When Is 'Nearest Neighbor' Meaningful: A Converse Theorem and Implications

Robert J. Durrant and Ata Kabán[1]

School of Computer Science, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK.
{R.J.Durrant, A.Kaban}@cs.bham.ac.uk

Abstract

Beyer et al. gave a sufficient condition for the high dimensional phenomenon known as the concentration of distances. Their work has pinpointed serious problems due to nearest neighbours not being meaningful in high dimensions. Here we establish the converse of their result, in order to answer the question as to when nearest neighbour *is* still meaningful in arbitrarily high dimensions. We then show for a class of realistic data distributions having non-i.i.d. dimensions, namely the family of linear latent variable models, that the Euclidean distance will not concentrate as long as the amount of 'relevant' dimensions grows no slower than the overall data dimensions. This condition is, of course, often not met in practice. After numerically validating our findings, we examine real data situations in two different areas (text-based document collections and gene expression arrays), which suggest that the presence or absence of distance concentration in high dimensional problems plays a role in making the data hard or easy to work with.

Key words: high dimensionality, distance concentration, latent variable models.

1 Introduction

In an influential paper, Beyer et al. [2] point out a serious threat for indexing and similarity-based retrieval in high dimensional databases, due to the following phenomenon, called the concentration of distances: As the dimensionality of the data space grows, the distance to the nearest point approaches the distance to the farthest one. Nearest neighbours become meaningless. The

[1] Corresponding author. Phone: +44 121 414 2792; Fax: +44 121 414 4281; E-mail: A.Kaban@cs.bham.ac.uk.

underlying geometry of this phenomenon was further studied in [11], strongly suggesting the detrimental effects often termed informally as the 'curse of dimensionality' are attributable to this phenomenon.

Beyond exponentially slowing down data retrieval [11], the problem of distance concentration is becoming a major concern more generally for high dimensional multivariate data analysis, and risks to compromise our ability to extract meaningful information from volumes of data [4,7]. This is because in many domains of science and engineering, the dimensionality of real data sets grows very quickly, while all data processing and analysis techniques routinely rely on the use of some notion of distance [7]. In particular, high impact application areas, such as cancer research, produce simultaneous measurements of the order of several thousands. As pointed out in [4], currently existing multivariate data analysis techniques were not designed with an awareness of such counter-intuitive phenomena intrinsic to very high dimensions. It is therefore imperative for this problem to be studied and better understood in its own right, before one can embark on trying to devise more appropriate computational techniques for high dimensional problems.

Despite its title "When is nearest neighbour meaningful" [2], the paper in fact answers a different question, namely "When nearest neighbour isn't meaningful". In formal terms, they give a sufficient condition for the concentration phenomenon. However, knowing the answer to the previous question would be very important and useful, since then one would have an objective to work towards in order to get round of the problem, in principle. This is what we address in this paper.

Although many previous authors mention, and admit on the basis of empirical evidence, that cases exist when the nearest neighbour is still meaningful in high dimensions [3,2,7], generally valid formal conditions are still lacking. All recent formal analyses have been conducted assuming data distributions with i.i.d. dimensions [1,7], which is unrealistic in most real settings. Yet, it has been observed that, if techniques for mitigating the concentration phenomenon are used carelessly, they may actually end up having a detrimental effect [7].

Here we make the following contributions: We establish the converse of Beyer et al.'s result, which gives us a generic answer to when nearest neighbour *is* meaningful in arbitrarily high dimensions. Then, we give a class of examples of realistic data distributions having non-i.i.d. dimensions, where we show the Euclidean distance will not concentrate when the dimensionality increases without bounds, as long as the amount of 'relevant' dimensions grows no slower than the overall data dimensions. Of course, this condition is not always met in practice; examples will follow later.

These results provide a formal explanation for previous informal and empirical

observations, such as [3] "increasing the input space dimension without enhancing the quantity of available information reduces the model's power and may give rise to the curse of dimension". Our theoretical result also provides a generic criterion that may be used as an objective to work towards in order to counter the problem when necessary.

2 Distance concentration

Let $F_m, m = 1, 2, \ldots$ be an infinite sequence of data distributions and $\boldsymbol{x}_1^{(m)}, \ldots, \boldsymbol{x}_n^{(m)}$ a random sample of n independent data vectors distributed as F_m. An arbitrary random vector distributed as F_m will be referred to as $\boldsymbol{x}^{(m)}$. For each m, let $||\cdot|| : dom(F_m) \to \mathbb{R}^+$ be a function that takes a point from the domain of F_m and returns a positive real value. Further, $p > 0$ will denote an arbitrary positive constant, and it is assumed that $\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]$ and $\mathrm{Var}[||\boldsymbol{x}^{(m)}||^p]$ are finite and $\mathrm{E}[||\boldsymbol{x}^{(m)}||^p] \neq 0$ throughout this section.

In the context of the problem at hand, the interpretation of the function $||\cdot||$ is that of a distance metric (or norm) — though the theory does not rely on this interpretation, i.e. there is no requirement for it to satisfy the properties of a metric. Similarly, the positive integer m may be interpreted as the dimensionality of the data space.

Theorem 1 (Beyer et al. [2]).
If $\lim_{m\to\infty} \frac{\mathrm{Var}[||\boldsymbol{x}^{(m)}||^p]}{\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]^2} = 0$, then $\forall \epsilon > 0$, $\lim_{m\to\infty} P\left[\max_{1\leq j\leq n} ||\boldsymbol{x}_j^{(m)}|| < (1+\epsilon) \min_{1\leq j\leq n} ||\boldsymbol{x}_j^{(m)}||\right] = 1$; where the operators $\mathrm{E}[\cdot]$ and $\mathrm{Var}[\cdot]$ refer to the theoretical expectation and variance of the distributions F_m, and the probability on the r.h.s. is over the random sample of size n drawn from F_m.

The proof can be found in [2].

