Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **60**(1), 2018, 4–19 doi: 10.1111/anzs.12212

A survey of high dimension low sample size asymptotics

Makoto Aoshima[1,*] **Dan Shen**[2] **Haipeng Shen**[3]
Kazuyoshi Yata[1] **Yi-Hui Zhou**[4] **and J. S. Marron**[5]

University of Tsukuba, University of South Florida, University of Hong Kong, North Carolina State University and University of North Carolina

Summary

Peter Hall's work illuminated many aspects of statistical thought, some of which are very well known including the bootstrap and smoothing. However, he also explored many other lesser known aspects of mathematical statistics. This is a survey of one of those areas, initiated by a seminal paper in 2005, on high dimension low sample size asymptotics. An interesting characteristic of that first paper, and of many of the following papers, is that they contain deep and insightful concepts which are frequently surprising and counter-intuitive, yet have mathematical underpinnings which tend to be direct and not difficult to prove.

Key words: canonical correlations; classification; geometric representation; hypothesis testing; principal component analysis

1. Introduction

The authors of this paper deeply mourn the loss of Peter Hall, who was the premier mathematical statistician of his era. While his body of work on the bootstrap and on statistical smoothing is widely known and appreciated, less well known is his work in many other areas. Nevertheless that work continues, and we believe will continue long into the future, to influence the course of mathematical statistics in many directions.

Here we survey one such area, which was designated *High Dimension Low Sample Size (HDLSS) Asymptotics* in Hall, Marron & Neeman (2005). The limits studied there

[*]Author to whom correspondence should be addressed.
[1]Institute of Mathematics, University of Tsukuba, Ibaraki, 305-8571, Japan
e-mail: aoshima@math.tsukuba.ac.jp
[2]Interdisciplinary Data Sciences Consortium, Department of Mathematics & Statistics, University of South Florida, Tampa, FL 33620, USA
[3]Innovation and Information Management, Faculty of Business and Economics, University of Hong Kong, Hong Kong City, Hong Kong
[4]Departments of Biological Sciences, Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27695, USA
[5]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27514, USA

Acknowledgements. The research of the first author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Exploratory Research, Japan Society for the Promotion of Science (JSPS), under Contract Numbers 15H01678 and 26540010. The research of the third author was partially supported by US NSF grant DMS-1407655 and HKU Stanley Ho Alumni Challenge Award. The research of the fourth author was partially supported by Grant-in-Aid for Young Scientists (B), JSPS, under Contract Number 26800078. We gratefully acknowledge the CF patients, the Cystic Fibrosis Foundation, the UNC Genetic Modifier Study and the Canadian Consortium for Cystic Fibrosis Genetic Studies, funded in part by Cystic Fibrosis Canada and by Genome Canada through the Ontario Genomics Institute per research agreement 2004-OGI-3-05, with the Ontario Research Fund-Research Excellence Program, and supported in part by R21HG007840, R01HL068890R01. The research of the fifth author was partially supported by US NSF Grant 1633074.

were taken as the dimension d tended to ∞, while the sample size n was held fixed. This was quite a departure at the time from the far more conventional asymptotics based on n tending to ∞ with d fixed. However, this became the foundation of research in mathematical statistics on very high dimensional data analysis, which is an important part of the currently fashionable area called *Big Data*. Scientific areas where these insights have proven to be very useful include genetics and other types of cancer research, neuroscience and also image and shape analysis.

In Section 2 we discuss some early surprises that were discovered in that first paper. A particularly insightful use of HDLSS asymptotics was made in the analysis of properties of various statistical methods, such as *principal component analysis* (PCA) and various approaches to the *classification* (i.e. discrimination) problem. An overview of these results appears in Section 3, together with results on canonical correlations and hypothesis testing. The insights gained from HDLSS mathematics have also been quite useful in the invention of new statistical methods, as discussed in Section 4. Section 5 describes the relationship between HDLSS and other domains of asymptotics.

An interesting historical note is that Hall, Marron & Neeman (2005) was not the first statistical paper to use HDLSS asymptotics. As far as we know, the first work in this area was done by Casella & Hwang (1982), in the special context of Stein estimation.

An ironical anecdote is that a draft of Hall, Marron & Neeman (2005) was first submitted to *Biometrika*. That submission was rejected on the grounds that the technical assumption that the distributions should satisfy a mixing condition along the coordinates was viewed as too artificial. The reviewers could not conceive of a situation where that type of assumption would be relevant to real data. Fortunately, the reviewers at the Journal of the Royal Statistical Society had the vision to understand the large scale implications of the mathematics involved. It was later discovered that precisely this assumption is very natural in the area of *Genome Wide Association Studies* (GWAS), because of the way that chromosomes are combined during procreation. Another irony is that the first publication on GWAS, Klein *et al.* (2005), appeared in the *same year* as Hall, Marron & Neeman (2005).

2. Geometric representation

An important component of the initial exploration of HDLSS asymptotics in Hall, Marron & Neeman (2005) was the revelation of a surprisingly rigid geometric structure that naturally emerges from the asymptotics in an increasingly random setting. This phenomenon was termed *geometric representation* in that paper. There are two fundamental concepts that arise here.

The first key idea is revealed by considering a single realisation of the standard Gaussian distribution in d dimensions, $\boldsymbol{Z} \sim \mathrm{N}(0, I)$. The length of the vector $\boldsymbol{Z} = (Z_1, \ldots, Z_d)^\top$ is $\|\boldsymbol{Z}\| = (\sum_{i=1}^{d} Z_i^2)^{1/2}$, whose probability distribution is the square root of the Chi Square distribution. The use of what Peter once termed 'a back of the envelope calculation' (based on the delta method, i.e. a simple Taylor expansion) reveals that in the limit as $d \to \infty$,

$$\|\boldsymbol{Z}\| = d^{1/2} + O_p(1). \tag{1}$$

Thus high dimensional Gaussian data tend to lie relatively near the surface of the sphere of radius $d^{1/2}$ in $\mathbb{R}^d$. This fact tends to be rather surprising to students whom have been trained to think in terms of the Gaussian probability density, which of course is highest

near the origin (the most likely point according to the likelihood principal), and goes to 0 quite rapidly in the tails. This seeming paradox can be resolved by recalling that the density is with respect to Lebesgue measure, which is the actual source of this perhaps surprising phenomenon. In particular, if one approaches the problem of finding the volume of the unit sphere in $\mathbb{R}^d$ by computing a multivariate integral with respect to polar coordinates, then the final integral with respect to the radius has an integrand which becomes extremely large right near the $r = 1$ boundary. Thus Lebesgue measure pushes probability mass outwards, while the Gaussian density simultaneously tries to hold mass inwards. The balance point between these competing effects is located at $d^{1/2}$. As an interesting aside, note that this makes a general statement about populations, e.g. populations of people: *no member of the population is at all likely to be close to the average*. While parents may wish for average children, their children are very unlikely to be average, and the same assertion applies to the children of anyone else.

