On high-dimensional principal component analysis in genomics: consistency and robustness

Kristoffer Herland Hellton

Dissertation presented for the degree of

Philosophiae Doctor (PhD)

NORWEGIAN **CANCER** SOCIETY

Oslo Centre for Biostatistics and Epidemiology

Department of Biostatistics

University of Oslo

Oslo, September 2014

Acknowledgments

First and foremost, I would like to thank my supervisors Magne Thoresen and Arnoldo Frigessi for all the excellent guidance, support and inspiration you have given me during the three previous years.

The Department of Biostatistics has been a great academic and social environment with the best possible colleagues, and for this I am very grateful. I would especially like to thank Kathrine F. Frøslie for all your kind help, both with statistics and life in general.

I would also like thank the Norwegian Cancer Society (grant number 744088) for funding the work, and Peter Bühlmann and the Seminar für Statistik research group, ETH Zürich for being very welcoming when I visited the autumn of 2013.

I want to direct a special thanks to my fantastic friends, Tonje, Vebjørn, Nikolai, Marita, Hannah, Anna, Agnes, Lauritz and many others, for filling my life with fun evenings. Finally, my deepest gratitude goes to my parents for their endless love and support and for always being there for me.

Kristoffer H. Hellton
Oslo, September 2014

List of papers

Paper I

Hellton, K. H. and Thoresen, M. (2014). Asymptotic distribution of principal component scores connected to pervasive, high-dimensional eigenvectors. Submitted to *Journal of Multivariate Analysis.*

Paper II

Hellton, K. H. and Thoresen, M. (2014). The Impact of Measurement Error on Principal Component Analysis. Published in *Scandinavian Journal of Statistics.*

Paper III

Hellton, K. H. and Thoresen, M. (2014). Integrative clustering of high-dimensional data with joint and individual clusters, with an application to the Metabric study. Submitted to *Biostatistics.*

Contents

Acknowledgments **iii**

List of papers **iv**

1 Introduction **1**

2 Principal component analysis **3**
2.1 Population and sample PCA 3
2.2 Interpretation of components 6
2.3 Large sample asymptotics 10

3 Challenges for PCA in genomics **11**
3.1 Genetic data and use of PCA 12
3.2 High-dimensional asymptotics 14
3.3 Measurement error 18
3.4 Clustering 19

4 Aims **21**

5 Summary **22**
5.1 Paper I 22
5.2 Paper II 23
5.3 Paper III 23

6 Discussion **24**
6.1 The role of sparsity 26
6.2 Selection of the number of clusters 28

References **30**

Papers I-III **37**

1 Introduction

The technological developments of the last decades have made us able to generate massive amounts of measurements, enhancing the need for data exploration. We often understand data more easily through visualization tools; such as boxplots, histograms and scatter plots, but these univariate approaches (together with classical statistical analyses) will be impossible to use, when the number of recorded variables becomes too large. Instead, we need a reduced representation of all variables, which should exhibit the typical variation of the whole data set. This is the aim of principal component analysis: to understand large complex data by identifying the main axes of variation and explore and analyze the observations along these axes.

Principal component analysis (PCA) was introduced by Hotelling (1933) as a way of constructing a few highly informative scores representing a larger data set. All the papers in this thesis are concerned with different aspects of these scores in a high-dimensional setting: asymptotic consistency, robustness against measurement error and as a tool for clustering.

PCA is the workhorse of variable reduction in applied data analysis (Jolliffe, 2002), and the low-dimensional scores are used as a visualization tool or as input in conventional classification, clustering and regression methods. Mathematically, the procedure is equivalent to finding the singular value decomposition of the data matrix, i.e. the eigendecompostion of the sample covariance matrix, giving PCA a clear foundation within linear algebra. Hotelling (1933) interpreted the principal components (PCs) as the uncorrelated combination of variables expressing the most variance, and introduced the difference between the population and sample components. Later, Girshick (1939) and Anderson (1963) established the consistency of the procedure, when the number of variables is fixed and the sample size increases.

With the rapid technological development the last decade, especially in genetics, a new framework has entered, where the number of variables is larger than the number of observations. In this high-dimensional setting, Paul (2007) and Johnstone and Lu (2009) have shown that PCA is in fact not consistent. However, it is used extensively with great success in a range of genetic applications (Wall et al., 2003; Price et al., 2006; Patterson et al., 2006), for instance in identifying ethnic populations or discriminating between cancer subtypes. A paradox therefore exists between the theoretical inconsistency and the applied success of the method, and an aim of this thesis has been to explain this situations, in particular in **Paper I**.

Another important, but often overlooked aspect of modern genomics is the inherent measurement error. Genetic variables, be it gene activity, base pairs or methylation differences, are difficult to measure accurately, but these difficulties are seldom taken into account from a statistical point of view. However, overlooking such measurement error can lead to biased parameter estimates and loss of power to detect significant differences (Buonaccorsi, 2010; Carroll et al., 2012). In the modeling of genetic measurement error, most work has been on microarrays and Rocke and Durbin (2001); Karakach and Wentzell (2007) concluded that a model combining additive and multiplicative errors is well-suited. **Paper II** examines the robustness of PCA under the influence of classical additive measurement error in a high-dimensional genetic setting.

The use of the principal component scores in further analyses is an effective way of reducing the data dimension, and has been used with success in clustering. Chang (1983); De Soete and Carroll (1994); Arabie and Hubert (1996) suggested to use a subset of PC scores for different clustering algorithms, while Ding and He (2004) later proved that PCA can be considered as a relaxation of the k-means clustering scheme; the scores are the continuous equivalent to the discrete cluster membership matrix. With this useful formulation, k-means clustering can accommodate several data types in an integrative fashion. **Paper III** uses the component scores to construct both common and specific clusters for multiple data types simultaneously, by utilizing the JIVE framework (Lock et al., 2013).

The general topic of this thesis is the use of principal component analysis in genomic applications. A specific aim of the thesis has been to explain the paradoxical situation between the theoretical inconsistency and the practical success of the method. Further aims have been to evaluate the robustness of the principal components under measurement error and explore the role of PCA in integrative methods and clustering.

The outline of the thesis is the following: In section 2, we introduce the methodology of PCA and the common interpretations of the method, together with the standard asymptotic theory. Section 3 gives an introduction to genetics and an overview of the data analytic challenges facing PCA in genomics: high-dimensional asymptotic theory, the measurement error in genetic technologies and the link between PCA and integrative clustering. Section 4 states the aims of the thesis and in Section 5 and 6, we give a summary of the three papers constituting the thesis and discuss their contributions, strengths and weaknesses, especially covering the role of sparsity and selection of the number of clusters.

2 Principal component analysis

2.1 Population and sample PCA

Principal component analysis is the go-to method for reducing data dimension in many fields of applied data analysis, e.g. meteorology, genomics and finance. PCA is used to construct a small number of highly informative scores or surrogate variables for each observation. These scores are further used to visualize the structure of the original data or to carry out classification, clustering or regression analyses. The definition of the principal components for a population is as follows:

Definition (Population PCA)**.** *Let $X = [x_1, \dots, x_p]^T \in \mathbb{R}^p$ be a p-dimensional random variable with expectation zero, $\mathbb{E}(X) = \mathbf{0}$, and covariance matrix $\operatorname{Cov}(X) = \Sigma$. The eigendecompostion of the covariance matrix of X is given by*

$$\Sigma = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T, \tag{1}$$

where $\mathbf{\Lambda}$ is a diagonal matrix of descending population eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ and $\mathbf{V}$ is the corresponding matrix of population eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_p$.

Then, the population principal component are defined by the eigenvectors as

$$S_j^T = \mathbf{v}_j^T X, \quad j = 1, \dots, p, \tag{2}$$

and standardized population principal component are defined as

$$Z_j^T = \frac{\mathbf{v}_j^T X}{\sqrt{\lambda_j}}, \quad j = 1, \dots, p. \tag{3}$$

S_j and Z_j are referred to as the jth principal component score and standardized score, respectively, and are the low-dimensional representative of the original data. The construction of the principal components is equivalent in a sample setting. Suppose $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ is a $p \times n$ data matrix with n independent observations of a p-dimensional random variable with expectation zero. Then the sample estimate of the covariance matrix Σ is given by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T.$$

If $\mathbb{E}(\mathbf{x}_i) = 0$ is not assumed, the observed data is centered by the estimated mean

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T,$$

where $\bar{\mathbf{x}} = [\bar{x}_1, \dots, \bar{x}_p]$ is the vector of all variable means. For simplicity, we will further assume the expectation to be zero, $\mathbb{E}(\mathbf{x}_i) = 0$. Then the eigendecomposition of the sample covariance matrix defines the sample principal components:

Definition (Sample PCA)**.** *Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$ be a $p \times n$ data matrix with n independent observations of a p-dimensional random variable with expectation zero. The eigendecompostion of the sample covariance matrix is given as*

$$\hat{\Sigma} = \hat{\mathbf{V}}\mathbf{D}\hat{\mathbf{V}}^T,$$

where $\mathbf{D}$ is a diagonal matrix of the sample eigenvalues $d_1 > \cdots > d_p$ and $\hat{\mathbf{V}}$ is the matrix of sample eigenvectors $\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_p$. The sample principal component scores are then defined

$$\hat{S}_{ij} = \hat{\mathbf{v}}_j^T \mathbf{x}_i, \quad j = 1, \dots, p, \quad i = 1, \dots, n, \tag{4}$$

and the standardized sample component scores are defined as

$$\hat{Z}_{ij} = \frac{\hat{\mathbf{v}}_j^T \mathbf{x}_i}{\sqrt{d_j}} \quad j = 1, \dots, p, \quad i = 1, \dots, n. \tag{5}$$

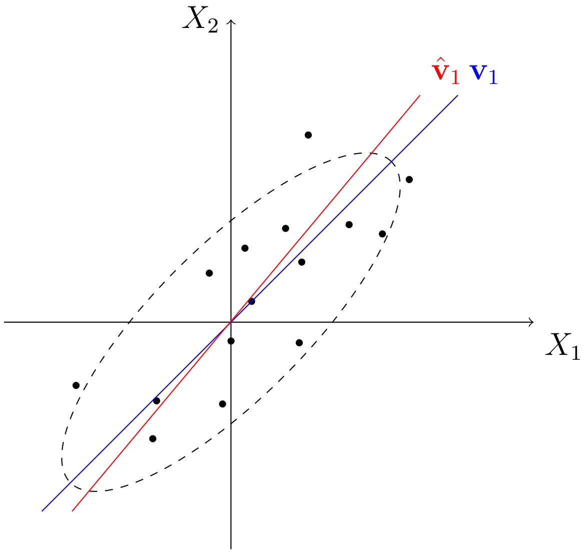

Figure 1: Visualization of the first population eigenvector (in blue), and first sample eigenvector (in red). The population covariance matrix is shown by the normal distribution ellipse.

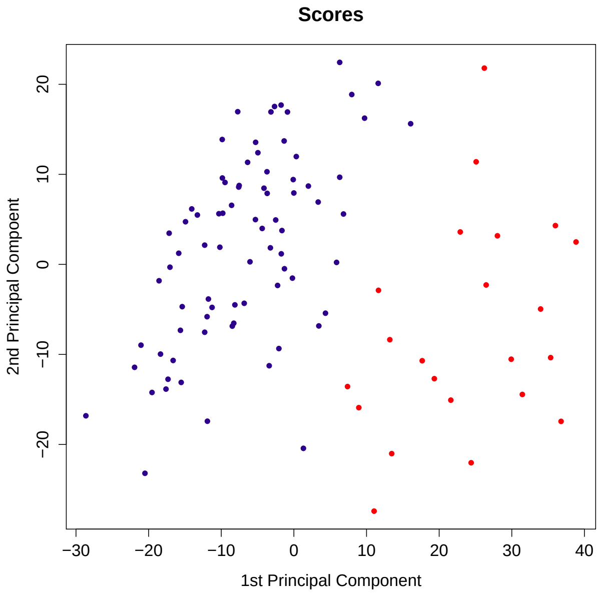

Figure 2: A plot of the first and second component scores of the genetic activity in 100 tumor samples, colored according to the Estrogen Receptor status, either positive (blue) or negative (red).

In Figure 1, we visualize the difference between the population and sample principal components. The dashed line shows the probability contour of a bivariate normal distribution, representing the population covariance matrix and the scattered dots show the observations, representing the sample covariance. The first population eigenvector (in blue) is given by the direction of the major axis of the distribution ellipse, while the first sample eigenvector (in red) fits to the observed data scatter.

In genomic applications, principal component analysis is widely used to visualize data through two-dimensional plots of the component scores. This can identify the main patterns of variability and help explore relationships between high-dimensional genetic variables, disease variables and other clinical covariates. A score plot of microarray gene expression data from the Metabric breast cancer study (Curtis et al., 2012) is displayed in Figure 2, showing the first and second component scores of 100 tumor samples. Each observation is colored according to the Estrogen Receptor (ER) status of the tumor, revealing this to be a main axis of variation in the genetic activity (Perou et al., 2000).

2.2 Interpretation of components

The principal components can be interpreted in three distinct frameworks:

- maximum explained variance
- optimal geometry with minimal reconstruction error
- factor model with homogeneous error

Earlier, there were clear lines between these different interpretations, but with the renewed interest in PCA in the high-dimensional setting, these boundaries have become blurred.

Explained variance

The most common interpretation of the principal components is in terms of maximum explained variance, as established by Hotelling (1933). Given p random variables $X = [x_1, \dots, x_p]^T$, the principal component scores can be defined as weighted linear combinations of the variables:

$$S_j = \boldsymbol{\alpha}_j^T X = \alpha_{j1}x_1 + \alpha_{j2}x_2 + \cdots + \alpha_{jp}x_p,$$

where $\boldsymbol{\alpha}_j = [\alpha_{j1}, \dots, \alpha_{jp}]^T$ is a set of coefficients. In each component, these coefficients will up-weight or down-weight the original variables, and $\boldsymbol{\alpha}_j^T$ is therefore referred to as loadings: the weight each original variable contributes within the component score. Hotelling (1933) sought the linear combination $\boldsymbol{\alpha}_1^T X$ with the maximum variance:

$$\max_{\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1=1} \mathrm{Var}\left(\boldsymbol{\alpha}_1^T X\right) = \max_{\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1=1} \boldsymbol{\alpha}_1^T \Sigma \boldsymbol{\alpha}_1, \tag{6}$$

and termed this the first principal component. To achieve an identifiable $\boldsymbol{\alpha}_1$, a normalization constraint, $\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1 = 1$, is imposed. The solution can be found by using the Lagrange multiplier λ, maximizing the expression

$$\boldsymbol{\alpha}_1^T \Sigma \boldsymbol{\alpha}_1 - \lambda(\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1 - 1).$$

By setting the derivative of the expression with respect to $\boldsymbol{\alpha}_1$ to zero, we obtain the eigenequation of Σ

$$\Sigma\alpha_1 - \lambda\boldsymbol{\alpha}_1 = 0,$$

Then $\boldsymbol{\alpha}_1$ is given by the eigenvector, $\mathbf{v}_1$, corresponding to the largest eigenvalue:

$$\arg\max_{\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1=1} \mathrm{Var}\left(\boldsymbol{\alpha}_1^T X\right) = \mathbf{v}_1.$$

In conclusion, the principal component based on the first eigenvector is therefore the linear combination of variables which explains the most variance.

Further, the second principal component $\boldsymbol{\alpha}_2^T X$ is the linear combination explaining the most variance, orthogonal to $\boldsymbol{\alpha}_1^T X$. This means the two components are uncorrelated

$$\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1 = 0, \quad \mathrm{Cov}(\boldsymbol{\alpha}_1^T X, \boldsymbol{\alpha}_2^T X) = 0.$$

The second principal component is then given by the eigenvector corresponding to the second largest eigenvalue. All consecutive principal components are defined in the same way: find the component $\boldsymbol{\alpha}_j^T X$ which maximizes the variance, being orthogonal to $\boldsymbol{\alpha}_{j-1}^T, \ldots, \boldsymbol{\alpha}_1^T$. Damon and Marron (2013) refer to this as the "forward" approach and shows the contrast with a "backwards" approach, generalizing the concept of PCA.

The principal component scores can therefore be interpreted as a representation of the original data explaining the most variance, the second-most variance and so on, such that they can be regarded as highly informative about the data structure.

Geometric interpretation

The principal components can also be interpreted purely in terms of their geometric properties, as first done by Pearson (1901). He posed the following question: For a set of p-dimensional observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, which line or q-dimensional subspace will best fit the data? This low-dimensional reconstruction, approximating the original data, gives a geometric and non-statistical interpretation of PCA. Figure 3 shows a simplified visualization of the situation, where observations in a two-dimensional plane ($p = 2$) are approximated by observations on one-dimensional line ($q = 1$). Each observation $\mathbf{x}_i$ is projected onto the line, where $\mathbf{m}_i$ denotes the position of the projection in the two-dimensional space and $\mathbf{r}_i = \mathbf{x}_i - \mathbf{m}_i$ denotes the orthogonal projection, the perpendicular distance from the observation to the subspace.

As defined by Pearson (1901), the optimal q-dimensional subspace minimizes the norm of the orthogonal projection, the sum of the squared perpendicular distances,

$$\sum_{i=1}^{n} \mathbf{r}_i^T \mathbf{r}_i = \sum_{i=1}^{n} \|\mathbf{r}_i\|_2^2.$$

The optimal subspace is then given by the first q eigenvectors of the sample covariance matrix, as shown by Jolliffe (2002, p. 34).

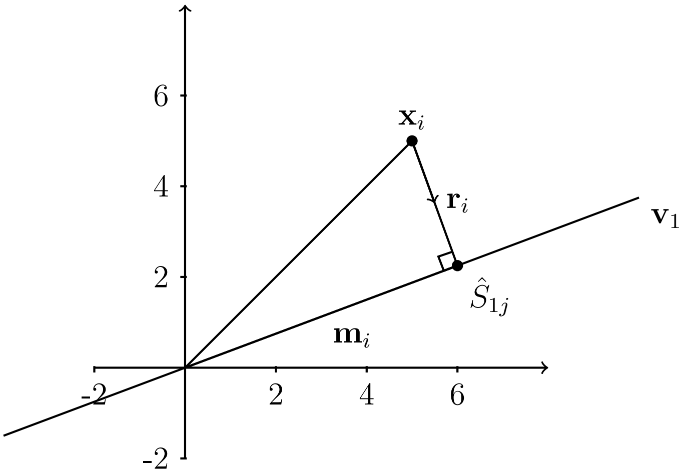

Figure 3: Geometric properties of eigenvectors and principal component scores.

The projection within the q-dimensional subspace, or along the line given by $\mathbf{v}_1$ in Figure 3, is then given by the corresponding sample score $\hat{S}_{1j}$, and the position of the projection in the p-dimensional space is given by

$$\mathbf{m}_i = \sum_{j=1}^{q} \hat{S}_{ij}\mathbf{v}_j,$$

the q first eigenvectors and component scores corresponding to the ith observation. The sum of the squared orthogonal projections is therefore equivalent to the *reconstruction error*:

$$\sum_{i=1}^{n} \|\mathbf{r}_i\|^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{m}_i\|^2 = \sum_{i=1}^{n} \left\|\mathbf{x}_i - \textstyle\sum_{j=1}^{q} \hat{S}_{ij}\mathbf{v}_j\right\|^2. \tag{7}$$

A geometric interpretation of the sample principal components is therefore given: the first q components express the optimal low-dimensional approximation of the original data. Given a set of observations, the q first eigenvectors and component scores will minimize the q-dimensional reconstruction error of the data matrix $\mathbf{X}$. The score plot of the first and second PC scores is therefore the best two-dimensional projection of the p-dimensional point cloud, in terms of minimal reconstruction error.

Latent variables

The third interpretation of PCA connects the component scores to latent variable estimation. PCA and factor analysis are closely linked, but the exact relationship is somewhat confusing. In factor analysis (Bartholomew et al., 2011), the observations are assumed to follow a latent variable model, illustrated in Figure 4, with m latent variables z_{ij} for $j = 1, \ldots, m$ given as

$$\mathbf{x}_i = \sum_{j=1}^{m} \mathbf{v}_j z_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n$$

where $\mathbf{v}_j$ is a vector of factor coefficients and ε_i denotes a general noise term, for instance normally distributed, $\varepsilon_i \sim N(0, \Sigma)$.

Tipping and Bishop (1999) showed that if the noise is assumed to be normally distributed with homogeneous variance, $\varepsilon_i \sim N(0, \sigma^2 I)$, the maximum likelihood estimates of the latent variables and the factor coefficients are equivalent to the principal component scores and the loadings. They further defined a model-based version of PCA, probabilistic PCA, where the principal components are equivalent to the latent variables found by an EM-algorithm. When the error is assumed homogeneous, the data are distributed as

$$\mathbf{x}_i \mid \mathbf{z}_i \sim N\left(\sum_{j=1}^{m} \mathbf{v}_j z_{ij}, \sigma^2 I\right),$$

where the maximum likelihood estimate of the noise parameter is given by

$$\sigma^2_{ML} = \frac{1}{p-m} \sum_{j=m+1}^{p} \lambda_j,$$

the mean of the $p - m$ remaining eigenvalues.

