



On the distance concentration awareness of certain data reduction techniques

Ata Kabán*

School of Computer Science, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

ARTICLE INFO

Article history:

Received 30 November 2009

Received in revised form

6 July 2010

Accepted 11 August 2010

Keywords:

Distance concentration

Dimensionality reduction

Feature selection

Projection pursuit

Sure independence screening

ABSTRACT

We make a first investigation into a recently raised concern about the suitability of existing data analysis techniques when faced with the counter-intuitive properties of high dimensional data spaces, such as the phenomenon of distance concentration. Under the structural assumption of a generic linear model with a latent variable and an additive unstructured noise, we find that dimension reduction that explicitly guards against distance concentration recovers the well-known techniques of Fisher's linear discriminant analysis, Fisher's discriminant ratio and a variant of projection pursuit. Extrapolation to regression uncovers a close link to sure independence screening, which is a recently proposed technique for variable selection in ultra-high dimensional feature spaces. Hence, these techniques may be seen as distance concentration aware, despite they have not been explicitly designed to have this property. Throughout our analysis, other than the dependency structure implied by the mentioned linear model, we make no assumptions about the distributions of the variables involved.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Creating univariate projections of high dimensional data in directions that satisfy some specified criteria is often the method of choice in high dimensional data analysis. For instance, Fisher's linear discriminant analysis (FLDA) provides the direction that optimally preserves the class structure. In the unsupervised setting, projection pursuit seeks projections that are maximally non-Gaussian [10], e.g. having maximal kurtosis.

Owing to the increasingly high dimensionality of data sets in a number of areas, most notably in cancer research, a serious concern has been raised recently [6] that questions the suitability of existing data analysis techniques when faced with the properties of high dimensional data spaces. In particular, in this study we are concerned with a specific aspect of the dimensionality curse, known as the phenomenon of *distance concentration*—that is, as the data dimensionality increases, all pairwise distances between points may become too similar to each other in certain cases [3,6,15]. Contrary to other problems of high dimensionality, like data sparseness and computational overhead, the phenomenon of distance concentration is rather counter-intuitive, and hence its effects are much less obvious.

Pattern recognition has been the basis for a series of generic methodologies with great potential in a wide range of domains, such as face recognition, spam filtering and multimedia applications. It has

been a key to computer-aided diagnosis systems to support the human expert's interpretations and findings. Data become higher and higher dimensional in all such application areas. Yet, the effects of distance concentration have, so far, not been examined in the context of pattern recognition methodologies. As pointed out in [6], the existing tools have not been designed with an awareness of phenomena that only occur in high dimensions. To what extent is their use appropriate in such high dimensional problems? Since the phenomenon of distance concentration has now been completely characterised [3,11,20], we may be able feasibly address the issue.

Here we proceed in a systematic, model-driven manner. We examine some commonly used linear data reduction techniques specifically from the point of view of their awareness of the distance concentration problem, by looking at their effect on the *relative variance* of the pairwise inter-point distances under the generic model under consideration. The sequence of relative variances (indexed by data dimensionality) was previously shown to be the key to describe the phenomenon of concentration of pairwise distances between points drawn from an arbitrary data distribution [3,11,20]. In addition, it has been shown [11] that this is governed by the (lack of) correlation structure among the data features relative to their noise content. Since pattern recognition would be impossible if the data had no structure, while a noise content is also an inevitable reality, we find it most useful for our analysis to consider a data model that captures both of these characteristics.

We ask the following questions:

- Which is the direction of maximal inter-distance relative variance?

* Tel.: +44 121 414 2792; fax: +44 121 414 4281.

E-mail address: A.Kaban@cs.bham.ac.uk

- Which of the data features have the maximal inter-distance relative variance?
- What statistical property should a latent variable have, so that its high dimensional expansion has maximal relative variance?

It turns out that, in these settings, the answer to the first two questions recovers Fisher's discriminant analysis, Fisher's discriminant ratio, and the recent technique of sure independence screening. The answer to the latter of our questions provides a new justification of the kurtosis index, frequently employed in projection pursuit for finding 'interesting structure' in the data.

2. Background and problem formulation

2.1. Distance concentration

Let F_m , $m=1,2,\dots$ be an infinite sequence of data distributions and $\mathbf{x}_1^{(m)}, \dots, \mathbf{x}_N^{(m)}$ a random sample of n independent data vectors distributed as F_m . For each m , let $\|\cdot\| : \text{dom}(F_m) \rightarrow \mathbb{R}^+$ be a function that takes a point from the domain of F_m and returns a positive real value. Further, $s > 0$ will denote an arbitrary positive constant, and it is assumed that $E[\|\mathbf{x}^{(m)}\|^s]$ and $\text{Var}[\|\mathbf{x}^{(m)}\|^s]$ are finite and $E[\|\mathbf{x}^{(m)}\|^s] \neq 0$.

Theorem 2.1 (Beyer et al. [3]). If $\lim_{m \rightarrow \infty} \text{Var}[\|\mathbf{x}^{(m)}\|^s] / E[\|\mathbf{x}^{(m)}\|^s]^2 = 0$, then $\forall \varepsilon > 0$, $\lim_{m \rightarrow \infty} P[\max_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\| < (1 + \varepsilon) \min_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|] = 1$; where the operators $E[\cdot]$ and $\text{Var}[\cdot]$ 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 .

Theorem 2.2 (Durrant and Kabán [11] Converse of Theorem 2.1). Assume N is large enough, so that $E[\|\mathbf{x}^{(m)}\|^s] \in [\min_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|^s, \max_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|^s]$. If $\lim_{m \rightarrow \infty} P[\max_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\| < (1 + \varepsilon) \min_{1 \leq j \leq N} \|\mathbf{x}_j^{(m)}\|] = 1, \forall \varepsilon > 0$, then $\lim_{m \rightarrow \infty} \text{Var}[\|\mathbf{x}^{(m)}\|^s] / E[\|\mathbf{x}^{(m)}\|^s]^2 = 0$.

For the analysis that follows, the function $\|\cdot\|^s$ will be instantiated as the squared Euclidean norm, $\mathbf{x}^{(m)}$ will signify differences between m -dimensional points, and the positive integer m is the dimensionality of the data space, i.e. the number of features. The expression $\text{Var}[\|\mathbf{x}^{(m)}\|^2] / E[\|\mathbf{x}^{(m)}\|^2]^2$ shall be referred to as the *relative variance*.

In words, the above two results say that having the largest distance no more than epsilon away from the smallest one is equivalent to having the relative variance of the distance distribution converge to zero as the dimensionality increases. This characterisation of the phenomenon of concentration of distances holds for data drawn from an arbitrary multivariate distribution, and for any positive valued dissimilarity function, however, this work is only concerned with the Euclidean distance.

2.2. The data model investigated

The study in this paper concerns the following generic data model. This may be instantiated as a linear regression or classification, as well as an unsupervised generative model:

$$x_i = a_i y + \delta_i, \quad \forall i = 1, \dots, m \quad (1)$$

This model asserts that the observed dimensions x_i are generated from a common systematic factor y embedded in a high dimensional space by the parameters a_i and additive noise δ_i . The noise term will be assumed zero-mean and independent from y , but no Gaussianity or other distributional form will be imposed on either δ_i or y .

In the sequel, \mathbf{x} will denote the column-vector with components x_i , $i=1, \dots, m$, and similarly \mathbf{a} and $\boldsymbol{\delta}$ are the column vectors

with components a_i , $i=1, \dots, m$, and δ_i , $i=1, \dots, m$ respectively. It can be easily verified that switching the notation from points to differences between points leaves the model unchanged, since from (1) we have $(\mathbf{x}' - \mathbf{x}'') = \mathbf{a}(y' - y'') + (\boldsymbol{\delta}' - \boldsymbol{\delta}'')$ of the same form. Hence, it is sufficient to use the simpler notation as in (1), without loss of generality.

2.3. The concentration of Euclidean distances under the model

Definition 2.3. The following will be referred to as the relative variance under the model (1):

$$RV_m(\mathbf{x}) = \frac{\text{Var}[\|\mathbf{a}y + \boldsymbol{\delta}\|^2]}{E[\|\mathbf{a}y + \boldsymbol{\delta}\|^2]^2} \quad (2)$$

Here, the upper indexes of data vectors are dropped for convenience, and instead the data dimensionality m is indicated in the index of RV_m .

Notice, this can be computed exactly, by replacing (1) (details are given in the Appendix), yielding the following:

$$\begin{aligned} RV(\mathbf{x}^{(m)}) &= \frac{\text{Var}[y^2] (\sum_{i=1}^m a_i^2)^2 + \sum_{i,j} 4E[y^2] a_i a_j E[\delta_i \delta_j] + 4E[y] a_i E[\delta_i \delta_j^2] + \text{Cov}(\delta_i^2, \delta_j^2)}{(E[y^2] \sum_{i=1}^m a_i^2 + \sum_{i=1}^m E[\delta_i^2])^2} \end{aligned} \quad (3)$$

We can see that unless the noise has a dense dependency structure, so that the sum of cross-statistics in the numerator is $\theta(m^2)$ —in which case there is no problem with distance concentration even in arbitrarily high dimensions [11]—then the first term is the only one that can possibly still achieve $\text{Var}[\|\mathbf{x}\|_2^2] \in \theta(m^2)$. However, most often the noise is unstructured, even if not necessarily i.i.d., making (36) of the order $o(m^2)$. Since the denominator is always of the order $\theta(m^2)$, it is the latter case that is prone to the distance concentration phenomenon, i.e. this is the situation when $RV_m \rightarrow 0$. For this reason, the remainder of the paper is concerned with the latter case only. In this case, we have

$$RV_m(\mathbf{x}) = \frac{\text{Var}[y^2] (\sum_{i=1}^m a_i^2)^2 + o(m^2)}{(E[y^2] \sum_{i=1}^m a_i^2 + \sum_{i=1}^m E[\delta_i^2])^2} \quad (4)$$

Remark 2.4. If $\lim_{m \rightarrow \infty} \|\mathbf{a}^{(m)}\|^2 / E[\|\boldsymbol{\delta}^{(m)}\|^2] = 0$ then the sequence in Eq. (4) converges to zero, i.e. $\lim_{m \rightarrow \infty} RV_m(\mathbf{x}) = 0$.

