving an overall $\sqrt{D}$ convergence. In practice, we expediently set $\alpha = 0$ and $\mathrm{P}(\{V_c \le \alpha\}) = 1$ when we use Theorem 4.4 to give bounds on probabilities in Section 5.

Proof. The strategy here is to separate out the first c components in (18) so their influence can be bounded and the Berry-Esseen theorem used to get (20). We begin by letting $\alpha > 0$ and writing

$$\begin{aligned} &\mathrm{P}(\mathsf{cost}(\mathbf{Y}, \mathcal{P}) \le \mathsf{cost}(\mathbf{Y}, \mathcal{Q})) \\ &= \mathrm{P}\left(\sum_{d=1}^{D} Z_d \le 0\right) \\ &= \mathrm{P}\left(\frac{1}{\sqrt{D-c}} \sum_{d=c+1}^{D} \frac{Z_d}{\sigma} \le -\frac{C + V_c}{\sigma\sqrt{D-c}}\right) \\ &= \mathrm{P}\left(\frac{1}{\sqrt{D-c}} \sum_{d=c+1}^{D} \frac{Z_d}{\sigma} \le -\frac{C + V_c}{\sigma\sqrt{D-c}} \Big| V_c > \alpha\right) \mathrm{P}(V_c > \alpha) \end{aligned} \tag{21}$$

$$+ \mathrm{P}\left(\frac{1}{\sqrt{D-c}} \sum_{d=c+1}^{D} \frac{Z_d}{\sigma} \le -\frac{C + V_c}{\sigma\sqrt{D-c}} \Big| V_c \le \alpha\right) \mathrm{P}(V_c \le \alpha). \tag{22}$$

First, we show that (21) is satisfactorily small, and so can be ignored. Observe that

$$\begin{aligned} 0 \le\ &\mathrm{P}\left(\frac{1}{\sqrt{D-c}} \sum_{d=c+1}^{D} \frac{Z_d}{\sigma} \le -\frac{C + V_c}{\sigma\sqrt{D-c}} \Big| V_c > \alpha\right) \\ &\times\ \mathrm{P}(V_c > \alpha) \le 1 \times \mathcal{O}\Big(e^{-\alpha/8}\Big). \end{aligned} \tag{23}$$

This follows from using standard bounds on tail probabilities of the form $P(X > \alpha)$ where X is either normal or Chi-square and recalling that V_q is a sum of finitely many normal and Chi-square distributed random variables. It is seen that as long as α increases slowly with D as in the rate hypothesis in the statement of the theorem, the upper bound in (23) will go to zero at a rate faster than $\mathcal{O}\Big(1/\sqrt{D}\Big)$.

Next, for (22) we use the same bounds on tail probabilities to assert that $P(V \le \alpha) \approx 1 - \mathcal{O}\big(e^{-\alpha/8}\big)$, in the sense that upper and lower bounds on $P(V \le \alpha)$ of tightness $\mathcal{O}\big(e^{-\alpha/8}\big) = o(D)$ can be given.

Now, it remains to bound the conditional probability in (22). Since V_c is independent of $\sum_{i=c+1}^{D} Z_d$, the conditioning can be treated point-wise in V_c for

$V_c < \alpha$. Specifically, fix α and define the events

$$W(V_c) = \{\frac{1}{\sqrt{D-c}} \sum_{d=c+1}^{D} \frac{Z_d}{\sigma} \leq -\frac{C+V_c}{\sigma\sqrt{D-c}}\} \quad (24)$$

$$B = \{V_c \leq \alpha\}. \quad (25)$$

Then, by definition,

$$\mathrm{P}\left(\frac{1}{\sqrt{D-c}} \sum_{d=c+1}^{D} \frac{Z_d}{\sigma} \leq -\frac{C+V_c}{\sigma\sqrt{D-c}} \Big| V_c \leq \alpha\right)$$
$$= \mathrm{E}\left(I_{W(V_c)}|B\right) = \frac{1}{P(B)} \int_B I_{W(V_c)} d\,\mathrm{P}(V_c = v_c),$$

where I_A indicates the indicator function for a set A. Let $\alpha_0 = \alpha$ and choose $\alpha_0 > \alpha_1 > \ldots > \alpha_m$ and set $\alpha_{m+1} = \infty$. Write $B_k = \{\alpha_k > V_c > \alpha_{k+1}\}$ so $\cup_{k=0}^{m} B_k = B$. Now,

$$\int_B I_{W(V_c)} d\,\mathrm{P}(V_c = v_c) \quad (26)$$
$$= \sum_{k=0}^{m} \int_{B_k} I_{W(V_c)} dP_{V_c}(v_c)$$
$$= \sum_{k=0}^{m} \mathrm{E}\left(I_{W(V_c)} \Big| \alpha_k > V_c > \alpha_{k+1}\right) P_{V_c}(\alpha_k > V_c > \alpha_{k+1})$$
$$\to \int_{-\infty}^{\alpha} \mathrm{E}\left(I_{W(V_c)} \Big| V_c = \alpha'\right) P_{V_c}(\alpha') d\alpha'$$
$$= \int_{-\infty}^{\alpha} \mathrm{P}\big(W(V_c = \alpha') \Big| V_c = \alpha'\big) P_{V_c}(\alpha') d\alpha',$$

in which the convergence follows by invoking a conditional dominated convergence theorem (since the indicator function for a set is bounded by one). The conditional probability in the integral is

$$\mathrm{P}\left(\frac{1}{\sqrt{D-c}} \sum_{d=c+1}^{D} \frac{Z_d}{\sigma} \leq -\frac{C+\alpha'}{\sigma\sqrt{D-c}} \Big| V_c = \alpha'\right), \quad (27)$$

for each $V_c = \alpha'$ where $\alpha' < \alpha$ and the Berry-Esseen theorem can be used to control it. Letting the distribution function of $\sum_{d=c+1}^{D} Z_d/(\sigma\sqrt{D-c})$ be F_{D-c}, (27) is $F_{D-c}((C+\alpha')/(\sigma\sqrt{D-c}))$. The Berry-Esseen theorem then gives

$$\left|F_{(D-c)}(t) - \Phi(t)\right| \leq \frac{\delta\rho}{\sigma^3\sqrt{D-c}},$$

and the theorem follows by integrating the approximation with respect to the marginal for $V_c = \alpha'$ over $\alpha' < \alpha$ and multiplying by $1/\mathrm{P}(B)$. □

Note that (19) amounts to $\alpha = o(\ln D)$, which can in principle swamp the effect of C. However, in calculating bounds on the cost curves in Sec. 5.1, we used $\alpha = 0$ and obtained reasonable results. This may mean the $o(\ln D)$ only takes effect for very large D, perhaps due to a very small multiplicative factor, or that the bound in which α appears is loose.

A final convergence rate of ξ_D can now be stated. Because of the bounds on the growth rate of the informative components, as mentioned in section (3.1), we expect this rate to be $\mathcal{O}\left(1/\sqrt{D}\right)$. This is borne out in the following.

Corollary 4.5. *The asymptotic convergence of* $\xi_D - 1/2$ *has rate at most* $\mathcal{O}\left(1/\sqrt{D}\right)$.

Proof. This result follows from Taylor expanding $\Phi^*(\cdot)$ around 0 for fixed α. Since the approximating normal Φ for each value of α' satisfies $\Phi(0) = 1/2$, we have that $\Phi^*(0) = 1/2$ and therefore

$$\Phi^*(-a_D) = \Phi^*\left(\frac{-C+\alpha}{\sigma\sqrt{D-c}}\right) = \frac{1}{2} - \frac{C-\alpha}{\sigma\sqrt{2\pi(D-c)}} + o\left(\frac{1}{D}\right)$$
$$\implies \sqrt{D-c}\left(\xi_D - \frac{1}{2}\right) \in \left[\frac{C-\alpha}{\sigma\sqrt{2\pi}} - \tau, \frac{C-\alpha}{\sigma\sqrt{2\pi}} + \tau\right]$$
$$\text{where } \tau = \frac{\delta\rho}{\sigma^3} + o\left(\frac{1}{\sqrt{D}}\right).$$

□

5 CLASSIFICATION AND CLUSTERING

Here we outline a strategy for establishing a sort of consistency theorem for clustering. This is in contrast to the impossibility of clustering established in the previous section, and generalized in the following section. Our strategy rests on the fact that clustering and classification are closely related and there are well-known consistency theorems for classification. Geometrically, when the clusters in a clustering are relatively concentrated, they often correspond to

the distinct modes in the distribution that generated the data. These modes may be plausibly regarded as the distinct classes in a classification problem. Thus, the clustering problem is, loosely, the classification problem where the number of classes is unknown and the knowledge of which class a data point came from is unavailable. This often means that clustering problems are harder than classification problems and must be solved with less information. However, if we can situate a clustering problem plausibly in a classification context, concepts from classification may become applicable.

For instance, VC dimension is a measure of the richness of a collection of sets $\mathcal{S}$ that often arises is classification. Loosely, the VC dimension of $\mathcal{S}$ is the maximum number of points that can be separated in all possible ways by the elements of $\mathcal{S}$. This definition will be made precise below. It is intuitive that VC dimension must play a large role in a proper characterization of clustering procedures since every clustering procedure has a collection of possible clusterings it can generate and these can be regarded as partitions of a space. Since VC-dimension is a way to measure the ability of a collection partitions to separate points, VC-dimension is a measure of the richness of the collection of clusterings a clustering method can generate. Accordingly, we expect VC dimension to play a large role in characterizing the complexity of clustering methods. In effect, we expect that controlling VC-dimension will ensure the collection of clusterings is small enough that an impossibility theorem will not hold.

To explore this, regard a data dependent partition, such as clustering generates, purely as a description of the distribution. That is, regard a clustering technique as a way to generate a partition of the space that reflects the relative positions of the data. Thus, if a partition is well chosen it may be more stable, heuristically, than the values used to obtain it.

Since each clustering defines a collection of sets it is reasonable to suggest that clustering techniques choosing from larger collections of partitions are more prone to impossibility-type results that clustering techniques choosing from smaller collections of partitions. This follows because, intuitively, the richer a collection of partitions is, the easier it is to choose a partition that is weakly reflective, or not reflective at all, of the structure of the data, purely by chance. For instance a clustering method that chooses any valid partition of D-dimensional real space is likely to be much more unstable than normal mixture model clustering that produces clusters based on the boundaries between modes.

Roughly, the VC dimension of a class of sets is the maximum number of points the sets can separate in all possible ways. That is, to test if the VC dimension of a class of sets is at least k it is enough to find a set of k points for which all 2^k possible subsets can be picked out by the class of sets. For a formal definition see [4], Chap. 12.4, which is used below. This is most natural in a classification context where the goal is to separate points based on class membership.

Theoretically, a clustering problem can be converted into a classification problem by indexing a set of candidate modes for a distribution by $1, \ldots, M$ and adjoining to each $\mathbf{y}_i$ the index of the mode (essentially $\mathbf{x}_i$) it 'belongs to' perhaps in the sense of being closest to it. We refer to this as the augmented data (which is usually unavailable). Binary classification therefore corresponds to $M = 2$. In real classification problems, M is known whereas in clustering problems M is unknown.

Now, assuming we know the correct M, we can compare clustering methods by the VC-dimensions of the collection of sets each induces. Intuitively, the VC dimension of a clustering method (as opposed to an individual clustering) is the maximal number of points that can be separated in all possible ways by using elements from the union of all sets the method can generate. This makes sense because each separation of the points in a data set corresponds to a partition of the data points into clusters. As a generality, each clustering method has a range of partitions it can generate. Often these have infinite VC dimension. However, restricting attention to clustering methods that have ranges of partitions with finite VC-dimension will permit nontrivial upper bounds as seen below.

To develop the link among clustering, classification and VC dimension, we consider asymptotics in n rather than D. This is the reverse from the results in Sections 3 and 4 in which D increases and n is fixed

(or grows very slowly with D). So, suppose for each fixed D we have a collection of partitions ostensibly defined by the clustering method and we have known modes; below the partitions are denoted are the $\mathcal{P}$'s with collections of elements $\mathcal{B}(\mathcal{P})$ all gathered into in $\mathcal{F}(D)$. Given this, we can set up an application of Theorem 21.2 in [4] for the case $M = 2$.