The second major concept that arises in the HDLSS geometric representation can be illustrated using two independent realisations from the standard Gaussian distribution, $\boldsymbol{Z}_1, \boldsymbol{Z}_2 \sim \mathrm{N}(0, I)$. The angle between these two vectors, with vertex at the origin, is the arc cosine of the inner product, and through a calculation very similar to that on which the first concept is based, is seen to satisfy

$$\text{angle}\left(\boldsymbol{Z}_1, \boldsymbol{Z}_2\right) = 90^\circ + O_p\left(d^{-1/2}\right). \tag{2}$$

Thus not only do random Gaussian data points tend to lie near the surface of a sphere, they also tend to be distributed on that sphere so that they are all about 90° apart from each other. An intriguing thought experiment is to try to imagine where more than three such data vectors can lie. The first three are easy, but for more than three, the structure becomes very hard to understand, likely because of the limitations of the human perceptual system. This limitation may be caused by the fact that our perceptual systems evolved in ancestors whose main task was finding food, which naturally lives in three dimensions.

Combining these two main ideas, we see that for a sample (i.e. a set of independent and identically distributed (i.i.d.) random variables) of size n drawn from the standard Gaussian distribution, the data are more and more random as $d \to \infty$, yet that randomness is almost entirely a random rotation. This is because a rotation is available whereby the data points will be close to $d^{1/2}$ times the first n unit vectors, i.e. to the vertices of the n-dimensional simplex in $\mathbb{R}^d$, a rigid deterministic structure. This simplex is the basis of the concept of geometric representation.

While the Gaussian distribution was used for simplicity of presentation in the foregoing discussion, the main ideas of geometric representation extend far beyond this distribution. This is demonstrated in Figure 1, using the GWAS data analyzed in Zhou, Marron & Wright (2016). This paper studies the pairwise angles between pairs of centered SNP vectors of dimension $d = 21,205$, for 3444 people. The data are far from Gaussian, because the vector entries are ternary (0, 1, or 2). Yet the data do satisfy the mixing condition assumed in Hall, Marron & Neeman (2005), due to rapid decay in correlations of successive vector values. This tendency is well-understood in human genetics, ultimately being caused by the crossing-over phenomenon in meiosis. The distribution of the angles is shown as both a jitter plot (each angle is displayed as a point, with random height for visual separation) overlaid with a kernel density estimate. Note that most of the pairwise angles (shown as black dots) are very close to 90°, as predicted by the HDLSS geometric representation. However there

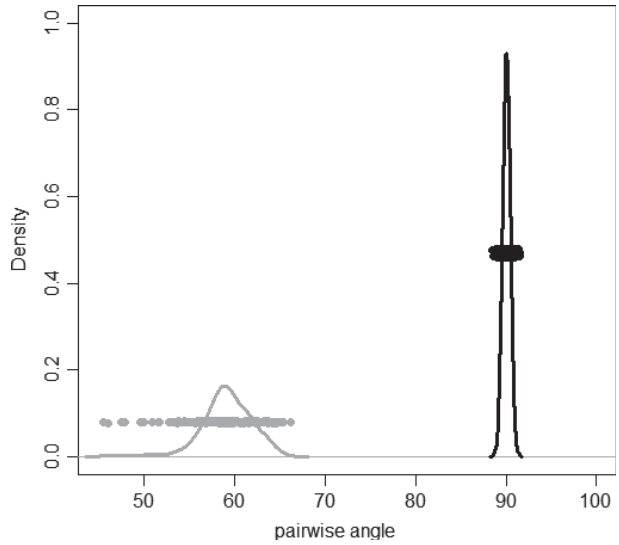

Figure 1 The distribution of pairwise angles between pairs of people in the Genome Wide Association Studies data, shown as a jitterplot overlaid with kernel density estimates. The black points and density are near 90° as predicted by the High Dimension Low Sample Size theory, while the gray points and density correspond to pairs of closely related people who share about half common SNPs, thus resulting in much smaller angles.

are a number of striking exceptions, shown as gray dots, which are due to a large number of first degree relatives (417 families of size 1, 20 of size 3 and 1 of size 4) in this study. Such relatives share approximately half of their genetic material, so these vectors are not expected to be orthogonal. The expected angle of first-degree relatives is calculated, and used to propose an appropriate adjustment in Zhou, Marron & Wright (2016).

In the second paper on geometric representation that we consider, Ahn *et al.* (2007), the mixing condition on vector components was replaced by a mild approximate sphericity condition. A weakening of the mixing type assumption was provided in Jung & Marron (2009), where only some permutation of the entries needs to satisfy a mixing condition. This weakened assumption is appealing in the context of genetic microarray data, because it corresponds well with actual chromosomal structure. Geometric representation was really taken beyond the Gaussian assumption by Yata & Aoshima (2010b, 2012), where the assumptions were focused in a manner designed to deal with non-Gaussianity. They determined two types of geometric representations in terms of the eigenspace of the sample covariance matrix.

In the papers referred to above, by combining (1) and (2), geometric representations of the sample covariance matrix were provided as follows: Let $\boldsymbol{X}=(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n)$ be a $d \times n$ data matrix, where $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are i.i.d. with a d-dimensional distribution having mean zero and positive definite covariance matrix $\boldsymbol{\Sigma}$. The eigen-decomposition of $\boldsymbol{\Sigma}$ is given by $\boldsymbol{\Sigma}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\top}$, where $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues, $\lambda_1 \geqslant \cdots \geqslant \lambda_d$, and $\boldsymbol{U}=(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_d)$ is an orthogonal matrix whose columns are the corresponding eigenvectors. Let us write the $(d \times d)$ sample covariance matrix as $\boldsymbol{S}=n^{-1}\boldsymbol{X}\boldsymbol{X}^{\top}$ and the $(n \times n)$ dual sample covariance matrix , also called the *Gram matrix*, as $\boldsymbol{S}_D=n^{-1}\boldsymbol{X}^{\top}\boldsymbol{X}$. Let $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_n \geqslant 0$ be the eigenvalues of $\boldsymbol{S}_D$. Note that

$\boldsymbol{S}$ and $\boldsymbol{S}_D$ share their non-zero eigenvalues. Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{W}$. Ahn *et al.* (2007) and Jung & Marron (2009) considered the following sphericity condition:

$$\frac{\mathrm{tr}(\boldsymbol{\Sigma}^2)}{\mathrm{tr}(\boldsymbol{\Sigma})^2} \to 0 \text{ as } d \to \infty. \tag{3}$$

When $\boldsymbol{W}$ is Gaussian (that is, all the elements of $\boldsymbol{W}$ are independent) or ρ-mixing, they showed the following geometric representation:

$$\frac{n}{\mathrm{tr}(\boldsymbol{\Sigma})}\boldsymbol{S}_D \xrightarrow{P} I \text{ as } d \to \infty. \tag{4}$$

where '$\xrightarrow{P}$' indicates convergence in probability. On the other hand, Yata & Aoshima (2012) showed that (4) follows from (3) and a certain regularity condition on $\boldsymbol{W}$ which is milder than either the Gaussianity or ρ-mixing conditions. From the geometric representation given by (4), which is called the *surface concentration* of the n-dimensional unit sphere, one can see that the eigenvalues of $\boldsymbol{S}_D$ become deterministic as $d \to \infty$: they converge to a common value and the corresponding eigenvectors have arbitrary directions. This is the reason why it is difficult to estimate eigenvalues and eigenvectors by using $\boldsymbol{S}_D$ (or $\boldsymbol{S}$) in classical PCA. For the setting in which $\boldsymbol{W}$ is non-Gaussian and non-ρ-mixing, a different geometric representation has been described in Yata & Aoshima (2012).