Proof. A simple rewriting of Eq. (4) gives

$$RV_m(\mathbf{x}) = \frac{\text{Var}[y^2] + o(1)}{\left(E[y^2] + \frac{E[\|\boldsymbol{\delta}^{(m)}\|^2]}{\|\mathbf{a}^{(m)}\|^2} \right)^2} \quad (5)$$

Since $E[y^2]$ and $\text{Var}[y^2]$ are independent of m , hence, under the precondition the sequence in Eq. (5) converges to zero. \square

Hence, in order for the sequence $RV_m(\mathbf{x})$ not to be convergent to zero, the expected norm of the additive noise must grow slower than the norm of the systematic component contribution, when the data dimensionality increases, i.e. we must have

$$\lim_{m \rightarrow \infty} \frac{\|\mathbf{a}^{(m)}\|^2}{E[\|\boldsymbol{\delta}^{(m)}\|^2]} = 0 \quad (6)$$

It should be noted, though, that in practice, even if the sequence RV_m will not converge to exactly zero, it might reach

some small positive value rather quickly as the unstructured noise accumulates with the increasing number of features.

Example 2.5. For an illustration, consider a simple case designed such that $\lim_{m \rightarrow \infty} a_m^2 = A$, and $\delta_i = \delta$, $\forall i = 1, \dots, m$, so the limit of RV_m can be easily computed analytically (using Cesaro–Stolz' theorem), yielding $\lim_{m \rightarrow \infty} RV_m = \text{Var}[y^2]/(E[y^2] + E[\delta^2]/A)^2$. Fig. 1 shows this analytical limit, as a function of the model parameters, when the model (1) is instantiated as $y \sim \text{uniform}[0, b]$. We see clearly that the limit decreases as the ratio $E[\delta^2]/A$ increases or as b decreases—i.e. exactly when the noise content increases relative to the systematic factor's contribution. Although none of these limits are exactly zero, as long as $E[\delta^2]/A$ is finite and b is non-zero, a value close to zero still means a poor contrast between large and small inter-point distances.

2.4. Problem formulation: distance concentration aware dimensionality reduction

Considering the model defined by Eq. (1), we ask the following question: seek a vector \mathbf{w} that projects the m -dimensional data onto a line,

$$\mathbf{w}\mathbf{x} = \mathbf{w}\mathbf{a} + \mathbf{w}\delta \quad (7)$$

such that in the projected space the relative variance, i.e. $RV_1(\mathbf{w}\mathbf{x})$ under the model is maximised.

Depending on whether y is known or unknown, we have the supervised or the unsupervised instance of this problem. Depending on whether \mathbf{w} has real-valued entries or 0/1 entries, we have two forms of data reduction, i.e. dimensionality reduction and feature selection, respectively.

Rather interestingly, as we shall see in the next section, some widely used data reduction methods turn out to satisfy this maximisation objective.

3. Results

3.1. A re-interpretation of Fisher's linear discriminant analysis

Consider the case when $y \in \{-1, 1\}$. Then, the model (1) is a two-class mixture density with shared covariance.

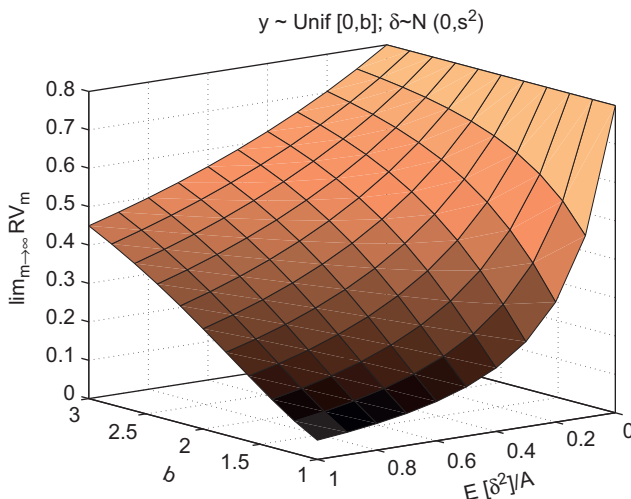


Fig. 1. The limit of $RV_m(\mathbf{x})$ as a function of model parameters, in the case of a generative model with univariate factor $y \sim \text{uniform}[0, b]$ (so $E[y^2] = b^2/3$ and $\text{Var}[y^2] = 4b^4/45$) embedded in high dimensions.

We seek a linear projection of the high dimensional data down to 1D, such that the relative variance of the projected data—w.r.t. the Euclidean distance and the above mentioned data model—is maximised. In other words, how can we down-project the data to make it least concentrated for a subsequent two-class mixture-based classification.

Denote $RV_m^{\text{lead}}(\mathbf{x})$ the leading term in (4), that is, $RV_m^{\text{lead}}(\mathbf{x}) = \text{Var}[y^2](\sum_{i=1}^m a_i^2)^2 / (E[y^2] \sum_{i=1}^m a_i^2 + \sum_{i=1}^m E[\delta_i^2])^2$. Further, we use $RV_m^{\text{lead}}(\mathbf{x}|y)$ to refer to the *conditional* relative variance, i.e. where the variable y is known. In practice, we are given a sample (\mathbf{x}_i, y_i) , $i = 1, \dots, m$, as in classification or regression, so sample estimates of the required statistics are available.

Theorem 3.1. Consider the data model (1) with $y \in \{-1, 1\}$ given. The vector $\mathbf{w} \in \mathbb{R}^{1 \times m}$ that maximises a sample estimate of $RV_1^{\text{lead}}(\mathbf{w}\mathbf{x}|y)$ is the projection obtained by Fisher's LDA.

Proof. Observe that the centres of the two classes are the following:

$$\mathbf{c}_1 \equiv E[\mathbf{x}|y = 1] = \mathbf{a}; \quad \mathbf{c}_2 \equiv E[\mathbf{x}|y = -1] = -\mathbf{a} \quad (8)$$

and they are situated at a distance of $2\|\mathbf{a}\|$ from each other. Hence, the numerator of the condition (6) we derived earlier corresponds to the squared distance between the centres:

$$\sum_{i=1}^m a_i^2 = \|\mathbf{a}\|^2 = \frac{1}{4}\|\mathbf{c}_1 - \mathbf{c}_2\|^2 \quad (9)$$

From the form of $RV_m^{\text{lead}}(\mathbf{x}|y)$, it is easy to see that maximising it is equivalent to maximising $\|\mathbf{a}\|_2^2 / \sum_{i=1}^m E[\delta_i^2]$.

Now, using (9) and recalling that $E[\delta_i] = 0$, $i = 1, \dots, m$, so $E[\delta_i^2] = \text{Var}[\delta_i]$, we can verify that, seeking a linear transform $\mathbf{w} \in \mathbb{R}^{1 \times m}$ of the data that maximises a sample estimate of this expression in the 1D projected space yields

$$\begin{aligned} \frac{|\mathbf{w}\mathbf{a}|^2}{\text{Var}[\mathbf{w}\delta]} &= \frac{1/4[\mathbf{w}(2\mathbf{a})]^2}{1/2(\text{Var}[\mathbf{w}\delta|\mathbf{a}, y = 1] + \text{Var}[\mathbf{w}\delta|\mathbf{a}, y = -1])} \\ &= \frac{1}{2} \frac{[\mathbf{w}(\mathbf{c}_1 - \mathbf{c}_2)]^2}{\text{Var}[\mathbf{w}\mathbf{x} - \mathbf{w}\mathbf{a}] + \text{Var}[\mathbf{w}\mathbf{x} + \mathbf{w}\mathbf{a}]} \\ &= \frac{1}{2} \frac{\mathbf{w}(\mathbf{c}_1 - \mathbf{c}_2)(\mathbf{c}_1 - \mathbf{c}_2)^T \mathbf{w}^T}{\mathbf{w} \left\{ \sum_{k=1}^2 E[(\mathbf{x} - \mathbf{c}_k)(\mathbf{x} - \mathbf{c}_k)^T] \right\} \mathbf{w}^T} \\ &= \frac{1}{2} \frac{\mathbf{w}\mathbf{S}_B\mathbf{w}^T}{\mathbf{w}\mathbf{S}_W\mathbf{w}^T} \end{aligned}$$

where \mathbf{S}_B and \mathbf{S}_W denote the between-class and within class variances (or scatters), respectively. This is exactly the objective of Fisher's linear discriminant analysis (FLDA). \square

It should be noted that FLDA [16,19] is one of the most enduring classical methods, which has seen numerous successful applications in dimensionality reduction and classification. In particular, the scheme proposed in [24] for high dimensional proteomic data classification, presents an interesting combination of FLDA data projection followed by mixture-based classification in the reduced space. Although this might appear a heuristic combination at first, supported by empirical evidence, it may now also be viewed, in the light of our analysis, as a principled optimal way of projecting the data to minimise distance concentration specifically for a subsequent mixture-based classification in the projection space.

On another note, it may be also interesting to observe that the criterion (6) implies that the distance between the centres, $\|\mathbf{c}_1 - \mathbf{c}_2\|$ should grow with the square root of the dimensions, $m^{1/2}$. Remarkably, this matches the requirement derived in [7] from entirely different considerations.

The next section further establishes the importance of the notion of optimality satisfied by Fisher's LDA in dealing with high dimensional data, by analysing another technique in the current setting, namely random projections—which also deals with pairwise distances and dimensionality reduction. We shall argue that these two techniques address two entirely different aspects of the dimensionality curse. Hence they should not be used interchangeably but based on a good understanding of the specific aspect of the curse that is more pronounced in the particular data set and task at hand. The aim of our analysis is to provide such understanding.

3.1.1. Fisher LDA vs. dimensionality reduction by a random projections

Random projections based dimensionality reduction has also been proposed as a means of dealing with the 'curse of dimensionality' [22]. It is of interest, therefore, to place this method in the context of our study.

The method of random projections is a non-adaptive technique. That is, it is independent of the particular data set being applied to. It works by projecting the high dimensional data down to a random subspace, and has the ability to approximately preserve pairwise distances from the high dimensional data space into the low dimensional projection space. Moreover, one of the proofs of this property is based on the phenomenon of concentration of distances [8].