Turning to formalities, we recall the following definitions from Chap. 21.1 in [4]. Let

$$\mathcal{F}(D) = \{\mathcal{P}(D) \,|\, \mathcal{P}(D) \text{ is a finite partition of } \mathcal{S}^D(R)\},$$

where $\mathcal{S}^D(R)$ is the ball of radius R centered at zero in D dimensions. Next, for a given $\mathcal{P}$, let $\mathcal{B}(\mathcal{P})$ be the collection of all $2^{|\mathcal{P}|}$ sets defined from $\mathcal{P}$ and let $\mathcal{A}(D) = \{A \in \mathcal{B}(\mathcal{P}) \text{ for some } \mathcal{P} \in \mathcal{F}\}$. It is seen that $\mathcal{A}$ is the collection of all sets that can be generated from $\mathcal{F}$ by taking unions of elements of partition in $\mathcal{F}$, Thus, $\mathcal{A}$ is the collection of sets whose VC-dimension characterizes the sensitivity of a clustering procedure.

Define $\Delta_n(\mathcal{F}) = s(\mathcal{A}, n)$ to be the shatter coefficient of $\mathcal{A}$. The shatter coefficient of a collection of sets such as $\mathcal{A}$ is the maximal number of different subsets of n points that can be 'picked out' by using its members. Formally, let $N_{\mathcal{A}}(z_1, \ldots, z_n)$ be the number of distinct sets in the collection of sets

$$\{z_1, \ldots, z_n\} \cap A$$

as A ranges over $\mathcal{A}$. Then,

$$s(\mathcal{A}, n) = \max_{z_1,\ldots,z_n \in (\mathcal{R}^D)^n} N_{\mathcal{A}}(z_1, \ldots, z_n)$$

The VC-dimension $V_{\mathcal{A}}$ of $\mathcal{A}$ is the largest k for which $s(\mathcal{A}, n) = 2^k$.

Next, to the augmented data, associate a partitioning rule π_n for each n. A partitioning rule is a function on the augmented data that gives a partition of the $\mathbf{Y}_i$'s. Write $\mathcal{F}_n$ to mean the family of partitions associated to π_n and denote $\pi_n(D^*) = \mathcal{P}_n$ where $D^* = D_n^*$ is the augmented data. For given $\mathbf{y}$, let $A_n(\mathbf{y})$ be the unique element of $\mathcal{P}_n$ that contains the point $\mathbf{y}$. Given this, we can define the 'classification rule'

$$g_n(\mathbf{y}) = \begin{cases} 0 & \text{if } \sum_{i=1}^n \mathbb{I}[\{]\mathbf{y}_i \in A_n(\mathbf{y}), u_i = 1\} \\ & \quad \leq \sum_{i=1}^n \mathbb{I}[\{]\mathbf{y}_i \in A_n(\mathbf{y}), u_i = 0\}, \\ 1 & \text{elsewhere.} \end{cases} \tag{28}$$

where $u_i = 1, 0$ according to whether $\mathbf{y}_i$ is in its correct modal class. Essentially, this classification rule assigns a future point $\mathbf{x}$ to the mode for which most of the previous $\mathbf{x}_i$s near it have been assigned, a sort of nearest neighbors condition based on the partition elements rather than on a distance measure explicitly.

Theorem 21.2 in [4] can now be applied. Its main hypotheses are

1. For each $R > 0$,
$$\lim_{n\to\infty} \frac{\log(\Delta_n(\mathcal{F}_n))}{n} = 0,$$
2. For any ball S_B and any $\gamma > 0$,
$$\lim_{n\to\infty} \mu(\{\mathbf{x}|\ \text{diam}(A_n(\mathbf{x}) \cap S_B) > \gamma\}) = 0$$
where $\mu = \mu_D$ is the dominating measure for the D-dimensional space $\mathbf{y}$ varies over.

When these are satisfied g_n is strongly consistent as a classifier, i.e., the classification loss associated with g_n converges in probability to its minimal value for all noise distributions some a specified class.

In the clustering context, this means that as $n \to \infty$ the clustering is consistent in the sense that it associates future $\mathbf{y}$'s to their correct modes. Note that this means good clustering would be achieved and therefore an impossibility theorem cannot hold. That is, the outline given here suggests conditions, quite different from those of the impossibility theorem, under which clustering can be consistent.

Item 1 means that the log of the shatter coefficient, effectively the log of the VC dimension, must be small enough, i.e., of order $o(n)$, and hence the partition cannot be too rich. Item 2 means that the diameter of the partition elements must shrink as n increases in the sense that there are few x's in measure μ for which their size is larger than a threshold., say γ. Taken together, the intuition this supports is that there is a threshold complexity, in a VC dimension sense, so that when the class of partition elements a clustering method can generate is restricted, but not so restricted as to be unable to partition neighborhoods, consistent clustering is possible.

By contrast, if the class of partition elements a clustering method can generate is too rich or unable

to partition sets with positive measure, it will be difficult for a clustering method to be consistent in the absence of other conditions. As a generality, if D increases, most clustering methods give a partition with elements drawn from a larger and larger class of possible partition elements. Heuristically, this means that as D increases it will be harder and harder to satisfy the conditions of Theorem 21.2 in [4] or similar results.

6 BEYOND SQUARED ERROR COST

One of the key limitations of the results presented in the previous sections is the reliance on squared error cost. This is so because squared error cost really only encapsulates a good notion of cost when the correct clustering consists of distinct, relatively spherical clusters. Other cost functions are therefore of interest both in their own right and as a way to evaluate how reasonable the results from squared error loss are. On the other hand, since distinct, spherical clusters are a paradigm case, the squared error cost is a useful benchmark for more general clustering problems. Moreover, K-means clustering, its variants such as Ward's method, and other clustering methods based on squared error are the most commonly occurring.

Before leaving squared error cost, we comment that some clustering problems that do not have distinct spherical clusters can be transformed into the setting of Theorem 3.2. For instance, there are cases where kernel methods can be used to transform clustering problems having non-linearly separable clusters into settings where a kernel based K-means or spectral clustering may be feasible. Instances of this can be found in [20] and [21]. In essence, transformation may make a clustering problem amenable to techniques best suited to separated, spherical clusters. So, the squared error cost function may be appropriate more generally that it might seem initially.

Nevertheless, there is real interest in variations on squared error cost to which we turn next.

6.1 $L^{r,s}$ Costs

Examination of the proof of Theorem 3.2 reveals that it rests on representing the difference in squared error costs as a sum over dimensions which can be decomposed into two terms. The first term is controlled by a central limit theorem; the second shows how far apart cluster centers can be and yet remain indistinguishable. This analysis seems to hold for more general norm-based cost functions as given below.

First, assume that the cost function can be separated by dimension:

$$\mathsf{cost}(\mathbf{Y}, \mathcal{P}) = \sum_{d=1}^{D} f_{\mathcal{P}}(\mathbf{Y}_d)$$

for some function $f_{\mathcal{P}}$ indexed by a partitioning $\mathcal{P}$. Second, assume that $f_{\mathcal{P}}(\mathbf{Y})$ is defined in terms of the rth norm to the power s:

$$f_{\mathcal{P}}(\mathbf{y}) = \sum_{P \in \mathcal{P}} \min_{y^* \in \mathbb{R}} \|\mathbf{y}_P - y^*\|_r{}^s \tag{29}$$

where $\mathbf{y}_P$ represents the $|P|$-dimensional vector of the components of $\mathbf{y}$ in partition element P, and we suppress the subscript d denoting dimension for readability. When $r = s = 2$, we have the case of squared error loss as given before; thus this represents an immediate generalization of this result. Note also that $r = \infty$, with $\|\cdot\|_\infty$ being defined as the componentwise maximum, is perfectly valid. For convenience, we refer to this as the $L^{r,s}$ cost function; in the next section, we prove impossibility results for the cases $s \in \{1, 2\}$ and $r \geq s$.

Now, let

$$Z_d = f_{\mathcal{P}}(\mathbf{Y}_d) - f_{\mathcal{Q}}(\mathbf{Y}_d) \tag{30}$$

Now, the key quantity in Theorem 3.2 becomes

$$\frac{1}{\sqrt{D}} \sum_{d=1}^{D} (Z_d - \mathrm{E}\, Z_d) + \frac{1}{\sqrt{D}} \sum_{d=1}^{D} \mathrm{E}\, Z_d. \tag{31}$$

Furthermore, equation (29) can be naturally generalized to include "soft" clusterings in which points have partial membership between the clusters, e.g.

the probabilistic membership found by mixture modeling. In this case, we could look at the more general "weighted" norm version:

$$f_{\mathcal{P}}(\mathbf{y}) = \sum_{k=1}^{K} \min_{y^* \in \mathbb{R}} \left(\sum_{i=1}^{n} (\alpha_{ik} y_i - y^*)^r \right)^{s/r} \tag{32}$$

In the case where α_{ik} is a binary indicator variable indicating membership with $\sum_i \alpha_{ik} = 1$, this reduces to equation (29). In the general case, with a few additional assumptions, an analogous impossibility result will still be shown to apply.

One of the reasons an impossibility theorem can often be established for separable cost functions as described by (29) is that they accumulate errors over all D dimensions. Consider if we combined the component-wise cost functions using an L^r norm, i.e. looked at

$$\xi_D = \mathrm{P}\left(\left(\sum_{d=1}^{D} f_{\mathcal{P}}^r(\mathbf{Y}_d) \right)^{1/r} \geq \left(\sum_{d=1}^{D} f_{\mathcal{Q}}^r(\mathbf{Y}_d) \right)^{1/r} \right) \tag{33}$$

In this case, for finite r, we can cancel the powers of $1/r$ and simply get

$$\xi_D = \mathrm{P}\left(\sum_{d=1}^{D} f'_{\mathcal{P}}(\mathbf{Y}_d) \geq \sum_{d=1}^{D} f'_{\mathcal{Q}}(\mathbf{Y}_d) \right) \tag{34}$$

for some new $f'_{\mathcal{P}}$ and $f'_{\mathcal{Q}}$. However, this does not apply if we let r go to infinity; in this case, the sum is replaced by a maximum over the dimensions, i.e.

$$\mathsf{cost}(\mathbf{Y}, \mathcal{P}) = \max_{d \in \{1,2,\ldots,D\}} f_{\mathcal{P}}(\mathbf{Y}_d) \tag{35}$$

A simple counterexample shows that an impossibility theorem of the manner given does not exist without additional assumptions. Suppose that the informative components $\mathbf{x}_d$ are non-zero only for $d = 1$, but suppose that the random noise on all dimensions is continuous but upper bounded by the value $\eta \max_{i,j} |\mathbf{x}_i - \mathbf{x}_j|$. Note that such an $\mathbf{x}$ and $\boldsymbol{\varepsilon}$ satisfy all the assumptions of Theorem 3.2; only the cost function is different. However, it is easy to see that, for η sufficiently small, the comparison of partitionings is determined entirely by dimension $d = 1$ as it will always be chosen by the max in (35).

However, using the maximum has its own problems, namely instability due to extreme values in the noise. Unless the noise is upper bounded, the law of large numbers guarantees that, asymptotically, the maximum will eventually be determined by noise. While such a result may be interesting, for the purposes of this paper we restrict ourselves to component-wise separable cost functions of the form given in (29).

6.2 A General Setting and Impossibility Theorem

Here we establish minimal conditions under which it seems a general impossibility theorem holds for separable cost functions as given in equation (29).