As mentioned, this result gives a sufficient condition for which the relative separation of points vanishes as m increases without bounds, though it says nothing when this condition does not hold. Therefore, we formulate and prove the following necessary condition. Before proceeding, it should be noted that nothing was said about the sample size n, so indeed Theorem 1 holds no-matter how large n is.

Theorem 2 (Converse of Theorem 1).
Assume the sample size n is large enough for $\mathrm{E}[||\boldsymbol{x}^{(m)}||^p] \in [\min_{1\leq j\leq n} ||\boldsymbol{x}_j^{(m)}||^p, \max_{1\leq j\leq n} ||\boldsymbol{x}_j^{(m)}||^p]$ to hold. Now, if $\lim_{m\to\infty} P\left[\max_{1\leq j\leq n} ||\boldsymbol{x}_j^{(m)}|| < (1+\epsilon) \min_{1\leq j\leq n} ||\boldsymbol{x}_j^{(m)}||\right] = 1$, $\forall \epsilon > 0$, then $\lim_{m\to\infty} \frac{\mathrm{Var}[||\boldsymbol{x}^{(m)}||^p]}{\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]^2} = 0$.

Proof. Denote $DMIN_m = \min_{1 \le j \le n} ||\boldsymbol{x}_j^{(m)}||$ and $DMAX_m = \max_{1 \le j \le n} ||\boldsymbol{x}_j^{(m)}||$.

Rewriting the precondition gives:

$$\lim_{m \to \infty} P[DMAX_m < (1+\varepsilon) DMIN_m] = 1 \Rightarrow \tag{1}$$

$$\lim_{m \to \infty} P\left[\frac{DMAX_m}{DMIN_m} - 1 < \varepsilon\right] = 1 \Rightarrow \tag{2}$$

$$\lim_{m \xrightarrow{P} \infty} \frac{DMAX_m}{DMIN_m} = 1 \tag{3}$$

using the definition of convergence in probability [2], and the fact that $\frac{DMAX_m}{DMIN_m} - 1 \ge 0$.

In the above, we must assume that $DMIN_m \neq 0, \forall m$. For this reason, we split the infinite sequence in two sub-sequences: one corresponding to the terms $DMIN_m = 0$ and the other to $DMIN_m > 0$, at least one of which is infinite. Now, since all infinite sub-sequences of a convergent sequence are convergent to the same limit, and finite sub-sequences may be dropped without altering convergence, it is enough to show that the statement of Theorem 2 holds for either case.

For economy of argument we deal first with the case of an infinite sub-sequence that corresponds to $DMIN_m = 0$. Now, if there were such an infinite sub-sequence having all terms $DMIN_m = 0$, by substituting these into the pre-condition we would have:

$$\lim_{m \to \infty} P\left[DMAX_m < (1+\varepsilon) DMIN_m\right] = 1 \text{ and so } \lim_{m \to \infty} P\left[DMAX_m < 0\right] = 1$$

But we know that $P[DMAX_m \ge 0] = 1, \forall m$, since $DMAX_m \ge 0, \forall m$; a contradiction! Therefore no infinite subsequence having $DMIN_m = 0$ exists under the given preconditions.

We now move on to the case of practical interest, namely the infinite sub-sequence that has $DMIN_m > 0, \forall m$. Using the fact that the functions $(.)^p$ and $1/(.)^p$ are continuous, we apply Slutsky's theorem ([9], pp. 119–120) to (3) twice, to yield:

$$\lim_{m \xrightarrow{P} \infty} \frac{DMAX_m^p}{DMIN_m^p} = 1; \quad \text{and} \quad \lim_{m \xrightarrow{P} \infty} \frac{DMIN_m^p}{DMAX_m^p} = 1 \tag{4}$$

Observe that here $DMAX_m > 0, \forall m$, since $DMIN_m > 0, \forall m$.

[2] To say that a sequence X_m of random variables converges in probability to X means that $\forall \epsilon > 0$, $\lim_{m \to \infty} P(|X_m - X| \ge \epsilon) = 0$, equivalently (and as used here) $\lim_{m \to \infty} P(|X_m - X| < \epsilon) = 1$. The short notation is $\lim_{m \xrightarrow{P} \infty} X_m = X$ [12], pp. 58–59.

Furthermore, using the precondition that $\mathrm{E}[||\boldsymbol{x}^{(m)}||^p] \in [DMIN_m^p, DMAX_m^p]$, and the fact that the power function $(\cdot)^p$ is a monotonically increasing function on the positive domain, we have that:

$$\frac{DMAX_m^p}{DMIN_m^p} \geq \frac{||\boldsymbol{x}_j^{(m)}||^p}{\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]} \geq \frac{DMIN_m^p}{DMAX_m^p}, \forall j \in \{1, \ldots n\} \tag{5}$$

Now, by the squeeze rule, it follows that the following limit exists and is equal to 1:

$$\lim_{m \xrightarrow{P} \infty} \frac{||\boldsymbol{x}^{(m)}||^p}{\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]} = 1 \tag{6}$$

In (6), we have a sequence of random variables that converges in probability to a constant. Noting that convergence in probability implies convergence in distribution (e.g. see [12] pp. 119–120), in this case to the probability function of a Dirac delta density, the required result follows i.e. the associated sequence of variances converges to zero:

$$\lim_{m \to \infty} \mathrm{Var}\left[\frac{||\boldsymbol{x}^{(m)}||^p}{\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]}\right] = \lim_{m \to \infty} \frac{\mathrm{Var}[||\boldsymbol{x}^{(m)}||^p]}{\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]^2} = 0 \quad \blacksquare \tag{7}$$

In the sequel, the value $RV_m = \frac{\mathrm{Var}[||\boldsymbol{x}^{(m)}||^p]}{\mathrm{E}[||\boldsymbol{x}^{(m)}||^p]^2}$ shall be referred to as the *relative variance*, and $DMAX_m/DMIN_m - 1$ is the *relative separation* of norms or distances. We should note, our use of the term 'relative variance' is a generalisation of that of [7], where it refers to the square root of RV_m with fixed $p = 1$.