Thus, it is quite interesting to follow that, on the one hand, we have distance concentration as a threatening aspect of the dimensionality curse, cf. literature on database theory [3,11,25,17], data engineering [15,20] and its important application domains [6]. On the other hand, we have the same phenomenon of distance concentration, as a remedy of the dimensionality curse, via the technique of random projections, cf. literature in theoretical computer science [8,22].

To reconcile this and avoid confusion, we ought to point out, though, that in these two branches of the literature, the term 'curse of dimensionality' actually refers to two entirely different aspects of the overall problematic issues associated with high dimensionality. The meaning referred to in the database and data engineering literature [3,25,17,15,20,11,6], as well as in this paper, is the phenomenon that the contrast between pairwise distances of data points can vanish in high dimensions—this is just one aspect of the 'curse'. In turn, the meaning of the 'curse of dimensionality' referred to in the theoretical computer science literature in the context of random projections [22,8] is the computational overhead associated with the processing of high dimensional data—this is an entirely different and complementary issue. It is the latter (not the former) that random projections have been advocated as a provably efficient means to bypassing the 'curse'.

Though, since our study concerns the distance concentration aspect of the curse, it is of interest to work out what is the effect of distance concentration on random projections based dimensionality reduction. This is important to elucidate, since high dimensional data may happen to suffer from both aspects of the dimensionality curse. To this end, in the sequel, we ask, within our model based framework, what is the target-conditional relative variance of the randomly projected data under the model, and how it compares to that of data reduced by Fisher's LDA.

We will make use of the Johnson–Lindenstrauss lemma (JLL), which is the following:

Theorem 3.2 (Johnson–Lindenstrauss). *Let $\mathbb{R}^{k \times m}$ a random projection matrix with entries $r_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1/m)$. Then, $\forall \varepsilon \in (0, 1)$, we have for*

any given point $\mathbf{x} \in \mathbb{R}^m$ that

$$\Pr \left\{ (1-\varepsilon) \frac{k}{m} \|\mathbf{x}\|^2 \leq \|\mathbf{R}\mathbf{x}\|^2 \leq (1+\varepsilon) \frac{k}{m} \|\mathbf{x}\|^2 \right\} \geq 1 - 2\exp(-k\varepsilon^2/4) \quad (10)$$

Comprehensive proofs of this important result may be found in [7] or [22].

Applying JLL to our problem, we have the following for a $k < m$ dimensional random projection of the data model in Eq. (1) (then simply instantiate $k=1$ to obtain the univariate projection version as in Eq. (7)).

Theorem 3.3. *Consider the data model (1) with $y \in \{0, 1\}$ given, and $\mathbf{W} \in \mathbb{R}^{k \times m}$ matrix with entries $w_{ij} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1/m)$, where m is the data dimensionality. Then, $\forall \varepsilon \in (0, 1)$ we have with probability at least $1 - 4\exp(-k\varepsilon^2/4)$ that*

$$\frac{\text{Var}[y^2]}{\left(E[y^2] + \frac{1+\varepsilon}{1-\varepsilon} \cdot \frac{E[\|\delta\|^2]}{\|\mathbf{a}\|^2}\right)} \leq RV_k^{\text{lead}}(\mathbf{W}\mathbf{x}|y) \leq \frac{\text{Var}[y^2]}{\left(E[y^2] + \frac{1-\varepsilon}{1+\varepsilon} \cdot \frac{E[\|\delta\|^2]}{\|\mathbf{a}\|^2}\right)} \quad (11)$$

Proof. Using the Johnson–Lindenstrauss lemma twice yields

$$\Pr \left\{ (1-\varepsilon) \frac{k}{m} \|\mathbf{a}\|^2 \leq \|\mathbf{W}\mathbf{a}\|^2 \leq (1+\varepsilon) \frac{k}{m} \|\mathbf{a}\|^2 \right\} \geq 1 - 2\exp(-k\varepsilon^2/4) \quad (12)$$

and

$$\Pr \left\{ (1-\varepsilon) \frac{k}{m} \|\delta\|^2 \leq \|\mathbf{W}\delta\|^2 \leq (1+\varepsilon) \frac{k}{m} \|\delta\|^2 \right\} \geq 1 - 2\exp(-k\varepsilon^2/4) \quad (13)$$

Taking expectations w.r.t. δ from all three sides of the event in (13), we have also that

$$\Pr \left\{ (1-\varepsilon) \frac{k}{m} E_\delta[\|\delta\|^2] \leq E_\delta[\|\mathbf{W}\delta\|^2] \leq (1+\varepsilon) \frac{k}{m} E_\delta[\|\delta\|^2] \right\} \geq 1 - 2\exp(-k\varepsilon^2/4) \quad (14)$$

Now, applying the union bound to ensure that both (12) and (14) hold simultaneously, and combining these two inequalities, yields that the following holds with probability at least $1 - 4\exp(-k\varepsilon^2/4)$:

$$\frac{1-\varepsilon}{1+\varepsilon} \frac{\|\mathbf{a}\|^2}{E_\delta[\|\delta\|^2]} \leq \frac{\|\mathbf{W}\mathbf{a}\|^2}{E_\delta[\|\mathbf{W}\delta\|^2]} \leq \frac{1+\varepsilon}{1-\varepsilon} \frac{\|\mathbf{a}\|^2}{E_\delta[\|\delta\|^2]} \quad (15)$$

Finally, noting that all sides in the above inequality are positive, we apply the function $\text{Var}[y^2]/(E[y^2] + 1/\cdot)^2$ to all sides. This function is monotonically increasing on the positive domain, hence the theorem follows. \square

That is, the leading term of the class-conditional relative variance is preserved with high probability. Hence, if there was a distance concentration problem in the high dimensional data space, it gets preserved with high probability.

In addition, it may be worth commenting that the form of the upper and lower bound on RV_k^{lead} as in Eq. (15) holds the same irrespective of the reduced dimension k . The only difference is that the probability with which this preservation is guaranteed gets larger with increasing the projection dimension k . The small (e.g. $k=1$ -dimensional) projections produce larger fluctuations around the class-conditional relative variance in the original data space. However, such a random perturbation is unlikely to be beneficial if the contrast between the nearest and farthest

neighbours was poor in the initial data space, since it may easily swap the nearest neighbour with the farthest one.

Furthermore, it should be pointed out that the situation of a severe distance concentration problem in the original data space fits the scenario given in [2] (data points lay on the vertexes of a high dimensional simplex), which is the worst-case for random projections in terms of the dimensionality required for approximate preservation of distances. Therefore accidental swaps of the nearest neighbour with the farthest one may occur even if k is relatively large.

All this means that the random projections are not suited to deal with the distance concentration aspect of the dimensionality curse. Indeed, intuitively, if the original high dimensional data has a poor contrast between large and small distances, then preserving this (lack of) structure will not cure the problem. By contrary, in this case the dimensionality reduction step should try to carefully enhance the contrast between the pairwise distances. This is what FLDA does, within the data model class under our study. On the other hand, if the original data does have a lot of 'structure', then it will not suffer from distance concentration despite of being high dimensional (see [11] for a more detailed formal treatment of this intuition). In that case, dimensionality reduction has the main role to reduce the computational aspect of the curse, while preserving the existing structure. This is the situation in which random projections are appropriate and beneficial.

3.2. A re-interpretation of Fisher's discriminant ratio

Under the data model under study, Eq. (1), we now ask the following question: Of all the m features, which is the one for which the relative variance of the pairwise distances is maximal? This could be used to rank the features w.r.t. their associated relative variance.

Theorem 3.4. *Let \mathcal{B} denotes the set of canonical basis vectors, i.e. the vectors that have one component 1 and all the others 0, and consider the data model in Eq. (1) with $y \in \{-1, 1\}$ given. Then, the feature selector $\mathbf{w} \in \mathcal{B}$ that maximises a sample estimate of $RV_1^{\text{lead}}(\mathbf{w}|\mathbf{x}|y)$ is Fisher's Discriminant Ratio.*

Proof. As previously, we have that the maximiser of $RV_1^{\text{lead}}(\mathbf{w}|\mathbf{x}|y)$ is

$$\arg \max_{\mathbf{w} \in \mathcal{B}} \frac{(\mathbf{w}\mathbf{a})^2}{\text{Var}[\mathbf{w}\delta]} = \arg \max_{\mathbf{w} \in \mathcal{B}} \frac{(\mathbf{w}(\mathbf{c}_1 - \mathbf{c}_2))^2}{\text{Var}[\mathbf{w}(\mathbf{x} - \mathbf{c}_1)] + \text{Var}[\mathbf{w}(\mathbf{x} - \mathbf{c}_2)]} \quad (16)$$

but this time \mathbf{w} is sought in the set \mathcal{B} . Since \mathbf{w} selects exactly one component, (16) can be rewritten as

$$\arg \max_{i \in \{1, \dots, m\}} \frac{(c_{i,1} - c_{i,2})^2}{\text{Var}[x_i - c_{i,1}] + \text{Var}[x_i - c_{i,2}]} \quad (17)$$

A sample estimate of (17) is indeed the Fisher discriminant ratio (FDR) criterion [13]. \square

FDR is a widely used criterion for feature ranking, and has the natural and intuitive interpretation that the features with largest separation between class centres and smallest sum of class variances receive the highest score.

3.3. Extension to regression: a connection with sure independence screening

The re-interpretation of FDR as a maximiser of the target-conditional relative variance in the reduced space, presented in the previous section, suggests that the scope of a FDR-like feature ranking method does not have to be restricted to the case when y

is discrete valued. As we shall see shortly, the same reasoning applied to the case when y is continuous valued (like in regression) recovers the recent and very promising technique called sure independence screening (SIS) [14]. SIS essentially performs correlation learning (or independence learning), and ranks standardised features x_i according to the magnitude of the sample estimate of $|E[x_i y]|$. It was shown to be particularly effective for reducing ultra-high dimensions to moderate, by its property that the top ranked $\mathcal{O}(N/\log(N))$ features contain all the ones relevant to the target y , with high probability [14]. To see this link, we will adopt the assumptions used by SIS in this section (standardised features, independent noise).