To simplify the theoretical machinery of the high-dimensional sample PCA, several authors (Johnstone and Lu, 2009; Paul, 2007; Nadler, 2008; Jung and Marron, 2009; Lee et al., 2010) assume the population eigenvalues and eigenvectors to follow from a model defined by m latent variables homogeneous and normally distributed noise:

$$\mathbf{x}_i = \sum_{j=1}^{m} \mathbf{v}_j z_{ij} + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2 I), \qquad i = 1, \ldots, n,$$

called the spiked covariance model, as introduced by Johnstone (2001).

This development contradicts the view of Jolliffe (2002, ch. 7), who stated: “PCA has often been dealt with in textbooks as a special case of factor analysis. This view is misguided since PCA and factor analysis are really quite distinct techniques.” For Jolliffe (2002), the main difference between the techniques was that factor analysis invoked a model relating the

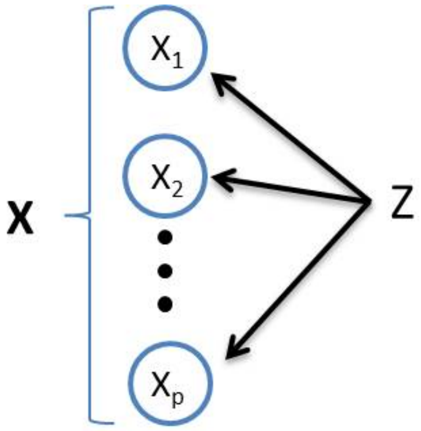

Figure 4: A latent unobserved variable j affects the observed variables $\mathbf{x}$.

observations to a set of latent variables, while PCA had no explicit model. However, for Tipping and Bishop (1999) and most authors dealing with PCA today, especially in the high-dimensional setting, the boundary between the two procedures is blurred.

The interpretation of the component scores as latent variables is utilized in **Paper I** to connect the asymptotic behavior of the population eigenvalues to the structure of the population eigenvector. Particularly, we utilized the fact that the eigenvector coefficients can be interpreted as an effect of the latent variable upon the observed variables. This is used to show that if the effect could be characterized as pervasive, the corresponding eigenvalue must scale linearly with the dimension.

2.3 Large sample asymptotics

From a theoretical perspective, we consider the large sample properties of PCA in terms of the consistency and asymptotic distribution of the sample components. Girshick (1939) and Anderson (1963) proved that, if p is fixed, the observations are normally distributed and the population eigenvalues are of multiplicity one

$$\lambda_1 > \cdots > \lambda_p,$$

all sample eigenvalues and -vectors converge to the population eigenvalues and -vectors as the sample size increases:

$$d_j \xrightarrow{p} \lambda_j, \quad \hat{\mathbf{v}}_j \xrightarrow{p} \mathbf{v}_j, \qquad j = 1, \ldots, p, \qquad \text{as } n \to \infty. \tag{8}$$

The sample eigenvalues and eigenvectors are therefore consistent estimators, which for Figure 1 means that $\hat{\mathbf{v}}_1$ converges to $\mathbf{v}_1$ as the sample size increases.

For the purpose of statistical inference, Anderson (1963) further established the asymptotic distribution of the sample eigenvalues and eigenvectors, $d_1, \ldots, d_p$ and $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p$, when the distribution of the observations $\mathbf{x}_i$ are multivariate normal. Then the sample eigenvectors and -values are also asymptotically normally distributed, and all sample eigenvalues are asymptotically independent of each other and independent of all eigenvectors.

Specifically, if all population eigenvalues are of multiplicity one, the sample eigenvector converge to

$$\sqrt{n}\ (\hat{\mathbf{v}}_j - \mathbf{v}_j) \xrightarrow{d} N(0, C);$$

a normal distribution with zero mean and covariance matrix

$$C = \sum_{l=1, l\neq j}^{p} \frac{\lambda_j \lambda_l}{(\lambda_j - \lambda_l)^2}\ \mathbf{v}_l \mathbf{v}_l^T.$$

The asymptotic covariance between two sample eigenvectors, $\hat{\mathbf{v}}_j$ and $\hat{\mathbf{v}}_k$, is given by

$$-\frac{\lambda_j \lambda_k}{(\lambda_k - \lambda_j)^2}\ \mathbf{v}_j \mathbf{v}_k^T.$$

Conversely for the sample eigenvalues: If all population eigenvalues are of multiplicity one, the sample eigenvalues converge to

$$\sqrt{n}\ (d_j - \lambda_j) \xrightarrow{d} N(0, 2\lambda_j), \quad n \to \infty,$$

and each pair of sample eigenvalues d_j and d_k are asymptotically independent. Anderson (1963) also derived the general distribution of the sample eigenvalues and -vectors in the case of any eigenvalue multiplicity, where some of the population eigenvalues can be equal.

3 Challenges for PCA in genomics

The last decades have seen remarkable advances in high-throughput genetic technology, enabling the exploration of the whole genome. The challenge for statistical genomics is therefore a vast amount of genetic variables combined with few observations, a situation where PCA has shown to be highly useful. The ability of PCA to reduce dimension and identify the axes of variation has made it popular in the analysis of many types of high-dimensional genetic data.

This section gives a brief introduction to genomics and the most important genetic examples where PCA is used. The section will also introduces the three analytic challenges investigated in this thesis; high-dimensionality, measurement error and integrative clustering.

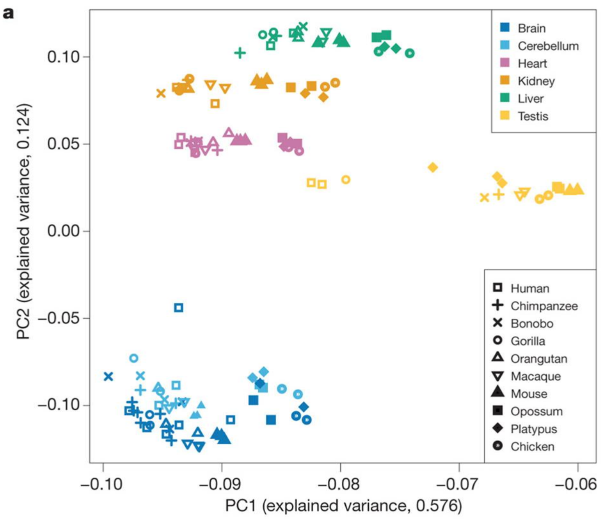

Figure 5: Example of principal component scores being used to visualize microarray expression levels in different mammalian tissues (Brawand et al., 2011).

3.1 Genetic data and use of PCA

Genomics is the field of studying all genomes, the complete set of chromosomes and genes, consisting of DNA. The DNA molecule is built up of four nucleotides or amino acids (denoted A, C, G and T), which form a double stranded helix of base pairs. The genetic information in the DNA is expressed by ribonucleic acid (RNA) through the process of transcription. Depending on the function of a gene, the transcribed RNA can form rRNA, tRNA or miRNA or an intermediate product called messengerRNA (mRNA), which creates protein. Although definitions vary, genes are often defined as a functional segment of the DNA that encodes a product, usually a protein. The longest genes are up to 10 000 base pairs and the human genome contains about 20 000 to 25 000 genes, making up about 1% of the total DNA (Ziegler et al., 2010).

The 3.3 billion base pairs in the human DNA sequence is about 99.9% identical across individuals, but the base pairs that do differ are highly important for understanding disease. A single nucleotide polymorphism (SNP) is a DNA sequence variation occurring commonly within a population (e.g. 1%) in which a single nucleotide differs between members.. A recent estimate of the total number of such genetic markers includes 17.8 million SNPs, in total 0.54% of the DNA, where 9.5 million have been validated (Ziegler et al., 2010).

Due the large number of genetic variables, dimension reduction is an important part of the analysis of genetic data. An early example is Cavalli-Sforza et al. (1994), who used PCA to infer human migration patterns by coupling component loadings of genetic markers with maps over sampled populations. Currently PCA is utilized in many different genomic measurements, and we will highlight its use in microarray data and genome-wide association studies (GWAS).

Microarray technology measures the quantity of mRNA, which can be interpreted as the activity of the corresponding genes. As all genes are measured simultaneously, microarray data give a snapshot of the genome-wide expression levels at a given time, and is therefore an important tool for identifying gene expression patterns in disease tissue. This has aided the understanding of the genetic influences in cancer (Perou et al., 2000), particularly by discovering novel subtypes, for instance in breast cancer (Sørlie et al., 2001) and confirming established histological differences (Golub et al., 1999). In data sets with the expression of ten-thousands of genes, PCA is highly useful for visualizing or identifying clusters in the main structures. The method was early adopted for the analysis of microarray data (Yeung and Ruzzo, 2001), and principal component scores are now commonly used to identify the patterns of variability in gene expression data, as seen in Figure 5. We have used microarray gene expression data as example data in both **Paper II** and **Paper III**.

Variant data, such as SNP markers, is another important way of identifying the genetic impact on disease. Genome-wide association studies have for identified numerous genetic loci linked to disease susceptibility. In the setting of identifying disease-related SNPs, PCA can be used to correct for confounding (Price et al., 2006; Patterson et al., 2006). Population stratification is a common issue in genome-wide association studies, because ethnicity can act as a confounder of the association between disease and the genetic markers. However, principal component scores have been shown to identify population strata (Yang et al., 2014) and can therefore express difference between ethnic subpopulations. This is seen in Figure 6, displaying the component scores derived from SNP markers recorded in different European populations. Component scores are therefore commonly used to correct for ethnicity confounding in genome-wide association studies.

Many other genetic data types are now possible to measure genome-wide; methylation data and copy number aberrations being two important examples. Diseases are often associated with several genetic and epigenetic layers, and an integrative approach to analysis can therefore be highly beneficial. With the fast technological development, an increasing number of genetic data types are be available. Integrative genomics is based on the principle

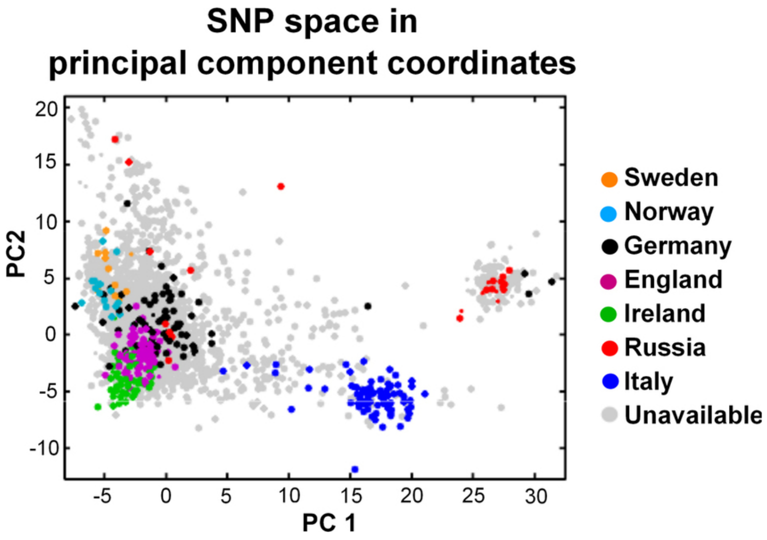

Figure 6: Example of principal component scores being used to visualize different patterns in SNP markers (Valente et al., 2012), where the components express ethnic difference.

that any biological mechanism builds upon multiple molecular phenomena, and a disease such as cancer can only be fully understood when considering the interplay between and within the different genomic layers (Kristensen et al., 2014). As the information content is higher in an integrative framework compared to the individual analyses, it is possible for such an approach to gain statistical power to detect relevant signals.

3.2 High-dimensional asymptotics

The emergence of high-dimensional data in genetics and other areas renewed the interest in asymptotic properties of PCA. It was initially proven by Lu (2002), then followed up by Paul (2007); Nadler (2008); Johnstone and Lu (2009), that principal component analysis in fact becomes inconsistent in the high-dimensional setting. Paul (2007) and Johnstone and Lu (2009) proved their results by using the asymptotic framework of random matrix theory, which allows $p > n$ under the assumption

$$p/n \to \gamma \geq 0,$$

as both the number of variables and observations increase, $p \to \infty$ and $n \to \infty$. The papers also introduced the spiked covariance model for the population eigenvalues, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_p)$, where the m first eigenvalues are substantially larger than the rest

$$\underbrace{\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m}_{Signal} \quad \gg \quad \underbrace{\lambda_{m+1} = \cdots = \lambda_p}_{Noise}. \tag{9}$$

The remaining eigenvalues are often assumed to be 1, for simplicity. Early results (Bai and Silverman, 2010) showed that the sample eigenvalues only converge to the population eigenvalues if $\gamma = 0$. In the high-dimensional setting where $\gamma > 0$, the sample eigenvalues are not consistent, and instead converge to

$$d_j \overset{\text{a.s.}}{\to} \lambda_j \left(1 + \frac{\gamma}{\lambda_j - 1}\right), \qquad j =, 1 \ldots, m, \qquad \text{if } \lambda_j > 1 + \sqrt{\gamma}.$$

Paul (2007) and Johnstone and Lu (2009) also showed that the sample eigenvectors are inconsistent. Lee et al. (2010) further developed the result to be valid for all γ, and showed that the inner product between the sample and population eigenvector converges to a constant, depending on λ_j and γ

$$|\langle \hat{\mathbf{v}}_j, \mathbf{v}_j \rangle| \overset{\text{P}}{\to} \sqrt{\left(1 - \frac{\gamma}{(\lambda_j - 1)^2}\right) / \left(1 + \frac{\gamma}{\lambda_j - 1}\right)}, \qquad j =, 1 \ldots, m,$$

when $\lambda_j > 1 + \sqrt{\gamma}$, assuming all eigenvalues to be of multiplicity one and the data to be normally distributed. When $\gamma > 0$ in the high-dimensional setting, the inner product cannot converge to 1 and the sample eigenvectors will be inconsistent.

Sparse PCA

Based on these results, Johnstone and Lu (2009) concluded:

> "The inconsistency asserts that ordinary PCA becomes confused in the presence of too many variables each with equal independent noise. If the principal components have a *sparse* representation, then selection of an appropriate subset of variables should overcome the inconsistency problem."

The inconsistency of the eigenvectors sparked an intensive research into sparse PCA, where penalization schemes are used to estimate sparse eigenvectors. Earlier attempts at finding sparse eigenvectors had been motivated by simplified interpretation, as the important variables are easier to identify when some loadings are exactly zero.

The first sparse PCA procedure, simplified component technique LASSO (SCoTLASS), was suggested by Jolliffe et al. (2003) and used the maximum-variance property of PCA in Equation (6) combined with an L_1 penalization of the eigenvectors:

$$\max_{\mathbf{v}_1^T\mathbf{v}_1=1} \mathbf{v}_1^T\hat{\Sigma}\mathbf{v}_1, \quad \text{subject to } \sum_{k=1}^{p}|v_{1k}| \leq t.$$

The absolute-value constraint will force some of the loadings to be exactly zero, depending on the value of t, and hence the estimated $\mathbf{v}_1$ will be sparse. Further components are found by requiring orthogonality between the kth component and the $k-1$ previous components. However, as the problem is not convex, computations become troublesome.

The next version of sparse PCA was given by Zou et al. (2006), using the reconstruction error property in Equation (7). Zou et al. (2006) defined the first sparse principal component to be given by the following minimization problem:

$$\min_{\mathbf{z},\mathbf{v}_1}\left\{\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{v}_1 z_{i1}\|_2^2+\lambda_1\|\mathbf{v}_1\|_1+\lambda_2\|\mathbf{v}_1\|_2^2\right\}, \quad \|\mathbf{z}_1\|_2=1,$$

where $\mathbf{z}_1 = [z_{11}, \ldots, z_{1n}]$. The first penalty on $\mathbf{v}_1$ will force the component loadings to be sparse. The criterion is not jointly convex in both $\mathbf{v}_1$ and $\mathbf{z}_1$, but it is convex in each parameter with the others fixed. The minimization over $\mathbf{v}_1$ can be efficiently solved as an elastic net problem, when $\mathbf{z}_1$ is fixed, and reversely, the minimization over $\mathbf{z}_1$ is solved by the singular value decomposition, when $\mathbf{v}_1$ is fixed. A set of sparse principal components can therefore be obtained by alternating these two steps until some convergence criterion is reached.

Other more sophisticated procedures for estimating sparse principal components have further been developed by d'Aspremont et al. (2008); Shen and Huang (2008); Witten et al. (2009); Journée et al. (2010); Ma et al. (2013) and several others.

Asymptotic behavior of component scores

The papers documenting the PCA inconsistency did however not prove any asymptotic results regarding the principal component scores. Lee et al. (2010) were the first to note that:

> "...inconsistency of the sample eigenvectors does *not necessarily* imply poor performance of PCA."

Principal component scores have in fact been applied with success in a number of high-dimensional genetic settings, e.g. genome-wide association studies or microarray studies (Price et al., 2006; Ma et al., 2006). This suggests that the component scores can be considered suitable for analysis, despite the inconsistency of the eigenvectors.

Lee et al. (2010) showed that the standardized sample principal component scores behave better than the eigenvectors, asymptotically. The inner product between the population standardized scores, $\mathbf{z}_j^T = \mathbf{v}_j^T\mathbf{X}/\sqrt{\lambda_j}$, and the sample standardized scores, $\hat{\mathbf{z}}_j^T = \hat{\mathbf{v}}_j^T\mathbf{X}/\sqrt{d_j}$, converges to

$$|\langle \mathbf{z}_j, \hat{\mathbf{z}}_j \rangle| \xrightarrow{\text{P}} \sqrt{1 - \frac{\gamma}{(\lambda_j - 1)^2}}, \quad \lambda_j > 1 + \sqrt{\gamma},$$

under the spiked covariance model defined in (9). The result shows that the asymptotic inner product of the scores is closer to 1 than that of the corresponding eigenvector. In conclusion, the principal component scores are not consistent, but the inconsistency for the corresponding eigenvectors is worse.

Later, Shen et al. (2012, 2013) proved an asymptotic result enlightening the paradoxical situation regarding the PC scores. Their result were in a different asymptotic framework (Jung and Marron, 2009; Ahn et al., 2007; Jung et al., 2012), where n is fixed and the population eigenvalues are assumed to grow with the dimension p. Jung and Marron (2009); Jung et al. (2012) defined the spiked covariance model with m fixed components as follows:

$$\lambda_1 = \sigma_1^2 p^\alpha, \quad \dots \quad \lambda_m = \sigma_m^2 p^\alpha,$$

with the growth rate parameter $\alpha > 0$. They showed that the behavior of the sample eigenvector depends on the value of α with distinct differences when α is smaller, larger or exactly equal to 1.

If $\alpha < 1$, the eigenvectors are strongly inconsistent and the sample and population eigenvector become asymptotically orthogonal. If $\alpha > 1$, the situation reverses as the eigenvectors are asymptotically consistent even for a single observation. The interesting case of $\alpha = 1$, covering the gap between the consistency and strong inconsistency, was explored by Jung et al. (2012). For $\alpha = 1$, the inner product between the sample and population eigenvectors will not degenerate, but converge to a random quantity depending on the sample size and the signal-to-noise ratio. We illustrate the situation for a single spike model with one important signal, where the first eigenvalue is given by $\lambda_1 = \sigma_1^2 p^\alpha$, while the rest are given $\lambda_2 = \cdots = \lambda_p = \tau^2$. For normally distributed data, the asymptotic limit of the inner product between the first sample and population eigenvector depends on α:

$$|\langle \hat{\mathbf{v}}_1, \mathbf{v}_1 \rangle| \xrightarrow{\text{P}} \begin{cases} 1, & \alpha > 1, \\ \left(1 + \frac{\tau^2}{\sigma^2 \chi_n^2}\right)^{-1/2}, & \alpha = 1, \\ 0, & \alpha < 1, \end{cases}$$

where χ_n^2 is a chi-squared distributed variable with n degrees of freedom.

In the case of $\alpha > 1$, Shen et al. (2012) further showed that the ratio between the individual sample and population scores converges to a random variable independent of the observation index k:

$$\left|\frac{\hat{z}_{ij}}{z_{ij}}\right| \xrightarrow{\text{P}} R_j, \qquad j = 1, \ldots, m, \quad i = 1, \ldots, n,$$

such that R_j is common for all scores within the same component. R_j is distributed as $\sqrt{n/\chi_n^2}$ based on the chi-squared distribution with n degrees of freedom (Shen et al., 2012, Theorem 1).

In terms of visualization, this means that the relative positions of the population scores are preserved in the sample scores, and the visual information conveyed by the score plot will be the same in the sample and population. In **Paper I**, we prove a corresponding result for the case of $\alpha = 1$, and show how this situation which can be interpreted as a pervasive signal, relevant in several genetic situations. The result can explain the paradoxical situation between the usefulness and theoretical inconsistency of PCA, as it demonstrates that the visual information is close to consistent even though the eigenvectors and scores are not.

3.3 Measurement error

An important, but often overlooked, aspect of genomic data is the inherent measurement error. Due to the nature of the genome, any genetic variable will be difficult to measure and technical errors will therefore always be present in the data. It is well-known that measurement error in covariates leads to biased parameter estimates and loss of power to detect significant differences in regression (Carroll et al., 2012; Buonaccorsi, 2010). Such problems can also be present in other multivariate techniques such as PCA, and they highlight the importance of considering measurement error in genetic data.