The significance of Theorem 2 is that we can now conclude that if the relative separation of distances tends to zero as the dimension of the data space grows to infinity, so does their relative variance. Equivalently, and most importantly, if the relative variance of the distances does not tend to zero, then neither does their relative separation.

2.1 Is it possible for the converse theorem to apply?

In [2], the authors demonstrate a large number of examples in which all L_p metrics fall prey to concentration, and the dimensionality may be of the order of tens for the problem to be of a practical concern already. Is there any room, then, for the converse theorem to apply? The only scenario previously identified formally not to produce a relative variance convergent to zero (and hence satisfying our converse) was the setting where all dimensions (data features) are identical [2]. Of course, to have all dimensions identical to each

other would be an unrealistic model, and in the sequel we identify a much larger class, where our converse theorem applies.

Consider the function $||\cdot||$ defined earlier, substantiated as the family of p-norms, as in the examples presented in [2]. Then, using definitions and making no assumption on the distribution structure, the relative variance of an m-dimensional random vector $\boldsymbol{x}^{(m)} = (x_1, \cdots, x_i, \cdots, x_m)^T$ may be written as the following:

$$RV_m = \frac{\mathrm{Var}[\sum_{i=1}^{m} |x_i|^p]}{\mathrm{E}[\sum_{i=1}^{m} |x_i|^p]^2} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{Cov}[|x_i|^p, |x_j|^p]}{\sum_{i=1}^{m}\sum_{j=1}^{m} \mathrm{E}[|x_i|^p]\mathrm{E}[|x_j|^p]}$$

Quite evidently, it is possible for RV_m not to converge to zero when m tends to infinity, provided that the numerator grows no slower with m than the denominator. Then, cf. the converse theorem, the p-norms remain spread-out despite m increasing to infinity.

One can verify for all examples of [2] that the problem is caused by a sparse correlation structure. Independent variables represent the most trivial case, but chain-like correlation structure is also unable to grow with dimensions at the rate of the denominator.

Thus, we are now able to make some formal sense of what 'structure' in the data means in the context of distance concentration, at least in principle. The next section details a concrete class of examples.

3 Examples: Linear latent variable models

Real data quite often exhibit a rich correlation structure, yet previous studies on distance concentration [7,1] assume data distributions with independent dimensions. The aim in these works has been to identify a non-Euclidean or even non-metric dissimilarity function that would concentrate more slowly with growing m, in data with i.i.d. features. Instead, here we re-examine the more intuitively appealing Euclidean distance in a fairly simple but still more realistic family of data distributions having dependent dimensions. In particular, we consider the family of linear latent variable data-models [6,13]. These models capture dependencies between dimensions with the use of a latent variable, and are known for their ability to describe a variety of real-world data sets. As such, they have been widely used for multivariate data analysis in numerous areas of science and engineering [6]. Hence analysing distance concentration effects in such models, rather than models with i.i.d. dimensions, will give us a better understanding of the concentration issues that one may expect to encounter in real high-dimensional data sets, and will reveal some

of the key causes that govern this problem. The analysis framework and ideas are also applicable, in principle, to non-linear data models [3], though this is outside the scope of this paper.

3.1 Finite latent dimensions

Let L denote the dimension of a latent linear subspace, and each observed dimension x_i is some linear combination of the L latent systematic factors y_l with additive noise.

$$x_i = \sum_{l=1}^{L} a_{il} y_l + \delta_i, \forall i \in \{1, .., m\} \tag{8}$$

In the above, the parameters a_{il} are real valued constants, independent of m. The noise term is assumed to be zero mean, i.i.d. and independent from the systematic factors – these are all standard assumptions in latent variable modelling – so that the latent space captures all the structure content of the data. We also assume that $\mathrm{Var}(y_l^2) \neq 0$.

The model (8) encompasses a number of instantiations that are widely used in practice, such as Probabilistic Principal Component Analysis, Factor Analysis, Independent Factor Analysis and mixture density models [13]. The special case when $L = 1$ includes linear classification and regression models.

Applying our result described in Section 2, we examine the convergence of RV_m derived from the model (8), in order to determine whether the L2 distance between m-dimensional points concentrates in data that follows this density. Straightforward calculations (detailed in the Appendix) and neglecting $\mathcal{O}(m)$ terms in the numerator yield the following expression for the limit of the relative variance, if this limit exists:

$$\lim_{m\to\infty} RV_m = \lim_{m\to\infty} \frac{\sum_{l,k,l',k'=1}^{L,L,L,L} \mathrm{Cov}[y_l y_k, y_{l'} y_{k'}] \sum_{i=1}^{m} a_{il} a_{ik} \sum_{j=1}^{m} a_{jl'} a_{jk'}}{(\sum_{l,k=1}^{L,L} \mathrm{E}[y_l y_k] \sum_{i=1}^{m} a_{il} a_{ik} + \sum_{i=1}^{m} \mathrm{E}[\delta_i^2])^2} \tag{9}$$

For the $L = 1$ case, we have the following simpler form,

$$\lim_{m\to\infty} RV_m = \lim_{m\to\infty} \frac{\mathrm{Var}[y^2]}{\left(\mathrm{E}[y^2] + \frac{\sum_{i=1}^{m} \mathrm{E}[\delta_i^2]}{\sum_{i=1}^{m} a_i^2}\right)^2} \tag{10}$$

Inspecting the obtained expressions, it is easy to see (in either example) that in the noiseless case we have terms of the same order w.r.t. m in both the numerator and denominator. So – unless the positive and negative correlation

[3] Nonlinear data models are often approximated by multiple locally linear models.

terms happen to cancel the numerator of (9) (impossible if e.g. the latent variables are uncorrelated, as the variances are non-zero) – we can conclude that, concentration of the L2 distance will not occur in the noiseless case in these models. Thus it is safe to use L2 distances in arbitrarily high dimensions in this case. Next, the noisy case of interest will be discussed.