Theorem 3.5. *Consider the data model (1) with $y \in \mathbb{R}$ given, and assume the components of the noise term δ in the model (1) are pairwise independent, and the data features are standardised. Then, the feature ranking produced by the magnitudes of the sample estimates of $RV_1^{\text{lead}}(\mathbf{w}|\mathbf{x}|y)$, $\mathbf{w} \in \mathcal{B}$ is equivalent to that produced by sure independence screening.*

Proof. We have, as previously

$$\arg \max_{i \in \{1, \dots, m\}} \frac{a_i^2}{\text{Var}[x_i - a_i y]} = \arg \max_{i \in \{1, \dots, m\}} \frac{a_i^2}{E[(x_i - a_i y)^2]} \quad (18)$$

where a sample estimate of \mathbf{a} may be used, as computed with the model (1)—this is in analogy with using sample estimates of \mathbf{c}_1 and \mathbf{c}_2 in Fisher's discriminant ratio, which is now the following:

$$a_i = E[x_i y] / E[y^2] \quad (19)$$

Next, we show that the argument of (18) is a monotonically increasing function of $|a_i|$. Indeed, we can rewrite

$$\frac{a_i^2}{E[(x_i - a_i y)^2]} = \frac{a_i^2}{E[x_i^2] + a_i^2 E[y^2] - 2a_i E[x_i y]} = \frac{a_i^2}{1 - E[y^2] a_i^2} \quad (20)$$

where we used (19) and that the features are standardised. The first order derivative of (20) w.r.t. $|a_i|$ is $2|a_i| / \{E[y^2] a_i^2 - 1\}^2$, which is always non-negative. Hence, (20) is monotonically increasing with $|a_i|$.

In consequence, the ranking produced by (18) is the same as the ranking of $|a_i|$. That is, (18) is equivalent to

$$\arg \max_{i \in \{1, \dots, m\}} |E[x_i y]| / E[y^2] = \arg \max_{i \in \{1, \dots, m\}} |E[x_i y]| \quad (21)$$

where in the last equality we used the fact that $E[y^2]$ is independent of i .

Now, (21) is exactly the SIS [14] method for feature ranking, and the theorem follows. \square

It may be worth commenting that SIS was derived from entirely different considerations from ours, namely as a component-wise regression method designed to avoid the problem of spurious correlation estimates caused by small sample effects. In contrast, our analysis framework did not consider the issues of estimation error and finite sample effects at all. However, the gist of the problem from both viewpoints lies with the scarceness of true correlations between features, i.e. the 'lack of structure' in the true underlying distribution of the data.

3.3.1. Subset selection

For both classification and regression, the relative variance maximisation procedure could also be formulated, in principle, for feature subset selection, by considering the class-conditional RV_k where k is the number of features to be retained. The maximisation of the leading term of RV_k is

$$\arg \max_{w_i \in \{0, 1\}, i = 1, \dots, m} \frac{\sum_{i=1}^m w_i a_i^2}{\sum_{i=1}^m \text{Var}[w_i(x_i - a_i y)]} \quad (22)$$

However, this is a combinatorial problem and therefore may be impractical. In fact, both the FDR and the SIS ranking criteria may be seen as a heuristic solution to this combinatorial search problem.

3.4. A re-interpretation of projection pursuit

We now turn our attention to the unsupervised setting. The exploratory analysis of high dimensional data is often performed through linear univariate projections, such as projection pursuit (PP). PP seeks a direction such that the projected data have certain statistical properties. The desirable properties of the projections need to be specified by the user as a measure of 'interestingness'. Maximal kurtosis is among the most popular ones [10,19].

As commonly done in PP, we take y as having 0-mean and unit variance in this section. This will also simplify the exposition. We ask the following question: Find a 1D projection of the data in the direction that maximises the relative variance of the pairwise distances.

Theorem 3.6. Consider the model (1) with y being a 0-mean unit variance hidden variable. Then, maximising $RV_1^{\text{lead}}(\mathbf{w}\mathbf{x})$ as a function of both \mathbf{w} and y is equivalent to minimising the mean square error of the observed variable and maximising the kurtosis of y .

Proof. This time, y is unknown. Hence, maximising the leading term of the relative variance in the model, $RV_1^{\text{lead}}(\mathbf{w}\mathbf{x})$ needs to be done w.r.t. both \mathbf{w} and y .

$$\arg \max_{\mathbf{w}, y} RV_1^{\text{lead}}(\mathbf{w}\mathbf{x}) = \arg \max_{\mathbf{w}, y} \frac{\text{Var}[y^2]}{\left\{ E[y^2] + \frac{E[(\mathbf{w}\delta)^2]}{(\mathbf{w}\mathbf{a})^2} \right\}^2} \quad (23)$$

Writing the numerator as $\text{Var}(y^2) = E(y^4) - E(y^2)^2$ and using the assumptions on y , we get

$$\arg \max_{\mathbf{w}, y} RV_1^{\text{lead}}(\mathbf{w}\mathbf{x}) = \arg \max_{\mathbf{w}, y} \frac{E[y^4] - 1}{\left\{ 1 + \frac{E[(\mathbf{w}\delta)^2]}{(\mathbf{w}\mathbf{a})^2} \right\}^2} \quad (24)$$

$$= \arg \max_{\mathbf{w}, y} \frac{\text{kurt}(y) + 2}{\left\{ 1 + \frac{E[(\mathbf{w}\mathbf{x} - \mathbf{w}\mathbf{a}y)^2]}{(\mathbf{w}\mathbf{a})^2} \right\}^2} \quad (25)$$

where $\text{kurt}(y)$ is the kurtosis of y . Further, manipulating the denominator we find

$$\arg \max_{\mathbf{w}, y} RV_1^{\text{lead}}(\mathbf{w}\mathbf{x}^{(1)}) = \arg \max_{\mathbf{w}, y} \frac{\text{kurt}(y) + 2}{\left\{ 1 + E\left[\left(\frac{\mathbf{w}}{\mathbf{w}\mathbf{a}} \mathbf{x} - y \right)^2 \right] \right\}^2} \quad (26)$$

$$= \arg \max_{\mathbf{u}, y} \frac{\text{kurt}(y) + 2}{\{1 + E[(\mathbf{u}\mathbf{x} - y)^2]\}^2} \quad (27)$$

This is equivalent to minimising the mean square error and maximising the kurtosis of the latent variable, and this is exactly what a 'one-unit' independent component analysis (ICA) [21] algorithm does. \square

Further, the noiseless case recovers a simple version of projection pursuit, as the following.

Corollary 3.7. Consider the model (1) with zero noise. Then, maximising $RV_1^{\text{lead}}(\mathbf{w}\mathbf{x})$ is equivalent to finding a direction in which the projection of the data has maximal kurtosis.

Proof. The projected noiseless model is

$$\mathbf{w}\mathbf{x} = \mathbf{w}\mathbf{a}y \quad (28)$$

Denoting $\mathbf{u} \equiv \mathbf{w}/\mathbf{w}\mathbf{a}$, we can equivalently write this as

$$y = \mathbf{u}\mathbf{x} \quad (29)$$

Hence, the previous result in Eq. (27) simplifies to

$$\arg \max_{\mathbf{w}, y} RV_1^{\text{lead}}(\mathbf{w}\mathbf{x}) = \arg \max_{\mathbf{u}} \text{kurt}(\mathbf{u}\mathbf{x}) \quad \square \quad (30)$$

3.4.1. Generative alternative

An alternative to the projection approach for unsupervised exploratory data analysis is the approach of generative modelling. This is to infer the posterior statistics of the latent variable y from the data, under the model (1). In this case, the problem of specifying a suitable projection criterion becomes that of specifying a suitable prior on the latent variable. Accordingly, our question may now be posed as follows: What statistical property should the latent variable have to ensure a large relative variance in the original data space under the model?

In this instance, we write the leading term of the relative variance under the model in the original m -dimensional data space:

$$RV_m^{\text{lead}}(\mathbf{x}) = \frac{\text{Var}(y^2)}{\left\{ E(y^2) + \frac{\sum_{i=1}^m E(\delta_i^2)}{\sum_{i=1}^m a_i^2} \right\}^2} = \frac{\text{kurt}(y) + 2}{\left\{ 1 + \frac{\sum_{i=1}^m \text{Var}(\delta_i)}{\sum_{i=1}^m \text{Var}(a_i y)} \right\}^2} \quad (31)$$

As we see from (31) that the leading term of the relative variance again depends on the kurtosis of the latent variable and the signal to noise ratio.

4. Experiments

This section presents a set of experiments that illustrate and validate the findings of the previous section, using synthetic data generated from various instantiations of the model Eq. (1) under study. We then also demonstrate the working of the identified methods, and their combinations, in benchmark real data experiments.

4.1. Validating the theoretical findings

4.1.1. Fisher's LDA

We start by demonstrating the finding of Section 3.1, using data generated from model (1) with $y \in \{-1, 1\}$ and standard Gaussian noise. We designed two sets of experiments on different settings of severity of the distance concentration problem. Each time we compare estimates of the class-conditional relative variances under the model in the original data, in Fisher's LDA projections of the data, and in random projections of the data. We estimate these from $N=200$ data points generated from this model (100 points in each class)—that is, using 19 900 pairwise distances. These results are given in Figs. 2 and 3, as follows.