In the classical measurement error model, we observe an error-prone covariate W instead of the true covariate X. The error U is usually modeled to follow an additive or multiplicative model. The effect of additive measurement error on parameter estimation is easily described in univariate regression. If the true model is

$$Y_i = \alpha_X + \beta_X X_i + \varepsilon_j, \quad \varepsilon_i \sim N(0, \sigma_X^2),$$

but instead of X, we observe

$$W_i = X_i + U_i, \quad U_i \sim N(0, \sigma_U^2),$$

the naïve approach is to calculate the regression coefficients using the error-prone data W, instead of the original data X. The relationship between the observed slope β_W and the true slope β_X is given by

$$\beta_W = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \beta_X,$$

where the ratio between the two parameters, given by the error and data variances, is called the attenuation coefficient. Due to the attenuation coefficient, the slope for the error-prone data will be smaller in absolute value than the true slope, resulting in a bias or attenuation towards zero. In multivariate regression, the bias is not necessarily towards zero, but instead depends on the covariance between explanatory variables.

The effect of measurement error on multivariate linear and logistic regression is well-established (Thoresen and Laake, 2000; Buonaccorsi, 2010), but this is not the case for PCA. Some attempts have been made by Sanguinetti et al. (2005) and Wentzell and Hou (2012), both suggesting new versions of PCA incorporating information about the measurement error. Wentzell and Hou (2012) assumed the covariance matrix of the measurement error to be known and incorporated this into the framework of probabilistic PCA (Tipping and Bishop, 1999). In chemometrics, where PCA is a widely used technique, several authors (Faber et al., 1995a,b; Wentzell and Lohnes, 1999; Narasimhan and Shah, 2008) have investigated the situation, but with a focus on eigenvalues and component selection and only in the case of homogeneous errors.

In **Paper II**, we theoretically characterize and explore the effect of general additive errors on principal component analysis, particularly the impact on loadings, component scores and the component selection.

3.4 Clustering

The aim of clustering is to divide a sample into K classes, where the observations are more similar within one class than between classes (Hastie et al., 2009). A classic procedure is the k-means clustering, where clusters are found by minimizing distances between the observations and the obtained set of cluster centroids in an iterative fashion (Hartigan and Wong, 1979). To aid the k-means procedure in the high-dimensional setting, it was early suggested to consider a low-dimensional projection of the data: first obtain a few principal components and then use k-means clustering on the component scores. This two-step procedure, termed "tandem clustering" by Arabie and Hubert (1996), was discouraged (Chang, 1983; De Soete and Carroll, 1994; Arabie and Hubert, 1996) due to the possibility that the chosen scores do not reflect the cluster structures of the entire data set.

However, in the machine learning literature, where tandem clustering is commonly used, Zha et al. (2001) and Ding and He (2004) explored the theoretical link between k-means clustering and PCA. Ding and He (2004) reformulated the cluster membership vectors with the observation indices C_k for $k = 1, \dots K$, as an $n \times K$ membership matrix $\mathbf{Z}^*$, where each row represents an observation and each column a cluster:

$$\mathbf{Z}^* = [n_1^{-1/2}\mathbf{z}_1^*, \dots, n_K^{-1/2}\mathbf{z}_K^*], \qquad \mathbf{z}_k^* = [0, \dots, 0, \underbrace{1, \dots, 1}_{n_k}, 0, \dots, 0]^T. \tag{10}$$

The single 1 in each row dictates which cluster the observation belongs to, as each column corresponds to a cluster. For simplicity, observations in the same cluster can be grouped together and each column is normalized by the square root of the cluster size. The k-means clustering criterion (Hartigan and Wong, 1979) is given as

$$\arg\min_{C_k} \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2,$$

where $\boldsymbol{\mu}_k$ is the kth cluster centroid vector. With the cluster representation $\mathbf{Z}^*$, this minimization problem can be reformulated as an equivalent maximization problem

$$\max_{\mathbf{Z}^* \text{in} (10)} \text{trace}\left(\frac{1}{n}\mathbf{Z}^{*T}\mathbf{X}\mathbf{X}^T\mathbf{Z}^*\right).$$

The key observation is that if the structure of $\mathbf{Z}^*$ is continuous, thus relaxing the discrete structure in (10), the maximization problem will be equivalent to the definition of PCA:

$$\max_{\mathbf{Z}^T\mathbf{Z}=I_K} \text{trace}\left(\frac{1}{n}\mathbf{Z}^T\mathbf{X}\mathbf{X}^T\mathbf{Z}\right).$$

The continuous relaxation of $\mathbf{Z}^*$ is therefore given by the K first principal component scores. In consequence, the K first component scores will span the subspace of the K centroids, such that the scores are optimal for finding K clusters. The discrete $\mathbf{Z}^*$ can then be reconstructed using k-means, thus reducing the dimensions handled by the algorithm from thousands to a handful. This property can explain the natural connection between PCA and clustering.

With the connection between k-means clustering, PCA and the latent factor modeling in Section 2.2, it is possible to construct integrative clustering procedures handling several data types. When PCA used on a single data type can be formulated as finding a latent variable j, this can be extended to several data types $X_1, \ldots, X_m$ by assuming the j to be shared between all data types:

$$\begin{aligned} X_1 &= \mathbf{w}_1^T Z + \varepsilon_1, \\ &\vdots \\ X_M &= \mathbf{w}_M^T Z + \varepsilon_M, \end{aligned}$$

This approach is used in the iCluster methodology (Shen et al., 2009, 2013), where the noise terms are assumed heterogeneous, $\varepsilon_m \sim N(0, \Psi_m), \Psi_m = \text{diag}(\sigma_1^2, \ldots, \sigma_{p_m}^2)$. The parameter estimates are obtained by maximum likelihood estimation using the EM-algorithm. If ε_m was homogeneous, the solution is analytically given by the singular value decomposition. In iCluster, one can also enforce sparsity on the loading matrices by penalizing the data log-likelihood (Shen et al., 2013). After convergence of the EM-algorithm, the rows of j are clustered by the k-means algorithm to obtain the group membership of each observation.

This approach is further developed in **Paper III**, where both joint and individual latent variables are introduced. This results in both common and data type-specific clusterings, which can give further biological insights into the different data layers. The latent variable model is assumed to have homogeneous noise, such that the JIVE methodology (Lock et al., 2013) is easily applicable.

4 Aims

The overall topic of this thesis is the use of principal component analysis in high-dimensional data, particularly in genomics. Specifically, the aims have been:

- to offer an explanation of the paradox of the theoretical inconsistency results and practical results of PCA in genomics.

- to characterize impact of the genetic measurement error on principal component analysis and evaluate the robustness of loadings and scores in the high-dimensional setting.

- to construct and expand high-dimensional integrative clustering for several genomic data types, utilizing the framework of PCA.

5 Summary

In this section, we will give a brief summary of the three papers comprising this thesis, highlighting the main contributions.

5.1 Paper I

Hellton, K. H. and Thoresen, M. (2014). Asymptotic distribution of principal component scores connected to pervasive, high-dimensional eigenvectors. Submitted to *Journal of Multivariate Analysis*.

The aim of this paper is to give a possible explanation to the paradox of PCA in the high-dimensional setting, where the method shows great success in certain genetic applications, despite being inconsistent. Theoretical results by Johnstone and Lu (2009) and Paul (2007) have shown that both eigenvectors and component scores are not asymptotically consistent when $p > n$. The initial response to these results was to introduce sparsity constraints on the eigenvectors to obtain consistency.

Our contribution has been to investigate the asymptotic behavior of the component scores, in a setting mimicking the structure of genetic data. Results are derived in a specific asymptotic framework, where relevant population eigenvalues scale linearly with the dimension. This is shown to correspond to a pervasive signal structure, where asymptotically a non-zero proportion of the variables are informative regarding the latent structure. Pervasive signal structures are reasonable in several genetic examples, for instance in recovering ethnic population stratification in SNP markers or in identifying gene expression patterns in different types or subtypes of cancer. In both these settings, PCA have shown to be successful.

We prove that under a pervasive signal structure, the ratio between sample and population PC scores converges asymptotically to a random variable approximately equal within each component. For reasonable values of the sample size and signal strength, the deviation from an identical scaling within each component is negligible. As a consequence, the visual information present in the population component scores will be preserved asymptotically in the sample scores. If the signal in the data is pervasive, classical PC scores will be a good way of visualizing the population differences, even though the eigenvectors and the scores are not consistent.

This is particularly relevant when discussing the use of sparsity constraints and sparse modeling in PCA, as it demonstrates that in the case of non-sparse structures, information can still be extracted.

5.2 Paper II

Hellton, K. H. and Thoresen, M. (2014). The Impact of Measurement Error on Principal Component Analysis. Published in *Scandinavian Journal of Statistics.*

The aim of this paper is to characterize the effect of measurement error on PCA. We describe the difference between the eigenvalues and -vectors derived from the original error-free data matrix and the eigenvalues and -vectors based on the data with error. The measurement error is assumed to follow a classical additive and normally distributed stochastic model, such that the expectation and variance of the difference in eigenvalues and eigenvectors are characterized by the covariance structure of the error. These expressions are obtained by conditioning on the original data and assuming the ratio between the error variance and the data eigenvalues to be small, thus allowing the difference to be approximated by a Taylor expansion of the eigenvalues and -vectors.

The expectation and variance of the differences in eigenvalues and -vectors are used to interpret the effect of independent, homogenous and heterogeneous measurement error. For both cases, we observed that the loadings, or eigenvector coefficients, are not robust against measurement error as the induced variance is substantial compared to loading values themselves. The situation is different for the scores, where the robustness depends heavily on the eigenvalues. In genomics, where the largest eigenvalues are usually very large, the component scores corresponding to the largest eigenvalues will be robust against independent measurement error.

5.3 Paper III

Hellton, K. H. and Thoresen, M. (2014). Integrative clustering of high-dimensional data with joint and individual clusters, with an application to the Metabric study. Submitted to *Biostatistics*.

The aim of this paper is to extend the framework of integrative clustering utilizing the principal component scores. Earlier methodologies, such as iCluster (Shen et al., 2009, 2013), cluster patient samples of several genetic data types using latent variables or factor scores, but under the assumption that joint clustering is identical for each data type. Our contribution is a cluster scheme taking into account both joint and data type-specific cluster structures by using the JIVE methodology (Lock et al., 2013). JIVE decomposes the data matrix into an additive set of latent variables with joint and individual components by an iterative procedure. The ranks of the latent structures are directly connected to the number of clusters, as proven by Ding and He (2004). For K clusters, our method uses the $K-1$ rank singular value decomposition.

The selection of the numbers of clusters, in total $M+1$ for M data types, are fixed before iteratively calculating the latent components. Our proposed method evaluates the presence of cluster structures in each component separately by comparing the distribution of scores to the normal distribution. If the score distribution in a component, either joint or individual, does not deviate significantly from the normal distribution (as evaluated by a normal quantile-quantile plot), there is no evidence of any cluster structure.

To evaluate the total numbers of clusters, we first find the total number of relevant components in all data types jointly, when not allowing for any individual structures. Then, the total number of relevant components is assessed individually in each data-type. The ranks of the joint and individual structures are then calculated using these estimates.

6 Discussion

In this thesis, we have explored different issues regarding PCA in the high-dimensional setting: consistency, robustness and clustering. We will first briefly discuss the practical impact of our results, and then go more deeply into two issues important in **Paper I** and **Paper III**.

In **Paper I**, we proved asymptotic results enlightening why classical principal component scores can be used to visually explore pervasive signal structures. The consequences are two-fold: firstly it highlights the need for a conscious use of the sparsity assumption and secondly it shows that the use of classical PCA in certain genetic data examples is not flawed, despite the theoretical shortcomings. The former is important when considering the motivations behind sparse PCA and this will be discussed further in Section 6.1. The latter can reassure data analysts and geneticists using PCA, that the several papers highlighting the eigenvector inconsistency do not doom PCA in every situation. An issue is, however, the applicability of our asymptotic framework. Asymptotic results are not automatically transferable to the finite sample situation, as the number of variables cannot be infinite in practice. But the distributional expressions for finite p and n will in our case be too complicated, such that asymptotic result in **Paper I** will be a useful approximation and simplification. This is analogous to the use of the central limit theorem.

Further in **Paper II**, we investigated the impact of measurement error on PCA. The results suggest that independent, additive measurement error is not a problem for component scores connected to the largest eigenvalues. The loadings are however highly affected, such that we discourage the interpretation of loadings in situations with large measurement errors. A remaining question is, however, if we can correct for error in practice. To be able to do so, one will need an estimate of the error covariance matrix Σ_e. To estimate a high-dimensional error covariance matrix using replicates, will be very difficult due to the large number of parameters. High-dimensional covariance matrix estimation is challenging even in the error-free case, and must utilize penalization schemes (Pourahmadi, 2013). The proper estimation of the error covariance matrix has therefore been avoided in earlier works, either by alternative statistical procedures (Sørensen et al., 2014) or by using external information, such as technical probe statistics in microarrays, as an proxy for the correct measurement error (Sanguinetti et al., 2005; Sørensen et al., 2012). The latter approach was used in **Paper II**, in addition to assuming independent errors, a strong assumption often not properly fulfilled in genetic examples.

Finally in **Paper III**, we proposed a joint and individual integrative procedure for clustering patient samples based on several genomic data types. This approach can be highly useful in situations with heterogeneous data types, where a cluster structure only present in a single data layer might confound the joint structure. However, a disadvantage is that the joint and individual clusterings have to be orthogonal in terms of the estimated latent variables, meaning that the clusterings must be completely uninformative about each other. The realism of this restriction is difficult to assess. The current version of the method can also only handle homogeneous noise, but a future extension could allow for heterogeneous noise using an EM-algorithm. The approach to the integrative analysis used in **Paper III** does not assume any specific relationships between the data layers, such that the integration of the different data types can be said to be unstructured. In an exploratory or hypotheses-generating setting, such as clustering, this may be seen as an advantage. However, if the goal is to perform integrative inference, further assumptions about the structure and relationships between the data layers could strengthen the statistical analysis.

In the following, we will go more deeply into two issues important for **Paper I** and **Paper III**.

6.1 The role of sparsity

Within the statistical community, there is currently an enormous interest in sparse modeling and penalized regression, an interest also influencing the methodological development of PCA. This has especially been the case after Johnstone and Lu (2009) concluded that the eigenvector inconsistency made it necessary to sparsify PCA. **Paper I** adjusts this view, as it shows that classical PCA can extract visual information about the population structure, if the signal is pervasive. Yang et al. (2014) substantiated this conclusion by showing that the classical component scores can be used to discover population strata. Thus, if the component scores are the main aim of PCA, a conscious attitude towards sparsity is needed. This discussion will present the main motivations behind the sparsity assumption in PCA and regression. The aim is to understand how the motivations relate to each other and how one could conclude about the validity of the assumption.

The sparsity assumption expresses the idea that only few variables are relevant for analysis. From a theoretical perspective, sparsity is most commonly defined as a small number of non-zero coefficients (Bühlmann and Van De Geer, 2011). The different motivations behind the sparsity assumption can in general be divided in two; either based on the belief that the real world is in fact sparse or as a way of improving the analysis.

In examples from signal processing, image analysis or astronomy, it is well-established that the signal is sparse. This might also be argued in genomics; either as an exact description of the genetic mechanisms or as an appropriate approximation. On the other hand, sparsity can be motivated by a wish for optimal prediction and parameter consistency or by simplified interpretation. In high-dimensional regression, the sparsity assumption will aid in ensuring optimal prediction and consistency of the estimated regression coefficients. Bühlmann et al. (2014) state that "Reasonable prediction and estimation can be achieved if the underlying truth is sparse. If the true underlying model is not sparse, then high-dimensional statistical inference is ill posed and uninformative." Their argument also applies to PCA, where assuming sparsity will ensure consistent estimation of eigenvectors. However, this should not be the most important motivation, if the component scores are the main output of the analysis. In addition, sparsity can simplify the interpretation of parameters, as many variable effects are estimated to be exactly zero. Such parsimonious estimates were the main motivation of Jolliffe et al. (2003), when introducing sparse PCA.

We find the divide between these two types of justifications, the aim of describing the true world versus ensuring optimal methodological properties, also in the nature of modeling. Cox (1990) differentiated between the substantive and empirical role of a statistical model. A substantive model explains observations through detailed mechanisms (e.g. the Poisson process), while the empirical model represents dependencies in an idealized form (e.g. regression or ANOVA). The wish for simple interpretation and parsimony fits within the empirical role, while a correct reflection of the real world is the ideal of the substantive role. It can be argue that within the empirical role one can assume sparsity, even though the world is not sparse. Instead it is seen as an appropriate idealization. This view was expressed by Box and Draper (1987, p. 424), when formulating the quintessential nature of the empirical model: "Essentially, all models are wrong, but some are useful." The sparsity assumption is in practice untestable and the usefulness of the assumption, in terms of simplified interpretation, can be a valid argument in its own right.

Another fault line in the discussion, particularly relevant for regression, exists between describing the true world and achieving optimal prediction properties: the contrast between a predictive and explanatory framework. According to Breiman (2001), the statistical mind set can be divided into two worlds: When prediction is the main aim, optimality of a method is measured by the predictive accuracy, while if explanation is the main aim, optimality is measured by the goodness-of-fit of a specified model. The former does not utilize external information, while the latter utterly rely on expert knowledge about the phenomenon. The role of model assumptions is therefore inherently different in the two frameworks: in prediction, an assumption should improve accuracy, while for explanation, the assumptions need to reflect expert knowledge or be a correct interpretation or approximation of the real world.

From a predictive point of view, sparsity is needed to achieve optimal prediction, as stated by Bühlmann et al. (2014). When the true world is not sparse, utilizing a few of the largest effects may be better for prediction. From an explanatory point of view, one needs to consider if the true signal is sparse. If the true structure is in fact non-sparse, correct estimation of the individual effects might be impossible. But through classical PCA one can still extract information about the population structure and overall differences, as shown in **Paper I**. External knowledge is therefore needed to determine the appropriate assumptions. If explanation is the aim, sparsity should be assumed when knowledge dictates this to be appropriate.

6.2 Selection of the number of clusters

A key point in clustering is the selection of the number of clusters, and during the work with **Paper III**, this arose as a particularly difficult and troublesome factor. Both because several established procedures did not yield good results within the proposed method, but also because the novel solutions did not comply perfectly with the data examples. In the following, we discuss some the problems and solutions found in **Paper III**.

The classical approach to cluster selection (Kaufman and Rousseeuw, 2009) is to evaluate a measure of the distances between the clusters for different numbers of clusters K, and choose the cluster arrangement with the optimal value. There exists a range of different distance measures; based on the mean of distances, the distances between means or different combination of maximum and minimum distances. Milligan and Cooper (1987, 1985) compared a range of criteria and concluded that, for well-separated clusters the Calinski-Harabasz criterion (Caliński and Harabasz, 1974) or the Dunn index Dunn (1974) seem to preform best in a low-dimensional setting. Different criteria will emphasize different aspects of the data and give different conclusions, making it difficult to objectively decide on an optimal criterion.

Beyer et al. (1999) showed that the difference between the minimum and maximum distance between an observation and its neighbors converges to zero as the dimensions increase. The ability of Euclidean distances to discriminate between clusters will therefore decrease in high-dimensional data. The results of Beyer et al. (1999) are, however, asymptotic in nature and one could still try to compare distances in the finite sample situation. From the initial simulations and data analyses in **Paper III**, we observed that the lack of stable distance measures was a challenge, regardless of the criterion. All distance measures were highly dependent on the data, such that removing a small (random) subset of the data had a substantial impact on the optimal number of clusters. This prompted the search for alternative selection procedures.

Shen et al. (2012) and Newell et al. (2013) used an approach related to the prediction strength criteria of Tibshirani and Walther (2005), where cluster reproducibility is used to define clusters. The connection between k-means clustering and PCA enables easy cluster prediction in new data, and Shen et al. (2012) therefore evaluated the predictive power for different K by randomly splitting the data in discovery and validation sets and measuring the similarity between the prediction and validation clusterings. This approach utilizes the same concepts as cross-validation, evaluating how clusters can be reproduced in subsets of the data. However, as the procedure chooses the number of clusters to give good cluster reproducibility, the cluster separation is not necessarily taken into account. In the Metabric analysis in **Paper III**, it was observed that the predicted scores are very stable in the high-dimensional setting, resulting in perfect reproducibility regardless of sub-sampling. This has the unintended consequence that good reproducibility does not ensure any cluster separation, and the implicit assumption that good prediction equals cluster differences is not necessarily true. If we optimize K for reproducibility, we can only ensure that the clusters are stable, not separated.

To avoid this problem in **Paper III**, we instead used a procedure based on the work of Hamerly and Elkan (2003). Here, deviations from normality in the score distribution are interpreted as an indication of a cluster structure. This introduces a factor of subjectivity into the selection, as normality is best evaluated with normal quantile-quantile-plots or with a critical use of normality tests. A subjective choice of clusters seems to contradict the usual approach to selection, where an optimal value is chosen in an objective fashion. However, in practice the optimal choice is never straight forward. For instance, in the original Metabric analysis (Curtis et al., 2012), several cluster combinations were explored before a final choice was made. In this regard, our procedure can give a more transparent framework for the selection of the number of clusters.

Another objection to our choice is that the approach of Hamerly and Elkan (2003) requires the noise to be continuous, normally distributed, a difficult assumption in applied genomics. The Metabric study analyzes copy number aberrations, which are inherently not continuous and therefore not properly normally distributed. The procedure still worked, but made the choices particularly difficult and ambiguous. A possible future direction is to model the copy number aberrations as specifically tailored random variables, and evaluate deviations from the resulting distribution.

References

Ahn, J., J. Marron, K. M. Muller, and Y.-Y. Chi (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika 94*(3), 760–766.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics*, 122–148.