Since both $\mathrm{E}[\delta_i^2]$ and $a_{il}^2, i = 1, ..., m$ would be positive constants in a finite-dimensional formulation, it is reasonable to take their infinite sequences to be bounded. So, $\sum_{i=1}^{m} E[\delta_i^2] \in \mathcal{O}(m)$ and likewise [4], $\forall l, \sum_{i=1}^{m} a_{il}^2 \in \mathcal{O}(m)$. Further, it is unlikely for the data features to have a negligible noise contribution, therefore the sequence $\mathrm{E}[\delta_i^2]$ may also be assumed to be bounded from below by a non-zero positive value. Hence the case of practical interest is when $\sum_{i=1}^{m} \mathrm{E}[\delta_i^2] \in \Theta(m)$.

Now, the denominator has the order $\Theta(m^2)$, therefore, to ensure RV_m will not converge to zero, the numerator must also be of the order $\Theta(m^2)$ (it cannot be of higher order anyway). Since the covariance terms when $l = k = k' = l'$ are definitely non-zero ($\mathrm{Var}(y_l^2) \neq 0$) and L is finite, it is enough to require from the factor coefficients that:

$$\neg \left[\lim_{m \to \infty} \frac{\sum_{i=1}^{m} a_{il}^2}{m} = 0 \right], \text{for some } l \tag{11}$$

In words, at least one systematic factor must generate data features of the order $\Theta(m)$ — i.e. the number of features (regarded as random variables) that receive contribution from a generative systematic factor y_l must be of this order. We may conclude therefore that an abundance of features with content from the i.i.d. noise but no content from the systematic factors is a key cause of distance concentration in this model. A less likely cause, as mentioned earlier, would be if the weighted covariance and variance terms in the numerator happen to cancel each other.

Summarising, the main conclusions of this section are the following: (1) In noiseless linear latent-variable data models, the L2 distance does not concentrate. (2) In linear latent variable models with additive i.i.d. noise, the key requirement for the L2 distance not to concentrate is that the cumulative contribution of the systematic component(s) must grow no slower than that of the noise. When the noise variance is bounded away from zero, this may be restated as: The cumulative contribution of the systematic component(s) must grow no slower than the data dimensionality.

[4] With a_m and b_m two real-valued sequences,
$a_m \in \mathcal{O}(b_m)$ (or a_m is of the order $\mathcal{O}(b_m)$) stands for: $\exists C > 0, m_0 : \forall m > m_0, |a_m| \leq C|b_m|$.
$a_m \in \Theta(b_m)$ (or a_m is of the order $\Theta(b_m)$) stands for: $\exists C, C' > 0, m_0 : \forall m > m_0, C|b_m| < |a_m| < C'|b_m|$.

It has been conjectured [2,7] that the underlying 'intrinsic dimension' or the 'actual degree of freedom' of the data needs to be small, otherwise the concentration phenomenon would reappear. Since these terms are often defined in different ways we will refer to L as the 'latent dimension' instead. We will give examples that show this conjecture does not hold in general.

Before proceeding, note that for the case $L \to \infty$, the requirement analogous to (11) now becomes: $\neg \left[\lim_{m\to\infty, L\to\infty} \frac{\sum_{l=1}^{L} \sum_{i=1}^{m} a_{il}^2}{m} = 0 \right]$, which no longer implies that any one factor must have of the order $\Theta(m)$ contribution to the data features, but still, at least of the order $\Theta(m)$ contribution from some of the (infinitely many) latent factors is required. For the remainder of the section, it will be assumed that there exists an $\Theta(m^2)$ term in the numerator of eq. (9), so we can assess the effects of increasing L separately.

I.i.d., zero-mean latent variables. Knowing that i.i.d. data dimensions imply the concentration of L2 distances when $m \to \infty$ as an immediate consequence of the weak law of large numbers [2], one may be somewhat surprised to find that i.i.d. latent dimensions do not necessarily have this effect when $L \to \infty$. Indeed, in this case, we have:

$$\lim_{m\to\infty, L\to\infty} RV_m = \lim_{m\to\infty, L\to\infty} \frac{\sum_{l}^{L} \sum_{k=1}^{L} \mathrm{Var}[y_l y_k] \sum_{i=1}^{m} a_{il}^2 \sum_{j=1}^{m} a_{jk}^2}{(\sum_{l=1}^{L} \mathrm{E}[y_l^2] \sum_{i=1}^{m} a_{il}^2 + \sum_{i=1}^{m} \mathrm{E}[\delta_i^2])^2} \tag{12}$$

The leading terms of both the numerator and denominator are $\mathcal{O}(L^2)$, so in general, it is again feasible for the relative variance not to converge to zero.

Orthogonal coefficients. For the sake of another example, let us now assume that in (9) all pairs of coefficient vectors $\boldsymbol{a}_l$ and $\boldsymbol{a}_k, l \neq k$ are orthogonal. Then, the terms of the form $\sum_{i=1}^{m} a_{il} a_{ik}$ are zero except when $l = k$. Let the latent factors be non-i.i.d. this time. Then, eq (9) becomes:

$$\lim_{m\to\infty, L\to\infty} RV_m = \lim_{m\to\infty, L\to\infty} \frac{\sum_{l=1}^{L} \sum_{l'=1}^{L} \mathrm{Cov}[y_l^2, y_{l'}^2] \sum_{i=1}^{m} a_{il}^2 \sum_{j=1}^{m} a_{jl'}^2}{(\sum_{l=1}^{L} \mathrm{E}[y_l^2] \sum_{i=1}^{m} a_{il}^2 + \sum_{i=1}^{m} \mathrm{E}[\delta_i^2])^2} \tag{13}$$

Again both the numerator and the denominator have leading terms of the order $\mathcal{O}(L^2)$. Therefore, in general (with suitable covariance, i.e. no excessive cancellations in the numerator), converge to 0 when $L \to \infty$ is still not implied.