An analogue of geometric representation for data lying on a manifold was developed in Sen (2008). That work revealed that geometric representation also holds for data lying in (sequences of) compact sets, such as the unit cube $[0,1]^d$. Perhaps surprisingly at first, the same $d^{1/2}$ type of limiting distributions holds (e.g. the distance between the center of the cube and a random point grows at the rate $d^{1/2}$). This observation is consistent with the fact that human intuition about high dimensions tends to be very poor.

Of course there are limitations on the set of distributions which admit a geometric representation. A thought provoking example of this type, pointed out by John Kent during an informal discussion, is a scale mixture of Gaussians. A simple example is

$$f(x) = \frac{1}{2}\mathrm{N}(0, I) + \frac{1}{2}\mathrm{N}\left(0, 4 \times I\right),$$

where about half of the data vectors will lie near the surface of the sphere of radius $d^{1/2}$, and the other half will be out at twice that distance from the origin, whence the standard geometric representation is not possible. Another interesting fact about this distribution is that it provides an excellent example of why 0 covariance is a weaker condition than independence. Although the covariance matrix of this distribution is a multiple of the identity, the entries of a vector having this distribution are far from independent of each other. In particular when one entry is large, the data vector is likely to have come from the second mixture component, and consequently all entries will tend to be large. For general high-dimensional mixture models, some geometric representations were found by Yata & Aoshima (2015). Their work revealed that HDLSS data can be classified by PCA in a surprisingly explicit way.

Some earlier results on geometric representation in probability theory can be found in a sequence of papers on the topic called *concentration of measure* (see e.g. Talagrand 1991, 1995). For a good historical overview, which includes discussion of closely related ideas appearing even earlier in other contexts, see Ledoux (2005). The interesting phenomenon

revealed in Hall, Marron & Neeman (2005) has motivated many other studies. For example, Lv (2013) investigated the impact of high dimensionality in HDLSS settings from two different perspectives: a probabilistic one and a nonprobabilistic (geometric) one.

3. Analysis of statistical methods

In addition to providing the general insights about population structure discussed in Section 2, HDLSS asymptotics have provided some new tools for one of the most commonly performed tasks of modern mathematical statistics: the analysis and validation of existing statistical methods.

3.1. Classical PCA

A good introduction to several viewpoints on, and many important properties of, PCA can be found in Jolliffe (2005). The first HDLSS consideration of its properties can be found in Ahn *et al.* (2007). More complete studies, under a variety of assumptions can be found in Yata & Aoshima (2009, 2013b), and Jung & Marron (2009). Those papers are based on the insightful notion of a *spike model*, first developed in the related context of random matrix theory (asymptotics in the limit as both $n, d \to \infty$, with $n/d \to C > 0$) by Johnstone (2006). The main idea is that most eigenvalues of the underlying theoretical covariance matrix are constant, representing the white noise component of the distribution, while one (or perhaps finitely many) takes on a much larger value(s). In the case of one large eigenvalue, great insight comes from parametrizing it as d^{α} for $\alpha > 0$. Then the angle between the sample and theoretical first eigenvectors has the interesting property that for $\alpha > 1$, the angle converges to 0 (i.e. the eigenvector is *consistent*), while for $\alpha < 1$ the angle converges to 90° (i.e. the eigenvector is *strongly inconsistent*). This strong dichotomy can be understood at an intuitive level using geometric representation ideas. Recall the principle of geometric representation that random data tend to lie near the surface of a sphere of radius $d^{1/2}$. When the spike is large enough, the major axis of the ellipsoidal contours of the density will lie outside that sphere, giving consistency of the first eigenvector. When the spike is small, no data will lie outside the sphere of natural variation, so the first eigendirection will be random. Now recall that random directions tend to be 90° apart, thus leading to strong inconsistency. The reason the boundary is $\alpha = 1$, is that eigenvalues are on the scale of variance, not standard deviation. They can be put on the scale of standard deviation by taking the square root, which results in a boundary given by $\alpha = 1/2$, thus producing the familiar $d^{1/2}$.

As noted in Shen *et al.* (2016), an anonymous reviewer once made the interesting observation that the foregoing results hold even for a sample of size $n = 1$. Consistency sounds like a strange notion in such a case, but the assertion is mathematically true. In the large spike case, a single observation will with high probability lie in essentially the direction of the first theoretical eigenvector. PCA will produce only one non-zero eigenvalue, and in the limit that direction will tend towards the theoretical eigen-direction. The reviewer went on to conclude that this shows the assumption of $\alpha > 1$ is too strong to be meaningful in any practical situation. In any situation where the assumption $\alpha > 1$ fails, the mathematics shows PCA must be inconsistent, thus resulting in random, and in particular practically useless directions. PCA is so widely used in practice because it has a well established history of providing meaningful exploratory analyses, even in HDLSS cases, see for example Benito

et al. (2004) and Kimes *et al.* (2014). These lead to the conclusion that while $\alpha > 1$ may seem like a strong assumption, it is actually satisfied in real data analysis settings. See Hellton & Thoresen (2014a,b) for similar insights from a different perspective.

Note that the boundary case of $\alpha = 1$ was not mentioned in the foregoing results. Asymptotics in that case have been established by Jung, Sen & Marron (2012) whose deeper analysis discovered a limiting distribution for the angle, which is essentially a mixture of distributions, with components corresponding to the $\alpha > 1$ and $\alpha < 1$ cases. Related HDLSS results can be found in Lee (2012).

The application of the concept of 'sparsity' is a currently popular approach to high dimensional data in general, and sparse approaches to PCA have been proposed by Moghaddam, Weiss & Avidan (2005), d'Aspremont *et al.* (2007), Shen & Huang (2008), and Johnstone & Lu (2009). Asymptotic properties of sparse PCA have been studied by Johnstone & Lu (2009) from the perspective of random matrix theory. The HDLSS asymptotic properties of sparse PCA have been investigated by Shen, Shen & Marron (2013), who showed that for distributions whose first eigenvectors have a vanishingly small number of non-zero entries, sparse PCA approaches can break through the $\alpha = 1$ consistency barrier described above for classical PCA.

A less fortunate property of PCA is that it can be strongly affected by outliers in the data. An important goal in the field of robust PCA is to find analogues which are less sensitive to outliers. There is a large literature on this, but as noted in Locantore *et al.* (1999) most of the early proposed methods were strongly challenged by HDLSS situations. The Locantore *et al.* (1999) paper proposed a simple method called *spherical PCA*, which has proven to be quite effective in many circumstances, although many improvements have been proposed since. The HDLSS asymptotics of spherical PCA were worked out in the paper by Zhou & Marron (2015), who showed that spherical PCA has the same $\alpha = 1$ consistency properties discussed above in the presence of no outliers, but has much stronger consistency properties when outliers are present. Zhou & Marron (2016) adapted the L1 PCA methodology of Brooks, Dulá & Boone (2013) to obtain an improved robust version of PCA.