The setting in Fig. 2 was designed as a severe distance concentration case: It has only two relevant features, i.e. only two entries of \mathbf{a} are non-zero, while the data dimensionality is increased progressively. That is, an increasing number of features bring only unstructured noise content. We see the sequence of class-conditional RV_m converges to zero quite rapidly in the original data space, and becomes indistinguishable from zero to cca. 50 dimensions onwards. In turn, the values of the class-conditional RV_1 estimates of Fisher's LDA projections decrease a lot more slowly. Also superimposed on this figure, we see the estimates of the class-conditional RV_k of the random projections of the same data, onto $k=1$ and 5 dimensions, respectively. Since this is a randomised dimensionality reduction, we repeated each

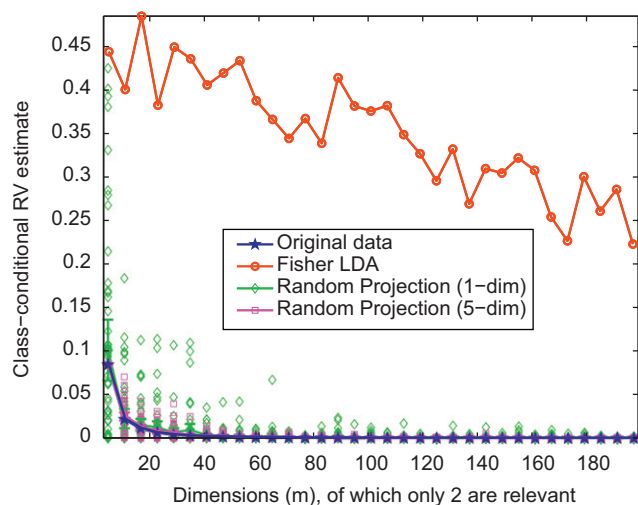


Fig. 2. Experiment with $y \in \{-1, 1\}$ given, standard Gaussian additive noise, in a severe distance concentration setting: $a_1 = a_2 = 1$; $a_{i>2} = 0$, i.e. an increasing number of irrelevant noise features is present. We compare the concentration of distances as it occurs in the original space, in Fisher's LDA projected space, and in random projection spaces of $k=1$ and 5 dimensions, respectively. Fisher's LDA is clearly most able to keep the distances from becoming meaningless.

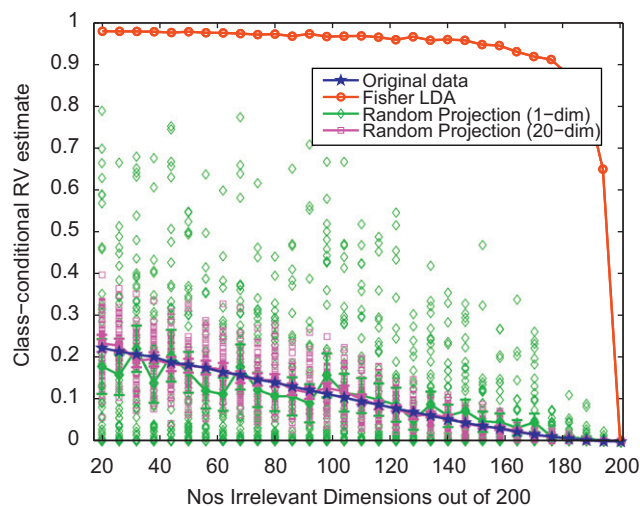


Fig. 3. Experiment in light-to-moderate-to-severe distance concentration settings, with $y \in \{-1, 1\}$ given, and standard Gaussian additive noise. The degree of severity is varied by having as many entries a_i set to 0 as the x-axis indicates, and the remaining ones are set to 1. We compare the concentration of distances occurring in the original space, in Fisher's LDA projected space, and in random projection spaces on $k=1$ and 20 dimensions. Again, Fisher's LDA is clearly most able to keep the distances from becoming meaningless, and succumbs to distance concentration only when no feature contains anything but noise.

experiment 50 times, and show their average (solid line), as well as the individual class-conditional relative variances (markers without line). We see that, for both values of k tested, the average values of RV_k of the random projections follow closely the sequence of RV_m of the original data—hence preserving the distance concentration problem as such into the reduced space. Observe the random fluctuations of the individual RV_k values concentrate tighter and tighter around their average (very close to zero) as the dimensionality m increases. We can conclude therefore that, of the class of methods tested, Fisher's LDA is most able to transform the nearly meaningless distances in high dimensions into more meaningful ones in the reduced space.

The settings depicted in Fig. 3 are very different from the previous, and have been designed to show a wider range of

severity of the distance concentration phenomenon, from light to moderate to severe. Here we have $m=200$ dimensions in all experiments, and the number of relevant features is varied. That is, in the setting at the leftmost end of the figure, we have all the 200 features relevant (they all contain contribution from the systematic factor y , i.e. have all associated entries of \mathbf{a} non-zero). Progressing on the x-axis from left to right, the number of features switched to irrelevant (i.e. their associated entry a_i is set to 0, so that the feature x_i no longer has any contribution from y) increases. All other characteristics are the same as in the previous set of experiments, namely $y \in \{-1, 1\}$ are given with the sample, and the additive noise is standard Gaussian. We estimate the sequences RV_m in the data space and compare with their counterpart in the reduced spaces produced by the competing methods. We see from the figure once again that the sequence of class-conditional RV_m estimates in the original data space are preserved by the average class-conditional RV_k values of the random projections for both values of k tested. The fluctuations around these values, made by the individual random projections, are now larger than previously, since in the case of light and mild conditions of distance concentration there is still some contrast between small and large distances in the original space. However, Fisher's LDA most apparently stands out in all settings, by producing projections of the data that increases the class-conditional relative variance of the pairwise distances. Indeed, we see that it only succumbs to concentration when all features turn into just noise (see the rightmost end of the plot). This demonstrates once more the abilities of Fisher's LDA projections to enhance the contrast of the pairwise distances.

An interesting open issue is to find out how does distance concentration affect classification performance? Does the new-found ability of Fisher's LDA translate into good classification performance in difficult settings? Although answering this through theoretical analysis needs further research, we find it enlightening to give empirical results for the 0–1 generalisation error of Fisher's LDA, as estimated on a test set that we generated from the same models as in the previous experiments, and using the 200 points with given class labels as a training set. These results are given in Figs. 4 and 5, respectively. The results suggest that, in severe cases of distance concentration, the use of Fisher's LDA in the original space is a better choice, even if its use may be computationally demanding. This is because in such settings, where the distances in the data space are close to meaningless, Fisher's LDA helps to enhance the contrast between small and large distances—whereas random projections' strength is to preserve the original distances. Alternatively, when there is reason to believe that some features only contain noise, then the use of feature ranking and filtering may be also beneficial—this will be the subject of the next section, and will be also demonstrated on real data, in combination with Fisher's LDA, shortly.

In turn, in cases where distance concentration issues are not severe, so the distances are meaningful in the high dimensional original space, then the use of random projections bring computational savings without sacrificing the classification performance. We see this convincingly in Fig. 5. This also agrees with the analysis in [12].

4.1.2. Fisher's DR and SIS

We now demonstrate our findings regarding the distance concentration awareness of the feature ranking methods identified in the Sections 3.2 and 3.3. We generate data from the model (1), first in the classification setting, i.e. the targets $y \in \{-1, 1\}$ (100+100 points), and then with continuous targets $y \in \mathbb{R}$ (200 points). Fig. 6 summarises the results, for the two regimes of

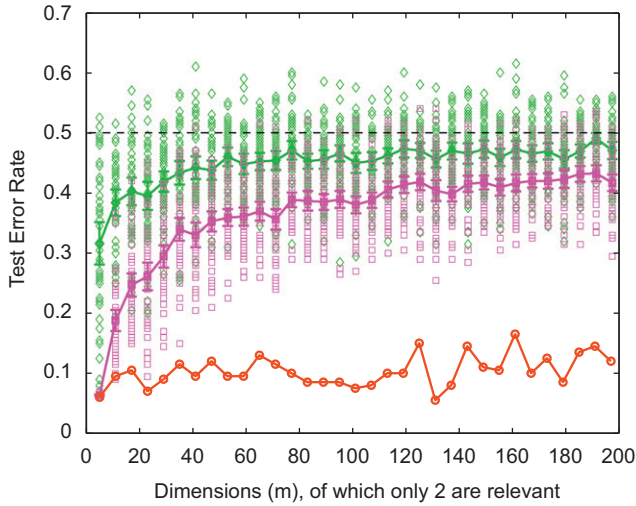


Fig. 4. Misclassification error rates in severe distance concentration conditions. The data are the same as in the experiments for Fig. 2. 'o': Fisher's LDA in the original data space; \diamond : Fisher's LDA on $k=1$ dimensional random projections of the data; \square : Fisher's LDA on $k=5$ dimensional random projections of the data. We see that Fisher's LDA is much more effective in the original data space in these settings. There is little scope for random projections in such data conditions.

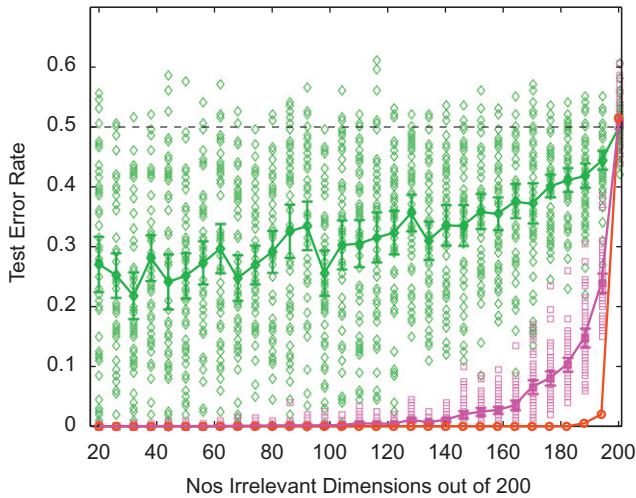


Fig. 5. Misclassification error rates in light-to-moderate-to-severe distance concentration settings. The data are the same as in the experiments for Fig. 3. 'o': Fisher's LDA in the original data space; \diamond : Fisher's LDA on $k=1$ dimensional random projections of the data; \square : Fisher's LDA on $k=20$ dimensional random projections of the data. In not so severe distance concentration conditions Fisher's LDA works well not only in the original space, but also on random projections of the data. This is where random projections are most advantageous, since they save computation time.

severity of the problem in the same spirit as previously (see caption for more details). We see from these plots that, in each case, the 1-dimensional class-conditional relative variance estimates of the individual features very clearly tell apart the relevant features from the irrelevant ones. We also see, as expected from the theoretical analysis, that Fisher's discriminant ratio method (for classification targets) and the sure independence screening method (for continuous targets) have the same behaviour.

4.1.3. The influence of the kurtosis of the latent variable

The last set of synthetic data experiments demonstrates the findings of Section 3.4, namely the effect of the kurtosis of y for

the target-conditional relative variance of the data in the model under study.