Arabie, P. and L. Hubert (1996). Advances in cluster analysis relevant to marketing research. In *From Data to Knowledge*, pp. 3–19. Springer.

Bai, Z. and J. W. Silverman (2010). *Spectral analysis of large dimensional random matrices.* New York: Springer.

Bartholomew, D. J., M. Knott, and I. Moustaki (2011). *Latent variable models and factor analysis: a unified approach*, Volume 899. John Wiley & Sons.

Beyer, K., J. Goldstein, R. Ramakrishnan, and U. Shaft (1999). When is "nearest neighbor" meaningful? In *Database Theory—ICDT'99*, pp. 217–235. Springer.

Box, G. E. and N. R. Draper (1987). *Empirical model-building and response surfaces.* John Wiley & Sons.

Brawand, D., M. Soumillon, A. Necsulea, P. Julien, G. Csárdi, P. Harrigan, M. Weier, A. Liechti, A. Aximu-Petri, M. Kircher, et al. (2011). The evolution of gene expression levels in mammalian organs. *Nature 478*(7369), 343–348.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science 16*(3), 199–231.

Bühlmann, P., M. Kalisch, and L. Meier (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application 1*, 255–278.

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications.* Springer.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications.* CRC Press.

Caliński, T. and J. Harabasz (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods 3*(1), 1–27.

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2012). *Measurement error in nonlinear models: a modern perspective.* CRC press.

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza (1994). *The history and geography of human genes.* Princeton university press.

Chang, W.-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 267–275.

Cox, D. R. (1990). Role of models in statistical analysis. *Statistical Science*, 169–174.

Curtis, C., S. P. Shah, S.-F. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. Lynch, S. Samarajiwa, Y. Yuan, et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature 486*(7403), 346–352.

Damon, J. and J. Marron (2013). Backwards principal component analysis and principal nested relations. *Journal of Mathematical Imaging and Vision*, 1–8.

d'Aspremont, A., F. Bach, and L. E. Ghaoui (2008). Optimal solutions for sparse principal component analysis. *The Journal of Machine Learning Research 9*, 1269–1294.

De Soete, G. and J. D. Carroll (1994). K-means clustering in a low-dimensional euclidean space. In *New approaches in classification and data analysis*, pp. 212–219. Springer.

Ding, C. and X. He (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 29. ACM.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics 4*(1), 95–104.

Faber, N., M. Meinders, P. Geladi, M. Sjöström, L. Buydens, and G. Kateman (1995a). Random error bias in principal component analysis. part i. derivation of theoretical predictions. *Analytica chimica acta 304*(3), 257–271.

Faber, N., M. Meinders, P. Geladi, M. Sjöström, L. Buydens, and G. Kateman (1995b). Random error bias in principal component analysis. part ii. application of theoretical predictions to multivariate problems. *Analytica chimica acta 304*(3), 273–283.

Girshick, M. (1939). On the sampling theory of roots of determinantal equations. *The Annals of Mathematical Statistics 10*(3), 203–224.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science 286*(5439), 531–537.

Hamerly, G. and C. Elkan (2003). Learning the k in k-means. In *NIPS*, Volume 3, pp. 281–288.

Hartigan, J. A. and M. A. Wong (1979). Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, 100–108.

Hastie, T., R. Tibshirani, J. Friedman, T. Hastie, J. Friedman, and R. Tibshirani (2009). *The elements of statistical learning*, Volume 2. Springer.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology 24*(6), 417.

Johnstone, I. and A. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104*(486), 682–693.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, 295–327.

Jolliffe, I. (2002). *Principal component analysis.* New York: Springer.

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics 12*(3), 531–547.

Journée, M., Y. Nesterov, P. Richtárik, and R. Sepulchre (2010). Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research 11*, 517–553.

Jung, S. and J. S. Marron (2009). Pca consistency in high dimension, low sample size context. *Ann. Statist. 37*(6B), 4104–4130.

Jung, S., A. Sen, and J. S. Marron (2012). Boundary behavior in high dimension, low sample size asymptotics of pca. *J. Multivariate Anal.*.

Karakach, T. and P. Wentzell (2007). Methods for estimating and mitigating errors in spotted, dual-color dna microarrays. *Omics: a journal of integrative biology 11*, 186–199.

Kaufman, L. and P. J. Rousseeuw (2009). *Finding groups in data: an introduction to cluster analysis*, Volume 344. John Wiley & Sons.

Kristensen, V. N., O. C. Lingjærde, H. G. Russnes, H. K. M. Vollan, A. Frigessi, and A.-L. Børresen-Dale (2014). Principles and methods of integrative genomic analyses in cancer. *Nature Reviews Cancer 14*(5), 299–313.

Lee, S., F. Zou, and F. Wright (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist. 38*(6), 3605.

Lock, E. F., K. A. Hoadley, J. Marron, and A. B. Nobel (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics 7*(1), 523.

Lu, A. Y. (2002). *Sparse principal component analysis for functional data.* Ph. D. thesis, Stanford University.

Ma, S., M. R. Kosorok, and J. P. Fine (2006). Additive risk models for survival data with high-dimensional covariates. *Biometrics 62*(1), 202–210.

Ma, Z. et al. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41*(2), 772–801.

Milligan, G. W. and M. C. Cooper (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika 50*(2), 159–179.

Milligan, G. W. and M. C. Cooper (1987). Methodology review: Clustering methods. *Applied Psychological Measurement 11*(4), 329–354.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Statist. 36*(6), 2791–2817.

Narasimhan, S. and S. L. Shah (2008). Model identification and error covariance matrix estimation from noisy data using pca. *Control Engineering Practice 16*(1), 146–155.

Newell, M. A., D. Cook, H. Hofmann, J.-L. Jannink, et al. (2013). An algorithm for deciding the number of clusters and validation using simulated data with application to exploring crop population structure. *The Annals of Applied Statistics 7*(4), 1898–1916.

Patterson, N., A. L. Price, and D. Reich (2006). Population structure and eigenanalysis. *PLoS genetics 2*(12), e190.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica 17*(4), 1617.

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2*(11), 559–572.

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

Pourahmadi, M. (2013). *High-dimensional covariance estimation: with high-dimensional data*. John Wiley & Sons.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics 38*(8), 904–909.

Rocke, D. M. and B. Durbin (2001). A model for measurement error for gene expression arrays. *Journal of computational biology 8*(6), 557–569.

Sanguinetti, G., M. Milo, M. Rattray, and N. D. Lawrence (2005). Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics 21*(19), 3748–3754.

Shen, D., H. Shen, H. Zhu, and J. Marron (2013). Surprising asymptotic conical structure in critical sample eigen-directions. *arXiv preprint arXiv:1303.6171*.

Shen, D., H. Shen, H. Zhu, and J. S. Marron (2012). High dimensional principal component scores and data visualization. *arXiv preprint arXiv:1211.2679*.

Shen, H. and J. Z. Huang (2008). Sparse principal component analysis via regularized low rank matrix approximation. *Journal of multivariate analysis 99*(6), 1015–1034.

Shen, R., A. B. Olshen, and M. Ladanyi (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics 25*(22), 2906–2912.

Shen, R., S. Wang, and Q. Mo (2013). Sparse integrative clustering of multiple omics data sets. *The annals of applied statistics*.

Sørensen, Ø., A. Frigessi, and M. Thoresen (2012). Measurement error in lasso: Impact and correction. *arXiv preprint arXiv:1210.5378*.

Sørensen, Ø., A. Frigessi, and M. Thoresen (2014). Covariate selection in high-dimensional generalized linear models with measurement error. *arXiv preprint arXiv:1407.1070*.

Sørlie, T., C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences 98*(19), 10869–10874.

Thoresen, M. and P. Laake (2000). A simulation study of measurement error correction methods in logistic regression. *Biometrics 56*(3), 868–872.

Tibshirani, R. and G. Walther (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics 14*(3), 511–528.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.

Valente, A. X., J. H. Shin, A. Sarkar, and Y. Gao (2012). Rare coding snp in dzip1 gene associated with late-onset sporadic parkinson's disease. *Scientific reports 2*.

Wall, M. E., A. Rechtsteiner, and L. M. Rocha (2003). Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pp. 91–109. Springer.

Wentzell, P. and S. Hou (2012). Exploratory data analysis with noisy measurements. *Journal of Chemometrics 26*(6), 264–281.

Wentzell, P. D. and M. T. Lohnes (1999). Maximum likelihood principal component analysis with correlated measurement errors: theoretical and practical considerations. *Chemometrics and Intelligent Laboratory Systems 45*(1), 65–85.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, kxp008.

Yang, F., K. Doksum, and K.-W. Tsui (2014). Principal component analysis (pca) for high-dimensional data. pca is dead. long live pca (perspectives on big data analysis: Methodologies and applications).

Yeung, K. Y. and W. L. Ruzzo (2001). Principal component analysis for clustering gene expression data. *Bioinformatics 17*(9), 763–774.

Zha, H., X. He, C. Ding, M. Gu, and H. D. Simon (2001). Spectral relaxation for k-means clustering. In *NIPS*, Volume 1, pp. 1057–1064.

Ziegler, A., I. R. König, and F. Pahlke (2010). *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an e-learning platform.* John Wiley & Sons.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.

Asymptotic distribution of high-dimensional principal component scores connected to pervasive eigenvectors

Kristoffer Hellton, Magne Thoresen
Department of Biostatistics, University of Oslo,
P.O.Box 1122 Blindern N-0317, Oslo, Norway
k.h.hellton@medisin.uio.no

September 26, 2014

Abstract

Principal component analysis (PCA) is a widely used technique for dimension reduction, also for high-dimensional data. In the high-dimensional framework, PCA is not asymptotically consistent, as sample eigenvectors do not converge to the population eigenvectors. However, in this paper it is shown that for a pervasive signal, the visual content of the sample principal component (PC) scores will be the same as for the population PC scores. The asymptotic distribution of the ratio between the individual sample and population scores is derived, assuming that eigenvalues scale linearly with the dimension. The distribution of the ratio consists of a main shift and a noise part, where the main shift does not depend on the individual scores. As a consequence, all sample scores are affected by an approximate common scaling, such that the relative positions of the population scores are kept. Simulations show that the noise part is negligible for the purpose of visualization, for small to moderate sample sizes depending on the signal strength. The realism of the eigenvalue assumption is supported by introducing the pervasive signal structure, where the number of non-zero effects is a non-vanishing proportion of the total number of variables. If an eigenvector is pervasive with fixed values, we show that the corresponding eigenvalue will scale linearly with the dimension. Two data examples from genomics, where pervasiveness is reasonable, are discussed.

Keywords: Consistency, Asymptotic distribution, High-dimensional data, Principal component analysis, Principal component scores, Visualization.

1 Introduction

Principal component analysis (PCA) is the workhorse of variable reduction in applied data analysis. It is used to construct a small number of informative scores from the original data, and these scores are then used further in visualization or in conventional classification, clustering or regression methods. This is highly useful in the context of modern high-dimensional data analysis, where the number of measured variables p exceeds the sample size n. Genomics is an application area are where the first step in exploring data is often to visually investigate the first few principal component (PC) scores.

The asymptotic behavior of high-dimensional PCA has attracted a substantial amount of attention the last few years. It has been shown, by Paul (2007) and Johnstone and Lu (2009), that the population eigenvalues and -vectors in PCA are not consistently estimated by the sample eigenvalues and -vectors under the finite γ regime, where $p/n = \gamma$ as $p, n \to \infty$. The inconsistency of the eigenvectors is quantified in terms of the inner product between the sample and the population eigenvectors, which then does not converge to 1. In view of this, Johnstone and Lu (2009) suggest that one could either conduct an initial dimension reduction, from the original number of variables to a value less than n, before applying PCA, or introduce a sparse penalty on the eigenvectors, giving rise to the sparse PCA methodology (Witten et al., 2009; Zou et al., 2006). However, in an applied setting, the behavior of the principal component scores is also of interest, in addition to the eigenvectors and eigenvalues.

Until now, only few papers have focused on the asymptotic behavior of the principal component scores. An exception is Lee et al. (2010), who note that: "Inconsistency of the sample eigenvectors does not necessarily imply poor performance of PCA". The success of applied PC scores in genomics suggests that they can be considered suitable for an analysis, in spite of the inconsistency of the eigenvectors. In this paper, we explore this somewhat paradoxical situation further, and try to bridge the gap between the theoretical problems of PCA and the practical usefulness of the method.

We first review the main structure of high-dimensional PCA and earlier asymptotic results. Next, we introduce the concept of pervasive effects and demonstrate that this leads to population eigenvalues scaling linearly with the data dimension. Under this assumption about the eigenvalues, we derive the asymptotic limiting distribution of the ratio between the estimated and the true principal component scores. The implications of our findings are explored theoretically and by simulations. We show that all sample PC scores are subject to a common scaling and a small noise term, such that the relative positions of the sample and population scores are essentially unchanged by the scaling.

2 Principal component analysis

2.1 Methods and notation

PCA reduces the data dimension by constructing orthogonal linear combinations of variables, which explain their variability. The first component is the normalized linear combination of variables with the highest variance, while the second component will be the linear combination, orthogonal to the first, with the highest variance, and so on. The mathematical basis of PCA is the eigendecomposition of the sample covariance matrix.

Let $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ be a $p \times n$ data matrix, where $\mathbf{x}_i = [x_{i1}, \dots, x_{ip}]^T$ are independent and identically distributed with $\mathrm{E}\,\mathbf{x}_i = \mathbf{0}$ and $\operatorname{var}\mathbf{x}_i = \Sigma$. The eigendecomposition of the covariance matrix is given by

$$\Sigma = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T,$$

where $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ and $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_p]$ is the matrix of eigenvectors. The weights of the orthogonal linear combinations are given by the eigenvectors, usually referred to as loadings. We denote the vector of the resulting population component scores by

$$\mathbf{s}_j^T = \mathbf{v}_j^T\mathbf{X} = [\mathbf{v}_j^T\mathbf{x}_1, \dots, \mathbf{v}_j^T\mathbf{x}_n]. \tag{1}$$

The eigenvalues express the variance of the component scores, such that the vector of standardized population component scores is given by

$$\mathbf{z}_j^T = \frac{\mathbf{v}_j^T\mathbf{X}}{\sqrt{\lambda_j}},$$

where the jth vector of scores is $\mathbf{z}_j^T = [z_{j1}, \dots, z_{jn}]$ and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_p]^T$.

An applied data analysis is based on the sample covariance matrix denoted by $\hat{\Sigma} = \frac{1}{n}\mathbf{X}\mathbf{X}^T$, with the eigendecomposition

$$\hat{\Sigma} = \hat{\mathbf{V}}\mathbf{D}\hat{\mathbf{V}}^T,$$

Here, $\mathbf{D} = \operatorname{diag}(d_1, \dots, d_p)$ contains the sample eigenvalues and $\hat{\mathbf{V}} = [\hat{\mathbf{v}}_1, \dots, \hat{\mathbf{v}}_p]$ the corresponding sample eigenvectors. Following the earlier notation, we construct the sample component scores as

$$\hat{\mathbf{s}}_j^T = \hat{\mathbf{v}}_j^T\mathbf{X},$$

and the sample standardized scores as

$$\hat{\mathbf{z}}_j^T = \frac{\hat{\mathbf{v}}_j^T \mathbf{X}}{\sqrt{d_j}}.$$

We further assume that the population eigenvalues follow the spiked eigenvalue model introduced by Johnstone (2001), where the first m population eigenvalues are substantially larger than the remaining non-spiked eigenvalues.

2.2 Brief summary of earlier results

The question of consistency is central in statistics, as the sample estimates should converge to the population parameters when the sample size increases. Anderson (1963) showed that the sample eigenvectors and -values, $\hat{\mathbf{v}}$ and d, will consistently estimate the population eigenvectors and -values, $\mathbf{v}$ and λ, when p is fixed and $n \to \infty$.

However, this is not true in the high-dimensional setting, where $p > n$. Starting with Paul (2007) and Johnstone and Lu (2009), it has been shown that the sample eigenvalues and -vectors are not asymptotically consistent when $p, n \to \infty$ at a constant ratio $p/n = \gamma > 0$ and the population eigenvalues are fixed. Paul (2007) showed that the inner product between the sample and the population eigenvector converges, when $\lambda_j > 1 + \sqrt{\gamma}$, to

$$|\langle \hat{\mathbf{v}}_j, \mathbf{v}_j \rangle| \to \sqrt{\left(1 - \frac{\gamma}{(\lambda_j - 1)^2}\right) / \left(1 + \frac{\gamma}{\lambda_j - 1}\right)} \quad j = 1, \ldots, m.$$

A different asymptotic setting starts from the geometrical structure of the data in a high-dimensional space (Ahn et al., 2007; Hall et al., 2005). Jung and Marron (2009) introduced the high dimension, low sample size (HDLSS) setting, where n is fixed and the spiked eigenvalues grow with the dimension p, according to $\lambda_i = \sigma_i^2 p^\alpha, i = 1, \ldots, m$. In this asymptotic setting, the consistency of PCA depends on α, as $p \to \infty$. Eigenvectors are estimated consistently when $\alpha > 1$, while the estimates are strongly inconsistent when $\alpha < 1$. In the boundary case $\alpha = 1$, a situation explored by Jung et al. (2012), the sample eigenvectors are neither consistent nor strongly inconsistent, but reach a limiting distribution depending on n. In the case $m = 1$, where there is a single spiked eigenvalue, we have the following:

$$|\langle \hat{\mathbf{v}}_1, \mathbf{v}_1 \rangle| \xrightarrow{\mathrm{d}} \begin{cases} 1 & \alpha > 1, \\ \left(1 + \frac{\tau^2}{\sigma^2 \chi_n^2}\right)^{-1/2} & \alpha = 1, \\ 0 & \alpha < 1. \end{cases}$$

The main focus of the above-mentioned papers has been on the eigenvector inconsistency, and few results are concerned with principal component scores. On exception is Lee et al. (2010), which established the asymptotic limit of the inner product between the sample and the population scores. They extended the result to prediction and found a theoretical asymptotic shrinkage factor for predicted scores. This can be applied as a bias adjustment, which turns out to be useful in the context of genetic population stratification problems. Also Yata and Aoshima (2009, 2012) explore the consistency of scores as $p \to \infty$ and $n \to \infty$. Leek (2011) showed that the estimated right singular vectors, which corresponds to the PC scores, in a low-dimensional conditional factor model converge (fixed n and $p \to \infty$) to a set of vectors which span the same column space as the true factors.

Further, Shen et al. (2013, 2012) investigated the ratio between the individual sample and the population scores $\hat{z}_i/z_i, i = 1, \ldots, n$, instead of the inner product between the score vectors. Following the regime of Jung and Marron (2009), they showed that for $\alpha > 1$ the ratio converges to a random variable independent of i. This implies that a two-dimensional plot of the sample scores is asymptotically only a scaled version of the population score plot. The visual information contained in the samples scores will therefore remain the same as in the population scores. In this paper, we investigate the same problem as Shen et al. (2012), but in the situation where $\alpha = 1$. We first motivate why this is an interesting assumption to make, and then prove the asymptotic behavior of the sample scores under this assumption.

3 Data structure and eigenvalues

The initial results concerning the consistency of sample eigenvalues and -vectors were derived on the basis of random matrix theory (Bai and Silverman, 2010). This requires the ratio p/n to remain constant as $p, n \to \infty$ and the population eigenvalues to be fixed. In contrast, the HDLSS regime of Jung and Marron (2009) considers situations where the population eigenvalues depend asymptotically on the dimension p, according to $\lambda_i \sim p^\alpha, \alpha > 0$.

Which of these two settings is the more appropriate remains an open question. As Lee et al. (2010) conclude:

> "It may be argued that for real data where p/n is "large," we should follow the paradigm of [the HDLSS regime]. However, for any real study, it is unclear how to test whether p increases at a faster rate than λ_i or vice versa, making the application of [the HDLSS regime] difficult in practice."

Establishing a natural connection between the eigenvalue model and real data problems is not an easy task. We believe it is appropriate to assume the eigenvalues to scale linearly with the dimension p, corresponding to the case of Jung et al. (2012) where $\alpha = 1$. We argue in favor of this by introducing an assumption on the eigenvector coefficients. Our aim is to translate the assumption about the eigenvalues into an assumption regarding the latent structure and the data-generating mechanism, as this is generally easier to relate to.