However, if both restrictions considered earlier (i.e. orthogonal coefficient vectors and also i.i.d. 0-mean latent variables) are simultaneously present, then the resulting data distribution becomes too sparsely correlated and the concentration effect appears: Indeed, in that case the numerator grows slower

than the denominator, since all terms of the form $\mathrm{Cov}[y_l^2, y_{l'}^2]$ other than $l = l'$ are zero:

$$\lim_{m\to\infty} RV_m = \lim_{m\to\infty} \frac{\sum_{l=1}^{L} \mathrm{Var}[y_l^2](\sum_{i=1}^{m} a_{il}^2)^2}{(\sum_{l=1}^{L} \mathrm{E}[y_l^2] \sum_{i=1}^{m} a_{il}^2 + \sum_{i=1}^{m} \mathrm{E}[\delta_i^2])^2} = \frac{\mathcal{O}(L)}{\mathcal{O}(L^2)} \underset{L\to\infty}{\longrightarrow} 0 \quad (14)$$

From the examples given above, we can conclude that, contrary to a naïve intuition, a very large latent dimension, on its own, does not automatically imply the concentration phenomenon. Instead, it is the richness of correlations between the data features that governs the concentration effect.

4 Numerical validation

In this section, we numerically validate our findings and also examine real data sets from two different areas (text-based documents and gene expression arrays). The results suggest that the presence or absence of distance concentration is a major cause for the success or failure of automated data analysis.

4.1 Validating theoretical results

An example of numerical simulation is demonstrated in Figure 1 as m increases, for an instantiation of the model (8): $L = 1$, $y \sim \mathrm{Uniform}[0, 2]$, the noise terms δ_i were sampled from a 0-mean spherical Gaussian with variance varied in $[0, 1]$ and a_i were designed such that $\lim_{m\to\infty} a_m^2 = 1$. Empirical estimates are superimposed with the corresponding analytical limits. We see the sequence of the RV_m estimates converge in agreement with the analytical limits. As predicted from the theory, the limit of the sequence RV_m gets smaller as the noise level increases, but neither the relative variance nor the $\log[DMAX_m/DMIN_m]$ get arbitrarily close to zero.

Figure 2 shows an example with increasing latent dimensionality alongside of increasing data dimensionality. Here, the underlying factors y_l were drawn from i.i.d. standard normal distributions $N(0, 1)$, and $\delta_i \sim N(0, 1)$. The coefficients a_{il} were chosen randomly from [0,3] (all positive and non-orthogonal). As expected cf. the results in Sec. 3.2, concentration does not show up, despite both the data dimensionality m and the latent dimension L increase. Also, as expected from Theorem 2, the picture is similar both in terms of the relative variance and the relative separation of norms.

Contrariwise, in the example shown in Figure 3, we have the same i.i.d. zero-mean factors as before (y_l are drawn i.i.d. from $N(0, 1)$), but we have also

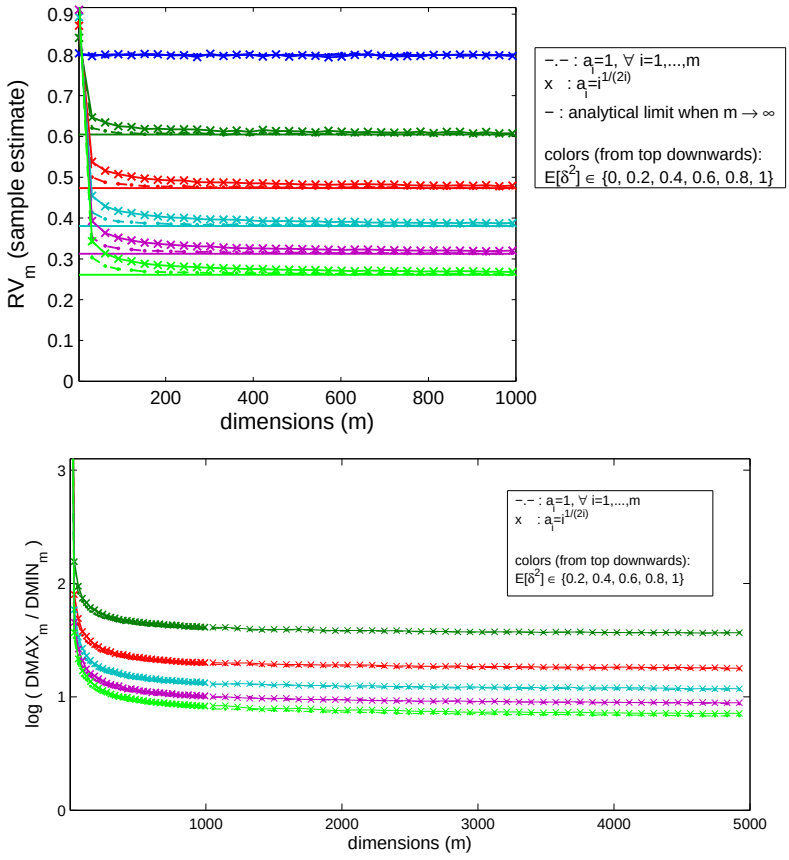

Fig. 1. Example showing the behaviour of RV_m and the $\log[DMAX_m/DMIN_m]$ as m increases. $L = 1$, $y \sim$ Uniform$[0, 2]$ (so, $\mathrm{E}[y^2] = 4/3, \mathrm{Var}[y^2] = 64/45$) and $\delta_i \sim N(0, \sigma^2)$ where σ^2 is varied in $[0, 1]$. For each m, the estimation is based on 15,000 points generated from the model, and the estimates are averaged over 10 repeats.