First, as in section 2, define the difference in component-wise cost between two partitionings $\mathcal{P}$ and $\mathcal{Q}$ as

$$Z_d = Z_d(f_{\mathcal{P}}, f_{\mathcal{Q}}) = f_{\mathcal{P}}(\mathbf{x}_d + \boldsymbol{\varepsilon}_d) - f_{\mathcal{Q}}(\mathbf{x}_d + \boldsymbol{\varepsilon}_d). \tag{36}$$

However, it is necessary to introduce an additional term to handle "bias" in the cost function, i.e. the difference in cost on pure noise:

$$U_d = U_d(f_{\mathcal{P}}, f_{\mathcal{Q}}) = f_{\mathcal{P}}(\boldsymbol{\varepsilon}_d) - f_{\mathcal{Q}}(\boldsymbol{\varepsilon}_d). \tag{37}$$

When $f_{\mathcal{P}}$ is squared error cost, $\mathrm{E}\, U_d = 0$ for partitions $\mathcal{P}$ and $\mathcal{Q}$ with the same number of elements. An impossibility theorem would hold if

$$\frac{1}{\sqrt{D}} \sum_{d=1}^{D} (Z_d - \mathrm{E}\, U_d) \xrightarrow{D} \mathcal{N}\big(0, V^2\big) \tag{38}$$

for some $V^2 > 0$. Note that the comparison of two partitionings is done in terms of $\sum_{d=1}^{D} Z_d$ using (36). This means that if the expected value of the bias correction term, U_d, is not zero, then the comparison between $\mathcal{P}$ and $\mathcal{Q}$ is eventually dominated by the bias, independently from the informative components. If the bias is zero, as in the squared error cost, then the comparison reduces to a purely random result with $\xi_D = P(\sum_{d=1}^{D} Z_d \geq 0) \to \frac{1}{2}$ as in the proof of Theorem 3.2.

We comment that Berry-Esseen bounds to control the behavior of ξ_D under general, separable cost functions can often be found. More specifically, if an

impossibility result holds and the third moment of $f_{\mathcal{P}}(\mathbf{Y}_d)$ exists, theorem 4.4 applies, with the quantities a_D, b_D, ρ_d, and σ_d^2 determined using the more general differences of cost functions. These quantities can be determined empirically; we do this for the illustrative plots in section 7.

To establish the sort of impossibility result suggested by (38), we first formalize the technical requirements on Z_d, the class of component-wise cost functions, and the distribution of the noise. Denote a class of component-wise cost functions by $\mathcal{F}$, with $f_{\mathcal{P}}, f_{\mathcal{Q}} \in \mathcal{F}$, and consider the following list of six conditions.

A1. *VC-subgraph condition*: $\mathcal{F}$ is a VC-subgraph class of functions.

A2. *Envelope function*: $\mathcal{F}$ has a square integrable envelope function.

A3. *Uniform Lindeberg condition*: Write $\|U_d\|_{\mathcal{F}^2}$ to mean the supremum norm of the cost difference on pure noise, $U_d = U_d(f, g)$, in $\mathcal{F}^2 = \mathcal{F} \times \mathcal{F}$. We require

$$\frac{1}{D}\sum_{d=1}^{D} \mathrm{E}\, \|Z_d\|_{\mathcal{F}^2} \mathbf{1}_{\{\|Z_d\|_{\mathcal{F}^2} > \varepsilon\sqrt{D}\}} \;\to\; 0 \;\forall\; \varepsilon > 0.$$

A4. *Covariance process convergence*: The covariance process $Z_d(f_1, g_1)Z_d(f_2, g_2)$ converges point-wise on $\mathcal{F}^2 \times \mathcal{F}^2$.

A5. *Existence of a semi-metric*: There exists a semi-metric $\rho\,[(f_1, g_1), (f_2, g_2)]$ defined on pairs of functions $(f_1, g_1), (f_2, g_2) \in \mathcal{F}^2$, such that, for every sequence $\delta_D \downarrow 0$,

$$\sup_{\rho[(f_1,g_1),(f_2,g_2)]\le\delta_D} \sum_{d=1}^{D} \mathrm{E}\,(Z_d(f_1, g_1) - Z_d(f_2, g_2))^2$$

goes to zero.

A6. *Measurability*: The completion of the product space on which $(Z_1, Z_2, ..., Z_D)$ is defined is measurable.

Since some of these conditions may be unfamiliar we provide some details. First, for A1, the VC condition, the subgraph of a function f is the set $\{(x, t) : t < f(x)\}$. A collection of functions is called a VC-subgraph class if the collections of all subgraphs from functions $f \in \mathcal{F}$ forms a VC class of sets, i.e., has finite VC dimension. For instance, if f is indexed by all the possible hard partitionings of the data, the class of functions is finite and condition A1 is trivially satisfied. Instead, if f is indexed by all soft partitionings of the data, additional work must be done to verify this assumption. Sufficient conditions for a class of functions being a VC-subgraph class is found in [22] or [23].

Second, A2 is a standard assumption from the theory of empirical processes commonly used to ensure bounds can be taken uniformly over the functions within the envelope. This means that the theorem we state below will be uniform over classes of cost functions rather than for individual cost functions as in Theorem 3.2; for non-finite classes of cost functions, a uniform bound is required to ensure the result holds even under optimization.

Third, condition A3 is stated for the difference in costs, but in fact it is enough for it to hold for individual cost functions. This is analogous to the the condition used in Theorem 3.1.

Items A4, A5, and A6 are largely technical. Covariance processes must converge so that the normal limit can be identified. A metric, not just a semi-metric, can be defined for hard clusterings by any one of a number of clustering indices such as the Rand index [24] or the Jacard index. In these cases, ρ in item 5 can be phrased in terms of such a divergence m as

$$\rho\,[(f_{\mathcal{P}_1}, f_{\mathcal{P}_2}), (f_{\mathcal{Q}_1}, f_{\mathcal{Q}_2})] = m(\mathcal{P}_1, \mathcal{Q}_1) + m(\mathcal{P}_2, \mathcal{Q}_2)$$

In the general case, which includes soft partitioning, such a semi-metric depends on the context. Finally, the measurability is a requirement for all the quantities to be well-defined. Note that weaker but more technical assumptions could replace the ones given; we refer the reader to [22] or [23] for the details. Now, we can state our general theorem.

Theorem 6.1. *Suppose that assumptions A1–A6 are*

satisfied. Then

$$\frac{1}{\sqrt{D}}\sum_{d=1}^{D}(Z_d - \mathrm{E}\,U_d) \xrightarrow{D} \mathcal{N}\big(0, V^2\big) \qquad (39)$$

where $V > 0$*, provided that* $\mathrm{Var}\,(U_d) > 0$ *and*

$$\frac{1}{\sqrt{D}}\sum_{d=1}^{D}\mathrm{E}\,(Z_d - U_d) \xrightarrow{\mathrm{P}} 0 \text{ as } D \to \infty.$$

Furthermore, let $T_D = (1/\sqrt{D})\sum_{d=1}^{D}\mathrm{E}\,U_d$*. If* $T_D \to \alpha \in [-\infty, +\infty]$ *as* $D \to \infty$*,*

$$\xi_D = P\left[D^{-1/2}\sum_{d=1}^{D} Z_d > 0\right] \xrightarrow{\mathrm{P}} \begin{cases} 0 & T_D \to +\infty \\ \Phi(\alpha/V) & T_D \to \alpha \\ 1 & T_D \to -\infty, \end{cases} \qquad (40)$$

where Φ *is the distribution function of a* $\mathcal{N}(0,1)$*. In particular, if* $\mathrm{E}\,U_d = 0$*, then* $\xi_D \to \frac{1}{2}$*.*

Note that (40) is a generalization of Theorem 3.2 to account for bias. The three limiting cases can be understood as follows. If T_D converges to α then the event in ξ_D gives a normal percentile because it is a 'central' value of $D^{-1/2}\sum_{d=1}^{D} Z_d$. However, if T_D diverges to $\pm\infty$, the limiting distribution will be overwhelmed by bias making $P\Big(D^{-1/2}\sum_{d=1}^{D} Z_d > 0\Big)$ go to zero or one. This theorem characterizes the behavior of the difference in cost functions. However, these cases depend on the asymptotic behavior of T_D. While this can often be determined empirically, as we do in section 7, it is usually difficult to characterize explicitly. The most important special case of this result is $\alpha = 0$, where $\Phi(0) = 1/2$ gives clustering impossibility.

Proof. Dealing formally with the technical details is beyond the scope of this paper; we therefore sketch a proof and refer the reader to [22] for a treatment of the technical aspects.

First write equation (39) as

$$\frac{1}{\sqrt{p}}\sum_{d=1}^{D}(Z_d - \mathrm{E}\,U_d) = \frac{1}{\sqrt{D}}\sum_{d=1}^{D}(Z_d - \mathrm{E}\,Z_d) + \frac{1}{\sqrt{D}}\sum_{d=1}^{D}\mathrm{E}\,(Z_d - U_d). \qquad (41)$$

Now, consider the first term. Under A1 - A6, the assumptions of the uniform central limit theorem for stochastic processes, as given in Theorem 2.11.1 of [22], are satisfied. (The entropy condition in Theorem 2.11.1 follows from A1 and A2; Theorem 2.6.7 gives sufficient conditions for a VC-class of functions to satisfy the hypotheses of Lemma 2.11.6 which gives the entropy condition. The other conditions of Theorem 2.11.1 are items A3 – A6) Thus, $D^{-1/2}\sum_{d=1}^{D}(Z_d - \mathrm{E}\,Z_d)$ converges in distribution to a mean zero Gaussian random variable and the nonzero variance of U_d implies that this distribution is non-degenerate. Thus, (39) follows.

More generally, it is seen that the second term in (41) determines the location of the overall limit. So, (40) follows immediately. □

Note that Theorem 6.1 is a variant on (31) and again, Berry-Esseen bounds to control the behavior of ξ_D can be obtained by computing the quantities a_D, b_D, ρ, and σ^2 in Theorem 4.4.

6.3 Impossibility for $L^{r,s}$ Cost Functions

We illustrate Theorem 6.1 using the class of cost functions defined in (29),

$$f_{\mathcal{P}}(\mathbf{x}) = \sum_{P\in\mathcal{P}} \min_{x^*\in\mathbb{R}} \|\mathbf{x}_P - x^*\|_r{}^s \qquad (42)$$

with the restriction $s \in \{1, 2\}$. Recall that $\mathbf{x}_P$ represents the $|P|$-dimensional vector of the components of $\mathbf{x}$ in partition element P.

We will derive bounds on $\mathrm{E}\,(Z_d - U_d)$ in Theorem 6.1, analogous to (10) in Theorem 3.2, then show some computational results for how the bias $\mathrm{E}\,U_d$

behaves for 3 different cost function structures. Our first result bounds the difference in costs of clustering one component of real data and the cost of clustering one component of noise when $s = 1$ in (42).

Lemma 6.2. *Let $r \in [1, \infty]$ and suppose $s = 1$ in (42). Then, we have that*

$$\mathrm{E}\,|f_{\mathcal{P}}(\mathbf{x}+\boldsymbol{\varepsilon}) - f_{\mathcal{P}}(\boldsymbol{\varepsilon})| \leq C f_{\mathcal{P}}(\mathbf{x}) \tag{43}$$

where C is a constant independent of $\mathbf{x}$.

Proof. See Appendix A4. □

To obtain an analogous result for $s = 2$, we use an adapted form of Lemma 2.1 in [25].

Lemma 6.3. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$. Then, for $r \in [2, \infty)$,*

$$\begin{aligned}\left| \|\mathbf{a}+\mathbf{b}\|_r{}^2 - \|\mathbf{a}\|_r{}^2 \right| &\leq 2\left|\mathbf{a}^T\mathbf{b}\right| + (r-1)\|\mathbf{b}\|_r{}^2 \\ &\leq 2d\,\|\mathbf{a}\|_1\,\|\mathbf{b}\|_r + (r-1)\|\mathbf{b}\|_r{}^2\end{aligned}$$

Proof. See Appendix A5. □

Now, we can state bounds parallel to (43) but for $s = 2$.

Lemma 6.4. *Suppose $r \in [2, \infty)$ and*

$$f_{\mathcal{P}}(\mathbf{x}) = \sum_{P\in\mathcal{P}} \min_{x^*\in\mathbb{R}} \|\mathbf{x}_P - x^*\|_r{}^2,$$

where $\mathbf{x}_P$ denotes the components in $\mathbf{x}$ restricted to the partition element P. Then

$$\begin{aligned}&\mathrm{E}\,|f_{\mathcal{P}}(\mathbf{x}+\boldsymbol{\varepsilon}) - f_{\mathcal{Q}}(\boldsymbol{\varepsilon})| \\ &\leq \sum_{P\in\mathcal{P}} \min_{x^*\in\mathbb{R}} \left(C_P\,\|\mathbf{x}_P - x^*\|_r + D_P\,\|\mathbf{x}_P - x^*\|_r{}^2\right) \\ &\leq C\sqrt{f_{\mathcal{P}}(\mathbf{x})} + D f_{\mathcal{P}}(\mathbf{x})\end{aligned}$$

where C_P, D_P, C, and D are identifiable constants depending on $\mathcal{P}$, $\mathrm{E}\,\|\tilde{\varepsilon}_P\|_1$ and r but independent of $\mathbf{x}$.