We considered three different distributions with 0-mean and unit variance, but with differing kurtosis to instantiate the latent variable y : Laplace (kurt=3), Gaussian (kurt=0), and uniform (kurt=-1.2) distributions. We tested the effect of this in the model (1), again with standard normal noise, in two different scenarios: (a) first, by setting $a_i=1, \forall i=1, \dots, m$ all features are relevant; In this case the analytic limits of RV_m as $m \rightarrow \infty$ are $(\text{kurt}(y)+2)/4$ (cf. Section 3.4). (b) Second, by setting $a_i=1, i=1, \dots, 20$, and $a_i=0, i=21, \dots, m$ we have 20 relevant and an increasing number of irrelevant features. In this case, the analytic limit of RV_m is zero.

Fig. 7 shows the sequences of RV_m estimates as the data dimension m increases, superimposed with the analytical limits for the above two settings. We see the sequence of RV_m estimates converges to the analytical limit in all experiments. As predicted by the theory, we see the RV_m sequence stays non-zero to arbitrary high dimensions when the features are all relevant, and the limit is higher when the kurtosis of the generator variable is larger (left hand plot). On the right hand plot we have the case with a small number of relevant and an increasing number of irrelevant features. The analytical limits at $m \rightarrow \infty$ are all zero in this case, irrespectively of the distribution or kurtosis of y . However, we see the estimated sequences of RV_m approach zero slower when the kurtosis of y is larger. Hence, maximising the kurtosis is still beneficial in finite dimensional data sets.

4.2. Benchmark tests on high dimensional real data

In high dimensional real data settings, several aspects of the dimensionality curse come into play, and distance concentration is most likely not the only source of problems. It is therefore difficult, if not impossible, to verify the effects of these factors in isolation. Moreover, the data models that serve as a simplified abstraction for formal analysis are also unlikely to hold in practice. It is therefore of interest, in addition to the previous tests, to see how well the methods—that we have identified as distance concentration aware—do actually perform in comparison with the state of the art. We have chosen benchmark data sets from computational biology for our experiments, namely gene expression arrays, since this is an area where the concern of distance concentration has been most explicitly raised [6]. As we shall see, the links uncovered through the theoretic analysis also suggest advantageous combinations of these techniques.

4.2.1. Classification of microarray data

Fisher's LDA has been the core of many successful classifiers. Here we investigate and demonstrate this further, combining it with feature pre-selection using the ranking approach described in Sections 3.2 and 3.3. As already observed in the previous sections, for binary targets in $\{-1, 1\}$, SIS coincides with Fisher's discriminant ratio. Our rational is that features (genes) that rank low in terms of their relative variance are better eliminated early, since otherwise they contribute noise and the problem of distance concentration appears more severe. After this phase, the remaining features may still contain considerable noise, mixed with useful information. This can no longer be treated by eliminating features, however, FLDA's capability to project the data such that the relative variance increases, i.e. the contrast between small and large distances is enhanced, can be exploited. As we shall see, the final classification performance can improve by this simple approach. Of course, we should note that besides our rational that, in this paper, comes solely from the objective of increasing the relative variance, there is also other theoretical

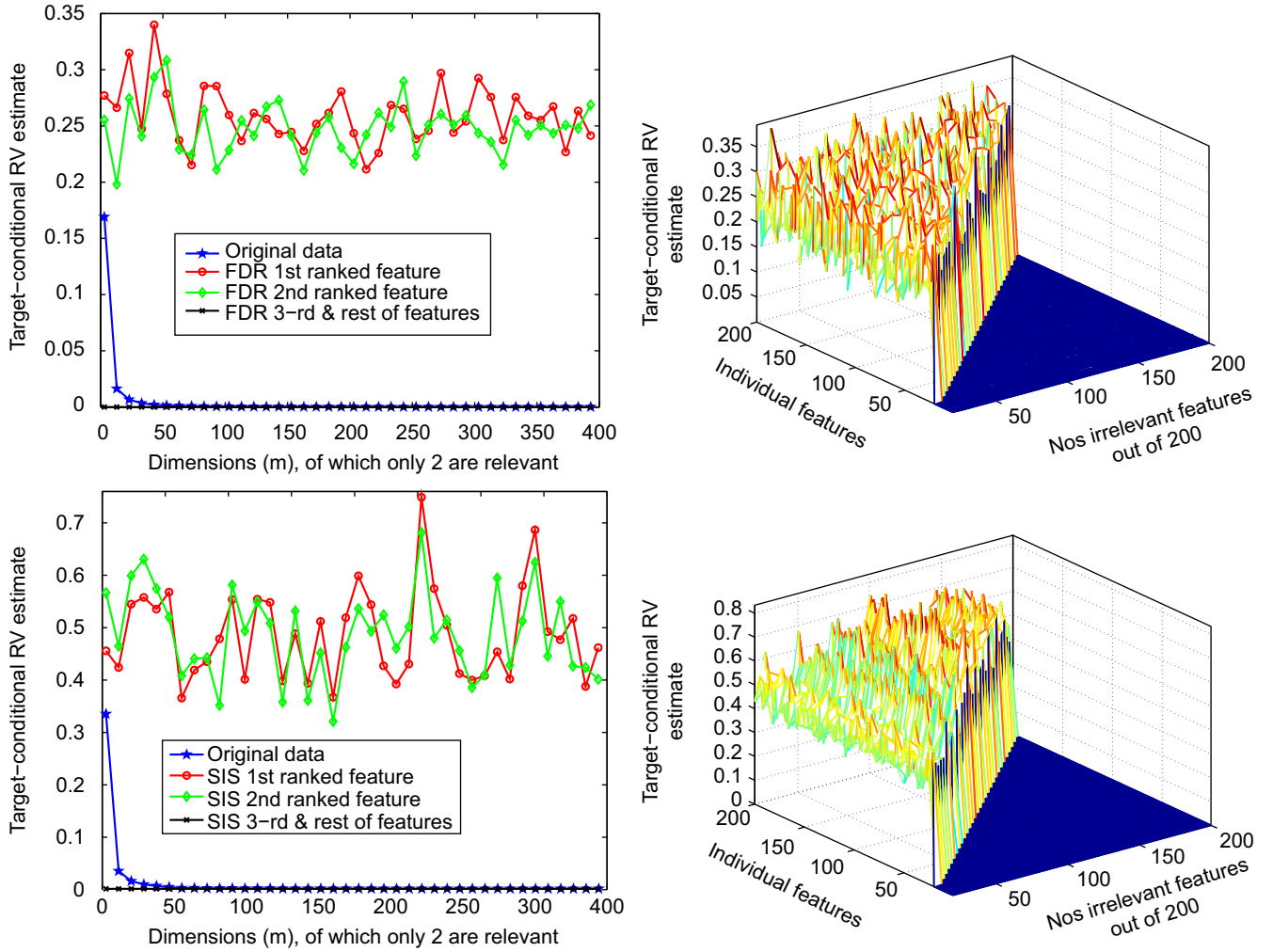


Fig. 6. Experiments with feature ranking. *Top row:* The data follows the model (1) with classification targets $y \in \{-1, 1\}$, and standard Gaussian noise. Cf. Section 3.2, in this case the rank order of the target-conditional relative variance estimates coincides with the rank order given by Fisher's discriminant ratio. *Bottom row:* The data follow the model (1) with continuous targets $y \sim N(0, 1)$, and standard Gaussian noise. Cf. Section 3.3, in this case the rank order of the target-conditional relative variance estimates coincides with the rank order given by sure independent screening. *Left hand plots* (in both rows): Severe distance concentration settings, with only two relevant features ($a_1 = a_2 = 1$; $a_i > 2 = 0$) and an increasing number of irrelevant noise features. We see the sequence RV_m of the original data goes to zero rather quickly. Superimposed, we have the sequences RV_i of top ranked individual features. The first two, corresponding to the two relevant features maintain values of relative variance that are well away from zero. The third ranked, and all the others have a relative variance that are practically zero. *Right hand plots* (in both rows): The total number of features is constantly 200, of which the number of relevant ones is varied linearly by switching to zero their corresponding coefficient in \mathbf{a} . The third dimension depicts the target-conditional RV_i estimates for each individual feature. We see that the relevant ones have their relative variances consistently away from zero, whereas those of the irrelevant features are practically zero.

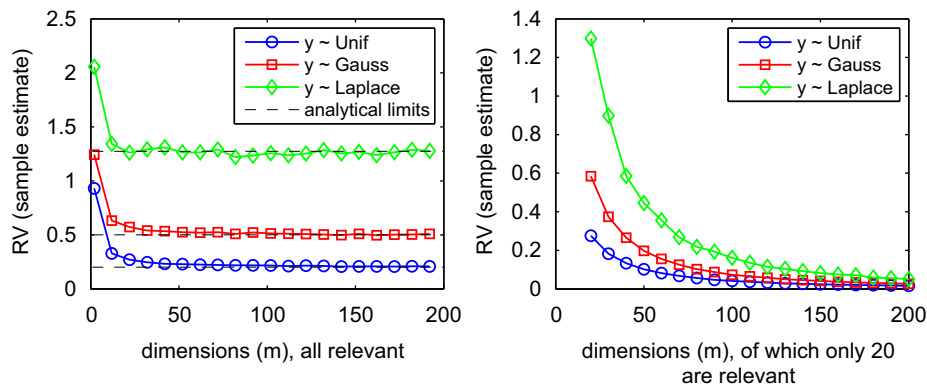


Fig. 7. Experiments showing the effect of the kurtosis of y . *Left:* All features relevant; $a_i = 1$, $\forall i = 1, \dots, m$. We see the sequences RV_m converge to the analytic limits and the values of these are larger when the kurtosis of y is larger. *Right:* 20 relevant and an increasing number of irrelevant features; $a_i = 1$, $i = 1, \dots, 20$, $a_i = 0$, $i = 21, \dots, m$. The analytic limits of RV_m as $m \rightarrow \infty$ are all zero in this case, irrespectively of the distribution of y . We see the estimated sequences of RV_m approach zero slower when the kurtosis of y is larger.

justification for pre-filtering features in this way, cf. the analysis of SIS [14]. However, the combination of SIS with FLDA has not been considered previously.

We use four publicly available benchmark data sets of gene arrays from previous studies [2,30,18,29,9], so that we can compare our results with state-of-the-art performance. The features are genes, and the samples correspond to various experimental conditions. The task is to classify a previously unseen gene array into one of the conditions or classes. The characteristics of the data sets that we use, and their original sources, are given in Table 1.