As for the case when $\alpha < 1$, classical PCA cannot have the consistency property stated above. However the case $\alpha < 1$ arises commonly in the context of high-dimensional data, and new theories and methodologies are expected to be developed to handle that important case. Actually, Yata & Aoshima (2013b) showed that microarray data sets often satisfy $0.5 < \alpha \leqslant 1$. In Section 4 we introduce new PCAs developed by Yata & Aoshima (2010a, 2012, 2013b) to handle the $\alpha < 1$ case.

3.2. Classification

Classification, also called discrimination, is an important statistical task that lies at the heart of machine learning. The canonical example is automated disease diagnosis, where one starts with a set of patients previously labeled sick or healthy, and uses that set to construct a rule for assigning the labels to new patients. See Duda, Hart & Stork (2012) for access to the large literature on this topic. It was first studied from the HDLSS viewpoint by Hall, Marron & Neeman (2005), who provided the first mathematical quantification of the intuitive idea that distance weighted discrimination (DWD), Marron, Todd & Ahn (2007), makes more efficient use of the available data than the popular Support Vector Machine

(SVM), Cortes & Vapnik (1995). Deeper theory in that direction has been developed by Pham (2010) and Bolivar-Cime & Marron (2013). HDLSS asymptotics were used to analyze a loss of efficiency for DWD in the unbalanced case by Qiao *et al.* (2010), who went on to use those ideas to improve upon the DWD algorithm.

A hybrid classification method, which shares the resistance to over-fitting that characterises linear methods such as SVM and DWD, while simultaneously giving some of the flexibility of full kernel machine methods (see Schölkopf & Smola 2002), consists of *bidirectional* versions of SVM and DWD proposed in Huang, Liu & Marron (2012). This method gives added flexibility by working not with one, but instead two direction vectors whose projections are combined in a product operation to yield more flexible classification regions. HDLSS asymptotics were used to make precise the relative advantages of this new hybrid method over existing approaches.

A simple HDLSS classification with robustness properties was proposed by Jiang, Marron & Jiang (2009). Distance-based classification methods were proposed by Hall, Pittelkow & Ghosh (2008), Chan & Hall (2009) and Aoshima & Yata (2014). All of these authors studied a variety of HDLSS asymptotic properties. In particular, Aoshima & Yata (2014) showed that the distance-based classification is quite flexible for HDLSS non-Gaussian data and provided a multiclass, high-dimensional classification called the *misclassification rate adjusted classification* in which misclassification rates are controlled to be no more than specified thresholds. On the other hand, Aoshima & Yata (2011, 2015b) provided quadratic classification methods based on a geometric representation of HDLSS data. In a manner similar to (1), the high dimensional data from one class (the ith class, say) tend to lie relatively near the surface of the sphere of radius $\mathrm{tr}(\mathbf{\Sigma}_i)^{1/2}$, where $\mathbf{\Sigma}_i$ denotes the covariance matrix of the ith class. The geometric classification method draws information about heteroscedasticity through the fact that the geometric representation incorporates different radii, $\mathrm{tr}(\mathbf{\Sigma}_i)^{1/2}$s, for the two expanding spheres. Aoshima and Yata showed that the misclassification rates of the classifier tend to 0 as d increases even when the mean vectors are common. Recently, Aoshima & Yata (2015c) considered a class of quadratic classifiers and discussed their optimality under high-dimensional and non-sparse settings.

Another unexpected HDLSS phenomenon is the Maximal Data Piling direction, studied by Ahn & Marron (2010). Given a set of data with two labeled classes, projections on this direction have the property that all points in one class project to a single point, while all points in the other class project to a different single point. It turns out that within the subspace generated by the data there is a great circle of such directions, with two antipodal directions maximising the difference between the projected points. Interestingly the formula for this direction is very close to that for the Fisher Linear Discriminant (Fisher 1936), wherein the pooled within class covariance matrix is replaced by the overall covariance matrix. In addition to being a useful example for the motivation of DWD, this direction actually performs better than the standard linear classifiers when the errors are auto-correlated. This surprising phenomenon was explained using HDLSS asymptotics by Miao (2015). New HDLSS insights into Maximal Data Piling can be found in Lee, Ahn & Jeon (2013). An interesting combination of sparsity with data piling ideas was proposed by Ahn & Jeon (2015).

HDLSS asymptotics were also used to study *sparse bi-clustering* in Lee *et al.* (2010), *predictive learning* by Liang, Cai & Cherkassky (2009), and even survival analysis by Hochstein *et al.* (2013).

3.3. Canonical correlations

To understand the relationships between two data sets on a common set of subjects, *Canonical Correlation* and *Partial Least Squares* are quite useful. Canonical Correlation was proposed by Hotelling (1936), and finds directions with maximise correlation between the two data types. Partial Least Squares is due to Wold (1966), and gives directions that maximise the related criterion of covariance. In low dimensional situations, Canonical Correlation is usually preferred for most purposes. However, in HDLSS situations, Canonical Correlation often results in a direction which gives perfect correlation, but results in a level of over-fitting which is too extreme to be useful. This phenomenon is due in part to the fact that Canonical Correlation can be driven by small scale noise artifacts. In such situations, Partial Least Squares can give a quick and simple solution with much better performance. HDLSS analysis of these methods, together with the proposal of a whole indexed family of methods that constitute a series of compromises between them called *continuum canonical correlations* can be found in Lee (2007). Related HDLSS asymptotics can be found in Samarov (2009).

A more advanced method for tackling this type of data is the JIVE methodology due to Lock *et al.* (2013) and Feng, Hannig & Marron (2015). This method is based on components that separately capture joint and individual variation. The development of HDLSS asymptotics for JIVE is an interesting open problem.

Sen (2008) use HDLSS ideas to advocate for the application of Kendall's tau in high dimensional contexts, and Tsai & Sen (2010) study categorical variables from an HDLSS viewpoint.

3.4. Hypothesis testing

Hypothesis tests are among the most widely used inference procedures in science and engineering. The practical need for such tests arises in several areas, in particular in GWAS. Hypothesis tests involving high-dimensional mean vectors were studied by Bai & Saranadasa (1996), Chen & Qin (2010) and Aoshima & Yata (2011, 2015a, 2017). From (1) high dimensional data tends to lie relatively near the surface of a sphere in $\mathbb{R}^d$. The asymptotic normality of the sample mean on the surface was proven by Aoshima & Yata (2011, 2015a). Explicitly their result is that

$$\frac{\left\|\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right\|^2 - \mathrm{tr}(\boldsymbol{\Sigma})/n}{\left(\left\|\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}\right\|^2\right)^{1/2}} \xrightarrow{D} \mathrm{N}(0,1)$$

as $d \to \infty$ under some regularity conditions. Here, $\overline{\boldsymbol{X}}_n$ is the sample mean of samples of size n from a d-dimensional distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and '$\xrightarrow{D}$' indicates convergence in distribution. By using the asymptotic distribution on the surface, several two-sample tests were established and their asymptotic properties studied by Aoshima & Yata (2011, 2015a). Recently, Aoshima & Yata (2017) developed high-dimensional two-sample tests under two disjoint models: the *strongly spiked eigenvalue* (SSE) model whose spike index is $\alpha \geqslant 1/2$ and the *non-SSE* (NSSE) model whose spike index is $\alpha < 1/2$, thus creating new high-dimensional inference methods.