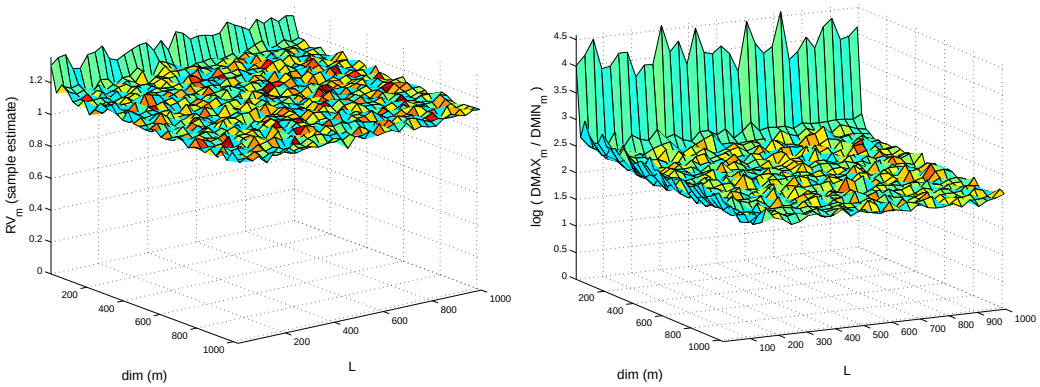

Fig. 2. Example where increasing L does not lead to concentration. $y_i \sim N(0, 1)$, $\delta_i \sim N(0, 1)$, $a_{il} \in [0, 3]$, and the pairs of vectors $\boldsymbol{a}_l, \boldsymbol{a}_k$ are non-orthogonal. Each estimate is based on 20,000 points generated from the model.

pairwise orthogonal coefficients. In this case we see that RV_m decreases with increasing L, and tends to zero eventually, as predicted by our theoretical

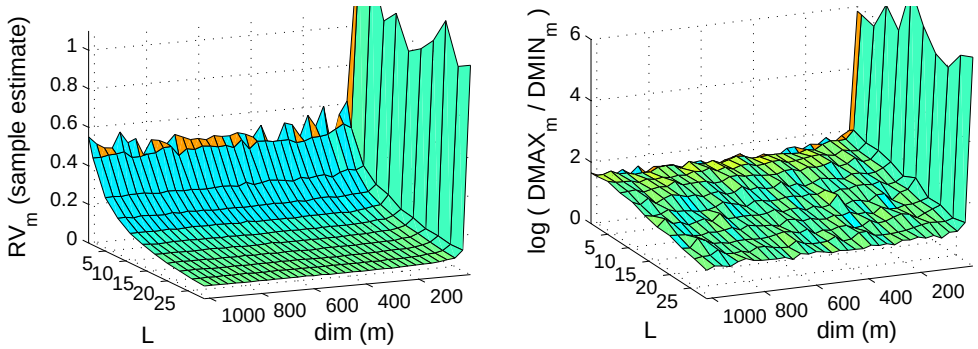

Fig. 3. Example where increasing L does lead to concentration: $y_i \sim N(0,1)$, $\delta_i \sim N(0,1)$ (as before), and pairwise orthogonal coefficient vectors $\boldsymbol{a}_l$. Each estimate is based on 20,000 points generated from the model.

analysis, eq (14). As we have seen in the previous section, this is because the pairwise correlations between features becomes too sparse in this setting.

4.2 Understanding the effect of 'irrelevant' dimensions

We call the i-th feature 'irrelevant' (from the point of view of its systematic structure content), if all its coefficients $a_{il}^2, l = 1, 2, ...$ are zero. Thus, an irrelevant dimension will only contain the contribution of the independent noise term.

Condition (11) says the cumulative contribution of an underlying systematic factor must grow no slower than the data dimensionality. Assuming the coefficients a_{il} are bounded away from zero, we may say, the number of 'relevant features' must grow no slower than the data dimensionality. This notion appears to be more close to what has been termed the 'intrinsic dimensionality' in [7], in the sense of the independent degrees of freedom that describes the data.

To see an example, Figure 4 demonstrates the effect of irrelevant dimensions in a regression model, i.e. $L = 1$. We have $y \sim N(0,1), \delta_i \sim N(0,1)$, and the proportion of relevant dimensions is varied on a grid. As expected, we see the presence of a large fraction of irrelevant dimensions triggers the concentration phenomenon, while in the case of predominantly relevant dimensions, distance concentration does not occur. This insight provides us a more concrete understanding of the nature of the concentration problem in data that obeys a distribution that can be well described by the considered family of models. Again, as expected, the picture is similar both in terms of the relative variance and the relative separation. The latter estimates are of course more

noisy, while the former is easier to estimate from the data sample.

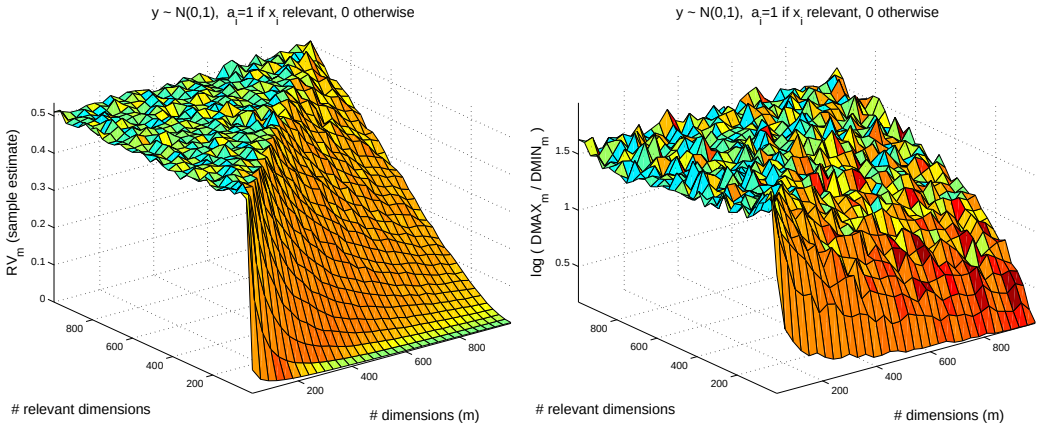

Fig. 4. Example showing that irrelevant features ($a_i = 0$) trigger the concentration phenomenon. $L = 1, y \sim N(0,1), \delta_i \sim N(0,1)$. The estimates are based on 25,000 sample points generated from the model.