Proof. See Appendix A6. □

Now, bounds on the rates, comparable to those in Theorem 3.2 are as follows.

Theorem 6.5. A) *Let $s = 1$ and $r \in [s, \infty]$. If $|f_{\mathcal{P}}(\mathbf{x})| \in o\left(\sqrt{D}\right)$ and $|f_{\mathcal{Q}}(\mathbf{x})| \in o\left(\sqrt{D}\right)$ i.e.,*

$$\frac{1}{\sqrt{D}}\sum_{d=1}^{D} f_{\mathcal{P}}(\mathbf{x}) \;\to\; 0 \quad \text{and} \quad \frac{1}{\sqrt{D}}\sum_{d=1}^{D} f_{\mathcal{Q}}(\mathbf{x}) \;\to\; 0,$$

then $P(\mathsf{cost}(\mathbf{x}, \mathcal{P}) > \mathsf{cost}(\mathbf{x}, \mathcal{Q}))$ either has no limit or converges to a constant determined entirely by the bias of the cost functions. If the bias is 0, then this constant is ½.

B) *Let $s = 2$ and $r \in [s, \infty)$. If, in addition,*

$$\frac{1}{\sqrt{D}}\sum_{d=1}^{D} \sqrt{f_{\mathcal{P}}(\mathbf{x})} \;\to\; 0 \;\text{ and }\; \frac{1}{\sqrt{D}}\sum_{d=1}^{D} \sqrt{f_{\mathcal{Q}}(\mathbf{x})} \;\to\; 0$$

then, again, $P(\mathsf{cost}(\mathbf{x}, \mathcal{P}) > \mathsf{cost}(\mathbf{x}, \mathcal{Q}))$ either has no limit or converges to a constant determined entirely by the bias of the cost functions. If the bias is 0, then this constant is ½.

Proof. Note that

$$\begin{aligned}\mathrm{E}\,|Z_d - U_d| \leq\; &\mathrm{E}\,|f_{\mathcal{P}}(\mathbf{x}_d + \boldsymbol{\varepsilon}_d) - f_{\mathcal{P}}(\boldsymbol{\varepsilon}_d)| \\ &+ \mathrm{E}\,|f_{\mathcal{Q}}(\mathbf{x}_d + \boldsymbol{\varepsilon}_d) - f_{\mathcal{Q}}(\boldsymbol{\varepsilon}_d)|\end{aligned}$$

The rest follows immediately from lemmas 6.2 and 6.3 and theorem 6.1. □

6.4 Computing the Bias

It remains to demonstrate the behavior of the bias term and bias correction on a few illustrative examples. Here we consider three $L^{r,s}$ cost functions, $(r, s) = (1, 1), (2, 2),$ and $(5, 2)$, and evaluate

$$\mathrm{E}\,U_d = \mathrm{E}\,(f_{\mathcal{P}}(\boldsymbol{\varepsilon}_d) - f_{\mathcal{Q}}(\boldsymbol{\varepsilon}_d)) \tag{44}$$

empirically to show their relative behavior. Notationally, we drop the d since our noise is i.i.d. and we are looking at the bias correction component-wise.

Let $n = 100$ and generate random noise from a one-dimensional $\mathcal{N}(0, 1)$ distribution. Let $\mathcal{Q}$ be the partition of these outcomes into two equisized sets randomly. We compare this $\mathcal{Q}$ to the partitions $\mathcal{P}_i$ formed by randomly assigning i points to one partition element and the other $n - i$ points to a second

partition element. Random assignment is reasonable since we are looking only at noise. It is enough to d this for $i = 1, \ldots, 50$ since $i = 51, \ldots, 100$ will giv symmetric results.

Panel (a) in Figure 1 shows the results of averag ing (44) over 10^7 runs. It is seen that, as expectec the $L^{2,2}$ cost function has no bias. Surprisingly, th bias curve for the $L^{5,2}$ cost function is negative, in dicating $\mathrm{E}\, f_{\mathcal{P}_i}(\boldsymbol{\varepsilon}_d) < \mathrm{E}\, f_{\mathcal{Q}}(\boldsymbol{\varepsilon}_d)$. This means that th equisized partition is strongly disfavored on average Informally, note that the spread of a random sam pling of half the elements is far more representativ of the spread of the full data. Since the $L^{r,s}$ cost function for larger r puts more weight on the outer points, one would expect a more compact partition to have significantly lower cost. Thus what we see is the expected behavior.

The $L^{1,1}$ cost function has quite different behavior. First, it has both positive and negative bias depending on whether the i in $\mathcal{P}_i$ is odd or even. When i is odd, $\mathcal{P}_i$ is favored; when i is even, the equisized partition $\mathcal{Q}$ is favored.. The reason this occurs is that the optimization to find a median is not necessarily uniquely defined. For n even, any value between the middle two values is a valid median; when n odd the middle value is unique. (We ignore the case of multiplicity since this happens with zero probability.) So, the cost of a partition with an even number of elements may be identical in cost to the same partition with one additional element added within the range of possible medians. Thus, on average, partitions with an odd number of elements are slightly favored.

As a second evaluation of the effect of the bias correction, we compare the behavior of ξ_D when it is not included and when it is. Suppose we generate a one-dimensional data set with $n = 100$ data points by combining 50 data points from each of a $\mathcal{N}(-0.5, 0.25^2)$ and a $\mathcal{N}(0.5, .25^2)$. Then, we augment the data by adding $D-1$ extra noise dimensions using outcomes of a $\mathcal{N}(0,1)$ so that the two dimensional data points can be regarded as D dimensional data. In fact, there is only one informative dimension since the other dimension merely has a tighter distribution. Again, let $\mathcal{P}$ be the equisized partition with two partition elements obtained by random assignment, but this time let $\mathcal{Q}$ be the random par-

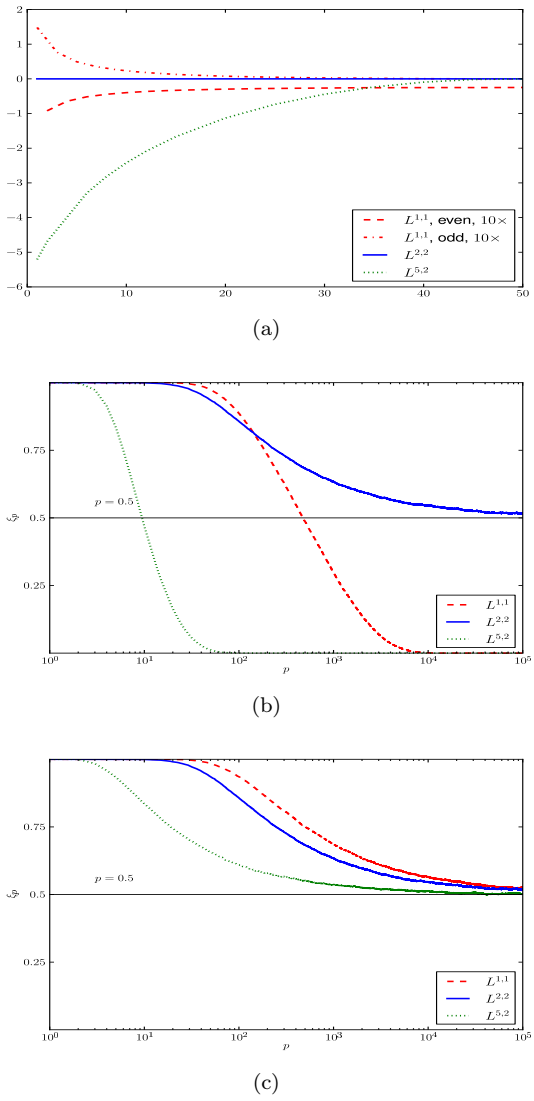


Figure 1: Panel (a) plots (44) for $\mathcal{P}$ and $\mathcal{Q}_i$. The dotted line is for $(r, s) = (2, 2)$ and the dot-dash line is for $(r, s) = (5, 2)$. The upper solid line is for $(r, s) = (1, 1)$ and i odd and the lower solid line is for i even. Panel (b) shows the behavior of ξ_D as D increases when the bias is not taken into account and panel (c) shows ξ_D when the bias is taken into account.

tition into a 65-35 split. Since our goal is to show impossibility, the specific choice of partitions is not important.

Panel (b) in Figure 1 does not correct for the bias and effectively assumes it is zero, simply comparing the cost of the two clusterings. Panel (c), however, shows the results of plotting

$$\xi_D = \mathrm{P}(\mathsf{cost}(\mathbf{Y}, \mathcal{P}) - \mathsf{cost}(\mathbf{Y}, \mathcal{Q}) \geq \mathrm{E}\, U_d),$$

where the $\mathrm{E}\, U_1$ accounts for the bias. The costs are as in (42) for $(r, s) = (1, 1), (2, 2),$ and $(5, 2)$. In the presence of bias, the curve for $(r, s) = (5, 2)$ drops to zero very quickly; it is clear that the bias has a significant affect. The unequal partitioning is favored, but this is due only to the noise and properties of the cost function. The curve for $(1, 1)$, which has significantly less bias, drops to zero more slowly. Note that we are comparing the equisized partitioning against one with two partitions that each have an odd number of data points; as expected, the latter is increasingly favored by the $L^{1,1}$ cost function. The $(2, 2)$ case, which has no bias, decays to $\frac{1}{2}$ as predicted.

7 COMPUTATIONS

In this section we present two computed examples. Both illustrate how the phenomenon established in Theorem 3.2 and Theorem 6.1 comes into effect for different choices of clustering. That is, the dimension D of a set of n vectors is allowed to grow and the difference in costs of one clustering over another is calculated repeatedly so that a curve $\xi = \xi_D$ can be given. In both cases, the number of informative dimensions is much smaller than the apparent D, a sort of sparsity common in many data classes. The first example uses simulated data and the second uses gene expression data analyzed in [3].

7.1 Simulations

As a first example of how clustering impossibility can be seen, consider the following simple scenario. Suppose a 2-dimensional data set of size $n = 120$ is generated by taking 40 i.i.d. data points from each of three bivariate normal distributions. These normals have means $(-0.5, 1)$, $(0.5, 1)$ and $(0, -0.75)$ with covariance matrices given by $\mathrm{diag}(.2^2, .25^2)$, $\mathrm{diag}(.15^2, .25^2)$ and $\mathrm{diag}(.45^2, .35^2)$, respectively. The top panels of Fig. 1 show three ways to cluster the data. Panel (a) shows the correct clustering, $\mathcal{P}_{\mathrm{best}}$. Panel (b) shows $\mathcal{P}_{\mathrm{bad}}$, an incorrect clustering: The bottom cluster is correct, but the top two clusters are split along a line at $y = 1$ rather than on any vertical line which would visibly separate them. Panel (c) shows an even worse clustering, $\mathcal{P}_{\mathrm{random}}$, a random assignment of the data into three clusters.

These bivariate data are extended to data of dimension $D = 3, 4, \ldots$ by adding $D - 2$ coordinates that were pure noise and hence uninformative. This was done in three ways. The simplest is to add coordinates that are $\mathcal{N}(0, 1)$. However, we also used two other noise distributions, a mean-shifted χ^2_2, i.e., $\chi^2_2 - 2$, and a Student's-t with 4 degrees of freedom, to see any effect from heavier tails.

Fig. 2 shows our computation of ξ_D for six scenarios: We look at two different comparisons of partitions under the three types of noise terms for the squared error cost function. Figs. 3 and 4 show the same six scenarios for two more cost functions, L^1 and L^5 for which cases we used the bias correction procedure discussed in subsections 6.2 and 6.4.