High-dimensional inference on covariance matrices was studied by Aoshima & Yata (2011) and Ishii, Yata & Aoshima (2016). In particular, Ishii, Yata & Aoshima (2016) applied the noise reduction methodology given by Yata & Aoshima (2012) to the test for the equality

of two covariance matrices in HDLSS settings. Likewise, high-dimensional inference on covariance structures was studied by Yata & Aoshima (2013a, 2016a). In particular, Yata & Aoshima (2013a) developed the *extended cross-data-matrix (ECDM) methodology* that is an extension of the CDM method given by Yata & Aoshima (2010a). Yata & Aoshima (2016a) applied the ECDM methodology to pathway analysis used on GWAS data.

Other high dimensional hypothesis tests, which have been analyzed using HDLSS asymptotics, include those developed by Biswas & Ghosh (2014), Biswas, Mukhopadhyay & Ghosh (2014) and Ghosh & Biswas (2016).

4. Motivation of new statistical methods

While insight into the performance of existing statistical methods as discussed in Section 3 is important, HDLSS ideas have also led to substantial advances in the direction of motivating new statistical methods. Results of this type are reviewed here.

HDLSS ideas have motivated several improvements over classical PCA, based on an eigenanalysis of the sample covariance matrix. Yata & Aoshima (2010a) proposed the use of a cross data matrix approach, which broke through the $\alpha = 1$ spike barrier of classical PCA discussed in detail in Section 3, and created a new PCA called the *cross-data-matrix (CDM) methodology*. On the other hand, Yata & Aoshima (2012) proposed another new PCA called the *noise-reduction (NR) methodology* that is based on a geometric representation of the sample covariance matrix. We emphasize that the new PCAs can provide consistent estimators of both the eigenvalues and PC directions together with PC scores in HDLSS situations even when $\alpha < 1$. We describe the NR methodology in more detail. For simplicity, we assume a simple spiked model for the eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_d$ as follows:

$$\lambda_1 = d^{\alpha} \text{ and } \lambda_2 = \cdots = \lambda_d = 1. \tag{5}$$

We note that the surface concentration (4) appears when $\alpha < 1$ since then (3) is satisfied. Thus classical PCA is strongly inconsistent when $\alpha < 1$. When the NR method is applied, λ_1 is estimated by

$$\tilde{\lambda}_1 = \hat{\lambda}_1 - \frac{\mathrm{tr}(\boldsymbol{S}) - \hat{\lambda}_1}{n-1},$$

where $\hat{\lambda}_1$ is the largest eigenvalue of $\boldsymbol{S}$. If the population is Gaussian distributed and $\alpha \in (1/2, 1)$,

$$\frac{\tilde{\lambda}_1}{\lambda_1} \xrightarrow{D} \frac{\chi^2_n}{n} \text{ and } \frac{\hat{\lambda}_1}{\lambda_1} \xrightarrow{P} \infty$$

as $d \to \infty$, where χ^2_n denotes a random variable having a χ^2 distribution with n degrees of freedom. Thus $\tilde{\lambda}_1$ has an asymptotic distribution even when $\alpha < 1$, while $\hat{\lambda}_1$ is strongly inconsistent when $\alpha < 1$. See also Jung & Marron (2009) and Ishii, Yata & Aoshima (2016). For general spiked models in non-Gaussian settings, see Yata & Aoshima (2013b). In a related development, Yata & Aoshima (2016b) applied the NR method to a problem of recovering a high-dimensional, low-rank signal matrix for general spiked models.

Principal component analysis has been a workhorse method for HDLSS data visualization. It derives its effectiveness by giving several important one and two dimensional projections that can be usefully organized into a *scatterplot matrix*. See Carr *et al.* (1987) for insights that such plots reveal about the relationships between members of a population. When there are two or more components, *brushing* (Becker & Cleveland 1987), can provide

additional insights into how sub-populations compare. More direct study of subgroups can be effected by replacing the PCA direction in the visualization with a direction such as DWD which is expressly designed to provide a notion of best separation of the data. While such techniques can be very useful in providing insight, the putative insights can also be quite spurious and just artifacts of over-fitting the data. Wei *et al.* (2013) proposed the *DIrection PROjection PERMutation* hypothesis test to provide an objective understanding of when visual differences actually reveal important underlying structure, and are not mere sampling artifacts. A challenge in developing that methodology was the choice of summary statistic, which was well resolved using HDLSS asymptotics.

In a virus hunting application, Xiong *et al.* (2015) used the HDLSS geometric representation to motivate a new classification method called *Radial DWD*. This method is generally useful for classification in situations where one class is fairly centrally located, while the other is spread widely around in random directions. In parallel with the compact manifold discussion in Section 2, the data analysed by Xiong *et al.* (2015) were on the unit simplex. Nevertheless the HDLSS considerations were the same.

5. Connections with other asymptotic domains

An area of asymptotics related to HDLSS is *High Dimension Medium Sample Size*, where both $d, n \to \infty$, but n grows more slowly in the sense that $n/d \to 0$. Important papers studying this type of asymptotics include Fan & Lv (2008) and Yata & Aoshima (2010a). A noteworthy difference of viewpoint between these two papers is that in the former paper, it is assumed that the driving asymptotics are the classical $n \to \infty$, with d a rapidly growing (say exponential) function of n. In contrast, in the latter paper it is assumed that $d \to \infty$, with n a slowly growing function of d (e.g. logarithmic). It may appear that these approaches are equivalent, and for the most part they are, but there is an exception in that the latter approach has a much more direct interface with HDLSS asymptotics, whereas the former encounters a form of singularity.

In related work, the consistency property of classical PCA when $d, n \to \infty$ was studied by Yata & Aoshima (2009, 2013b), Shen, Shen & Marron (2016) and Shen *et al.* (2016). In particular, Yata & Aoshima (2009) derived certain conditions for consistency of classical PCA. Let us now write $n = d^{\beta}$ with $\beta > 0$. For simplicity, we consider the spiked model (5). Yata & Aoshima (2009) showed that if $\beta > 1 - \alpha$, it follows that $\hat{\lambda}_1 / \lambda_1 = 1 + o_p(1)$ as $d \to \infty$ under a certain regularity condition. The boundary case in which $\alpha = 1$ was studied by Jung, Sen & Marron (2012), and the case in which $\alpha > 1$ was studied by Jung & Marron (2009). In the context of random matrix theory, the special case of $\alpha = 0$ (i.e. the eigenvalue is bounded) was studied by several authors (e.g. Baik & Silverstein 2006; Paul 2007) when $n/d \to C > 0$ (i.e. $\beta = 1$). Note that Yata & Aoshima (2009) developed the random matrix theory to cover the case in which $n/d \to 0$. We summarize the results in Figure 2. The boundary case in which $\beta = 1 - \alpha$ was studied by Shen *et al.* (2016). In more detail, assume that the population eigenvalues for the multiple spike model have the following properties as $n \to \infty$:

$$\begin{cases} \lambda_1 > \lambda_2 > \lambda_3 \gg \lambda_4 = \cdots = \lambda_d = 1, \\ \frac{d}{n\lambda_j} \to c_j, \quad j = 1, 2, 3, \quad \text{with} \quad 0 \leqslant c_1 < c_2 < c_3 \leqslant \infty. \end{cases} \tag{6}$$