4.3 Examining real data sets

In real data sets, examples of measurements that contain many fewer relevant than irrelevant features are frequently encountered in biomedical research [4]. Indeed, some of these data manage to break the best classifiers [10]. The primary cause identified for the unusual effects reported in [10] is the mismatch between the proximity relations in the data space and those in the target space. We may add the abundance of irrelevant dimensions, on its own, may easily destroy the proximities in the data, even if an underlying systematic relation exists between some of the observed features and the target.

We find it instructive to contrast the kind of data associated with this problem domain with that of another area with equally large dimensionality, such as text categorisation. For text-based documents, the data dimensions are dictionary words and the dimensionality equals the size of the dictionary used — typically of the order of tens of thousands. Despite this, many successes have been reported in this problem domain [8]. The question as to what makes the difference in difficulty has never been addressed. Though a possible answer would greatly enhance our understanding of the practical side of the rather vaguely defined 'curse of dimensionality' problem.

Relating these observations to our earlier results, we conjecture that the distance concentration phenomenon plays a role in making the data hard / or easy to work with. One should note that in the case of text irrelevant words (termed 'stop-words' in statistical text analysis, e.g. 'the', 'and', etc.) are relatively few (and typically filtered out based on a well-known list).

To test this conjecture, Figure 5 demonstrates the percentage of datum instances for which half of the remaining points are within some factor of the nearest neighbour, for a number of real data sets drawn from these two application areas. Plotting this quantity for varying percentages and factors was previously used in [2] and gives a suggestive visual representation of the degree of concentration. We have chosen three gene expression data sets of different

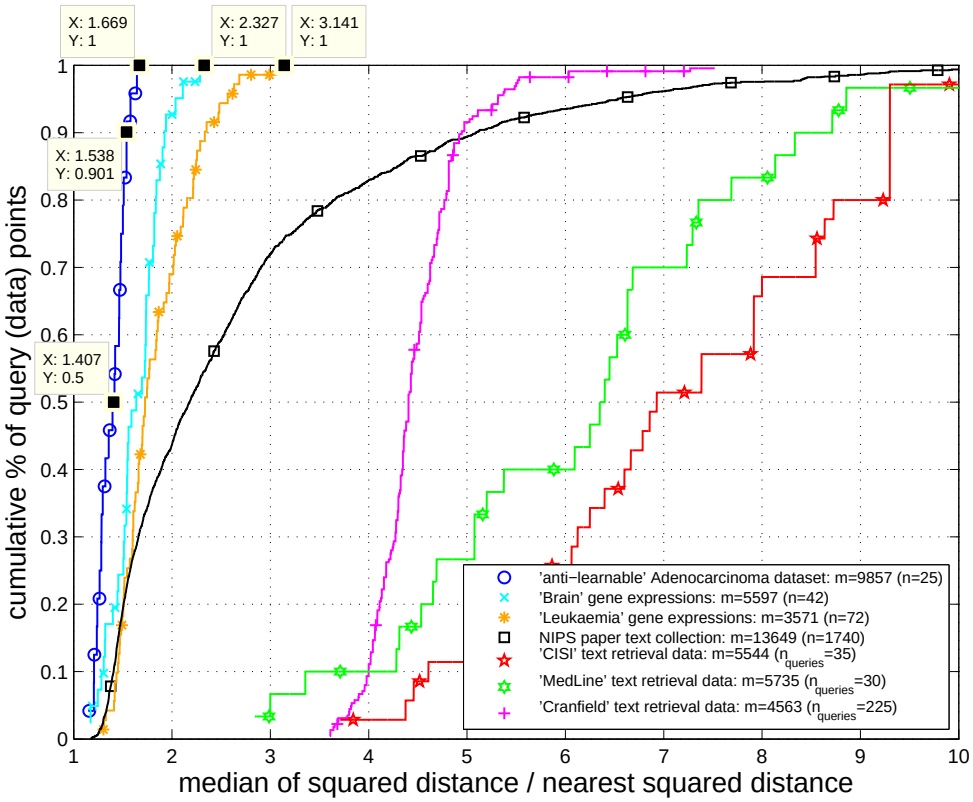

Fig. 5. Examining distance concentration in real data sets. Gene expression data are much more concentrated than text document data.

difficulty and four text data sets. The Adenocarcinoma gene expression arrays [10] were deemed 'anti-learnable' by means of classification methods in [10]. We see in Figure 5, this data is the most concentrated among all others tested. All arrays have half of the rest of the data within a factor of just 1.669 of the square distance from their nearest neighbour. 90% of the arrays have half of the data within a factor of just 1.538 from their nearest neighbours. That is an extremely poor relative spread. The next steepest curve belongs to the Brain tumour gene expressions (available from [5]). This is a data set that led consistently to the highest error rate among five other data sets tested in [5], by seven different classifiers considered in the comparative study of [5]. (E.g. SVM obtained an error rate of 28.29%, KNN 29.71%.) The data contains 5 different classes, yet we see in Figure 5 that all arrays have half of the rest of the data within a factor of 2.327 of the square distance from their nearest neighbour. So the concentration is still quite pronounced.

Next, the Leukaemia gene expressions represent a benchmark on which many studies with different methods reported reasonably good performance (e.g. a KNN achieved 3.83% error [5]). We observe the cumulative percentage curve is slightly less steep than for the previous two sets, which indicates a slightly better relative spread-out of the pairwise distances.

The remaining four data sets represent text-based document data. The NIPS conference paper collection [5] has the highest dimensionality, and we did not do any stop-word removal. Still, the pairwise distances are fairly well spread-out, as indicated by the considerably less steep cumulative percentage curve compared to the gene expression data sets. Further, CISI, MedLine and Cranfield [6] represent benchmark data sets used in many successful information retrieval studies, and we have chosen these for their comparable dimensionality to the gene expression data sets considered earlier. We can see in all these data the distances are well spread out, so the nearest neighbour is indeed meaningful.