First, let the observations $\mathbf{x}_i$ be generated by a Gaussian latent variable model with an additive, isotropic error and a single factor:

$$\mathbf{x}_i = \mathbf{v}z_i + \epsilon_i, \quad i = 1, \dots, n,$$

where the scalar $z_i \sim N(0, 1)$ and the noise vector $\epsilon_i \sim N(0, \sigma^2 I)$. Tipping and Bishop (1999) established the connection between the Gaussian latent variable model under isotropic noise and PCA. The population covariance matrix of $\mathbf{x}_i$ is then given by

$$\Sigma = \mathbf{v}\mathbf{v}^T + \sigma^2 I.$$

As $\mathbf{v}$ is an eigenvector of $\mathbf{v}\mathbf{v}^T$ when normalized, it is also an eigenvector of Σ, such that the corresponding eigenvalue of $\mathbf{v}\mathbf{v}^T$ is given by the normalizing constant $\lambda_1(\mathbf{v}\mathbf{v}^T) = \sum_{j=1}^{p} v_j^2$. Thus the largest eigenvalue of Σ is given by

$$\lambda_1 = \sum_{j=1}^{p} v_j^2 + \sigma^2.$$

The relationship between the largest population eigenvalue of Σ and the dimension can therefore be determined by the structure of $\mathbf{v}$. For instance will the following three structures set $\sum_{j=1}^{p} v_j^2$ and the eigenvalue to scale linearly with p, as $p \to \infty$:

i) the values of the v_j are fixed, and the number of non-zero effects scales with p

ii) the number of non-zero effects is fixed, and some values of the v_j scale linearly in p

iii) a combination of i) and ii), where the combined rate is linear

It can be difficult to find realistic examples which would fulfill the settings ii) or iii). However the first situation can be interpreted in terms of pervasiveness (Fan et al., 2011):

Definition 1 (Pervasiveness)**.** *A sequence of p-dimensional vectors $\mathbf{v} = [v_1, \dots, v_p]^T$ is pervasive, if the proportion of non-zero entries $r_p = \frac{1}{p}\sum_{i=1}^{p} I_{\{v_i^2 > 0\}}$ fulfills:*

$$\lim_{p \to \infty} r_p > 0,$$

In the field of high-dimensional approximated factor models, Fan et al. (2011) refer to Definition 1 as the pervasiveness assumption. Here, the number of non-zero entries in $\mathbf{v}$ is a non-vanishing proportion of the dimension p, as p increases. This stands in contrast to a sparse signal, where the number of non-zero effects is fixed, such that the proportion converges to zero.

If we assume the vector $\mathbf{v}$ to be pervasive with fixed values, meaning that the latent factor z_i will have a pervasive effect on the observed variable $\mathbf{x}_i$, we have the following:

Result 1. *If $\mathbf{x}_i = \mathbf{v}z_i + \epsilon_i$, where $z_i \sim N(0,1)$, $\epsilon_i \sim N(0, \sigma^2 I)$ and $\mathbf{v}$ is assumed pervasive with fixed values, the largest population eigenvalue λ_1 of the population covariance matrix Σ fulfills the bound*

$$c_1 p + \sigma^2 \leq \lambda_1 \leq c_2 p + \sigma^2.$$

The remaining population eigenvalues are given by $\lambda_i = \sigma^2$ for $i = 2, \ldots, p$.

The result follows from the existence of two constants $0 < c_1 \leq c_2 < \infty$, depending on r_p and the minimum and maximum of the non-zero square loadings v_j^2 respectively, such that the following bound is satisfied

$$c_1 p \leq \sum_{j=1}^{p} v_j^2 \leq c_2 p.$$

If the observations are given by m components

$$\mathbf{x}_i = \sum_{k=1}^{m} \mathbf{v}_k z_{ik} + \sigma \mathbf{u}_i,$$

where the $\mathbf{v}_k, k = 1, \ldots, m$ are orthogonal and pervasive with fixed values and $\sum_{j=1}^{p} v_{1j}^2 \geq \cdots \geq \sum_{j=1}^{p} v_{mj}^2$, the covariance matrix of $\mathbf{x}_i$ will have m eigenvalues, which scale linearly with the dimension

$$\lambda_i \sim p, \quad i = 1, \ldots, m.$$

It is also possible to interpret the pervasiveness in terms of the covariance matrix

$$\Sigma = \mathbf{v}\mathbf{v}^T + \sigma^2 I = \begin{bmatrix} v_1^2 + \sigma^2 & v_1 v_2 & \cdots & v_1 v_p \\ v_1 v_2 & v_2^2 + \sigma^2 & & \\ \vdots & & \ddots & \end{bmatrix}.$$

We can group the non-zero v_i together into blocks, where the dimension of the blocks depends on the proportion r. We illustrate the situation in Example 1, where all effects are equal and the population covariance matrix consists of separate clusters, where all variables within each cluster are equally correlated, while the different clusters are independent. Each cluster corresponds to an eigenvector proportional to $\mathbf{v}_j = [0, \dots, 0, 1, \dots, 1, 0, \dots, 0]$ with ones for the variables within the cluster and zeros for others. If the cluster sizes are not fixed, but a proportion of the total number of variables, the eigenvectors will be pervasive.

Example 1. *Assume Σ to be divided into different independent sub-matrices*

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 & \dots & 0 \\ 0 & \Sigma_2 & 0 & \vdots \\ \vdots & 0 & \Sigma_3 & 0 \\ 0 & \dots & 0 & \sigma^2 I \end{bmatrix}, \quad \textit{where} \quad \Sigma_j = \sigma^2 \begin{bmatrix} 1 & \rho_j & \dots & \rho_j \\ \rho_j & 1 & & \\ \vdots & & \ddots & \rho_j \\ \rho_j & & \rho_j & 1 \end{bmatrix},$$

and $\rho_1 > \rho_2 > \rho_3$. The dimension of Σ is $p \times p$ and the size of Σ_j is $r_j p$ with $1 \geq r_1 + r_2 + r_3 > 0$. Then there will be three top eigenvalues of Σ:

$$\lambda_1 = \sigma^2 \rho_1 r_1 \, p + \sigma^2 (1 - \rho_1),$$
$$\lambda_2 = \sigma^2 \rho_2 r_2 \, p + \sigma^2 (1 - \rho_2),$$
$$\lambda_3 = \sigma^2 \rho_3 r_3 \, p + \sigma^2 (1 - \rho_3),$$

where all scale linearly with the dimension, $\lambda_i \sim p$. The other eigenvalues $\lambda_4, \dots, \lambda_p$ are constant, where there are $p(1 - r_1 - r_2 - r_3)$ eigenvalues equal to

$$\lambda_i = \sigma^2,$$

while the remaining eigenvalues $\lambda_i = \sigma^2(1 - \rho_j)$ have multiplicity $pr_j - 1$ for $j = 1, 2, 3$.

Each of the three largest eigenvalues represents one cluster and the importance of the cluster is determined by the proportion r_j, how many variables that are represented, and the degree of correlation within the cluster ρ_j. More strongly correlated variables will exhibit a clearer signal, while the remaining eigenvalues represent the noise. The pervasive eigenvectors can therefore be interpreted as variable clusters, where the cluster size is a percentage of the total number of variables.

3.1 Realistic examples from genomics

We present two situations in genomics, an area with several types of high-dimensional data, where the biological processes suggest the pervasiveness assumption to be reasonable. One example is genetic markers such as SNPs, single-basepair polymorphic genetic loci, i.e. having at least two alleles with an associate allelic frequency in a population. The neutral theory of molecular evolution states that allele frequencies at most loci (SNPs) change due to two stochastic processes; mutation and random drift.

If the main variation in the data sample stems from differences between ethnic populations, random allelic drift is the main driver behind changes in the genetic markers. This will give many and randomly distributed differences and when new markers are included, we expect a certain proportion to be informative with respect to the ethnicity. This corresponds to our notion of pervasive effects, and if the effects are fixed, the corresponding eigenvalue will scale linearly with total number of included variables. The longer two populations have been separated, the larger degree of SNPs expressing differences we expect, as observed by Yamaguchi-Kabata et al. (2008) when comparing Europeans and Japanese to subgroups within the Japanese population.

Another example is microarray expression data, which quantify the amount of a gene product called mRNA, whose expression is necessary for making proteins. Cancer is a relevant disease in this respect, as it can be considered to have a systemic effect on gene expression (Perou et al., 2000). We therefore expect to observe many differentially expressed genes between groups of cancer patients and healthy individuals. The meta-analysis of Kondrakhin et al. (2008) showed that around 5% of 24726 genes are differentially expressed between cases of breast cancer and controls. This situation also corresponds to our notion of pervasive effects.

4 Asymptotic results

In the following, we present two results regarding the consistency of PC scores of high-dimensional data. The asymptotic framework follows the high-dimension low sample size regime for the case where $\alpha = 1$ as considered by Jung et al. (2012), and we state the same general conditions for the distribution of the component scores and for the structure of the population eigenvalues.

Firstly, the assumption of independent and normally distributed z_{ij} can be relaxed, as $\mathbf{x}_i$ has zero mean and covariance matrix Σ, to the following distributional condition:

Condition 1. *The standardized principal component scores $\mathbf{z}_i$ have finite fourth moments and are uncorrelated but possibly dependent fulfilling the ρ-mixing condition.*

The ρ-mixing condition is satisfied if the maximal correlation coefficient approach zero, $\rho(m) \to$

0, as $m \to \infty$, where

$$\rho(m) = \sup_{j,f,g} |\operatorname{cor}(f,g)|, \quad f \in L_2(\mathcal{F}_{-\infty}^{j}), g \in L_2(\mathcal{F}_{j+m}^{\infty}),$$

and $\mathcal{F}_K^L$ is the σ-field of events generated by the variables $\mathbf{z}_i, K \leq i \leq L$.

Secondly, the structure of the non-spiked eigenvalues $\lambda_{m+1}, \dots, \lambda_p$ in the spiked covariance model can be generalized by the following condition:

Condition 2. *For the eigenvalues $\lambda_{m+1}, \dots, \lambda_p$, it must hold that*

$$\frac{\sum_{i=m+1}^{p} \lambda_i^2}{\left(\sum_{i=m+1}^{p} \lambda_i\right)^2} \to 0, \quad \frac{1}{p} \sum_{i=m+1}^{p} \lambda_i \to \tau^2, \qquad \text{as } p \to \infty.$$

Condition 2 insures that the non-spiked eigenvalues do not decrease too fast and that the mean converges to τ^2. The constant non-spiked eigenvalues $\lambda_{m+1} = \cdots = \lambda_p = \tau^2$ in the spiked covariance model is the simplest situation which fulfils condition 2.

Finally, we assume the spiked eigenvalues to scale linearly with the dimension:

Assumption 1 (Linearity)**.** *For the m spiked components, the eigenvalues depend on the dimension p according to*

$$\lambda_1 = \sigma_1^2 p, \quad \lambda_2 = \sigma_2^2 p \quad \cdots \quad \lambda_m = \sigma_m^2 p,$$

where $\sigma_1^2 \geq \cdots \geq \sigma_m^2 > 0$ represent the signal strength.

Theorem 1 determines the asymptotic limiting distributions of the ratio between the sample and the population principal component scores under the Conditions 1 and 2, and Assumption 1. The results depend on the stochastic behavior of the eigenvalues and the eigenvectors of an $m \times m$ matrix $\mathbf{W} = \tilde{\mathbf{Z}}_{1:m}^T \tilde{\mathbf{Z}}_{1:m}$, where $\tilde{\mathbf{Z}}_{1:m} = [\sigma_1 \mathbf{z}_1, \dots, \sigma_m \mathbf{z}_m]$. We denote the jth eigenvalue of $\mathbf{W}$ by $\phi_j(\mathbf{W})$ and the jth eigenvector by $\mathbf{v}_j(\mathbf{W})$.

Theorem 1. *Under Conditions 1 and 2, and Assumption 1 for $m \geq 1$, the ratio between the sample and the population principal component scores converges in distribution to the following limit, as $p \to \infty$:*

$$\left| \frac{\hat{z}_{ij}}{z_{ij}} \right| \xrightarrow{\text{d}} R_j + \varepsilon_{ij} \quad i = 1, \dots, n; j = 1, \dots, m,$$

where the ratio R_j is distributed as

$$R_j \sim \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\, \sigma_j v_{jj}(\mathbf{W}),$$

and ε_{ij} is distributed as

$$\epsilon_{ij} \sim \sqrt{\frac{n}{\phi_j(\mathbf{W})}} \sum_{k=1,k\neq j}^{m} \sigma_k \frac{z_{ik}}{z_{ij}} v_{jk}(\mathbf{W}).$$

Remark 1. *If the standardized component scores are assumed to be* iid *normally distributed*

$$z_{ij} \sim \mathcal{N}(0,1), \quad i = 1,\dots,n, j = 1,\dots,p,$$

$\mathbf{W}$ *will be an $m \times m$ Wishart distributed matrix*

$$\mathbf{W} \sim W_m \left(\operatorname{diag}(\sigma_1^2,\dots,\sigma_m^2), n-1\right),$$

and $\phi_j(\mathbf{W})$ and $\mathbf{v}_j(\mathbf{W})$ will be asymptotically (as $n \to \infty$) independent and normally distributed (Jolliffe, 2002).

Proof of Theorem 1. Theorem 1 follows from Lemmas 1 and 2, which are given by the results of Jung et al. (2012).

Lemma 1 (Jung et al. (2012))**.** *Under Conditions 1 and 2, and Assumption 1, the sample eigenvalues converge in distribution*

$$p^{-1}d_j \xrightarrow{\mathrm{d}} \begin{cases} \phi_j(\mathbf{W})/n + \tau^2/n, & j = 1,\dots,m, \\ \tau^2/n, & j = m+1,\dots,p, \end{cases}$$

and for all $k = 1,\dots,p$, the sample eigenvectors satisfy

$$\hat{\mathbf{v}}_j^T \mathbf{v}_k = \sqrt{\frac{\lambda_k}{nd_j}}\, \mathbf{z}_k^T \hat{\mathbf{u}}_j, \quad j = 1,\dots,m. \tag{2}$$

Here, the $\hat{\mathbf{u}}_j$ are the sample eigenvectors of $p^{-1}\mathbf{X}^T\mathbf{X} = p^{-1}\sum_{k=1}^{p} \lambda_k \mathbf{z}_k \mathbf{z}_k^T$ which converge in distribution to

$$\hat{\mathbf{u}}_j \xrightarrow{\mathrm{d}} \frac{\tilde{\mathbf{Z}}_{1:m}\mathbf{v}_j(\mathbf{W})}{\sqrt{\phi_j(\mathbf{W})}},, \quad p \to \infty.$$

Lemma 2 (Jung et al. (2012))**.** *Under Condition 1 and Condition 2, it follows that*

$$\frac{1}{p} \sum_{i=m+1}^{p} \lambda_i \mathbf{z}_i \mathbf{z}_i^T \xrightarrow{\mathrm{P}} \tau^2 I_n,$$

in probability.

The normalized sample principal component scores can be decomposed by the expression $\hat{z}_{ij} = d_j^{-1/2} \hat{\mathbf{v}}_j^T \sum_{k=1}^{p} \lambda_k^{1/2} \mathbf{v}_k z_{ik}$. By using the expression in (2), the ratio between the sample PC scores and the population PC scores can be decomposed and we can insert the m spiked population eigenvalues to give:

$$\begin{aligned}\frac{\hat{z}_{ij}}{z_{ij}} &= d_j^{-1/2}\sum_{k=1}^{p}\lambda_k^{1/2}\frac{z_{ik}}{z_{ij}}\;\hat{\mathbf{v}}_j^T\mathbf{v}_k = \frac{1}{\sqrt{n}d_j z_{ij}}\left(\sum_{k=1}^{m}\lambda_k z_{ik}\mathbf{z}_k^T + \sum_{k=m+1}^{p}\lambda_k z_{ik}\mathbf{z}_k^T\right)\hat{\mathbf{u}}_j \\ &= \frac{1}{\sqrt{n}p^{-1}d_j z_{ij}}\left(\sum_{k=1}^{m}\sigma_k^2 z_{ik}\;\mathbf{z}_k^T + \frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}\mathbf{z}_k^T\right)\hat{\mathbf{u}}_j\end{aligned}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, m$.

When $p \to \infty$, the scaled sample eigenvalue converges according to Lemma 1 to $p^{-1}d_i \xrightarrow{\text{d}} \phi_i(\mathbf{W})/n + \tau^2/n$, while the term consisting of the non-spiked eigenvalues can be rewritten as the vector

$$\frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}\mathbf{z}_k^T = \left[\frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}z_{1k}, \ldots, \frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}^2, \ldots, \frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}z_{nk}\right]^T.$$

By Lemma 2, a version of the law of large numbers, we have for a fixed m that $\frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}z_{lk} \to 0$ for $l \neq i$ and $\frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}^2 \to \tau^2$. Therefore this vector converges to the unit vector $\mathbf{e}_i$ multiplied by τ^2 at position i and zero everywhere else:

$$\frac{1}{p}\sum_{k=m+1}^{p}\lambda_k z_{ik}\mathbf{z}_k^T \to \tau^2\mathbf{e}_i^T.$$

Then, according to Jung et al. (2012), the results in Lemma 1, the ratio between the sample and population scores converges to

$$\begin{aligned}\frac{\hat{z}_{ij}}{z_{ij}} &\xrightarrow{\text{d}} \frac{\sqrt{n}}{(\phi_j(\mathbf{W})+\tau^2)z_{ij}}\left(\sum_{k=1}^{m}\sigma_k^2 z_{ik}\mathbf{z}_k^T + \tau^2\mathbf{e}_i^T\right)\frac{\tilde{\mathbf{Z}}_{1:m}\mathbf{v}_j(\mathbf{W})}{\sqrt{\phi_j(\mathbf{W})}} \\ &= \frac{\sqrt{n}}{(\phi_j(\mathbf{W})+\tau^2)\sqrt{\phi_j(\mathbf{W})}z_{ij}}\left(\sum_{k=1}^{m} z_{ik}\sigma_k^2\mathbf{z}_k^T\tilde{\mathbf{Z}}_{1:m}\mathbf{v}_j(\mathbf{W}) + \tau^2\mathbf{e}_i^T\tilde{\mathbf{Z}}_{1:m}\mathbf{v}_j(\mathbf{W})\right). \qquad (3)\end{aligned}$$

The expression $\sigma^2\mathbf{z}_k^T\tilde{\mathbf{Z}}_{1:m}\mathbf{v}_j(\mathbf{W})$ in the first term corresponds to the kth row of $\mathbf{W}$, and due to the eigen-equation $\mathbf{W}\mathbf{v}_j(\mathbf{W}) = \phi_j(\mathbf{W})\mathbf{v}_j(\mathbf{W})$, this term can be rewritten as

$$\sigma_k\mathbf{z}_k^T\tilde{\mathbf{Z}}_{1:m}\mathbf{v}_j(\mathbf{W}) = \phi_j(\mathbf{W})v_{jk}(\mathbf{W}),$$

while the unit vector in the second term gives

$$\mathbf{e}_i^T\tilde{\mathbf{Z}}_{1:m}\mathbf{v}_j(\mathbf{W}) = [\sigma_1 z_{i1}, \ldots, \sigma_m z_{im}]\mathbf{v}_j(\mathbf{W}) = \sum_{k=1}^{m}\sigma_k z_{ik}v_{jk}(\mathbf{W}).$$

This simplifies expression (3) to

$$\frac{\sqrt{n}\left(\phi_j(\mathbf{W})+\tau^2\right)}{\left(\phi_j(\mathbf{W})+\tau^2\right)\sqrt{\phi_j(\mathbf{W})}z_{ij}}\sum_{k=1}^{m}\sigma_k z_{ik} v_{jk}(\mathbf{W}) = \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\sum_{k=1}^{m}\sigma_k \frac{z_{ik}}{z_{ij}} v_{jk}(\mathbf{W})$$

By splitting the sum, we get the result

$$\frac{\hat{z}_{ij}}{z_{ij}} \xrightarrow{\mathrm{d}} \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\,\sigma_j v_{jj}(\mathbf{W}) + \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\sum_{k=1,k\neq j}^{m}\sigma_k \frac{z_{ik}}{z_{ij}}\,v_{ik}(\mathbf{W}) = R_j + \epsilon_{ij},$$

and we have used the simple notations

$$R_j \sim \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\,\sigma_j v_{jj}(\mathbf{W}),$$

and

$$\epsilon_{ij} \sim \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\sum_{k=1,k\neq j}^{m}\sigma_k \frac{z_{ik}}{z_{ij}}\,v_{jk}(\mathbf{W}).$$

□

5 Implications for the visualization of scores

In the application of PCA, the first few sample scores are used for visualization and in conventional classification and regression methods. Besides the in itself valuable ability to visualize high-dimensional data in a two-dimensional (or 3D) fashion, the score plot can be useful for comparing observations, detecting subgroups, and for identifying outliers and bad data quality. PCA is often viewed as the canonical first step in an applied high-dimensional analysis. However, when it is known that eigenvectors are inconsistently estimated, will a plot of the sample scores give valid information about the population scores? We use Theorem 1 to answer this question in two steps. Firstly, we show by simulation and by providing a supporting theoretical argument, that ε_{ij} is considerably smaller than R_j. Therefore we refer to ε_{ij} as noise. Secondly, we highlight the fact that the R_j are independent of i. As the R_j express ratios, the relative positions of the sample scores will be more or less the same as for the population scores. To illustrate these two points, we take a detailed look at the situation with two components, $m = 2$, in Example 3.

Example 2. *If there is only one component, $m = 1$, $\mathbf{W}$ is a scalar such that the estimated eigenvalue is given by $\phi_1(\mathbf{W}) = \sigma_1^2 \mathbf{z}_1^T \mathbf{z}_1$ and the eigenvector is constant $\mathbf{v}_1(\mathbf{W}) = 1$. Therefore, as ϵ_{i1} is zero, the limiting distribution of the ratio between the normalized sample and population scores is according to Theorem 1 given as*

$$R_1 \sim \sigma_1\sqrt{\frac{n}{\phi_1(\mathbf{W})}}.$$

	$\sigma_2^2/\sigma_1^2 = 0.5$				$\sigma_2^2/\sigma_1^2 = 0.3$				$\sigma_2^2/\sigma_1^2 = 0.1$			
n	R_1	SD	ϵ_{i1}	SD	R_1	SD	ϵ_{i1}	SD	R_1	SD	ϵ_{i1}	SD
40	0.97	(0.17)	0.02	(0.2)	1.02	(0.12)	0.00	(0.08)	1.029	(0.12)	0.00	(0.02)
80	0.99	(0.09)	0.00	(0.13)	1.01	(0.08)	0.00	(0.05)	1.015	(0.08)	0.00	(0.01)
150	1.00	(0.06)	0.00	(0.09)	1.01	(0.06)	0.00	(0.04)	1.008	(0.06)	0.00	(0.01)
300	1.00	(0.04)	0.00	(0.06)	1.00	(0.04)	0.00	(0.02)	1.004	(0.04)	0.00	(0.01)

Table 1: Mean and standard deviation from 5000 realizations of the distribution of R_1 and the noise ε_{i1} with $p = 6000$ and $\sigma_1^2 = 1$ for different sample size n and values of σ_2^2. For the ε_{i1}, the ratio $z_{i1}/z_{i2} = 1$ is fixed.