Shen *et al.* (2016) showed that the sample eigenvectors $\hat{\boldsymbol{u}}_1$, $\hat{\boldsymbol{u}}_2$ and $\hat{\boldsymbol{u}}_3$ respectively lie in the three cones shown in Figure 3, whereas the angles are respectively $\theta_1 = \arccos(1/\sqrt{1 + c_1})$,

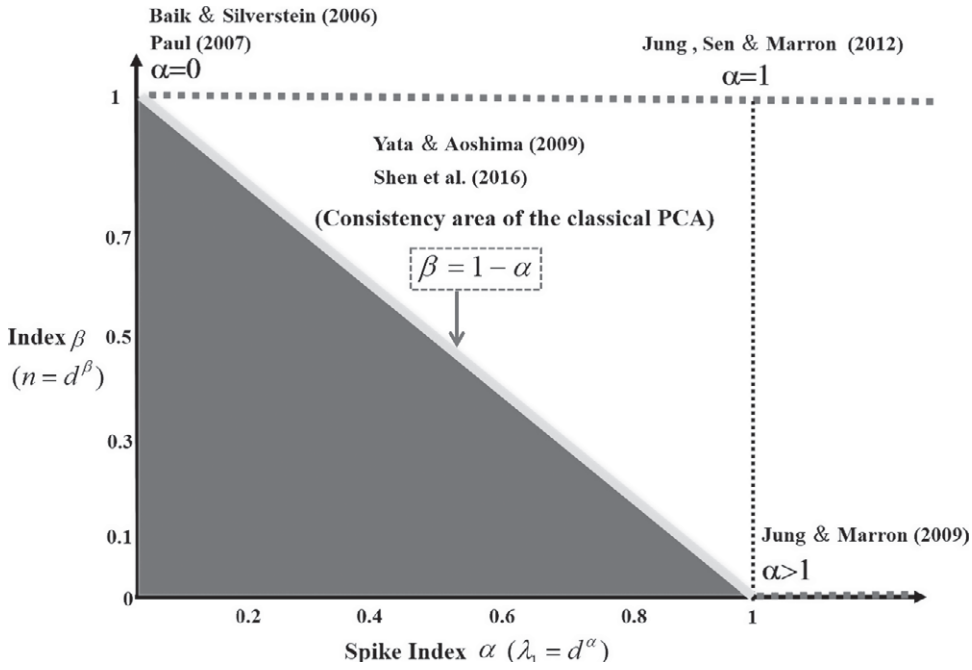

Figure 2 Consistency area of classical principal component analysis (PCA) for $n = d^{\beta}$ under the spiked model (5). When $\beta > 1 - \alpha$ (the white area), classical PCA is consistent. When $\beta \leqslant 1 - \alpha$ (the black area), it is inconsistent.

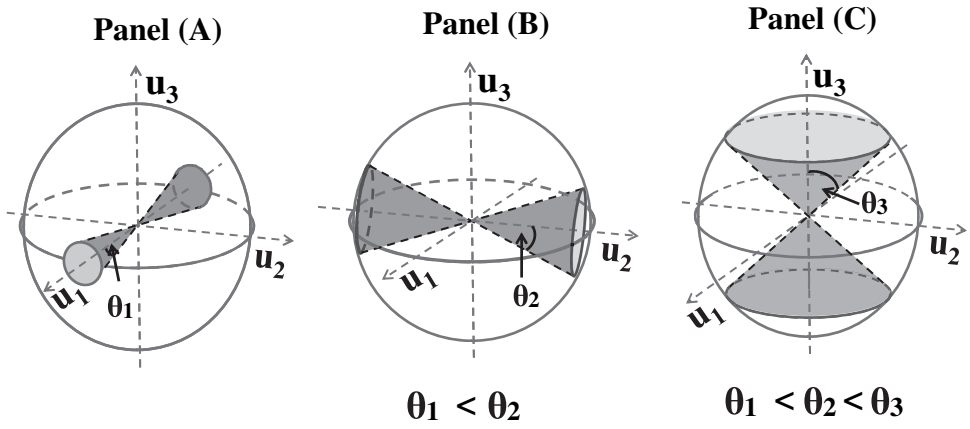

Figure 3 Geometric representation of PC directions in the multiple spike model (6). The sphere represents the space of possible sample eigenvectors. Panel (a) shows that the first sample eigenvector tends to lie in the cone, with the θ_1 angle. Similarly, Panels (b) and (c) show that the second and the third sample eigenvectors respectively tend to lie in the two cones, whose angles are θ_2 and θ_3.

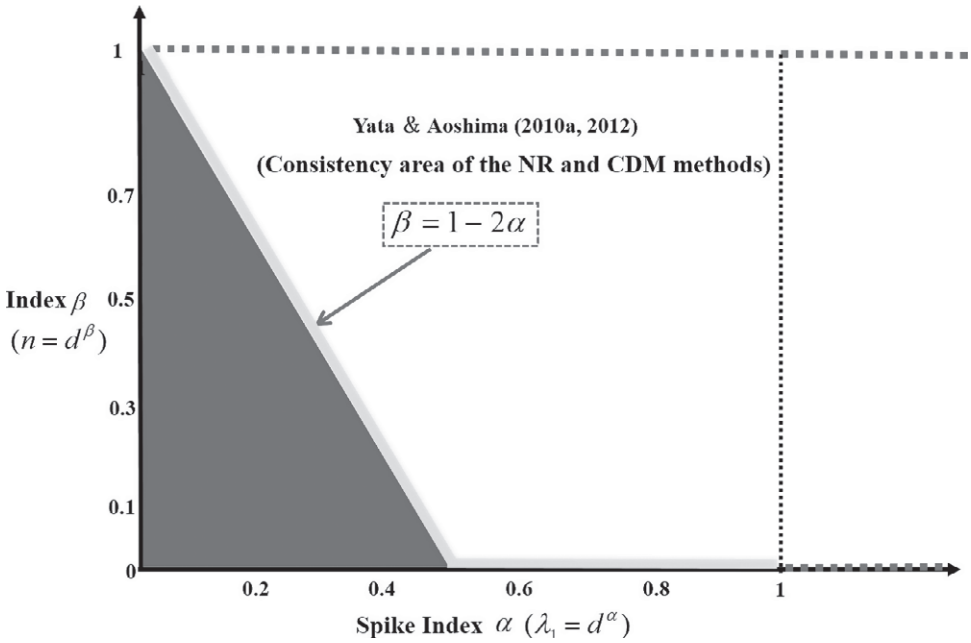

Figure 4 Consistency area of the noise-reduction and cross-data-matrix methods for $n = d^{\beta}$ under the spiked model (5). When $\beta > 1 - 2\alpha$ (the white area), these new methods are consistent. When $\beta \leqslant 1 - 2\alpha$ (the black area), they are inconsistent.

$\theta_2 = \arccos(1/\sqrt{1+c_2})$ and $\theta_3 = \arccos(1/\sqrt{1+c_3})$, shown in Figure 3. For the non boundary case in the multiple spike model (6), in which $\theta_i = 0$ or 90°, Shen, Shen & Marron (2016) showed the corresponding consistency and inconsistency convergence rates. Fan & Wang (2015) established the convergence rate for the boundary case in model (6).