We followed the experimental protocol in [9], which also allows us to have a fair comparison with results quoted from that work. We split the data randomly in two thirds for parameter estimation (training) and use the remaining one third for testing. Table 2 reports the medians and inter-quartile-ranges of the misclassification error rates from 500 independent repeats of the random splits. The compared methods are: diagonal Fisher's LDA (also tested in [9]), Fisher's LDA on a reduced set of just the top $N/\log(N)$ genes cf. the ranking provided by FDR (equivalently SIS), and thirdly the same using the top $N-1$ genes cf. the same ranking. These two cut-offs were chosen as suggested for SIS in [14], where it was shown that SIS has the property that its top ranked $\mathcal{O}(N/\log(N))$ features contain all the relevant ones with high probability. For comparison with the state of the art, we use a recently developed technique called Bayesian logistic regression (BLOGREG) by [5], and the best results quoted from [9], obtained with an SVM.

We can see from Table 2 that the combination SIS/FDR + Fisher's LDA achieves an accuracy comparable with the state of the art, and it is in fact the best performing method on three out of the four data sets tested. Of these, the difference is significantly superior on 'colon' and 'breast'. SVM wins over on one of the data sets ('prostate'). We also see that the choice of cut-off ($N/\log(N)$ or $N-1$) has a little influence on the final result. This was also observed for SIS in [14], and it is because the data indeed has many irrelevant features. In addition, in our case, the possibly remaining noise features are dealt with successfully by Fisher's LDA. We should also note in passing, that BLOGREG [5] could be shown to be distance concentration aware, through a reasoning similar to [23].

4.2.2. Reconstruction of gene regulatory networks

Here we demonstrate the advantage of the ideas of previous sections in an application to the problem of reconstructing gene regulatory networks from experimental data. This is a central issue in computational biology, and both (sparse) linear regression [28,27], and linear correlation learning [26] are widely used strategies at present. Such linear models provide a good approximation of an ODE system when linearised around the steady state. However, so far, correlation learning and sparse regression have been seen as competing approaches, and the combination of these two complementary techniques has not yet been investigated in the context of this problem.

In analogy with our reasoning in the previous subsection, we shall proceed in two stages: The first stage ranks the features (genes) in terms of their relative variance, and retains the top ranked ones only. As we have seen earlier in Section 3.3, this screening approach is equivalent to correlation learning, and the order of the ranks is the same as that of SIS. For this reason, since it has been shown in SIS that the top $\mathcal{O}(N/\log(N))$ ranked features of SIS contain all the important features with high probability [14], we will again use the cut-offs suggested in SIS, and retain only the top $N/\log(N)$, or the top $N-1$ of the features, where N is the number of training points. As pointed out in [14], and as we shall see once again shortly, the exact cut-off will have a little impact on the final outcome.

To assess the effectiveness of this feature reduction for the task at hand, we have chosen to employ compressed sensing [4] (CS) in the second phase of our analysis. This is a sparse recovery technique that is close to sparse regression, but focusing on the recovery of a sparse vector rather than on prediction, which makes it well suited to the task. CS comes with well-known guarantees on the reconstruction of a sparse vector (these represent gene connections in this application) as a function of its dimensionality and number of non-zero components. These rest on certain preconditions (that of course, may or may not be satisfied in practice), nevertheless, CS has been found quite robust to these. Therefore, if the feature reduction was a good one in the first phase, then we may expect a drastic reduction of the computation time in the second phase without sacrificing, by contrary, slightly enhancing the accuracy of the recovered network.

Table 1
Data characteristics.

Name	Source	#samples (N)	#features (m)	Description
Colon	Alon et al. [1]	62	2000	Tumour vs. normal tissue
Breast	West et al. [30]	49	7129	ER+ vs. ER-tumour
Leukaemia	Golub et al. [18]	72	3571	AML vs. ALL types
Prostate	Singh et al. [29]	102	6033	Tumour vs. normal tissue

Table 2
Misclassification rates.

	Colon	Breast	Leukaemia	Prostate
FLDA on all features	16.66 \pm 0.34	23.07 \pm 0.59	1.87 \pm 0.14	15.18 \pm 0.49
SIS/FDR+LDA ($N-1$ f.)	11.11 \pm 0.29	15.38 \pm 0.44	4.76 \pm 0.14	10.00 \pm 0.27
SIS/FDR+LDA ($N/\log N$ f.)	11.11 \pm 0.28	15.38 \pm 0.43	1.22 \pm 0.12	10.00 \pm 0.23
BLOGREG [5]	16.66 \pm 0.39	23.07 \pm 0.49	4.76 \pm 0.24	10.00 \pm 0.22
SVM (quoted from [9])	15.5	–	1.83	7.88

Median \pm iqr computed from 500 random partitions into training sets (two thirds of the data) and test sets (one third of the data)—except the SVM results, which are taken from [9] for reference purpose only, and which represent the mean over 50 random splits in the same proportion.

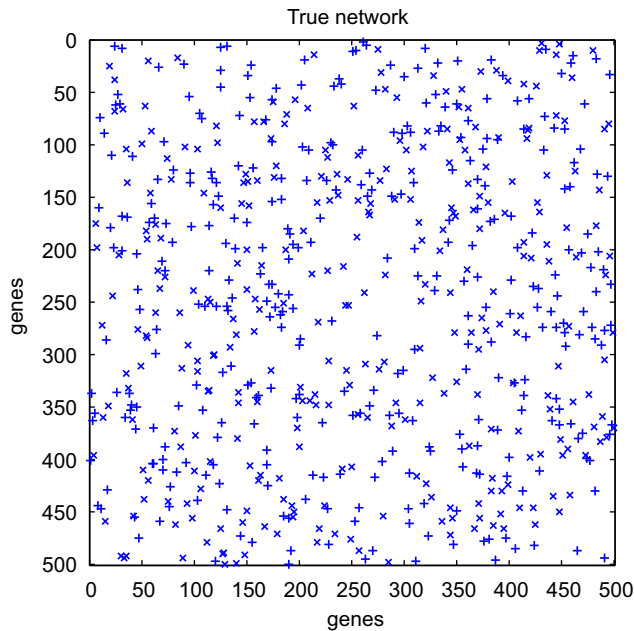


Fig. 8. The 'ground truth' network underlying our experiments. It contains 351 excitatory (+) and 317 inhibitory (×) connections.

To test this idea, we carried out experiments on a realistic network of 500×500 genes, synthesised in exactly the same manner as described in [26]. This setting allows us to objectively quantify our results by comparing the recovered graph of gene connections with the ground truth. We created data as follows (for full details on data generation see [26] or the supplementary material of [27]; note, however, that their studies were conducted on much smaller, 30×30 networks). An underlying 'ground truth' network (a directed graph) of 500×500 genes was first created by generating outward connections randomly from a power-law distribution on the out-degree for each gene, and assigning +1 (excitatory) or -1 (inhibitory) to the resulting connections uniformly. This network is depicted in Fig. 8. This is an instance of the so-called scale-free networks, which is a type of network that is known to describe many complex natural systems. One of the striking properties of such networks, that plays in favour of reverse-engineering it from relatively little data, is the sparsity of their edge structure.

From this network we then generate data by evaluating the steady state of a system of differential equations [26] through integrating it numerically. This gives 'wild type' arrays. In addition, mutants are created by holding the expression of a knocked-out gene at zero during the integration. To have a comprehensive set of tests in our experiments, we knocked out each gene in turn, so we have a mutant and a wild type array for each gene. Note, however that, as in [26], we do not assume the availability of mutants of the other genes while solving the recovery task for the connection into a particular gene. The design matrix will consist of a small number, R , of replicates of these two experiments for the gene of interest. The replicates are created by sampling R times from the steady state, and adding a Gaussian noise of standard deviation σ . So we have $2R$ expression levels per gene (500-dimensional each), from which our task is to recover the regulatory connections into the gene. This is repeated 500 times (i.e. for each gene in turn), and the full 500×500 graph is then assembled from the recovered sparse 500-dimensional vectors that represent the incoming edges into each of the genes.

Table 3 summarises the results, across several problem settings (by varying the noise level and the number of replicates

available), for the two cut-off values suggested in [14] in the screening stage, as well as for recovery from all the data. In each case, the performance is measured by matching the whole 500×500 network, as recovered from the data, against the ground truth network, from which the relevant criteria are calculated. These criteria are the sensitivity—i.e. the number of true connections recovered divided by the total number of true connections—and the complementary specificity—i.e. the number of false recoveries divided by the total number of non-existing connections. Both are values between 0 and 1. For the sensitivity, highest value is best, since a low sensitivity would mean that true connections are missed. For the complementary specificity, lowest value is best, since a high complementary specificity would mean that non-existing connections are being falsely discovered.

We see from Table 3 the beneficial effect of the SIS stage: It drastically reduces the overall computation time, and in terms of accuracy it is no worse, but even slightly better than the version that uses all the features. We also see that the exact number of features retained in the SIS stage has a little impact on the accuracy. This is because the vector of true connections is very sparse in a scale-free network, the genes whose relative variance is low are those that receive a little (no) contribution from the target gene, hence zeroing out connections from such irrelevant genes in the first stage leaves a smaller pool of candidates to deal within the second stage. Since the second stage employs a sparse recovery technique, this will further zero out the remaining irrelevant genes/connections. Indeed, we clearly see from the table that retaining just $N/\log N$ genes, i.e. zeroing out all but this number of connections before performing the sparse recovery on the remaining ones slightly improves the complementary specificity, whereas retaining $N-1$ genes enhances the sensitivity.

5. Conclusions

We made a first investigation into the question of whether or not certain existing data analysis techniques would be still suitable when faced with the counter-intuitive phenomenon of distance concentration in high dimensional data spaces. Our analysis was made under the structural assumption of one generating latent variable swamped by unstructured noise, and has examined several dimensionality reduction scenarios that would maximise the relative variance of the transformed data. We found that dimensionality reduction that maximises a sample estimate of the leading term of the relative variance under the model recovers the well-known techniques of Fisher's linear discriminant analysis, Fisher's discriminant ratio and a variant of projection pursuit. Hence, these techniques may be seen as distance concentration aware within the linear model class considered, despite they have not been explicitly designed to have this property. In addition, the same analysis in the regression setting has uncovered a link with sure independent screening, a recently proposed technique to reduce up to ultra-high dimensions down to moderate. Finally, we note that our analysis is more general than the named existing techniques in that, other than the structural assumption imposed by the model that we studied, which impacts the dependencies between the features, throughout our theoretic analysis we made no assumptions about the distributional form of the variables involved.