For the squared error cost, we calculated ξ_D for each value of D using the Monte Carlo simulation identified in Theorem 4.1 and 4.3. For speed of computation, we treated the difference in costs for the noisy components, $Z_d = \mathsf{cost}(\mathcal{P}, \varepsilon) - \mathsf{cost}(\mathcal{Q}, \varepsilon)$ as a random variable and estimated its distribution using a pool of 10^7 samples, each one from an i.i.d. draw of ε. Once this empirical distribution function was computed, we could quickly sample N values of Z_d, $Z_{d,1}, \ldots, Z_{d,N}$ until $\frac{1}{N}\sum_{j=1}^{N} \mathbb{I}\Big[\sum_{d=1}^{D} Z_{d,j} \geq 0\Big]$ converged (we chose $N = 50000$); this gave our estimate of $\xi_D = \mathrm{P}\Big(\sum_{d=1}^{D} Z_d \geq 0\Big)$ for D between 1 and 10^5. The middle curves (solid lines) in panels (d)–(i) in Figure 2 are the curves of ξ_D that we found for a variety of scenarios using L^2. The curves in Figs. 1, 3 and 4 were obtained in a similar fashion.

In addition, we calculated bounds on the ξ_D curves using Theorem 4.4 by expediently taking $\alpha' = 0$ in

(20). That is, we found $\hat{\sigma}$ and $\hat{\rho}$ empirically and used them to estimate a_D and b_D. The use of empirical estimates let us get around the sixth moment constraint in Theorem 4.4 for the t_4 to see what the results would look like. The results are shown as the lighter lines bracketing the central line in Panels (d)–(i). In effect, the vertical distance between the two lines for any given D is a sort of 'confidence interval' for ξ_D.

For the L^1 and L^5 cost functions, we performed analogous simulations including obtaining the Berry-Esseen bounds as discussed briefly at the end of subsection 6.2.

Overall, Figure 2 suggests the following. While $\mathcal{P}_{\text{bad}}$ is clearly suboptimal, once the number of noise terms is large enough, around $D = 100$, it becomes unreasonable to declare $\mathcal{P}_{\text{bad}}$ as worse than $\mathcal{P}_{\text{best}}$ – unless the noise dimensions are removed. While it is easier to distinguish between $\mathcal{P}_{\text{random}}$ and $\mathcal{P}_{\text{best}}$, ξ_D still gets close enough to $\frac{1}{2}$ by the time $D = 1000$ to cause problems.

The comparison of the noise terms is broadly consistent with intuition. In the case of noise from a standard normal distribution (see panels d and g), ξ_D is essentially one until $D \approx 100$ and drops to $1/2$ slower than for the other two noise distributions. That is, with normal noise, the cost function is able to compare two clusterings quite well, up to several hundred dimensions. In the case of the mean-shifted χ^2 with 2 degrees of freedom, ξ_D drops very quickly (see panel (e)), possibly due to the asymmetry. The decay of ξ_D shows that we can distinguish between $\mathcal{P}_{\text{best}}$ and $\mathcal{P}_{\text{random}}$ better than between $\mathcal{P}_{\text{best}}$ and $\mathcal{P}_{\text{bad}}$, but for $D \approx 1000$ these two clusterings are indistinguishable.

If the noise comes from a student-t distribution, the corresponding results comparing $\mathcal{P}_{\text{bad}}$ to $\mathcal{P}_{\text{best}}$ and $\mathcal{P}_{\text{bad}}$ to $\mathcal{P}_{\text{random}}$ (see panels (f) and (i)) are a little worse than those for the normal (left column) but a little better than for the shifted χ^2_2 (middle column). We suggest this occurs because the asymmetry of the χ^2_2 with exponential tails provides more distortion than the symmetric t- distribution does even though it has heavier tails.

The Berry-Esseen bounds reflect the true value of ξ_D quite well for large D, particularly for normal noise. Indeed, in looking at Theorem 4.4, it can be seen that a_D gives the midpoint of the interval while b_D controls the interval width. It can be seen in Figure 2 that the midpoints track the solid line closely and that the interval width narrows as D increases. When the interval is narrow and the solid line is near $1/2$, we can be quite sure that the clusterings are essentially indistinguishable and this happens before $D = 1000$ when there are two informative components.

Figs. 3 and 4 give qualitatively similar conclusions. However, comparing the second row of panels in Fig. 2 to the second rows of Figs. 3 and 4 reveals that convergence to $1/2$ is a little slower for L^1 and a little faster for L^5. That is, the L^1 cost function does not provide as rapid deterioration to noninformativity of clustering with increasing D as L^2 does while L^5 provides faster deterioration to noninformativity than L^2 does. This is broadly consistent with the fact that L^1 treats all distances equally while L^5 tends to be more affected by large differences. Otherwise put, L^1 is more robust than L^2 or L^5 in the sense that it is less seems to be less sensitive to the noise.

It should be remembered that this example is highly favorable to cost function based evaluation (not least because $\mathcal{P}_{\text{bad}}$ and $\mathcal{P}_{\text{random}}$ are so far wrong). So, this example indicates about the best performance they can give. This best performance is actually very good because correctly responding, even coarsely, to two informative components out of 1000 apparent components is pretty impressive. However, as will be seen in the next section, real examples typically provide greater challenges.

7.2 Pleural Mesothelioma Transcription Profile Data

To compare with the simulated results, we examined how our results are borne out in the cluster analysis of a data set consisting of gene expression profiles for malignant pleural mesothelioma patients[1]. Briefly, pleural mesothelioma is a specific kind of lung cancer mostly caused by asbestos exposure. Some 3000

[1]This data set is publicly available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE2549, where it is described in detail.

cases per year occur in the US and it is difficult to provide a differential diagnosis before death. There are at least three subtypes of pleural mesothelioma, however, the distinct subtypes do not satisfactorily relate to patient survival.

In this data set, $n = 54$ and each subject has had an expression array of 22283 transcripts collected. Of the 54, 40 had the disease, the other 14 were controls. [3] used a variety of dimension reduction techniques (mostly thresholding of marginal SD's; this is common in the medical literature) to reduce the 22283 components to the 1405 components believed to be most relevant to predicting the presence of pleural mesothelioma. Then, [3] did a hierarchical cluster analysis on the 1405 components. Among the 40 subjects, they identified three categories of tumor expression profiles, call them C_1 with 17 data points, C_2 with 14 data points and C_3 with 9 data points. The first two were relatively clear-cut clusters; C_3 was 'everything else'. Compared to the controls, C_1 had 56 genes with elevated expression and C_2 has 57 genes with elevated expression; for present purposes we ignore genes with decreased expression. Denote the two sets of genes by G_1 and G_2. Since $G_1 \cap G_2 = \emptyset$, 113 components can be regarded as informative.

Next we consider what conclusions from clustering can be extracted from this dataset under a set of hypothetical but plausible investigations. We chose this data set because C_1 and C_2 are defined and easily separated clusters providing a practically relevant "worse case scenario" in which to apply our results in contrast to the theoretically relevant simulation in subsection 4.1.

We began our analysis by assuming the genes in $(G_1 \cup G_2)^c$ were sufficiently uninformative as to be regarded as noise. To ensure this assumption was satisfied, we examined each of the $1292 = 1405 - 113$ components using a Mann-Whitney test to ensure it had no power to separate C_1 from C_2. (We comment that this marginal test is only a little better than the t-test that is standard; as a generality, marginal tests will not detect collections of components that are individually weak but collectively highly predictive.) Thus, for each of the 1292 genes, we used a Mann-Whitney test to see if a nonzero shift parameter relationship between C_1 and C_2 existed; this is slightly more general than the usual two-sample t-test for a difference in means and is more appropriate in the presence of non-normal data since it is nonparametric. Among the 1292 components we found that about 12% had a p-value of less than .05, a standard cutoff; these we discarded. Since our goal was to find data with which to generate a suitable noise distribution we did not do a multiple comparisons version of the test. After throwing out the potentially informative components, we pooled the remaining data to create an empirical distribution function which we took as our noise distribution. We comment that there are numerous other selection procedures we might have used so our results are conditional on the choices we made. Other variable selection choices are possible, but the impossibility theorem would still hold.

Next, we limited our data analysis to the squared error cost because the results from subsection 7.1 indicated that changing the cost function within the L^r class did not make a big difference. So, we ordered the genes in $G_1 \cup G_2$ by their contribution to the squared error cost function, i.e., by how well they separated C_1 and C_2 in squared error. Then, we defined clusterings $\mathcal{P}_{\text{random}}$ which uses the same subjects as (C_1, C_2) but randomly re-assigns them to two clusters (as in subsection 7.1), and $\mathcal{P}_{\text{flip}}$ in which 5 randomly chosen subjects are flipped from one of C_1 or C_2 to the other.

Now, from the ordered 113 genes, we have 31 vectors. We augment these to increase their dimension by adding samples from the empirical of the noise distribution. Thus, we can consider a sequence of clustering problems, of increasing dimensions with decreasing informativity, in which the clusters themselves are defined throughout.

Figure 5 shows the graphs of ξ_D as a function of D for six cases for comparing the clustering (C_1, C_2) to the clusterings $\mathcal{P}_{\text{flip}}$ and $\mathcal{P}_{\text{random}}$. The six cases correspond to how many of the most informative components are used for the clustering; call this c. Values of $c = 5, 30$ and 113 were chosen. For instance, when $c = 5$, the cost function uses the 5 most informative components and outcomes from the estimate of the noise distribution are added one at a time in each component to generate the curve over D. The other values of c are similar and indicated by the vertical

lines in each panel.

It can be seen that as the number of informative components increases, the curve for ξ_D and the empirical Berry-Esseen bounds shifts to the right. Nevertheless, when $c = 5$, by the time $D \approx 300$, ξ_D has dropped nearly to .8 for both comparisons. When $c = 30$, ξ_D drops to .8 when $D \approx 1,800$. However, when $c = 113$, $\xi_D > .8$ when $D \approx 10^5$. That is, if we regard .8 as a minimal standard for the believability of a clustering – a very generous allowance – we see that c/D for these three cases is .0167, $30/1800 = .0167$ and $113/10^5 = .001$. Roughly, this suggests that higher values of D can tolerate a lower value of c/D and still give decent performance. That is, for a given value of ξ_D, the required value of c for good performance increases with D sub-linearly.

References

[1] Perou C, Sørlie T, Eisen M, van de Rijn M, Jeffrey S, Rees C, Pollack J, Ross D, Johnsen H, Akslen L, *et al.*. Molecular portraits of human breast tumours. *Nature* 2000; **406**(6797):747–752.

[2] Neve R, Chin K, Fridlyand J, Yeh J, Baehner F, Fevr T, Clark L, Bayani N, Coppe J, Tong F, *et al.*. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 2006; **10**(6):515–527.

[3] Gordon G, Rockwell G, Jensen R, Rheinwald J, Glickman J, Aronson J, Pottorf B, Nitz M, Richards W, Sugarbaker D, *et al.*. Identification of novel candidate oncogenes and tumor suppressors in malignant pleural mesothelioma using large-scale transcriptional profiling. *The American journal of pathology* 2005; **166**(6):1827–1840.

[4] Devroye L, Gyorfi L, Lugosi G. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

[5] Biau G, Devroye L, Lugosi G. On the performance of clustering in Hilbert spaces. *IEEE Transactions on Information Theory* 2008; **54**(2):781.

[6] Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J. R. S. S. Ser. B* 2005; **63-2**:411–423.

[7] Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When Is"Nearest Neighbor"Meaningful? *Lecture Notes in Computer Science* 1999; **1540**:217–235.

[8] Hinneburg A, Aggarwal CC, Keim DA. What is the nearest neighbor in high dimensional spaces? *The VLDB Journal*, 2000; 506–515.

[9] Steinbach M, Ertöz L, Kumar V. The challenges of clustering high dimensional data. *Applications in Econophysics, Bioinformatics, and Pattern Recognition* 2003; .

[10] Hall P, Marron J, Neeman A. Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society series B* 2005; **67**:427–444.

[11] Murtagh F. The remarkable simplicity of very high dimensional data. *J. Class.* 2008; **26**(3):249–277.

[12] Biau L G Devroye, Lugosi G. A graph based estimator of the number of clusters. *ESAIM: Probability and Statistics* 2007; **11**:272–280.

[13] Ding C, He X. K-means clustering via principal component analysis. *Proceedings of the twenty-first international conference on Machine learning*, ACM New York, NY, USA, 2004.

[14] Jin J. Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences* 2009; **106**(22):8859.

[15] Fan J, Fan Y. High Dimensional Classification Using Features Annealed Independence Rules. *Arxiv preprint math.ST/0701108* 2007; .