On the other hand, as for the new PCAs discussed in Section 4, the NR and CDM methods are consistent under a certain regularity condition which is milder than that for classical PCA. Comparing Figure 4 with Figure 2, the consistency area is much broader for the new methods. Yata & Aoshima (2010a, 2012) showed that if $\beta > 1 - 2\alpha$, then the consistency property holds. In particular, when $\alpha > 1/2$, the new methods enjoy consistency for any small $\beta > 0$.

As an interesting relevant side note, Koltchinskii & Lounici (2015, 2016) considered random variables on a separable Hilbert space and established theorems about non-asymptotic properties of sample eigenvalues/eigenvectors, under either Gaussian or centered subgaussian assumptions, which can be used to obtain related results.

References

AHN, J. & JEON, Y. (2015). Sparse HDLSS discrimination with constrained data piling. *Computational Statistics & Data Analysis* **90**, 74–83.

AHN, J. & MARRON, J.S. (2010). The maximal data piling direction for discrimination. *Biometrika* **97**, 254–259.

AHN, J., MARRON, J.S., MULLER, K.M. & CHI, Y.Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94**, 760–766.

AOSHIMA, M. & YATA, K. (2011). Two-stage procedures for high-dimensional data. *Sequential Analysis (Editor's special invited paper)* **30**, 356–399.

AOSHIMA, M. & YATA, K. (2014). A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data. *Annals of the Institute of Statistical Mathematics* **66**, 983–1010.

AOSHIMA, M. & YATA, K. (2015a). Asymptotic normality for inference on multisample, high-dimensional mean vectors under mild conditions. *Methodology and Computing in Applied Probability* **17**, 419–439.

AOSHIMA, M. & YATA, K. (2015b). Geometric classifier for multiclass, high-dimensional data. *Sequential Analysis* **34**, 279–294.

AOSHIMA, M. & YATA, K. (2015c). High-dimensional quadratic classifiers in non-sparse settings. *arXiv preprint arXiv:1503.04549*.

AOSHIMA, M. & YATA, K. (2017). Two-sample tests for high-dimension, strongly spiked eigenvalue models. *Statistica Sinica*, in press (*arXiv:1602.02491*).

BAI, Z. & SARANADASA, H. (1996). Effect of high dimension: by an example of a two sample problem. *Statistica Sinica* **6**, 311–329.

BAIK, J. & SILVERSTEIN, J.W. (2006). Eigenvalues of large sample covariance matrices of spiked population models. *Journal of Multivariate Analysis* **97**, 1382–1408.

BECKER, R.A. & CLEVELAND, W.S. (1987). Brushing scatterplots. *Technometrics* **29**, 127–142.

BENITO, M., PARKER, J., DU, Q., WU, J., XIANG, D., PEROU, C.M. & MARRON, J.S. (2004). Adjustment of systematic microarray data biases. *Bioinformatics* **20**, 105–114.

BISWAS, M. & GHOSH, A.K. (2014). A nonparametric two-sample test applicable to high dimensional data. *Journal of Multivariate Analysis* **123**, 160–171.

BISWAS, M., MUKHOPADHYAY, M. & GHOSH, A.K. (2014). A distribution-free two-sample run test applicable to high-dimensional data. *Biometrika* **101**, 913–926.

BOLIVAR-CIME, A. & MARRON, J.S. (2013). Comparison of binary discrimination methods for high dimension low sample size data. *Journal of Multivariate Analysis* **115**, 108–121.

BROOKS, J.P., DULÁ, J. & BOONE, E.L. (2013). A pure l1-norm principal component analysis. *Computational Statistics & Data Analysis* **61**, 83–98.

CARR, D.B., LITTLEFIELD, R.J., NICHOLSON, W. & LITTLEFIELD, J. (1987). Scatterplot matrix techniques for large n. *Journal of the American Statistical Association* **82**, 424–436.

CASELLA, G. & HWANG, J.T. (1982). Limit expressions for the risk of james-stein estimators. *Canadian Journal of Statistics* **10**, 305–309.

CHAN, Y.B. & HALL, P. (2009). Scale adjustments for classifiers in high-dimensional, low sample size settings. *Biometrika* **96**, 469–478.

CHEN, S.X. & QIN, Y.L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *The Annals of Statistics* **38**, 808–835.

CORTES, C. & VAPNIK, V. (1995). Support-vector networks. *Machine Learning* **20**, 273–297.

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

DUDA, R.O., HART, P.E. & STORK, D.G. (2012). *Pattern Classification*. New York: John Wiley & Sons.

FAN, J. & LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911.

FAN, J. & WANG, W. (2015). Asymptotics of empirical eigen-structure for ultra-high dimensional spiked covariance model. *arXiv preprint arXiv:1502.04733*.

FENG, Q., HANNIG, J. & MARRON, J. (2015). Non-iterative joint and individual variation explained. *arXiv preprint arXiv:1512.04060*.

FISHER, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

GHOSH, A.K. & BISWAS, M. (2016). Distribution-free high-dimensional two-sample tests based on discriminating hyperplanes. *Test* **25**, 525–547.

HALL, P., MARRON, J.S. & NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 427–444.

HALL, P., PITTELKOW, Y. & GHOSH, M. (2008). Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 159–173.

HELLTON, K. & THORESEN, M. (2014a). Asymptotic distribution of principal component scores for pervasive, high-dimensional eigenvectors. *arXiv preprint arXiv:1401.2781*.

HELLTON, K.H. & THORESEN, M. (2014b). The impact of measurement error on principal component analysis. *Scandinavian Journal of Statistics* **41**, 1051–1063.

HOCHSTEIN, A., AHN, H.I., LEUNG, Y.T. & DENESUK, M. (2013). Survival analysis for HDLSS data with time dependent variables: Lessons from predictive maintenance at a mining service provider. In *2013 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*. IEEE, Dongguan, China, pp. 372–381.

HOTELLING, H. (1936). Relations between two sets of variates. *Biometrika* **28**, 321–377.

HUANG, H., LIU, Y. & MARRON, J.S. (2012). Bidirectional discrimination with application to data visualization. *Biometrika* **99**, 851–864.

ISHII, A., YATA, K. & AOSHIMA, M. (2016). Asymptotic properties of the first principal component and equality tests of covariance matrices in high-dimension, low-sample-size context. *Journal of Statistical Planning and Inference* **170**, 186–199.

JIANG, J., MARRON, J.S. & JIANG, X. (2009). Robust centroid based classification with minimum error rates for high dimension, low sample size data. *Journal of Statistical Planning and Inference* **139**, 2571–2580.

JOHNSTONE, I.M. (2006). High dimensional statistical inference and random matrices. *arXiv preprint math/0611589*.

JOHNSTONE, I.M. & LU, A.Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104**, 682–693.

JOLLIFFE, I. (2005). *Principal Component Analysis*. Wiley Online Library.

JUNG, S. & MARRON, J.S. (2009). PCA consistency in high dimension, low sample size context. *The Annals of Statistics* **37**, 4104–4130.

JUNG, S., SEN, A. & MARRON, J.S. (2012). Boundary behavior in high dimension, low sample size asymptotics of PCA. *Journal of Multivariate Analysis* **109**, 190–203.