	$\sigma_2^2/\sigma_1^2 = 0.5$				$\sigma_2^2/\sigma_1^2 = 0.3$				$\sigma_2^2/\sigma_1^2 = 0.1$			
n	R_2	SD	ϵ_{i2}	SD	R_2	SD	ϵ_{i2}	SD	R_2	SD	ϵ_{i2}	SD
40	1.01	(0.21)	0.00	(0.38)	1.04	(0.12)	0.00	(0.25)	1.04	(0.12)	0.00	(0.19)
80	1.02	(0.09)	0.00	(0.24)	1.02	(0.08)	0.00	(0.17)	1.02	(0.08)	0.00	(0.13)
150	1.01	(0.06)	0.00	(0.17)	1.01	(0.06)	0.00	(0.12)	1.01	(0.06)	0.00	(0.09)
300	1.00	(0.04)	0.00	(0.12)	1.00	(0.04)	0.00	(0.08)	1.01	(0.04)	0.00	(0.06)

Table 2: Mean and standard deviation from 5000 realizations of the distribution of R_2 and the noise ε_{i2} with $p = 6000$ and $\sigma_1^2 = 1$ for different sample size n and values of σ_2^2. For the ε_{i2}, the ratio $z_{i1}/z_{i2} = 1$ is fixed.

If the scores are iid normally distributed, $z_{ij} \sim N(0,1)$, the eigenvalue $\phi_1(\mathbf{W})$ is $\sigma^2\chi_n^2$-distributed and $R_1 \sim \sqrt{n/\chi_n^2}$, which is the same distribution as was found by Shen et al. (2012) in the $\alpha > 1$-case.

Example 3. *For $m = 2$, the ratios between the normalized sample and population scores converge to a limiting distribution of the form $R_j + \varepsilon_{ij}$, where*

$$R_j \sim \sigma_j \sqrt{\frac{n}{\phi_j(\mathbf{W})}}\, v_{jj}(\mathbf{W}) \quad j = 1, 2, \tag{4}$$

and

$$\varepsilon_{i1} \sim \sigma_2 \sqrt{\frac{n}{\phi_1(\mathbf{W})}}\, \frac{z_{i2}}{z_{i1}}\, v_{12}(\mathbf{W}), \quad \varepsilon_{i2} \sim \sigma_1 \sqrt{\frac{n}{\phi_2(\mathbf{W})}}\, \frac{z_{i1}}{z_{i2}}\, v_{21}(\mathbf{W}) \tag{5}$$

The distributions depend on the two eigenvectors $\mathbf{v}_j(\mathbf{W}) = \begin{bmatrix} v_{j1} \\ v_{j2} \end{bmatrix}$ and eigenvalues $\phi_j(\mathbf{W})$, $j = 1, 2,$

of a 2×2 *Wishart distributed matrix*

$$\mathbf{W} \sim W_2\left(\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}, n-1\right).$$

First, we illustrate within Example 3 that ε_{ij} can be considered as noise compared to R_j. This is done by simulating from the distributions in (4) and (5) when the normalized scores are assumed to be standard normally distributed $z_{ij} \sim N(0,1)$. Tables 1 and 2 display the simulated mean and the standard deviation from the distributions of R_1, ε_{i1} and R_2, ε_{i2} respectively for different signal strength ratios $\theta = \sigma_2^2/\sigma_1^2$ and sample sizes n. The tables are shown graphically in Figure 1. We observe from Tables 1 and 2 that the expectation of ε_{ij} is zero, whereas the R_j are expected to be one. When taking the variability into account, we see that for moderate sample sizes (from 80 and upwards), a noise level of two SD is around 15-20 % of R_1, still quite small. For larger sample sizes, the error drops to 2-8 % of R_1. The noise also decreases when the separation between the signals increases. Only when σ_1^2 and σ_2^2 are close to each other and n is small, could ε_{ij} be comparable to R_j. The distribution of the noise is also illustrated graphically in Figure 3 by the 90 % probability contour for each observation.

To better understand the structure of the noise, we explore in Result 2 the distribution of ε_{ij} for large n, a different asymptotic setting then considered earlier.

Result 2. *If* $m = 2$*, the normalized scores are iid normally distributed,* $z_{ij} \sim N(0,1)$*, and the spiked eigenvalues are simple,* $\sigma_1^2 > \cdots > \sigma_m^2$*, the noise* $n^{1/2}\varepsilon_{ij}$ *will be asymptotically normally distributed as* $n \to \infty$*:*

$$n^{1/2}\epsilon_{i1} \xrightarrow{\text{d}} N\left(0, \frac{\theta^2}{(1-\theta)^2}\right), \quad n^{1/2}\epsilon_{i2} \xrightarrow{\text{d}} N\left(0, \frac{1}{\theta^2(1-\theta)^2}\right), \quad \theta = \frac{\sigma_1^2}{\sigma_1^2}.$$

The proof is given in the Appendix.

Three parameters have an effect on the noise distribution; the sample size n, the signal strength σ^2, and the number of components m. From Result 2, we see the role of these parameters for the case of $m = 2$. As the sample size increases, the asymptotic variance of ε_{ij} will decrease, and if we use the standard deviation as a measure of magnitude of the noise, it will decrease as $n^{-1/2}$. This is observed in Figure 1b), which displays graphically the simulated standard deviation of ε_{ij} (as circles) for increasing sample size n together with the asymptotic values (dashed lines). The fit of the simulated standard deviation to the asymptotic scaling of $n^{-1/2}$ necessarily becomes better as n increases.

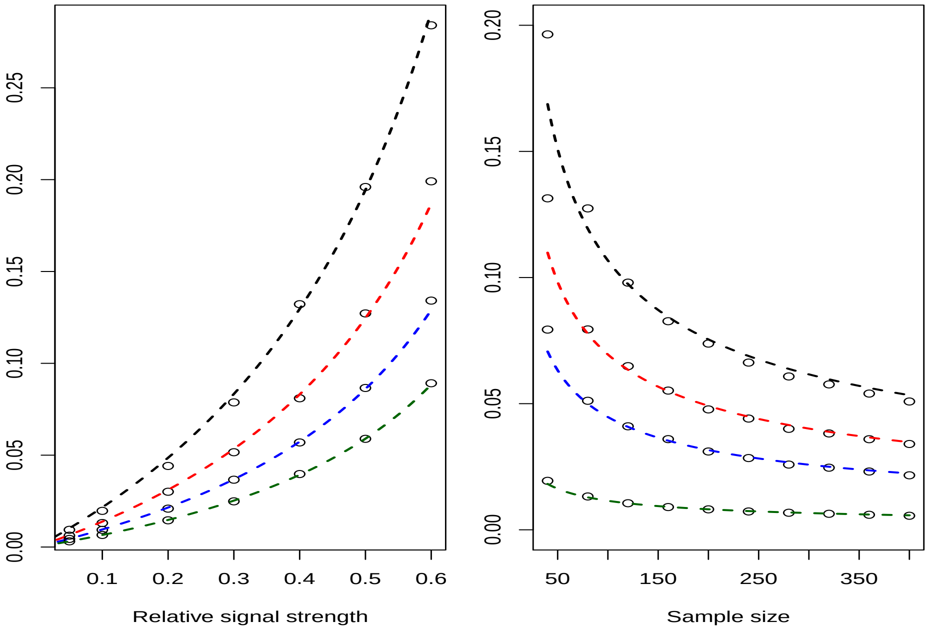

Figure 1: a) The simulated standard deviation of ε_{i1} for $m = 2$ for increasing θ and $n = 40$ (black), $n = 80$ (red), $n = 150$ (blue) and $n = 300$ (green). The dashed lines show the theoretical asymptotic standard deviation. b) The simulated standard deviation of ε_{i1} for $m = 2$ for increasing n and $\theta = 0.5$ (black), $\theta = 0.4$ (red), $\theta = 0.3$ (blue) and $\theta = 0.1$ (green). The dashed lines show the asymptotic standard deviation.

The impact of the two signal strengths σ_1^2 and σ_2^2 is, as shown by Result 2, in terms of the ratio $\theta = \sigma_2^2/\sigma_1^2$. The standard deviation of ε_{ij} scales with the relative signal strength as $\theta/(1-\theta)$ for ε_{i1}. This is seen in Figure 1a), which displays the simulated standard deviation (as circles) for increasing relative signal strength also together with the asymptotic standard deviation (dashed lines). If σ_1^2 and σ_2^2 are close, the variability of the noise increases sharply, which can be interpreted as an overlap or interaction between the signals, making them difficult to distinguish. As observed in Tables 1 and 2, when the strength of the first signal increases relative to the second, the noise decreases.

Because there are only two components, $v_{21}(\mathbf{W})$ and $v_{12}(\mathbf{W})$ will have the same absolute value, but with opposite signs; hence they have a perfect negative correlation and this is seen in Figure 3. The noise for the second component is larger due to the scaling of σ_1, which is reflected by greater extent of the contours in the vertical direction. Also the effect of the score ratios, z_{i1}/z_{i2}, on the

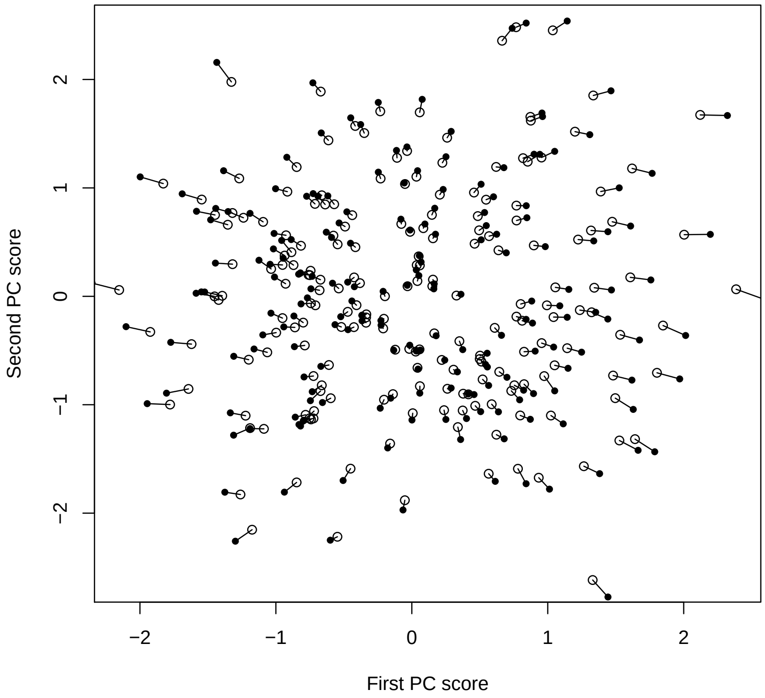

Figure 2: One simulation of the estimated set of first and second principal component scores (filled dots) compared to the population scores (circles) with $p = 6000, n = 200, \sigma_1^2 = 8$ and $\sigma_2^2 = 1$.

noise is seen as the contours are wider closer to the x-axis.

The second key observation is that R_1 and R_2 are independent of i and thereby common to all observations. As they are ratios, they express a common scaling for the scores, which can be seen graphically as a shift, outwards or inwards. The consequence is that the relative positions of sample scores, and thereby most of the visual information, will be consistent with the population scores. We observe this in Figure 2, which displays the sample and population first and second PC scores for one simulated sample. A radial shift, which is not exact due to the noise, is evident when comparing the sample and population scores. Also the fact that R_2 will be generally larger than R_1 can be observed in Figure 2. The score plot will therefore be an appropriate tool to explore the population features, even though eigenvectors and absolute score values are not correctly estimated, asymptotically.

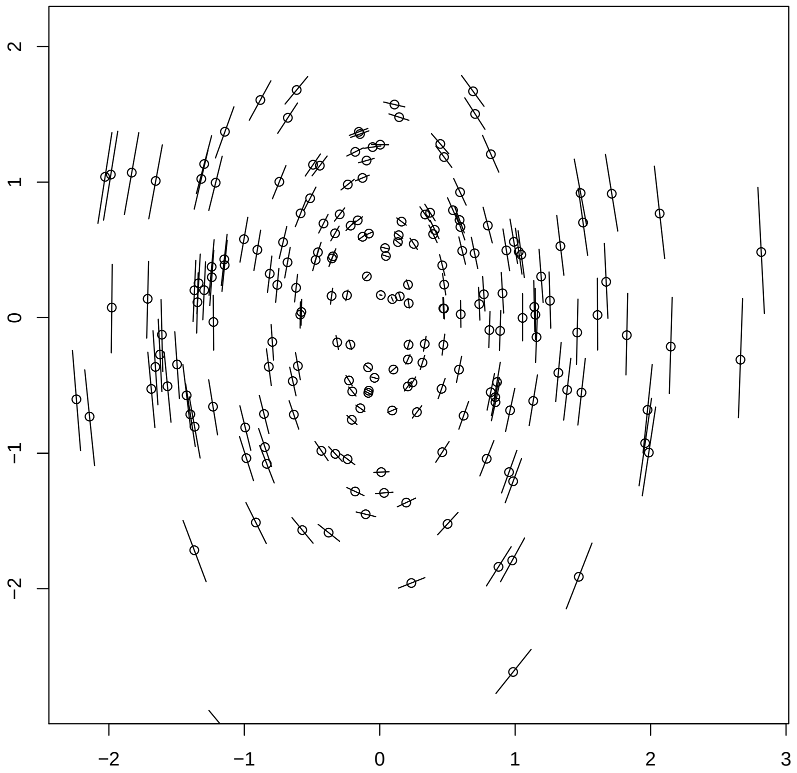

Figure 3: The 90 % probability contour of the distribution of ε_{ij} for the first and second principal component score with $p = 6000, n = 200, \sigma_1^2 = 0.1$ and $\sigma_2^2 = 0.03$.

6 Conclusion

The use of high-dimensional PCA suffers from the somewhat paradoxical situation that theoretically, the eigenvectors and -values are not correctly estimated, but the method is highly successful in practice. The results in this paper attempt to bridge this gap by showing that the relative positions, and thereby the visual content, in a PC score plot are more or less the same for the true and the estimated scores. The assumption is that the leading eigenvalues scale linearly with the dimension. This assumption is fulfilled if the variability is caused by a latent factor with pervasive effects on the variables. This situation is reasonable in genetic markers from different ethnic populations and in microarray expression data from cancer cases and controls.

Future work should consider the implication of these results, when using principal component scores in further analyzes. The same asymptotic framework can be considered for regression, clustering and classification. Especially the effect of the limiting distribution on regression coefficients and misclassification rates would be of interest.

Acknowledgement

The authors will like to thank Elja Arjas for discussions. This work was supported by the Norwegian Cancer Society.

Appendix

The proof of Result 2. If $z_{ij} \sim N(0,1)$ and $\sigma_1^2 > \cdots > \sigma_m^2$, then $\mathbf{W}$ is Wishart distributed

$$\mathbf{W} \sim W_m \left(\mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2), n-1 \right),$$

where $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ have the simple eigenvalues σ_j^2 and the eigenvectors, $\mathbf{e}_j$, the jth unit vectors. Then the result follows from the properties of the sample eigenvalues and eigenvectors given by Muirhead (2009, Corollary 9.4.1):

When $n \to \infty$, the $\mathbf{v}_j(\mathbf{W})$ and $\phi_j(\mathbf{W})$ are asymptotically independent and

$$n^{1/2} \left(\mathbf{v}_j(\mathbf{W}) - \mathbf{e}_j \right)$$

is asymptotically normally distributed with mean 0 and covariance matrix

$$\sum_{i=1, i \neq j}^{m} \frac{\sigma_i^2 \sigma_j^2}{(\sigma_j^2 - \sigma_i^2)} \mathbf{e}_i \mathbf{e}_i^T = \mathrm{diag} \left(\frac{\sigma_j^2 \sigma_1^2}{(\sigma_j^2 - \sigma_1^2)^2}, \ldots, 0, \ldots, \frac{\sigma_j^2 \sigma_m^2}{(\sigma_j^2 - \sigma_m^2)} \right).$$

The jth entry is zero due to not summing over $\mathbf{e}_j \mathbf{e}_j^T$.

For the sample eigenvalues, the $n^{1/2}(\phi_j(\mathbf{W})/n - \sigma_j^2)$ are asymptotically independently distributed as $N(0, 2\sigma_j^4)$. For $m = 2$, the multivariate Delta method gives that the $n^{1/2}\,\varepsilon_{ij}$ are asymptotically normally distributed as

$$n^{1/2}\,\varepsilon_{i1} \xrightarrow{\mathrm{d}} N\left(0, \frac{\sigma_2^4}{(\sigma_1^2 - \sigma_2^2)^2}\right), \quad n^{1/2}\,\varepsilon_{i2} \xrightarrow{\mathrm{d}} N\left(0, \frac{\sigma_1^4}{(\sigma_1^2 - \sigma_2^2)^2}\right).$$

□

References

Ahn, J., J. S. Marron, K. M. Muller, and Y. Chi (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika 94*(3), 760–766.

Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Ann. Statist. 34*(1), 122–148.

Bai, Z. and J. W. Silverman (2010). *Spectral analysis of large dimensional random matrices.* New York: Springer.

Fan, J., Y. Liao, and M. Mincheva (2011). Large covariance estimation by thresholding principal orthogonal complements. *Available at SSRN 1977673*.

Hall, P., J. S. Marron, and A. Neeman (2005). Geometric representation of high dimension, low sample size data. *J. R. Statist. Soc. B 67*(3), 427–444.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist. 29*, 295–327.

Johnstone, I. M. and A. Y. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Statist. Assoc. 104*(486), 682–693.

Jolliffe, I. T. (2002). *Principal component analysis.* New York: Springer.

Jung, S. and J. S. Marron (2009). Pca consistency in high dimension, low sample size context. *Ann. Statist. 37*(6B), 4104–4130.

Jung, S., A. Sen, and J. S. Marron (2012). Boundary behavior in high dimension, low sample size asymptotics of pca. *J. Multivariate Anal.*.

Kondrakhin, Y. V., R. Sharipov, A. E. Kel, and F. A. Kolpakov (2008). Identification of differentially expressed genes by meta-analysis of microarray data on breast cancer. *In Silico Biology 8*(5), 383–411.

Lee, S., F. Zou, and F. A. Wright (2010). Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Statist. 38*(6), 3605.

Leek, J. T. (2011). Asymptotic conditional singular value decomposition for high-dimensional genomic data. *Biometrics 67*(2), 344–352.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, Volume 197. Wiley. com.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statist. Sinica 17*(4), 1617.

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

Shen, D., H. Shen, H. Zhu, and J. Marron (2013). Surprising asymptotic conical structure in critical sample eigen-directions. *arXiv preprint arXiv:1303.6171*.

Shen, D., H. Shen, H. Zhu, and J. S. Marron (2012). High dimensional principal component scores and data visualization. *arXiv preprint arXiv:1211.2679*.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics 10*(3), 515–534.

Yamaguchi-Kabata, Y., K. Nakazono, A. Takahashi, S. Saito, N. Hosono, M. Kubo, Y. Nakamura, and N. Kamatani (2008). Japanese population structure, based on snp genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *The American Journal of Human Genetics 83*(4), 445–456.

Yata, K. and M. Aoshima (2009). Pca consistency for non-gaussian data in high dimension, low sample size context. *Communications in Statistics-Theory and Methods*, 2634–2652.

Yata, K. and M. Aoshima (2012). Effective pca for high-dimension, low-sample-size data with noise reduction via geometric representations. *Journal of multivariate analysis 105*(1), 193–215.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.

Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/sjos.12083

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics. Published by Wiley Publishing Ltd.

The Impact of Measurement Error on Principal Component Analysis

KRISTOFFER HERLAND HELLTON and MAGNE THORESEN

Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo

ABSTRACT. We investigate the effect of measurement error on principal component analysis in the high-dimensional setting. The effects of random, additive errors are characterized by the expectation and variance of the changes in the eigenvalues and eigenvectors. The results show that the impact of uncorrelated measurement error on the principal component scores is mainly in terms of increased variability and not bias. In practice, the error-induced increase in variability is small compared with the original variability for the components corresponding to the largest eigenvalues. This suggests that the impact will be negligible when these component scores are used in classification and regression or for visualizing data. However, the measurement error will contribute to a large variability in component loadings, relative to the loading values, such that interpretation based on the loadings can be difficult. The results are illustrated by simulating additive Gaussian measurement error in microarray expression data from cancer tumours and control tissues.