[16] Bickel P, Levina E. Some theory for Fisher's linear discriminant function,naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* 2004; **10**(6):989–1010.

[17] Dy J, Brodley C. Feature Selection for Unsupervised Learning. *The Journal of Machine Learning Research* 2004; **5**:845–889.

[18] Wolf L, Shashua A. Feature Selection for Unsupervised and Supervised Inference: The Emergence of Sparsity in a Weight-Based Approach. *The Journal of Machine Learning Research* 2005; **6**:1855–1887.

[19] Shiganov I. Refinement of the upper bound of the constant in the central limit theorem. *Journal of Mathematical Sciences* 1986; **35**(3):2545–2550.

[20] Dhillon I, Guan Y, Kulis B. A unified view of kernel k-means, spectral clustering and graph cuts. *Technical Report TR-04-25*, University of Texas at Austin, Department of Computer Sciences February 2005.

[21] Girolami M. Mercer kernel based clustering in feature space. *IEEE Trans. Neural Nets* 2002; **13**(4):780–784.

[22] van der Vaart A, Wellner J. *Weak Convergence and Empirical Processes*. Springer: New York, 1996.

[23] Dudley R. *Uniform Central Limit Theorems*. Cambridge University Press: Cambridge, 1999.

[24] Steinley D. Properties of the hubert-arabie adjusted rand index. *Psychol Methods* 2004; **9**(3):386–96.

[25] Duembgen L, van de Geer S, Veraar M. Nemirovski's inequalities revisited. *To appear: Amer. Math. Monthly* 2010; .

[26] Horn R, Johnson C. *Matrix Analysis*. Cambridge University Press: New York, 1985.

[27] Ash R, Doleans-Dade C. *Probability and Measure Theory*. Academic Press, 1999.

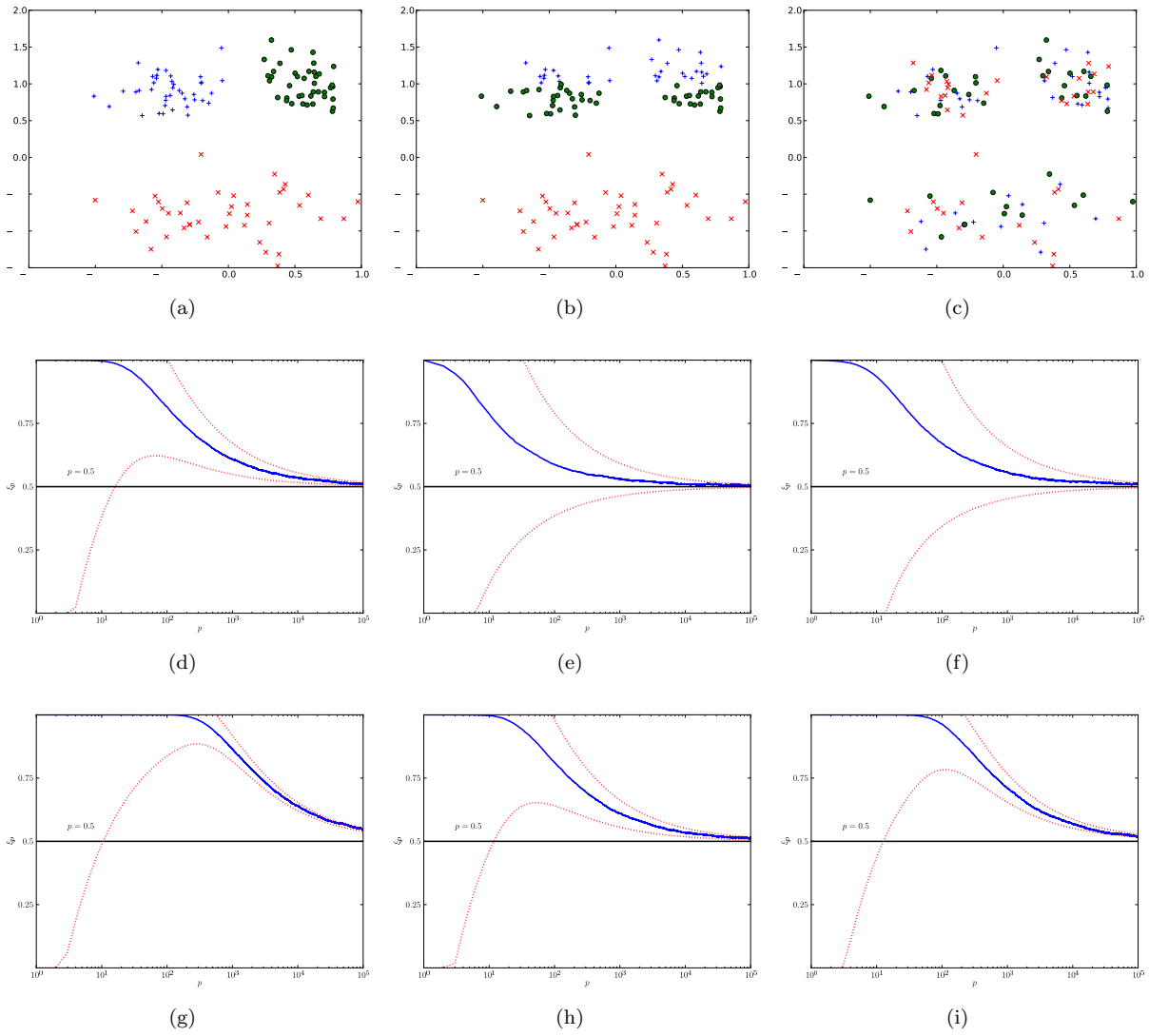
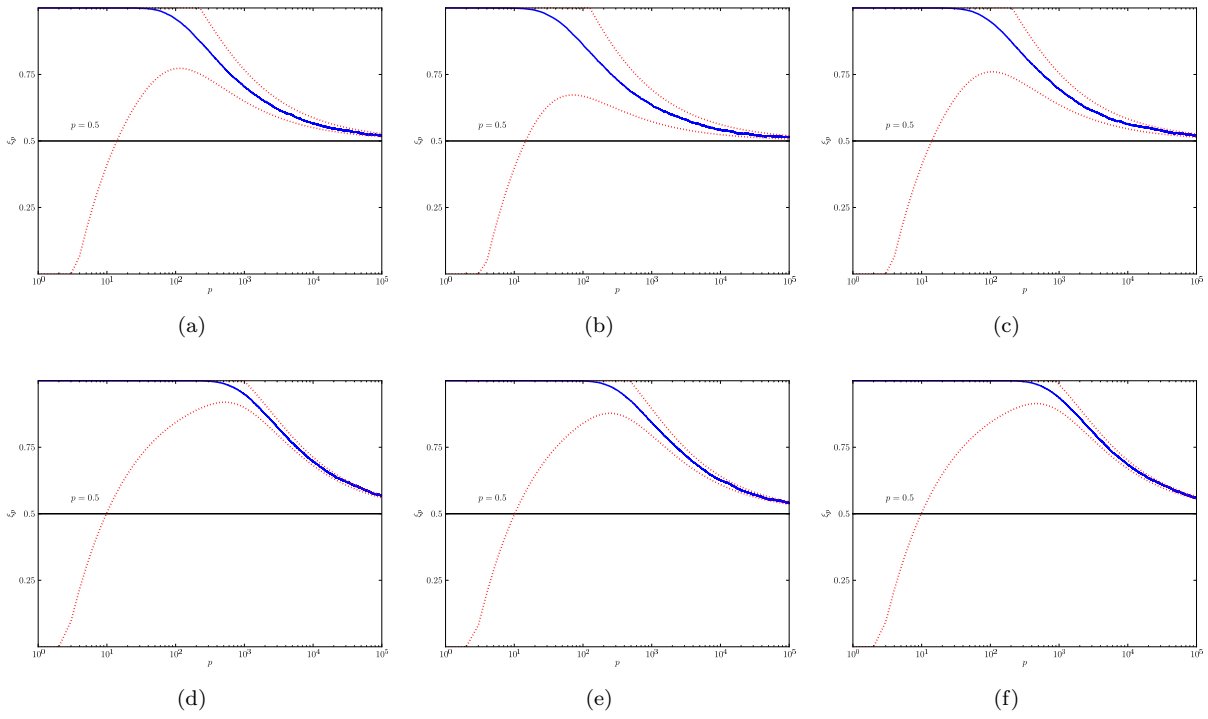


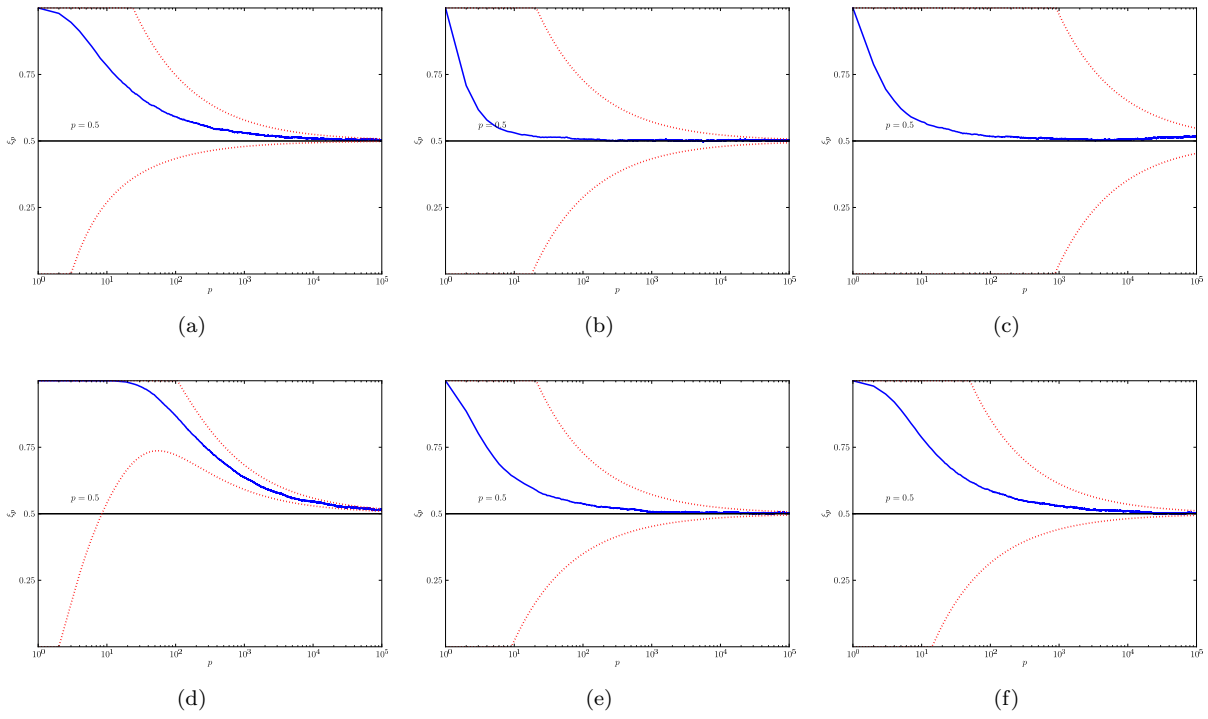
Figure 2: Bounds on how fast clusterings become impossible to compare as D increases. Two clusterings, $\mathcal{P}_{\text{bad}}$ (b) and $\mathcal{P}_{\text{random}}$ (c), are tested against the optimal clustering $\mathcal{P}_{\text{best}}$ (a). In the second and third rows, estimates of ξ_D as a function of D are plotted as a solid line. The corresponding bounds from theorem 4.4 are plotted as dashed lines. Panels (d), (e), and (f) show ξ_D for $\mathcal{P}_{\text{bad}}$ vs. $\mathcal{P}_{\text{best}}$ taking the standard normal, the mean-shifted chi-squared with 2 degrees of freedom and the student-t with 4 degrees of freedom as the noise models respectively. Panels (g), (h) and (i) show ξ_D for the "worst-case" scenario of $\mathcal{P}_{\text{random}}$ vs. $\mathcal{P}_{\text{best}}$ for the same noise distributions.



(a) (b) (c)

(d) (e) (f)

Figure 3: The comparison of Figure 1 extended to the $L^{1,1}$ cost function, with bias correction included.