Future work may extend this analysis to other data models. Also, having elucidated the complementary nature of the strengths of the dimensionality reduction techniques we identified and those of random projections, a worthwhile direction for further research would be to investigate ways in which their

Table 3

Results for the gene regulatory network recovery experiments, in terms of sensitivity = TP/(TP+FN) (higher is better); complementary specificity = FP/(FP+TN) (lower is better), and CPU time (in seconds).

Method	Eval. criteria	6 replicates (i.e. $N=12$)			10 replicates (i.e. $N=20$)		
		$\sigma = 0.01$	$\sigma = 0.02$	$\sigma = 0.05$	$\sigma = 0.02$	$\sigma = 0.05$	$\sigma = 0.1$
SIS+CS ($N/\log N$ f.)	Sens.	0.997	0.951	0.647	0.981	0.797	0.471
	C.spec.	0.007	0.007	0.008	0.011	0.012	0.013
	CPU time	1.06	1.25	1.37	1.28	1.07	1.09
SIS+CS ($N-1$ f.)	Sens.	0.998	0.971	0.717	0.987	0.854	0.610
	C.spec.	0.019	0.019	0.020	0.035	0.036	0.036
	CPU time	6.68	7.03	6.84	8.53	7.62	7.91
CS (all features)	Sens.	0.951	0.901	0.635	0.950	0.770	0.475
	C.spec.	0.019	0.020	0.020	0.035	0.036	0.037
	CPU time	128.87	149.47	148.17	170.54	165.77	162.25

It should be noted that correlation learning without a thresholding would have both its sensitivity and specificity close to 1.

combination could retain the best of both. Initial results in this direction [12] are quite promising.

Acknowledgement

This work was supported by an MRC Discipline Hopping Award (Grant G0701858).

Appendix

Derivations: The expectation and variance of the Euclidean norm is computed as follows, for the model defined in Section 2.

$$E[\|\mathbf{x}\|_2^2] = E\left[\sum_{i=1}^m (a_i y + \delta_i)^2\right] = E[y^2] \sum_{i=1}^m a_i^2 + \sum_{i=1}^m E[\delta_i^2] + 0 \quad (32)$$

where we used that $2E[a_i y \delta_i] = 2a_i E[y] E[\delta_i]$, since the noise is independent of the systematic component, and $E[\delta_i] = 0$ because the noise is zero-mean.

$$\text{Var}[\|\mathbf{x}\|_2^2] = \text{Var}\left[\sum_{i=1}^m (a_i y + \delta_i)^2\right] = \sum_{i=1}^m \sum_{j=1}^m \text{Cov}[(a_i y + \delta_i)^2, (a_j y + \delta_j)^2] \quad (33)$$

$$= \sum_{i=1}^m \sum_{j=1}^m \text{Cov}[a_i^2 y^2 + \delta_i^2 + 2a_i y \delta_i, a_j^2 y^2 + \delta_j^2 + 2a_j y \delta_j] \quad (34)$$

$$= \sum_{i=1}^m \sum_{j=1}^m a_i^2 a_j^2 \text{Var}[y^2] + a_i^2 \text{Cov}[y^2, \delta_j^2] + 2a_i^2 a_j \text{Cov}[y^2, y \delta_j] + a_j^2 \text{Cov}[\delta_i^2, y^2] + \text{Cov}[\delta_i^2, \delta_j^2] + 2a_j \text{Cov}[\delta_i^2, y \delta_j] + 2a_i a_j^2 \text{Cov}[y \delta_i, y^2] + 2a_i \text{Cov}[y \delta_i, \delta_j^2] + 4a_i a_j \text{Cov}[y \delta_i, y \delta_j] \quad (35)$$

In the above, $\text{Cov}[y^2, \delta_j^2] = \text{Cov}[\delta_i^2, y^2] = 0$, since by the model assumption δ_i are independent of y , therefore the second and the fourth terms give 0; $\text{Cov}[y^2, y \delta_j] = \text{Cov}[y \delta_i, y^2] = 0$, since $E[\delta_i] = 0$, so the third and the seventh terms also give 0. The remaining covariance terms can be written as

$$\sum_{i,j} \{4E[y^2] a_i a_j E[\delta_i \delta_j] + 4E[y] a_i E[\delta_i \delta_j^2] + \text{Cov}(\delta_i^2, \delta_j^2)\} \quad (36)$$

References

[1] N. Alon, Problems and results in extremal combinatorics, part I, Discrete Mathematics 273 (2003) 31–53.

[2] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Mack, J. Levine, Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proceedings of National Academy Sciences of the United States of America 96 (1999) 6745–6750.

[3] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is nearest neighbor meaningful? in: Proceedings of International Conference on Database Theory (ICDT), 1999, pp. 217–235.

[4] E. Candès, J. Romberg, ℓ_1 -magic: recovery of sparse signals via convex programming, <http://www.acm.caltech.edu/l1magic/>, 2005.

[5] G.C. Cawley, N.L.C. Talbot, Gene selection in cancer classification using sparse logistic regression with Bayesian regularisation, Bioinformatics 22 (19) (2006) 2348–2355.

[6] R. Clarke, H.W. Ransom, A. Wang, J. Xuan, M.C. Liu, E.A. Gehan, Y. Wang, The properties of high-dimensional data spaces: implications for exploring gene and protein expression data, Nature Reviews Cancer 8 (January) (2008) 37–49.

[7] S. Dasgupta, Learning mixtures of Gaussians, in: Proceedings of the 40th Annual Symposium on Foundations of Computer Science FOCS, 1999.

[8] S. Dasgupta, A. Gupta, An elementary proof of the Johnson–Lindenstrauss lemma, Random Structures and Algorithms 22 (2002) 60–65.

[9] M. Dettling, BagBoosting for tumor classification with gene expression data, Bioinformatics 20 (18) (2004) 3583–3593.

[10] P. Diaconis, D. Freedman, Asymptotics of graphical projection pursuit, Annals of Statistics 12 (1984) 793–815.

[11] R.J. Durrant, A. Kabán, When is ‘nearest neighbour’ meaningful: a converse theorem and implications, Journal of Complexity 25 (4) (2009) 385–397.

[12] R.J. Durrant, A. Kabán, Compressed Fisher linear discriminant analysis: classification of randomly projected data, in: Proceedings of the 16-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2010), 2010, pp. 1119–1128.

[13] I. Guyon, et al., Gene selection for cancer classification using support vector machines, Machine Learning 46 (1–3) (2002) 389–422.

[14] J. Fan, J. Lv, Sure independence screening for ultra-high dimensional feature space, Journal of the Royal Statistical Society Series B 70 (5) (2009) 849–911.

[15] D. François, V. Wertz, M. Verleysen, The concentration of fractional distances, IEEE Transactions on Knowledge and Data Engineering 19 (7) (2007) 873–886.

[16] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.

[17] C. Giannella, New instability results for high-dimensional nearest-neighbor search, Information Processing Letters 109 (2009) 1109–1113.

[18] T. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[19] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference and Prediction, Springer, Berlin, 2001.

[20] C.-M. Hsu, M.-S. Chen, On the design and applicability of distance functions in high-dimensional data space, IEEE Transactions on Knowledge and Data Engineering 21 (4) (2009) 523–536.

[21] A. Hyvärinen, E. Oja, Independent component analysis by general non-linear Hebbian-like learning rules, Signal Processing 64 (3) (1998) 301–313.

[22] P. Indyk, R. Motwani, Approximate nearest neighbors: towards removing the curse of dimensionality, in: STOC, 1998, pp. 604–613.

[23] A. Kabán, R.J. Durrant, A norm-concentration argument for non-convex regularization, in: ICML/UA/COLT Workshop on Sparse Optimization and Variable Selection, 9 July 2008, Helsinki, Finland.

[24] R.H. Lilien, H. Farid, B.R. Donald, Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum, Journal of Computational Biology 10 (6) (2003) 925–946.

[25] V. Pestov, On the geometry of similarity search: dimensionality curse and concentration of measure, Information Processing Letters 73 (2000) 47–51.

- [26] J. Rice, Y. Tu, G. Stolovitzky, Reconstructing biological networks using conditional correlation analysis, *Bioinformatics* 21 (6) (2005) 765–773.
- [27] S. Rogers, M. Girolami, A Bayesian regression approach to the inference of regulatory networks from gene expression data, *Bioinformatics* 21 (14) (2005) 3131–3137.
- [28] F. Steinke, M. Seeger, K. Tsuda, Experimental design for efficient identification of gene regulatory networks using sparse Bayesian models, *BMC Systems Biology* 1 (51) (2007) 1–15.
- [29] D. Singh, P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, et al., Gene expression correlates of clinical prostate cancer behavior, *Cancer Cell* 1 (2002) 203–209.
- [30] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, A.J. Olson Jr., J.R. Marks, J.R. Nevins, Predicting the clinical status of human breast cancer by using gene expression profiles, *Proceedings of National Academy Sciences of the United States of America* 98 (2001) 1146–1467.

Ata Kaban is a lecturer in the School of Computer Science of the University of Birmingham. Her current interests concern statistical machine learning, high dimensional data analysis, probabilistic modelling of data, and Bayesian inference. She received her B.Sc. degree with honours (1999) in Computer Science from the University "Babes-Bolyai" of Cluj-Napoca, Romania, and the Ph.D. degree in Computer Science (2001) from the University of Paisley, UK. She has been a visiting researcher at Helsinki University of Technology (June–December 2000 and in the summer of 2003) and at HIIT BRU, University of Helsinki (September 2005). Prior to her career in Computer Science, she received the B.A. degree in musical composition (1994) and the M.A. (1995) and the Ph.D. (1999) degrees in musicology from the Music Academy "Gh. Dima" of Cluj-Napoca, Romania.