Key words: eigenvalues, eigenvectors, high-dimensional data, measurement error, microarray data, perturbation theory, principal component analysis

1. Introduction

The last decades have seen an exploding production of complex, high-dimensional data in different fields, from genetics (Li & Xu, 2008) to finance (Fan *et al.*, 2011). Often in these examples the sample size can be quite small compared with the number of measured variables, thus an efficient strategy for dimension reduction is required. Principal component analysis (PCA) is a widely used technique, which reduces the high-dimensional data to a small set of component scores. The component scores can be used for visualization and as input in conventional methods, such as classification, clustering and regression. In practice, the principal components are often thought to represent underlying processes, accounting for the variability in the data, and the component loadings could be interpreted as the relative importance of the different variables in the unobserved processes.

In various high-dimensional data, we find that measurement error in the observed variables can be a severe problem. Examples include measurements of chemical spectra in chemometrics, functional magnetic resonance imaging brain scans or microarray expression data in genomics. In regression models, the presence of measurement error in covariates is known to cause bias in parameter estimates and loss of power to detect significant effects (Carroll *et al.*, 2006; Buonaccorsi, 2009).

To deal with the issue of measurement error in PCA within the setting of microarrays and chemometrics, Sanguinetti *et al.* (2005) and Wentzell & Hou (2012) constructed different variations of PCA where information about the measurement error is incorporated. Wentzell & Hou (2012) (based on Wentzell *et al.* (1997)) constructed a framework for maximum likelihood PCA, which incorporates an assumed known covariance matrix for the measurement error. Sanguinetti *et al.* (2005) extended the probabilistic PCA, which is solved by an expectation–maximization algorithm, to incorporate the technical precision connected to a microarray as a proxy for the measurement error in the data.

However, in practice, it is difficult to estimate the covariance matrix of the measurement error in a high-dimensional situation, and the PCA versions accounting for measurement error are not in common use. Analyses are often carried out naively, running standard PCA on the observed data without any correction for measurement error, and therefore, it will be useful to understand the impact of error on component loadings, scores and selection. In the framework of chemometrics, the effect of measurement error on eigenvalues was investigated by Faber *et al.* (1993, 1995), but only for homogeneous error and not considering the high-dimensional situation.

In this paper, we will derive the bias and variability in loadings and scores caused by a general, additive measurement error. This is performed by considering perturbations of the eigen decomposition, such that the bias and variability of the change in eigenvectors and values are given by the distribution of the errors.

2. Principal component analysis

Principal component analysis reduces the dimensionality of data by finding the low-dimensional linear subspaces where the projections of the data have the largest possible variability. Specifically, given a p-dimensional vector $\mathbf{x}$, the first principal component is a unit-length vector $\mathbf{v}_1 \in \mathcal{R}^p$, such that $\mathbf{v}_1^T\mathbf{x}$ has maximal variance. Because $\mathrm{Var}\,\mathbf{v}_1^T\mathbf{x} = \mathbf{v}_1^T\Sigma\mathbf{v}_1$, where Σ is the population covariance matrix, the first principal component is the eigenvector corresponding to the largest eigenvalue. The second principal component $\mathbf{v}_2$ is the unit-length vector with the largest variance orthogonal to $\mathbf{v}_1$ and is given by the eigenvector corresponding to the second largest eigenvalue and so on.

In practice, the principal components are given by the sample covariance matrix. Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n *iid* p-dimensional vectors and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]$ a $p \times n$ data matrix. When assuming for simplicity that $\mathbf{X}_r, r = 1, \ldots, n$ has a known zero expectation, the sample covariance matrix is $\mathbf{S}_X = \frac{1}{n}\mathbf{X}\mathbf{X}^T$. The principal components are then given by the eigen decomposition of $\mathbf{S}_X$ given by

$$\mathbf{S}_X = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T,$$

with eigenvalues $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ and eigenvectors $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$. The projections, denoted by $\mathbf{Z}_i = \mathbf{v}_i^T\mathbf{X}$ for $i = 1, \ldots, p$, are referred to as the ith component scores and represent the new data, which can be used in further analyses. As this linear projection can be seen as a weighted sum of the original variables, where the eigenvector gives the weight of each variable, the coefficients of the eigenvector are usually referred to as the loadings of the component. The dimensionality can be reduced by choosing the components corresponding to the largest eigenvalues to represent the data.

In a situation with measurement error, an error contaminated version of the data, $\mathbf{W}$, is observed instead of the original data $\mathbf{X}$. For the classical, additive measurement error model, $\mathbf{X}$ is observed by $\mathbf{W}$ with the errors $\sigma\mathbf{U}$, such that

$$\mathbf{W} = \mathbf{X} + \sigma\mathbf{U}.$$

The scaling σ controls the magnitude of the error matrix $\mathbf{U}$. Both the data $\mathbf{X}$ and the error $\mathbf{U}$ are for simplicity assumed to have known zero expectation; hence, the same is true for $\mathbf{W}$. Then the estimator for the population covariance matrix is $\mathbf{S}_W = \frac{1}{n}\mathbf{W}\mathbf{W}^T$, and when the error model is additive, the covariance matrix is given by

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

$$\frac{1}{n}\mathbf{W}\mathbf{W}^T = \frac{1}{n}(\mathbf{X}+\sigma\mathbf{U})(\mathbf{X}+\sigma\mathbf{U})^T = \frac{1}{n}\mathbf{X}\mathbf{X}^T + \frac{\sigma}{n}\mathbf{X}\mathbf{U}^T + \frac{\sigma}{n}\mathbf{U}\mathbf{X}^T + \frac{\sigma^2}{n}\mathbf{U}\mathbf{U}^T.$$

The covariance matrix $\mathbf{S}_W$ is decomposed into the covariance matrix $\mathbf{S}_X$ and the additive change depending on the scaling σ:

$$\mathbf{S}_W = \mathbf{S}_X + \sigma\Delta\mathbf{S}_1 + \sigma^2\Delta\mathbf{S}_2, \tag{1}$$

where $\Delta\mathbf{S}_1 = \frac{1}{n}\mathbf{X}\mathbf{U}^T + \frac{1}{n}\mathbf{U}\mathbf{X}^T$ and $\Delta\mathbf{S}_2 = \frac{1}{n}\mathbf{U}\mathbf{U}^T$.

Our aim is to assess the change in loadings, scores and component selection, when the PCA is carried out on $\mathbf{S}_W$ instead of $\mathbf{S}_X$. The eigenstructure of $\mathbf{S}_X$ is further assumed to be known in the sense that $\mathbf{X}$ is fixed. Then there will be p fixed eigenvectors $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_p]$ and eigenvalues $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

The spiked covariance model introduced by Johnstone (2001) considers the eigenvalues on a population level, where the m first eigenvalues are substantially larger than the remaining $p-m$ eigenvalues, which are all equal to some constant. We assume the eigenvalues to originate from a spiked covariance model on a population level, but fix them as a sample such that the non-zero eigenvalues are necessarily different from each other (Rao *et al.*, 2008). When $p > n$, there must also be at least $p-n$ zero eigenvalues, and we assume for simplicity that exactly $p-n$ eigenvalues are equal to zero, such that the eigenvalues of $\mathbf{\Lambda}$ fulfill the following:

$$\lambda_1 > \cdots > \lambda_m \gg \lambda_{m+1} > \cdots > \lambda_n > \lambda_{n+1} = \cdots = \lambda_p = 0.$$

2.1. *Perturbation problem*

Perturbation theory has been applied in several statistical settings, for instance by Kadane (1970) to investigate the effect of small errors on different estimators and restrictions for overidentification. Nadler (2008) used matrix perturbation theory to develop finite sample approximations for estimates of the leading eigenvalue and eigenvector in a single-spike model. As the sample estimation error in PCA can be modelled as an independent homogeneous measurement error, the results of the current paper will in this special case be similar to the results of Nadler (2008).

Our results have the following outline: First, the Taylor expansion of the eigenvalues and the eigenvectors of $\mathbf{S}_W$ are derived, giving the Taylor expansion of the principal component scores of $\mathbf{S}_W$. Then the expectation and variance of the difference between the eigenvectors and eigenvalues of $\mathbf{S}_W$ and $\mathbf{S}_X$ are derived on the basis of the Taylor expansions. For these results, we condition on the original data matrix $\mathbf{X}$, such that $\mathbf{X}$ represents n fixed realizations from a population distribution. The Taylor expansions of eigenvalues and eigenvectors have earlier been investigated by Wilkinson (1965) and Stewart & Sun (1990) for deterministic matrices in numerical perturbation analysis, whereas Stewart (1990) introduced a stochastic norm, which also allows random error matrices. We denote the ith eigenvalue and vector of $\mathbf{S}_X$ by λ_i and $\mathbf{v}_i$, and the ith eigenvalue and eigenvector of $\mathbf{S}_W$ by $\lambda_{W,i}$ and $\mathbf{v}_{W,i}$.

Lemma 1. *Assuming fixed $p \times n$ matrices* $\mathbf{X}$ *and* $\mathbf{U}$*, the Taylor expansion for the ith eigenvalue of* $\mathbf{S}_W$ *as $\sigma \to 0$ is given by*

$$\lambda_{W,i} = \lambda_i + \sigma\mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_i + \sigma^2\mathbf{v}_i^T\Delta\mathbf{S}_2\mathbf{v}_i + \sigma^2\sum_{j\neq i}\frac{\mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_j\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i}{\lambda_i - \lambda_j} + O(\sigma^3), \tag{2}$$

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

and the Taylor expansion for the ith eigenvector of $\mathbf{S}_W$ *up to a scaling constant as* $\sigma \to 0$ *is given by*

$$\begin{aligned}\mathbf{v}_{W,i} &= \mathbf{v}_i + \sigma \sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j}\, \mathbf{v}_j + \sigma^2 \sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_2 \mathbf{v}_i}{\lambda_i - \lambda_j}\, \mathbf{v}_j \\ &\quad + \sigma^2 \sum_{j\neq i}\sum_{k\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_k}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)}\, \mathbf{v}_j - \sigma^2 \sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i}{(\lambda_i - \lambda_j)^2}\, \mathbf{v}_j + O(\sigma^3),\end{aligned} \tag{3}$$

where $i = 1, \ldots, p$ *when* $p \leq n$ *and* $i = 1, \ldots, n$ *when* $p > n$. *The proof is found in Appendix A.1 of the Supporting Information.*

The first theorem establishes the Taylor expansion of the principal component scores. We denote the ith scores of $\mathbf{S}_X$ by $\mathbf{Z}_i$ and the ith scores of $\mathbf{S}_W$ by $\mathbf{Z}_{W,i}$. As the scores are given by $\mathbf{Z}_{W,i} = \mathbf{v}_{W,i}^T \mathbf{W}$, the result follows from the Taylor expansion of the eigenvectors combined with the observed data matrix $\mathbf{W}$.

Theorem 1. *Assuming fixed* $p \times n$ *matrices* $\mathbf{X}$ *and* $\mathbf{U}$, *the Taylor expansion for the ith component scores of* $\mathbf{S}_W$ *as* $\sigma \to 0$ *is given by*

$$\begin{aligned}\mathbf{Z}_{W,i} &= \mathbf{Z}_i + \sigma \sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j}\, \mathbf{v}_j^T \mathbf{X} + \sigma \mathbf{v}_i^T \mathbf{U} \\ &\quad + \sigma^2 \sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j}\, \mathbf{v}_j^T \mathbf{U} + \sigma^2 \sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_2 \mathbf{v}_i}{\lambda_i - \lambda_j}\, \mathbf{v}_j^T \mathbf{X} \\ &\quad + \sigma^2 \sum_{j\neq i}\sum_{k\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_k}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)}\, \mathbf{v}_j^T \mathbf{X} \\ &\quad - \sigma^2 \sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i}{(\lambda_i - \lambda_j)^2}\, \mathbf{v}_j^T \mathbf{X} + O(\sigma^3),\end{aligned} \tag{4}$$

where $i = 1, \ldots, p$ *when* $p \leq n$ *and* $i = 1, \ldots, n$ *when* $p > n$. *The proof is found in Appendix A.2 of the Supporting Information.*

The change induced by the measurement error can be quantified by the difference between the eigenvalues and eigenvectors of $\mathbf{S}_W$ and $\mathbf{S}_X$, denoted by $\Delta\lambda_i$ and $\Delta\mathbf{v}_i$:

$$\Delta\lambda_i = \lambda_{W,i} - \lambda_i, \quad \Delta\mathbf{v}_i = \mathbf{v}_{W,i} - \mathbf{v}_i. \tag{5}$$

We use the results from Theorem 1 and Lemma 1 to derive the expectation and variability of $\Delta\lambda_i$ and $\Delta\mathbf{v}_i$ under the assumption that $\sigma\mathbf{U}$ is normally distributed, and $\mathbf{X}$ and $\mathbf{U}$ are independent. Then the multivariate additive measurement error model for n samples $\mathbf{W} = [\mathbf{W}_1, \cdots, \mathbf{W}_n]$ is given by

$$\mathbf{W}_r = \mathbf{X}_r + \sigma\mathbf{U}_r, \quad \mathbf{U}_r \sim \mathcal{N}(\mathbf{0}, \Sigma_U), \quad r = 1, \ldots, n.$$

The covariance matrix of the error $\sigma\mathbf{U}_r$ is given as $\mathrm{Var}\,(\sigma\mathbf{U}_r) = \sigma^2\Sigma_U$, such that σ^2 controls the scaling of the variance. The expectation of $\Delta\lambda_i$ and $\Delta\mathbf{v}_i$ is the bias in $\lambda_{W,i}$ and $\mathbf{v}_{W,i}$.

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

Theorem 2 (Eigenvalues and eigenvectors). *Assume* $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_n]$ *to be independent and identically, normally distributed,* $\mathbf{U}_r \sim \mathcal{N}(\mathbf{0}, \Sigma_U)$ *for* $r = 1, \ldots, n$. *Then the expectation and variance of* $\Delta\lambda_i$ *as* $\sigma \to 0$, *conditional on* $\mathbf{X}$, *are given by*

$$\mathbb{E}\left(\Delta\lambda_i \mid \mathbf{X}\right) = \sigma^2 \mathbf{v}_i^T \Sigma_U \mathbf{v}_i + \frac{\sigma^2}{n} \sum_{j \neq i} \frac{\lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_j + \lambda_j \mathbf{v}_i^T \Sigma_U \mathbf{v}_i}{\lambda_i - \lambda_j} + O(\sigma^3), \tag{6}$$

$$\mathrm{Var}\left(\Delta\lambda_i \mid \mathbf{X}\right) = \frac{4\lambda_i \sigma^2}{n} \mathbf{v}_i^T \Sigma_U \mathbf{v}_i + O(\sigma^3). \tag{7}$$

The expectation of $\Delta\mathbf{v}_i$ *as* $\sigma \to 0$, *conditional on* $\mathbf{X}$, *is given by*

$$\begin{aligned}\mathbb{E}\left(\Delta\mathbf{v}_i \mid \mathbf{X}\right) = {} & \sigma^2 \sum_{j \neq i} \frac{\mathbf{v}_j^T \Sigma_U \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j - \frac{\sigma^2}{n} \sum_{j \neq i} \frac{2\mathbf{v}_j^T \Sigma_U \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j \\ & + \frac{\sigma^2}{n} \sum_{j \neq i} \sum_{k \neq i, j} \frac{\lambda_j \mathbf{v}_k^T \Sigma_U \mathbf{v}_i}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)} \mathbf{v}_j + O(\sigma^3),\end{aligned} \tag{8}$$

and the variance of the k*th coefficient of* $\Delta\mathbf{v}_i$ *is given by*

$$\begin{aligned}\mathrm{Var}\left(\Delta\mathbf{v}_{ik} \mid \mathbf{X}\right) = {} & \frac{\sigma^2}{n} \sum_{j \neq i} \frac{\lambda_j \mathbf{v}_i^T \Sigma_U \mathbf{v}_i + \lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_j}{(\lambda_i - \lambda_j)^2} \mathbf{v}_{jk}^2 \\ & + \frac{\sigma^2}{n} \sum_{j,l \neq i, j < l} \frac{2\lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_l}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_l)} \mathbf{v}_{jk} \mathbf{v}_{lk} + O(\sigma^3),\end{aligned} \tag{9}$$

where $i = 1, \ldots, p$ *when* $p \leq n$ *and* $i = 1, \ldots, n$ *when* $p > n$. *The proof is found in Appendix B.2 of the Supporting Information.*

The variance of $\Delta\mathbf{v}_{ik}$ is, to leading order, a weighted sum over the kth coordinate of all other eigenvectors, where the weights depend on the data and the error structure through the eigenvalues and the covariance matrix of the error.

Theorem 3 (Scores). *Assume* $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_n]$ *be independent and identically, normally distributed,* $\mathbf{U}_r \sim \mathcal{N}(\mathbf{0}, \Sigma_U)$ *for* $r = 1, \ldots, n$. *Then the expectation of* $\Delta\mathbf{Z}_i = \mathbf{Z}_{W,i} - \mathbf{Z}_i$ *as* $\sigma \to 0$, *conditional on* $\mathbf{X}$, *is given by*

$$\begin{aligned}\mathbb{E}\left(\Delta\mathbf{Z}_i \mid \mathbf{X}\right) = {} & \sigma^2 \sum_{j \neq i} \frac{\mathbf{v}_j^T \Sigma_U \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j^T \mathbf{X} + \frac{\sigma^2}{n} \sum_{j \neq i} \sum_{k \neq i, j} \frac{\lambda_j \mathbf{v}_k^T \Sigma_U \mathbf{v}_i}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)} \mathbf{v}_j^T \mathbf{X} \\ & - \frac{\sigma^2}{n} \sum_{j \neq i} \frac{2\mathbf{v}_j^T \Sigma_U \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j^T \mathbf{X} + \frac{\sigma^2}{n} \sum_{j \neq i} \frac{\mathbf{v}_j^T \Sigma_U \mathbf{v}_i \mathbf{v}_j^T \mathbf{X} + \mathbf{v}_j^T \Sigma_U \mathbf{v}_i \mathbf{v}_i^T \mathbf{X}}{\lambda_i - \lambda_j} \\ & + O(\sigma^3).\end{aligned} \tag{10}$$

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

The variance of the k*th coefficient of* $\Delta \mathbf{Z}_i$ *as* $\sigma \to 0$*, conditional on* $\mathbf{X}$*, is given by*

$$
\begin{aligned}
Var\,(\Delta \mathbf{Z}_{ik} \mid \mathbf{X}) =& \frac{\sigma^2}{n} \sum_{j \neq i} \frac{\lambda_j \mathbf{v}_i^T \Sigma_U \mathbf{v}_i + \lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_j}{(\lambda_i - \lambda_j)^2} \left(\mathbf{v}_j^T \mathbf{X}_k\right)^2 \\
&+ \frac{\sigma^2}{n} \sum_{j,l \neq i, j<l} \frac{2\lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_l \mathbf{v}_j^T \mathbf{X}_k \mathbf{v}_l^T \mathbf{X}_k}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_l)} + \sigma^2 \mathbf{v}_i^T \Sigma_U \mathbf{v}_i \\
&+ \frac{2\sigma^2}{n} \sum_{j \neq i} \frac{\mathbf{v}_j^T \Sigma_U \mathbf{v}_i \mathbf{v}_i^T \mathbf{X}_k + \mathbf{v}_i^T \Sigma_U \mathbf{v}_i \mathbf{v}_j^T \mathbf{X}_k}{\lambda_i - \lambda_j} \mathbf{v}_j \mathbf{X}_k + O(\sigma^3),
\end{aligned} \tag{11}
$$

where $i = 1, \ldots, p$ *when* $p \leq n$ *and* $i = 1, \ldots, n$ *when* $p > n$*. The proof is found in Appendix B.3 of the Supporting Information.*

Remark 1. If the measurement error is uncorrelated and homogeneous, such that $\mathbf{U}_r \sim N(0, \sigma^2 I)$ for $r = 1, \ldots, n$, the bias in the eigenvalues and eigenvectors simplifies. Because

$$\mathbb{E}\,(\mathbf{S}_W \mid \mathbf{X}) = \mathbf{S}_X + \sigma^2 I,$$

the expectation of the eigenvalues and eigenvectors of $\mathbf{S}_W$ are given as $\mathbb{E}\,(\lambda_{W,i} \mid \mathbf{X}) = \lambda_i + \sigma^2$ and $\mathbb{E}\,(\mathbf{v}_{W,i} \mid \mathbf{X}) = \mathbf{v}_i$, such that the bias is given exactly as

$$\mathbb{E}\,(\Delta\lambda_i \mid \mathbf{X}) = \sigma^2, \qquad \mathbb{E}\,(\Delta \mathbf{v}_i \mid \mathbf{X}) = \mathbf{0}.$$

3. Implications

We will now explore the implications of Theorems 2 and 3 for the loadings, scores and component selection, when the measurement error is assumed to be uncorrelated and either homogeneous or heterogeneous. In the case of uncorrelated, homogeneous measurement error, the variance of $\mathbf{U}_r$ is equal for all variables, such that $\Sigma_U = I$. Then the covariance matrix of the error is given as $\text{Var}\,(\sigma \mathbf{U}_r) = \sigma^2 I$. In the case of uncorrelated, heterogeneous measurement error, the variance of $\mathbf{U}_r$ is different in each variable, such that $\Sigma_U = \text{diag}(c_1, \ldots, c_p)$, where the constants c_k give the relative size of the variances. Then the covariance matrix of the error is $\text{Var}\,(\sigma \mathbf{U}_r) = \text{diag}(\sigma^2 c_1, \ldots, \sigma^2 c_p)$, where σ^2 controls the scaling.