(a) (b) (c)

(d) (e) (f)

Figure 4: The comparison of Figure 1 extended to the $L^{5,2}$ cost function, with bias correction included.

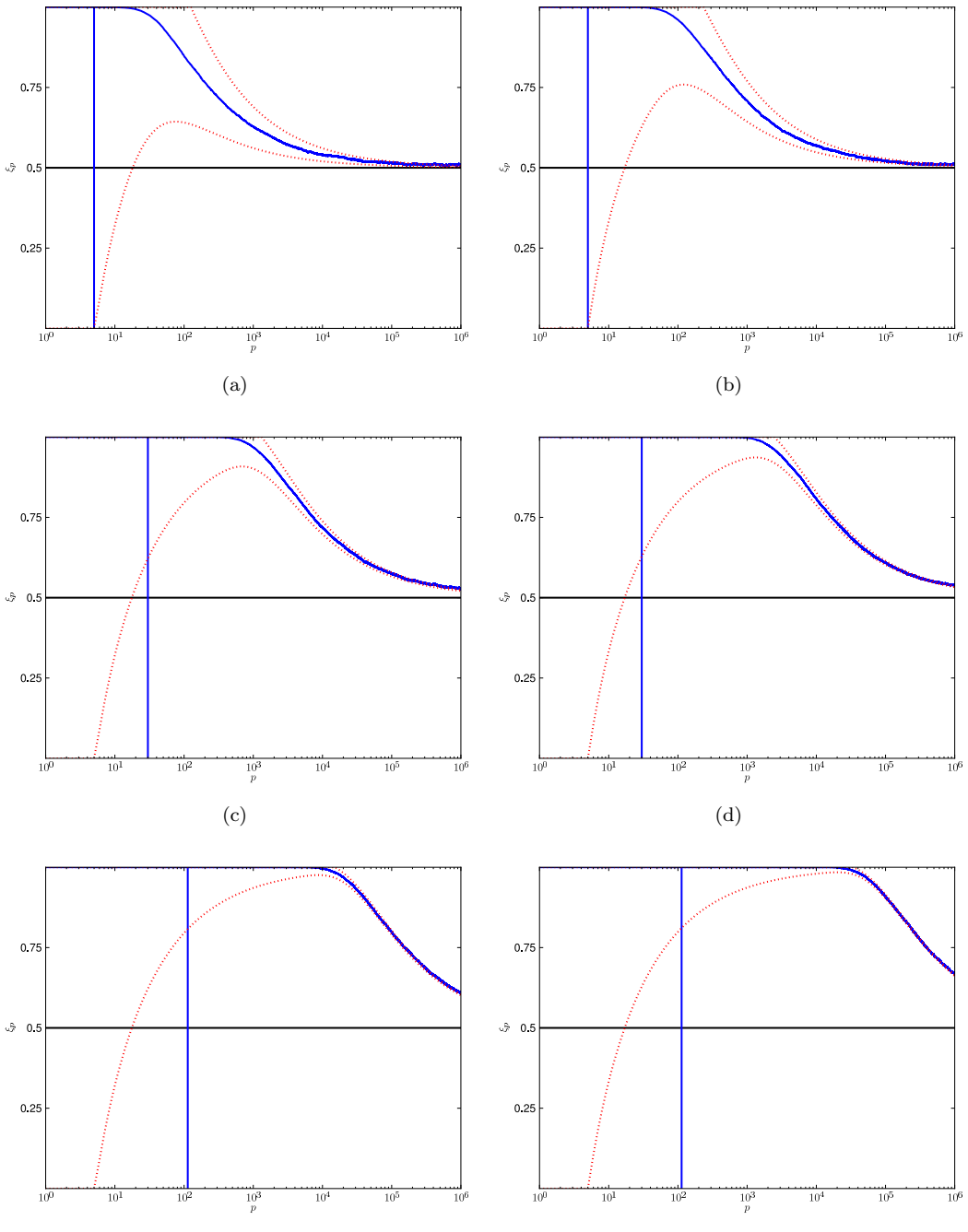


Figure 5: The probability of one clustering being detected as better than another under squared error cost for increasing number of included genes D. The clusterings being compared are the given (C_1, C_2) to $\mathcal{P}_{\text{flip}}$ (left column) and to $\mathcal{P}_{\text{random}}$ (right column). The three rows correspond to taking the first 5, 30 or 113 informative components and augmenting them by noise. In all plots, the solid blue curve denotes the empirically determined value of ξ_D, while the dotted red lines indicate the Berry-Esseen bounds from theorem 4.4.

A APPENDIX: ADDITIONAL PROOFS

A.1 Proof of theorem 2.4

Observe that $\mathbf{B}$ is symmetric and real. Thus, $\mathbf{B}$ has real eigenvalues and can be diagonalized with real (unitary) matrices. This proves part A. Now let $\mathbf{B}^{\mathcal{P}} = [b_{ij}^{\mathcal{P}}]$, where

$$b_{ij}^{\mathcal{P}} = \frac{1}{2\left|P_\ell\right|}\mathbf{1}_{\{i,j\in P_\ell\}}\,\mathbf{1}_{\{\nexists k \text{ s.t. } P_\ell = Q_k\}}\,,$$

and let $\mathbf{B}^{\mathcal{Q}}$ be similarly defined. Then from (8), we have that $\mathbf{B} = \mathbf{B}^{\mathcal{Q}} - \mathbf{B}^{\mathcal{P}}$. We first find the eigenvalues of $\mathbf{B}^{\mathcal{P}}$ and $\mathbf{B}^{\mathcal{Q}}$. $\mathbf{B}^{\mathcal{P}}$ can be permuted to form a block diagonal matrix of $K-M$ blocks $\mathbf{B}_1^\star, ..., \mathbf{B}_{K-M}^\star$, where M is the number of partitions in common between $\mathcal{P}$ and $\mathcal{Q}$. Now all the elements in block $\mathbf{B}_k^\star$ equal $(2\left|P_k\right|)^{-1}$. For the eigenvalue problem, we can treat each block separately. So

$$\begin{aligned}\nu_k^{\mathcal{P}}\mathbf{x}_k = \mathbf{B}_k^\star\mathbf{x}_k &= (2|P_k|)^{-1}\mathbf{1}\cdot\mathbf{1}^T\mathbf{x}_k \\ &= (2|P_k|)^{-1}(\Sigma\,\mathbf{x}_k, ..., \Sigma\,\mathbf{x}_k),\end{aligned}$$

where $\Sigma\,\mathbf{x}_k$ is the sum of its entries. It is seen that the eigenvalue equation is solved only when $\nu_k^{\mathcal{P}} = 0$ and $|\mathbf{x}_k| = 0$, or $\mathbf{x}_k = c$ for some constant c and

$$\nu_k^{\mathcal{P}}\,c = (2\left|P_k\right|)^{-1}(|P_k|\cdot c) = c \implies \nu_k^{\mathcal{P}} = 1/2.$$

Since the eigenvector corresponding to $\nu_k^{\mathcal{P}} = 1/2$ is constant, $\nu_k^{\mathcal{P}}$ has multiplicity 1. Thus there is exactly one nonzero eigenvalue associated with each block, for K total. The eigenvalues of $\mathbf{B}^{\mathcal{Q}}$ are similar, and it follows that $\text{rank}(\mathbf{B}^{\mathcal{P}}) = \text{rank}(\mathbf{B}^{\mathcal{Q}}) = K-M$. Now, $\text{rank}(\mathbf{B}) \le \text{rank}(\mathbf{B}^{\mathcal{Q}}) + \text{rank}(-\mathbf{B}^{\mathcal{P}}) = 2(K-M)$, so $\lambda_i = 0$ for $i = 2(K-M)+1, ..., n$ proving B and the second part of A. Furthermore, $\mathbf{B}^{\mathcal{P}}$ and $\mathbf{B}^{\mathcal{Q}}$ are symmetric and positive semi-definite so [26]

$$\max_i \lambda_i \le \left(\max_i(-\nu_i^{\mathcal{P}})\right) + \left(\max_i \nu_i^{\mathcal{Q}}\right) = 0 + 1/2$$

The analogous inequality for $-\mathbf{B} = \mathbf{B}^{\mathcal{P}} - \mathbf{B}^{\mathcal{Q}}$ gives us that $\max_i(-\lambda_i) \le 1/2 \implies \min_i \lambda_i \ge -1/2$, proving the interval bound in A. Finally,

$$\begin{aligned}\textstyle\sum_i \lambda_i &= \text{trace}\,(\mathbf{\Lambda}) = \text{trace}\left(\mathbf{U}^T\mathbf{B}\mathbf{U}\right) \\ &= \text{trace}\,(\mathbf{B}) = \text{trace}\left(\mathbf{B}^{\mathcal{Q}}\right) - \text{trace}\left(\mathbf{B}^{\mathcal{P}}\right) \\ &= \textstyle\sum_i (2\left|Q_\ell\right|)^{-1}\mathbf{1}_{\{i\in Q_\ell\}} - (2\left|P_k\right|)^{-1}\mathbf{1}_{\{i\in P_k\}} \\ &= K/2 - K/2 = 0\end{aligned}$$

which proves C, the final part of the theorem. □

A.2 Proof of theorem 2.6

To prove parts A and B, consider the following identity. Let $\mathbf{Z}$ be a random vector with $\text{E}\,(\mathbf{Z}) = \mathbf{m}$ and $\text{Var}\,(\mathbf{Z}) = \mathbf{V}$. Then, for any matrix $\mathbf{A}$,

$$\begin{aligned}\text{E}\left(\mathbf{Z}^T\mathbf{A}\mathbf{Z}\right) &= \text{E}\left(\text{trace}(\mathbf{A}\mathbf{Z}\mathbf{Z}^T)\right) \\ &= \text{trace}(\mathbf{A}\,\text{E}\left(\mathbf{Z}\mathbf{Z}^T\right)) \\ &= \text{trace}(\mathbf{A}\mathbf{V} + \mathbf{A}\mathbf{m}\mathbf{m}^T) \\ &= \text{trace}(\mathbf{A}\mathbf{V}) + \mathbf{m}^T\mathbf{A}\mathbf{m}.\end{aligned}$$

Now, let $\mathbf{Z} = \boldsymbol{\varepsilon}$ so that $\mathbf{m} = 0$ and set $\mathbf{V} = \mathbf{I}$, and $\mathbf{A} = \mathbf{B}$. By theorem 2.4 part E, $\text{trace}(\mathbf{B}) = 0$ giving $\text{E}\left(\boldsymbol{\varepsilon}^T\mathbf{B}\boldsymbol{\varepsilon}\right) = 0$, i.e. part A. So, $\text{E}\left(\mathbf{Y}^T\mathbf{B}\mathbf{Y}\right) = \mathbf{x}^T\mathbf{B}\mathbf{x}$ which is B..