These findings suggest that it is not the high dimensionality *per se* that causes problems for automated data analysis. Rather, it is the issue of distance concentration that, when present, appears to be a key source of serious problems. Further research is required to study the feasibility limits of existing feature selection methods and devising new ways of extending them based on the understanding gained in this study.

5 Conclusions

By establishing the converse of the theorem of [2], we formulated a necessary condition for the distance concentration phenomenon. We then examined a broad class of non-i.i.d. data models, known as linear latent variable models, and identified the settings where the Euclidean distance does not concentrate under reasonable conditions. This complements previous work that focused on non-Euclidean distances in data models with i.i.d. dimensions. Since latent variable models have a long and successful history in modelling dependencies in real data sets, our analysis provides guidance and explanation as to when and why distance concentration is or isn't a problem in high dimensional data settings. We gave numerical simulations that validated the theory, and we also examined several real data sets in two different application area. Our findings are in agreement with existing empirical observations, and our theory provides a novel explanation as to why and how data that exhibits structure suffers less from the curse of dimensionality.

[5] http://www.cs.toronto.edu/∼roweis/data.html
[6] http://scgroup6.ceid.upatras.gr:8000/wiki/index.php/Main_Page

The most foreseeable practical ramifications of these results include the following:

- For databases, the need for testing nearest neighbour processing techniques on 'meaningful' workloads (i.e. distributions in which the employed distance does not suffer from the concentration phenomenon) has been noted in [2]. The examples we provided in Sec. 3 can be directly used for this purpose. Moreover, for any distribution and distance function pair, one can test meaningfulness by using our generic theoretical result given in Sec. 2.
- For data analysis and learning from data, as noted in [4], existing techniques lack an awareness of the distance concentration phenomenon in high dimensional spaces. It is hoped that the understanding gained through analysis will pave the way towards a rigorous assessment of existing techniques and towards devising better ones. More research is needed in this area in order to produce and evaluate concrete techniques, however, a natural step may be to investigate the explicit use of RV_m as an objective to be maximised for feature selection and dimensionality reduction. A learning-theoretic study of how the distance concentration in the space of data features (as considered here) affects the generalisation [7] of learning methods in high dimensional problems is also a topic of further research.

Acknowledgements

We thank the referees for their valuable comments, which helped us improve the clarity of this paper. AK acknowledges support from an MRC Discipline Hopping Award (G0701858).

References

[1] C.C. Aggarwal, A. Hinneburg, and D.A. Keim. On the surprising behavior of distance metrics in high dimensional space. Proc. Int. Conf. Database Theory, pp. 420-434, 2001.

[2] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? Proc. Int. Conf. Database Theory, pp. 217-235, 1999.

[3] C. Bishop. Neural networks for pattern recognition. Oxford University Press, 1995.

[7] The same phenomenon, when considered across datum instances rather than their features in actually a blessing, which statistical learning theory builds upon.

[4] R. Clarke, H. W. Ressom, A. Wang, J. Xuan, M. C. Liu, E. A. Gehan, and Y. Wang. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. Nature Reviews Cancer, vol. 8, pp. 37-49, Jan. 2008.

[5] M. Dettling. BagBoosting for tumor classification with gene expression data. Bioinformatics 2004 20(18):3583-3593.

[6] B. Everitt. An introduction to latent variable models. Chapman and Hall, 1984.

[7] D François, V Wertz, and M Verleysen. The concentration of fractional distances. IEEE Trans. on Knowledge and Data Engineering, vol 19, no 7, July 2007.

[8] T. Joachims: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. ECML 1998: 137-142.

[9] K. Knight. Mathematical Statistics. CRC Press, 2000.

[10] A. Kowalczyk. Classification of anti-learnable biological and synthetic data, Proc. PKDD, pp. 176-187, 2007.

[11] V. Pestov. On the geometry of similarity search: Dimensionality curse and concentration of measure. Information Processing Letters 73, pp. 47–51, 2000.

[12] Jeffrey S. Rosenthal: A First Look at Rigorous Probability Theory, 2nd Edition. World Scientific Publishing, London, 2006.

[13] S.T. Roweis, and Z. Ghahramani: A Unifying Review of Linear Gaussian Models. Neural Computation 11(2): 305-345, 1999.

Appendix: Derivation details

To compute RV_m, the expectation of Euclidean distances is computed as:

$$\begin{aligned}\mathrm{E}[||\boldsymbol{x}||_2^2] &= \mathrm{E}[\sum_{i=1}^{m}|\sum_{l=1}^{L}a_{il}y_l+\delta_i|^2] \\ &= \sum_{l=1}^{L}\sum_{k=1}^{L}\mathrm{E}[y_l y_k]\sum_{i=1}^{m}a_{il}a_{ik}+\sum_{i=1}^{m}\mathrm{E}[\delta_i^2]+0\end{aligned}$$

and the variance is:

$$\begin{aligned}&\mathrm{Var}[||\boldsymbol{x}||_2^2] = \mathrm{Var}\left\{\sum_{i=1}^{m}|\sum_{l=1}^{L}a_{il}y_l+\delta_i|^2\right\} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m}\mathrm{Cov}[\sum_{l=1}^{L}\sum_{k=1}^{L}a_{il}a_{ik}y_l y_k, \sum_{l'=1}^{L}\sum_{k'=1}^{L}a_{jl'}a_{jk'}y_{l'}y_{k'}]+\mathcal{O}(m) \\ &= \sum_{l=1}^{L}\sum_{k=1}^{L}\sum_{l'=1}^{L}\sum_{k'=1}^{L}\mathrm{Cov}[y_l y_k, y_{l'}y_{k'}]\sum_{i=1}^{m}a_{il}a_{ik}\sum_{j=1}^{m}a_{jl'}a_{jk'}+\mathcal{O}(m)\end{aligned}$$