KIMES, P.K., CABANSKI, C.R., WILKERSON, M.D. *et al*. (2014). Sigfuge: single gene clustering of RNA-seq reveals differential isoform usage among cancer samples. *Nucleic Acids Research* **42**, e113–e113.

KLEIN, R.J., ZEISS, C., CHEW, E.Y. *et al*. (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389.

KOLTCHINSKII, V. & LOUNICI, K. (2015). Normal approximation and concentration of spectral projectors of sample covariance. *arXiv:1504.07333*.

KOLTCHINSKII, V. & LOUNICI, K. (2016). Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance. *Annals of Institute Henri Poincare* **52**, 1976–2013.

LEDOUX, M. (2005). *The Concentration of Measure Phenomenon*, vol. 89, Providence, Rhode Island: American Mathematical Soc.

LEE, M.H. (2007). Continuum direction vectors in high dimensional low sample size data. Ph.D. thesis, University of North Carolina at Chapel Hill.

LEE, M.H. (2012). On the border of extreme and mild spiked models in the HDLSS framework. *Journal of Multivariate Analysis* **107**, 162–168.

LEE, M., SHEN, H., HUANG, J.Z. & MARRON, J.S. (2010). Biclustering via sparse singular value decomposition. *Biometrics* **66**, 1087–1095.

LEE, M.H., AHN, J. & JEON, Y. (2013). HDLSS discrimination with adaptive data piling. *Journal of Computational and Graphical Statistics* **22**, 433–451.

LIANG, L., CAI, F. & CHERKASSKY, V. (2009). Predictive learning with structured (grouped) data. *Neural Networks* **22**, 766–773.

LOCANTORE, N., MARRON, J.S., SIMPSON, D.G. *et al*. (1999). Robust principal component analysis for functional data. *Test* **8**, 1–73.

LOCK, E.F., HOADLEY, K.A., MARRON, J.S. & NOBEL, A.B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *The Annals of Applied Statistics* **7**, 523.

LV, J. (2013). Impacts of high dimensionality in finite samples. *The Annals of Statistics* **41**, 2236–2262.

MARRON, J.S., TODD, M.J. & AHN, J. (2007). Distance-weighted discrimination. *Journal of the American Statistical Association* **102**, 1267–1271.

MIAO, D. (2015). Class-sensitive principal components analysis. Ph.D. thesis, University of North Carolina at Chapel Hill.

MOGHADDAM, B., WEISS, Y. & AVIDAN, S. (2005). Spectral bounds for sparse PCA: Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, pp. 915–922.

PAUL, D. (2007). Asymptotics of sample eigenstruture for a large dimensional spiked covariance model. *Statistica Sinica* **17**, 1617–1642.

PHAM, T.H. (2010). Some problems in high dimensional data analysis. Ph.D. thesis, University of Melbourne.
QIAO, X., ZHANG, H.H., LIU, Y., TODD, M.J. & MARRON, J.S. (2010). Weighted distance weighted discrimination and its asymptotic properties. *Journal of the American Statistical Association* **105**, 401–414.
SAMAROV, D.V. (2009). The analysis and advanced extensions of canonical correlation analysis. Ph.D. thesis, University of North Carolina at Chapel Hill.
SCHÖLKOPF, B. & SMOLA, A.J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press.
SEN, P.K. (2008). Kendalls tau in high-dimensional genomic parsimony. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*. Ohio, USA: Institute of Mathematical Statistics, pp. 251–266.
SHEN, H. & HUANG, J.Z. (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis* **99**, 1015–1034.
SHEN, D., SHEN, H. & MARRON, J.S. (2013). Consistency of sparse PCA in high dimension, low sample size contexts. *Journal of Multivariate Analysis* **115**, 317–333.
SHEN, D., SHEN, H. & MARRON, J.S. (2016). A general framework for consistency of principal component analysis. *Journal of Machine Learning Research* **17**, 1–34.
SHEN, D., SHEN, H., ZHU, H. & MARRON, J.S. (2016). The statistics and mathematics of high dimension low sample size asymptotics. *Statistica Sinica* **26**, 1747–1770.
TALAGRAND, M. (1991). A new isoperimetric inequality and the concentration of measure phenomenon. In *Geometric Aspects of Functional Analysis*. Berlin, Heidelberg: Springer, pp. 94–124.
TALAGRAND, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques de l'Institut des Hautes Etudes Scientifiques* **81**, 73–205.
TSAI, M.T. & SEN, P.K. (2010). Entropy based constrained inference for some HDLSS genomic models: Ui tests in a chen–stein perspective. *Journal of Multivariate Analysis* **101**, 1559–1573.
WEI, S., LEE, C., WICHERS, L., LI, G. & MARRON, J.S. (2013). Direction-projection-permutation for high dimensional hypothesis tests. *arXiv preprint arXiv:1304.0796*.
WOLD, H. (1966). Estimation of principal components and related models by iterative least squares. In *Multivariate Analysis*, ed. P.R. Krishnaiaah, pp. 394–420. NewYork: Academic Press.
XIONG, J., DITTMER, D.P., MARRON, J.S. *et al.* (2015). Virus hunting using radial distance weighted discrimination. *The Annals of Applied Statistics* **9**, 2090–2109.
YATA, K. & AOSHIMA, M. (2009). PCA consistency for non-gaussian data in high dimension, low sample size context. *Communications in Statistics-Theory and Methods* **38**, 2634–2652.
YATA, K. & AOSHIMA, M. (2010a). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate Analysis* **101**, 2060–2077.
YATA, K. & AOSHIMA, M. (2010b). Intrinsic dimensionality estimation of high-dimension, low sample size data with d-asymptotics. *Communications in Statistics-Theory and Methods* **39**, 1511–1521.
YATA, K. & AOSHIMA, M. (2012). Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of Multivariate Analysis* **105**, 193–215.
YATA, K. & AOSHIMA, M. (2013a). Correlation tests for high-dimensional data using extended cross-data-matrix methodology. *Journal of Multivariate Analysis* **117**, 313–331.
YATA, K. & AOSHIMA, M. (2013b). PCA consistency for the power spiked model in high-dimensional settings. *Journal of Multivariate Analysis* **122**, 334–354.
YATA, K. & AOSHIMA, M. (2015). Principal component analysis based clustering for high-dimension, low-sample-size data. *arXiv preprint arXiv:1503.04525*.
YATA, K. & AOSHIMA, M. (2016a). High-dimensional inference on covariance structures via the extended cross-data-matrix methodology. *Journal of Multivariate Analysis* **151**, 151–166.
YATA, K. & AOSHIMA, M. (2016b). Reconstruction of a high-dimensional low-rank matrix. *Electronic Journal of Statistics* **10**, 895–917.
ZHOU, Y.H. & MARRON, J.S. (2015). High dimension low sample size asymptotics of robust PCA. *Electronic Journal of Statistics* **9**, 204–218.
ZHOU, Y.H. & MARRON, J.S. (2016). Visualization of robust l1PCA. *Stat* **5**, 173–184.
ZHOU, Y.H., MARRON, J.S. & WRIGHT, F. (2016). Computation of ancestry scores with mixed families and unrelated individuals. *arXiv preprint arXiv:1606.08416*.