To prove C, note that Jensen's inequality gives us that $\text{E}\,\varepsilon^4 \ge \left(\text{E}\,\varepsilon^2\right)^2$. Then

$$\begin{aligned}\text{Var}\left(\boldsymbol{\varepsilon}^T\mathbf{B}\boldsymbol{\varepsilon}\right) &\le \left(\textstyle\sum_{i,j} b_{ij}^2\right)\text{E}\,\varepsilon^4 = \|\mathbf{B}\|_{\text{Frob}}\,\text{E}\,\varepsilon^4 \\ &= \left(\text{E}\,\varepsilon^4\right)\textstyle\sum_i \lambda_i^2 \\ &\le K\left(\text{E}\,\varepsilon^4\right)/2\end{aligned}$$

where we use theorem 2.4 the fact that the Frobenious norm is invariant to unitary transformations to rewrite it as the sum of the squared eigenvalues. The last inequality follows by use of part D in theorem 2.4. There are at most ℓ nonzero eigenvalues and they are bounded by 1/2. □

A.3 Proof of theorem 3.1

Let F_d be the distribution function of $Z_d - r_d$. The proof follows from the Lindeberg-Feller conditions on

the central limit theorem [27]. Let F_d be the distribution function of Z_d, and let

$$\mu_D = \sum_{d=1}^{D} r_d$$

If, $\forall\, \varepsilon > 0$,

$$LF_D = \frac{1}{c_D^2} \sum_{d=1}^{D} \int\limits_{\{x\,:\,|x-r_d| \geq \varepsilon c_D\}} (x - r_d)^2 \mathrm{d}F_d(x) \to 0 \text{ as } D \to \infty,$$

then Lindeberg's theorem [27] gives

$$(S_D - \mu_D) \,/\, c_D \xrightarrow{D} \varphi \text{ as } D \to \infty.$$

We later show that the μ_D term drops out in the limit. Now let $s_D^2 = \mathrm{mean}_{d \in \{1,2,\ldots,D\}}\, \sigma_d^2$. Then

$$\begin{aligned} LF_D &= \frac{1}{D s_D^2} \sum_{d=1}^{D} \int\limits_{\left\{x\,:\,|x-r_d| \geq \varepsilon s_D \sqrt{D}\right\}} (x - r_d)^2 \mathrm{d}F_d(x) \\ &\leq \frac{1}{D s_D^2} \left[D \max_{d \in \{1,2,\ldots,D\}} \int\limits_{\left\{x\,:\,|x-r_d| \geq \varepsilon s_D \sqrt{D}\right\}} (x - r_d)^2 \mathrm{d}F_d(x) \right] \\ &\leq \frac{1}{L^2} \left[\max_{d \in \{1,2,\ldots,D\}} \int\limits_{\left\{x\,:\,|x-r_d| \geq \varepsilon L \sqrt{D}\right\}} (x - r_d)^2 \mathrm{d}F_d(x) \right] \end{aligned}$$

where we use the fact that σ_d is bounded below by L and that the integral is always positive to get the final step.

Now $\forall\, r_d$, $\left\{x : |x - r_d| \geq \varepsilon L \sqrt{D}\right\} \searrow \emptyset$ as $D \to \infty$, so $\forall\, d$,

$$\int\limits_{\left\{x\,:\,|x-r_d| \geq \varepsilon L \sqrt{D}\right\}} (x - r_d)^2 \mathrm{d}F_d(x) \to 0$$

$$\implies \frac{1}{L^2} \left[\max_{d \in \{1,2,\ldots,D\}} \int\limits_{\left\{x\,:\,|x-r_d| \geq \varepsilon L \sqrt{D}\right\}} (x - r_d)^2 \mathrm{d}F_d(x) \right] \to 0$$

as $D \to \infty$. Thus

$$\frac{S_D - \mu_D}{s_D \sqrt{D}} \xrightarrow{D} \varphi \text{ as } D \to \infty. \tag{45}$$

Note that this is equivalent to (45). However, by assumption,

$$\frac{1}{s_D} \frac{\mu_D}{\sqrt{D}} \to 0 \text{ as } D \to \infty,$$

so (45) reduces to

$$\frac{S_D}{c_D} \xrightarrow{D} \varphi \text{ as } D \to \infty.$$

proving the theorem. □

A.4 Proof of Lemma 6.2

To begin, let

$$\begin{aligned} x_P^* &= \operatorname*{argmin}_{t \in \mathbb{R}} \|\mathbf{x}_P - t\|_r \,, \\ \varepsilon_P^* &= \operatorname*{argmin}_{t \in \mathbb{R}} \|\boldsymbol{\varepsilon}_P - t\|_r \,, \\ z_P^* &= \operatorname*{argmin}_{t \in \mathbb{R}} \|\tilde{\mathbf{x}}_P + \tilde{\boldsymbol{\varepsilon}}_P - t\|_r \,, \end{aligned}$$

and define the r-th norm centered versions of the $\mathbf{x}_P$'s and $\boldsymbol{\varepsilon}_P$'s by

$$\tilde{\mathbf{x}}_P = \mathbf{x}_P - x_P^* \quad \text{and} \qquad \tilde{\boldsymbol{\varepsilon}}_P = \boldsymbol{\varepsilon}_P - \varepsilon_P^*.$$

So, the triangle inequality gives

$$\begin{aligned} &\mathrm{E}\,|f_{\mathcal{P}}(\mathbf{x} + \boldsymbol{\varepsilon}) - f_{\mathcal{P}}(\boldsymbol{\varepsilon})| \\ &\quad \leq \sum_{P \in \mathcal{P}} \mathrm{E}\,|\,\|\tilde{\mathbf{x}}_P + \tilde{\boldsymbol{\varepsilon}}_P - z_P^*\|_r \;-\; \|\tilde{\boldsymbol{\varepsilon}}_P\|_r\,|. \end{aligned}$$

Now, using $\vee$ to denote the binary maximum operator, we have

$$\begin{aligned} &|\,\|\tilde{\mathbf{x}}_P + \tilde{\boldsymbol{\varepsilon}}_P - z_P^*\|_r \;-\; \|\tilde{\boldsymbol{\varepsilon}}_P\|_r\,| \\ &= \left[\|\tilde{\mathbf{x}}_P + \tilde{\boldsymbol{\varepsilon}}_P - z_P^*\|_r - \|\tilde{\boldsymbol{\varepsilon}}_P\|_r\right] \vee \left[\|\tilde{\boldsymbol{\varepsilon}}_P\|_r - \|\tilde{\mathbf{x}}_P + \tilde{\boldsymbol{\varepsilon}}_P - z_P^*\|_r\right] \\ &\leq \left[\|\tilde{\mathbf{x}}_P - z_P^*\|_r + \tilde{\boldsymbol{\varepsilon}}_P - \|\tilde{\boldsymbol{\varepsilon}}_P\|_r\right] \\ &\qquad \vee \left[\|\tilde{\boldsymbol{\varepsilon}}_P + \tilde{\mathbf{x}}_P - z_P^*\|_r + \|-\tilde{\mathbf{x}}_P + z_P^*\|_r - \|\tilde{\mathbf{x}}_P + \tilde{\boldsymbol{\varepsilon}}_P - z_P^*\|_r\right] \\ &\leq \|\tilde{\mathbf{x}}_P - z_P^*\|_r \end{aligned} \tag{46}$$

Since z_P^* is the shift required to recenter $\tilde{\boldsymbol{\varepsilon}}_P$ after $\tilde{\mathbf{x}}_P$ is added to it component-wise, $|z_P^*| \leq \|\tilde{\mathbf{x}}_P\|_\infty$ because the maximum absolute value in $\tilde{\mathbf{x}}_P$ upper bounds the distance of such a shift. Thus, we have

$$\begin{aligned} \|\mathbf{1} z_P^*\|_r &\leq |P|^{1/r}\, |z_P^*| \\ &\leq |P|^{1/r}\, \|\tilde{\mathbf{x}}_P\|_\infty \\ &\leq |P|^{1/r}\, \|\tilde{\mathbf{x}}_P\|_r \,, \end{aligned}$$

where $|P|$ is the cardinality of the partition element P. So (46) is bounded as

$$\begin{aligned}\|\tilde{\mathbf{x}}_P - z_P^*\|_r &\leq \|\tilde{\mathbf{x}}_P\|_r + \|\mathbf{1} z_P^*\|_r \\ &\leq (1 + |P|^{1/r}) \|\tilde{\mathbf{x}}_P\|_r\end{aligned}$$

and taking the sum over partition elements and replacing the constant with $\max_{P\in\mathcal{P}}(1+|P|^{1/r})$ gives the Lemma. QED

A.5 Proof of lemma 6.3

This lemma is a special case of Lemma 2.1 in [25]. For $r \in [2,\infty)$, define $h(\mathbf{a}) : \mathbf{a} thbb R^d \mapsto \mathbb{R}^d$ by

$$h_i(\mathbf{a}) = \begin{cases} 2\|\mathbf{a}\|_r{}^{2-r} |a_i|^{r-2} a_i & \mathbf{a} \neq 0 \\ 0 & \mathbf{a} = 0 \end{cases}.$$

Then, for arbitrary $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, this lemma gives

$$\begin{aligned}\|\mathbf{a}\|_r{}^2 + h(\mathbf{a})^T\mathbf{b} &\leq \|\mathbf{a}+\mathbf{b}\|_r{}^2 \\ &\leq \|\mathbf{a}\|_r{}^2 + h(\mathbf{a})^T\mathbf{b} + (r-1)\|\mathbf{b}\|_r{}^2\end{aligned}$$

When $\mathbf{a} = 0$, the result is trivial, so assume $\mathbf{a} \neq 0$. Now, since $(r-1)\|\mathbf{b}\|_r \geq 0$,

$$\begin{aligned}&\left|\|\mathbf{a}+\mathbf{b}\|_r{}^2 - \|\mathbf{a}\|_r{}^2\right|^2 \\ &\quad\leq \left|h(\mathbf{a})^T\mathbf{b}\right| + (r-1)\|\mathbf{b}\|_r{}^2 \\ &\quad\leq 2\left|\left(\max_i \frac{|a_i|}{\|\mathbf{a}\|_r}\right)^{r-2} \mathbf{a}^T\mathbf{b}\right| + (r-1)\|\mathbf{b}\|_r{}^2\end{aligned}$$

However, $\|\mathbf{a}\|_r \geq \max_i |a_i|$, so the inequality simplifies to

$$\begin{aligned}\left|\|\mathbf{a}+\mathbf{b}\|_r{}^2 - \|\mathbf{a}\|_r{}^2\right| &\leq 2\left|\mathbf{a}^T\mathbf{b}\right| + (r-1)\|\mathbf{b}\|_r{}^2 \\ &\leq 2\|\mathbf{a}\|_1 \|\mathbf{b}\|_1 + (r-1)\|\mathbf{b}\|_r{}^2,\end{aligned}$$

Now $\|\mathbf{b}\|_1 \leq d \max_i |b_i| \leq d\|\mathbf{b}\|_r$, yielding the final inequality. □

A.6 Proof of lemma 6.4

Let x_P^*, ε_P^*, $\tilde{\mathbf{x}}_P$, $\tilde{\boldsymbol{\varepsilon}}_P$, and z^* be defined as in lemma 6.2. We can then proceed as follows:

$$\begin{aligned}&\mathrm{E}\,|f_{\mathcal{P}}(\mathbf{x}+\boldsymbol{\varepsilon}) - f_{\mathcal{P}}(\boldsymbol{\varepsilon})| \\ &\leq \sum_{P\in\mathcal{P}} \mathrm{E}\left|\|\tilde{\mathbf{x}}_P + \tilde{\boldsymbol{\varepsilon}}_P - z_P^*\|_r{}^2 - \|\tilde{\boldsymbol{\varepsilon}}_P\|_r{}^2\right| \\ &\leq \sum_{P\in\mathcal{P}} \mathrm{E}\left(A_P\|\tilde{\mathbf{x}}_P - z_P^*\|_r + (r-1)\|\tilde{\mathbf{x}}_P - z_P^*\|_r{}^2\right)\end{aligned}$$

Where we applied lemma 6.3 to get the last step, with $A_P = 2|P|\,\|\tilde{\boldsymbol{\varepsilon}}_P\|_1$. Now, from the proof of lemma 6.2, $\|\tilde{\mathbf{x}}_P - z_P^*\|_r \leq (1+|P|^{1/r})\|\tilde{\mathbf{x}}_P\|_r$, giving an upper bound of

$$\mathrm{E}\,|f_{\mathcal{P}}(\mathbf{x}+\boldsymbol{\varepsilon}) - f_{\mathcal{P}}(\boldsymbol{\varepsilon})| \leq \sum_{P\in\mathcal{P}} C_P\|\tilde{\mathbf{x}}_P\|_r + D_P\|\tilde{\mathbf{x}}_P\|_r{}^2,$$

where

$$\begin{aligned}C_P &= 2|P|\,\mathrm{E}\,(\|\tilde{\boldsymbol{\varepsilon}}_P\|_1)(1+|P|^{1/r}) \\ D_P &= (r-1)(1+|P|^{1/r})^2.\end{aligned}$$

This gives the first inequality in the Lemma.

Letting $D = \max_{P\in\mathcal{P}} D_P$ and bringing it outside the sum gives the first term in the second inequality and letting $C = |\mathcal{P}|\max_{P\in\mathcal{P}}\sqrt{C_P}$ allows us to write

$$\sum_{P\in\mathcal{P}} C_P\|\tilde{\mathbf{x}}_P\|_r \leq C\sqrt{\sum_{P\in\mathcal{P}}\|\tilde{\mathbf{x}}_P\|_r{}^2}.\square$$