A key element in the bias and variance expressions of Theorems 2 and 3 is the quantity $\mathbf{v}_j^T \Sigma_U \mathbf{v}_i$, which captures the relationship between the error and the original data. For $j = i$, this corresponds to a projection of the error covariance matrix Σ_U onto the eigenvector space spanned by $\mathbf{v}_i$. For uncorrelated, homogeneous error, the projection of Σ_U is either $\mathbf{v}_i^T \Sigma_U \mathbf{v}_i = 1$ or $\mathbf{v}_j^T \Sigma_U \mathbf{v}_i = 0$ for $j \neq i$, which simplify the bias and variance expressions. For uncorrelated, heterogeneous error with covariance matrix $\Sigma_U = \text{diag}(c_1, \ldots, c_p)$, the projections are given as weighted sums, $\mathbf{v}_i^T \Sigma_U \mathbf{v}_i = \sum_{k=1}^p c_k \mathbf{v}_{ik}^2$ and $\mathbf{v}_j^T \Sigma_U \mathbf{v}_i = \sum_{k=1}^p c_k \mathbf{v}_{jk} \mathbf{v}_{ik}$, where the variances are weighted by the corresponding loadings. Because the sum of the weights $\mathbf{v}_{ik}^2$ are normalized to 1, the $\mathbf{v}_i^T \Sigma_U \mathbf{v}_i$ will be a weighted average of the error variances in the heterogeneous case.

3.1. Loadings

The impact of measurement error on the principal component loadings can be assessed through the bias and variance in the eigenvectors. If the measurement error structure is uncorrelated and homogeneous, the bias in the loadings will be zero due to the orthogonality of the eigenvectors. However, heterogeneous error or error structures with dependencies will introduce a bias.

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

We can illustrate this effect through a simple heterogeneous structure with measurement error in only one variable, $\Sigma_U = \mathrm{diag}(1, 0, \ldots, 0)$. The bias in the first loading of the ith component is given by

$$\mathbb{E}\left(\Delta \mathbf{v}_{i1} \mid \mathbf{X}\right) = \sigma^2 \mathbf{v}_{i1} \left(1 - \frac{2}{n}\right) \sum_{j \neq i} \frac{\mathbf{v}_{ij}^2}{\lambda_i - \lambda_j} + \frac{\sigma^2 \mathbf{v}_{i1}}{n} \sum_{j \neq i} \sum_{k \neq i, j} \frac{\lambda_j \mathbf{v}_{ij} \mathbf{v}_{ik}}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)} + O(\sigma^3). \tag{12}$$

As $p > n$, we have assumed the fixed eigenvalues to be zero for $j = n + 1, \ldots, p$. If p is much larger than n, the first two sums in (12) are approximated by $(1/\lambda_i) \sum_{j \neq i} \mathbf{v}_{ij}^2$ and due to the unit length of the eigenvectors, we have $\sum_{j \neq i} \mathbf{v}_{ij}^2 \simeq 1$, such that expression (12) is approximated by

$$\mathbb{E}\left(\Delta \mathbf{v}_{i1} \mid \mathbf{X}\right) \simeq \frac{\sigma^2}{\lambda_i} \mathbf{v}_{i1}.$$

The bias in the first loading in this simplified model depends, to leading order, on the loading value itself, thus the larger loadings have larger bias. It also depends on σ^2/λ_i, the ratio between the variance of the error $\sigma \mathbf{U}$ and the ith eigenvalue of $\mathbf{S}_X$. This ratio expresses an inverse signal-to-noise relationship, as the eigenvalues represent the overall structure or signal in the data. When the eigenvalues are large compared with the error variance, the inverse signal-to-noise ratio is close to zero, resulting in a very small bias in the eigenvector coefficients. The ratio σ^2/λ_i is always positive, such that the loading is overestimated in absolute value and thereby also the importance of the variable in question. This is natural as PCA is constructed to interpret high variability as important structure. As errors increase variability, the variables affected by error will erroneously be assigned an increased importance. For a general uncorrelated, heterogeneous error structure, that is, $\Sigma_U = \mathrm{diag}(c_1, \ldots, c_p)$, the bias will depend on whether the corresponding error variance $\sigma^2 c_k$ is smaller or larger than the average error variance over all variables:

$$\mathbb{E}\left(\Delta \mathbf{v}_{ik} \mid \mathbf{X}\right) \simeq \frac{\sigma^2 (c_k - \bar{c})}{\lambda_i} \mathbf{v}_{ik}.$$

where $\bar{c} = (1/p) \sum_{i=1}^{p} c_p$ is the mean of individual variances.

The induced variation in a loading is characterized by $\mathrm{Var}\left(\Delta \mathbf{v}_{ik} \mid \mathbf{X}\right)$ in (9). If the error is uncorrelated and homogeneous, the variance is given by

$$\mathrm{Var}\left(\Delta \mathbf{v}_{ik} \mid \mathbf{X}\right) = \frac{\sigma^2}{n} \sum_{j \neq i} \frac{\lambda_i + \lambda_j}{(\lambda_i - \lambda_j)^2} \mathbf{v}_{jk}^2 + O(\sigma^3). \tag{13}$$

The variance is, to leading order, a weighted sum of the eigenvalues, where the weights $\mathbf{v}_{jk}^2$ are the kth square coefficients of all other eigenvectors, and this makes it difficult to assess the magnitude of the variance. But due to the unit length of $\mathbf{v}_i$, the mean value of $\mathbf{v}_{jk}^2$ over the jth component is $1/p$, such that we have approximately $\sum_{j=1, j \neq i}^{p} \mathbf{v}_{jk}^2 \simeq 1$ for large p. When $p \gg n$ and most eigenvalues are zero, the sum in (13) can be approximated $\sum_{j \neq i} \frac{\lambda_i + \lambda_j}{(\lambda_i - \lambda_j)^2} \mathbf{v}_{jk}^2 \simeq (1/\lambda_i) \sum_{j \neq i} \mathbf{v}_{jk}^2$. We can therefore approximate the variability in each loading of the ith component by

$$\mathrm{Var}\left(\Delta \mathbf{v}_{ik} \mid \mathbf{X}\right) \simeq \frac{\sigma^2}{\lambda_i} \frac{1}{n}.$$

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

The variability in the loadings within the same component will therefore be of the same magnitude, and the variation should be seen relative to the loading value. From the example presented in Section 4, we will see that the variation will be small compared with the largest loadings, but large enough to be problematic for the average or small loadings. The large variability around the true value induced by the error may cause an interpretation based on the loadings to be incorrect.

3.2. Scores

The projections of the original data onto the eigenvector space, the component scores, are often used in other types of analyses, such that the measurement error is propagated further. In the case of uncorrelated and homogeneous error, $\Sigma_U = I$, the bias in the scores will, to leading order, be 0,

$$\mathbb{E}\left(\Delta \mathbf{Z}_{ik} \mid \mathbf{X}\right) = O(\sigma^3),$$

whereas the variance in the kth score of the ith component is given by

$$\mathrm{Var}\left(\Delta \mathbf{Z}_{ik} \mid \mathbf{X}\right) = \sigma^2 + \frac{\sigma^2}{n} \sum_{j \neq i} \frac{3\lambda_i - \lambda_j}{(\lambda_i - \lambda_j)^2} \left(\mathbf{v}_j^T \mathbf{X}_k\right)^2 + O(\sigma^3),$$

by collecting the second and last term in (11). The first term in the variance expression is the largest, such that the variance in the scores is mainly given by the error term $\sigma \mathbf{U}$. It is however difficult to assess the contribution of the second term without the specified scores. The induced variability in the scores can be compared with the error-induced variability in the observed data $\mathbf{W}$, which is given by $\mathrm{Var}\left(\mathbf{U}_{rk} \mid \mathbf{X}\right) = \sigma^2$. We see that the error variance in the scores is larger, due to the erroneously estimated eigenvectors.

It is also possible to quantify the impact of the error in terms of the overall variance of the scores, as this is given by the eigenvalues $\mathrm{Var}\left(\mathbf{Z}\right) = \Lambda$. The difference in the overall score variability is given by the bias in the eigenvalues, which for a homogeneous error is given by

$$\mathbb{E}\left(\Delta \lambda_i \mid \mathbf{X}\right) = \sigma^2 + \frac{\sigma^2}{n} \sum_{j \neq i} \frac{\lambda_i + \lambda_j}{\lambda_i - \lambda_j} + O(\sigma^3).$$

This expression can, when $p \gg n$, by approximated by $\mathbb{E}\left(\Delta \lambda_i \mid \mathbf{X}\right) \simeq \sigma^2 \left(1 + \frac{p}{n}\right)$, in the case of uncorrelated and homogeneous error. To assess the relative increase in the variance of the component scores, we compare the bias in the eigenvalues to the original eigenvalues, λ_i. If the eigenvalues are large, the relative increase in variability introduced by the error will be small.

3.3. Component selection

Dimension reduction can be achieved by selecting a subset of the components with the largest eigenvalues. Ferré (1995) performed an extensive comparison of different selection methods and concluded that there is no ideal selection criterion. However, the criteria most often used in practice, the percentage rule, the Kaiser rule and Scree plot, all specify a cut-off based on the eigenvalues, where only the components corresponding to the eigenvalues previously the cut-off value are kept. Our aim is to look into the effects of measurement error on these commonly used criteria. It should be mentioned that more recent work on component selection in situations with $p \gg n$ exits (Kritchman & Nadler, 2008).

According to the percentage rule, the chosen components will explain a specified proportion of the total data variability. Because the eigenvalues give the variance of the components, the

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

proportion is given by the sum of the eigenvalues of the chosen components divided by the sum of all eigenvalues. As the bias in the eigenvalues is approximately equal when $p \gg n$, the relative difference between the large and the small eigenvalues becomes smaller, such that additional components are needed to explain the same proportion of the variability. The eigenvalues of the additional components must outweigh the difference between the sum of the bias in the chosen eigenvalues and the sum of the bias in all eigenvalues. The fact that additional components must be chosen is a result of the error obscuring the original data structure.

With the Kaiser rule, the cut-off is the mean of the eigenvalues $\bar{\lambda}$ (Jolliffe, 2002). Simulations show that too few variables will be selected under this rule, and Jolliffe (2002) suggested a modified Kaiser rule with $0.7\bar{\lambda}$ as the cut-off. The Kaiser rule will, as opposed to the percentage rule, adapt to the introduced bias. If the bias is approximately equal in all eigenvalues, the increase in $\bar{\lambda}$ will be the same as in the individual eigenvalues, such that the number of components over the cut-off value remains the same.

A Scree plot is a graphical procedure to determine a cut-off value, where the eigenvalues are plotted in decreasing order. The cut-off is set where the slope of the eigenvalues shifts from steep to shallow (Jolliffe, 2002), and the components above this break point are retained. As the bias in the eigenvalues is approximately equal, the break point should not appear to move, but a graphical procedure can be difficult to assess.

4. Example – microarray expression data

We illustrate our results with microarray expression data from lung cancer patients available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-GEOD-10072. The data set consists of 107 samples in total with 58 adenocarcinoma tumor tissue samples and 49 non-tumor samples. In all samples, 22,284 genes are analysed using a HG-U133A Affymetrix GeneChip (Affymetrix, Santa Clara, CA, USA).

Research into measurement error in microarray expression data has suggested a combination of additive and multiplicative errors (Rocke & Durbin, 2001; Karakach & Wentzell, 2007). For the purpose of illustration, we will only assume an additive measurement error. With the Affymetrix chip technology, it is possible to use probe information to estimate the technical uncertainty in expression values, for instance by the Bayesian Gene Expression (BGX) methodology (Hein *et al.*, 2005; Turro *et al.*, 2007). BGX uses Bayesian hierarchical models to produce *a posteriori* distributions of the gene expressions by utilizing probe information. The probe set in an Affymetrix GeneChip consists of 11–20 probe pairs of perfect match probes and mismatch probes, which accounts for different sources of noise (Hein *et al.*, 2005). The method supplies *a posteriori* distributions for two parameters, the gene expression μ_k and the technical variability σ_k^2 for each sample.

We use the mean of the *a posteriori* distribution of μ_k as an estimate of the kth gene expression, and we use the mean of the distribution of σ_k^2 as an estimate of the gene-specific and sample-specific technical variance. However, we assume the technical variability to be equal for each sample and use the mean over all samples as our estimate of the measurement error.

The R package for BGX is highly labor-intensive, and our analysis is restricted to the 3000 genes with the highest variance. The estimated gene expression is seen as the original data, and the measurement error structure is assumed to be uncorrelated and heterogeneous with variance equal to the mean of the estimated technical variability over samples. To illustrate the effects of measurement error, we add a simulated Gaussian error to the data, and the principal components of the original and the error-prone data are compared. The robustness of the component loadings against error is explored with the aim of biological interpretation,

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

whereas the robustness of component scores is explored with the aim of classification and logistic regression.

The simulated additive measurement error is assumed to be normally distributed, with an uncorrelated and heterogeneous variance structure given by

$$\mathbf{W} = \mathbf{X} + \mathbf{U}, \quad U_{rk} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_k^2\right), \quad r = 1, \dots, n, k = 1, \dots, p,$$

where the error variance σ_k^2 is the estimate supplied by the BGX methodology.

Table 1 displays the 15 genes corresponding to the largest loadings (in absolute value) in the first principal component. The second column displays the original loadings, and the third column displays the difference in the loadings, when the simulated error is added to the data. The fourth and fifth columns display the theoretical bias and standard deviation in each loading. The last two columns display the genetic variance in the original data, and the ratio between the measurement variance estimated by BGX and the genetic variance. The last column therefore shows the degree of uncertainty in the measurements.

We observe that changes in the loadings in Table 1 are much larger than the theoretical bias, and this is due to the variability in the scores. The theoretical variability, in terms of the standard deviation, is substantially larger than the theoretical bias, as seen in Table 1. This illustrates that, when focusing on the loadings, the main impact of uncorrelated errors is increased variability and not bias. Biologically, the loadings can be interpreted as the relative importance of each gene in the underlying processes represented by the component, and the random fluctuations in the loading values can undermine the biological interpretation.

The impact of measurement error on the scores is illustrated graphically in Fig. 1, which displays the first and second principal component of the original data and the data with a simulated, additive error. An arrow indicates the change in scores, when the simulated error is introduced. We observe that the changes are very small compared with the overall positions of the scores, and this is due to the large first and second eigenvalues, $\lambda_1 = 1255.01$ and $\lambda_2 = 969.95$. The variance of the error ranges from 0.03 to 1.90 with a mean of 0.86. Even though the error variance can be substantial compared with the genetic variability, it is very

Table 1. *The 15 genes corresponding to the largest coefficients in absolute value in the first eigenvector in decreasing order, together with the difference induced by the simulated error, the theoretical bias, the theoretical standard deviations, the variance in the variable and the ratio between the variance of measurement error and variable variance*

Gene annotation	$\mathbf{v}_{1k}$	$\Delta\mathbf{v}_{1k}$	Bias	St. dev.	σ_X^2	σ_U^2/σ_X^2
214387_x_at	−0.0545	0.0008	0.000011	0.0026	8.361	0.05
205982_x_at	−0.0539	−0.0001	0.000012	0.0025	7.830	0.05
211735_x_at	−0.0531	0.0038	0.000013	0.0026	8.030	0.04
209612_s_at	−0.0528	0.0026	0.000001	0.0026	4.839	0.15
219230_at	−0.0502	0.0048	−0.000010	0.0026	4.912	0.19
209074_s_at	−0.0501	0.0064	−0.000013	0.0025	5.276	0.18
209613_s_at	−0.0498	−0.0029	−0.000006	0.0026	4.757	0.18
203980_at	−0.0496	0.0043	−0.000013	0.0026	5.397	0.18
205200_at	−0.0490	0.0006	−0.000007	0.0025	4.715	0.18
213317_at	−0.0477	0.0048	−0.000004	0.0026	4.420	0.18
204719_at	−0.0476	0.0012	−0.000011	0.0025	3.755	0.26
215454_x_at	−0.0474	−0.0009	−0.000001	0.0026	5.463	0.12
209763_at	−0.0469	0.0029	−0.000001	0.0025	3.659	0.22
212713_at	−0.0468	0.0012	−0.000008	0.0025	3.905	0.23
206488_s_at	−0.0463	−0.0002	−0.000012	0.0025	3.773	0.27

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

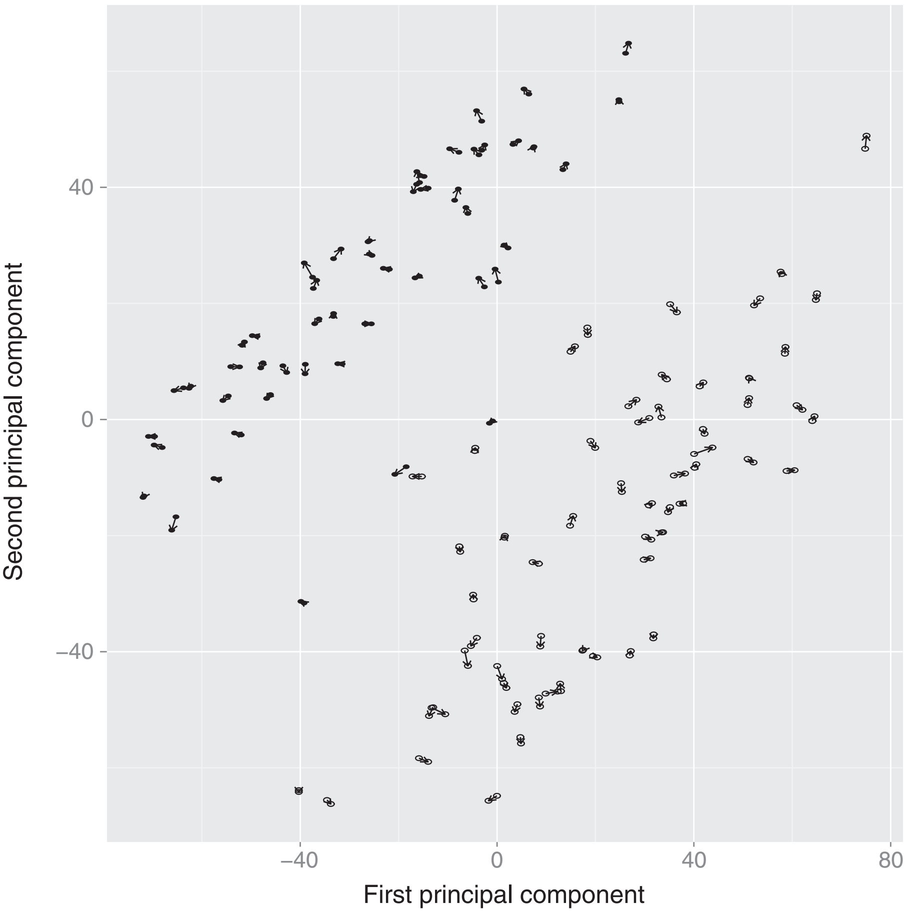

Fig. 1. Plot of first and second component scores from original data and data with simulated error based on the estimated error structure. An arrow indicates the change in scores from the original value. Black dots indicate adenocarcinoma tumour tissue, and open circles indicate non-tumour tissues.

small compared with the first two eigenvalues, and the relative impact of the error on the scores is determined by the ratio between the error variance and the eigenvalue. The plot of the first two components can be used to classify the tissues by cancer status, adenocarcinoma tumour or non-tumour, which are indicated by black and open circles, respectively. The arrows illustrate that both groups experience a slightly increased variability. This is only a problem for classification if the change causes the groups to overlap, but this does not occur in our example. The key point is the small relative change in the overall positions of the scores.

The classification can also be performed by logistic regression, where measurement error will often cause attenuation in estimated regression coefficients (Carroll *et al.*, 2006; Buonaccorsi, 2009). We illustrate this effect by using the first component scores $\mathbf{Z}_1$ without and with error in a logistic regression. For logistic regression, we assume

$$y_i \sim \text{Bernoulli}(p_i), \quad \text{logit}(p_i) = \beta_0 + \beta_1 \mathbf{Z}_1.$$

The binary outcome y_i is the cancer status, lung cancer or normal tissue. The estimated coefficients from the logistic regression based on the scores from the original data are $\hat{\beta}_{0,\mathbf{Z}_1} = -0.292$ and $\hat{\beta}_{1,\mathbf{Z}_1} = -7.224 \times 10^{-2}$, whereas the coefficients based on the scores from the data with error are $\hat{\beta}_{0,\mathbf{Z}_{W,1}} = -0.287$ and $\hat{\beta}_{1,\mathbf{Z}_{W,1}} = -7.161 \times 10^{-2}$. There is a slight underestimation of the slope coefficient, consistent with the well-known attenuation effect. The increased variability in the component scores causes the estimated slope $\beta_{1,\mathbf{Z}_{W,1}}$ to decrease in

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

absolute value. The attenuation factor gives the expected decrease as $\beta_{1,\mathbf{Z}_{W,1}} = \psi\beta_{1,\mathbf{Z}_1}$ with $\psi = \mathrm{Var}\,\mathbf{Z}_1/\mathrm{Var}\,\mathbf{Z}_{W,1}$, (Carroll *et al.*, 2006). As the variances of the scores are the eigenvalues, the factor is approximately $\lambda_1/\lambda_{1,W} = 0.981$ in our data, consistent with the effect seen in $\beta_{1,\mathbf{Z}_{1,W}}$.

5. Discussion

Our aim is to understand the effect of measurement error on PCA, motivated by applications in high-dimensional error-prone data. The impact of the error is characterized by the bias and variance of eigenvalues and eigenvectors based on second-order Taylor approximations. The results are given for additive errors with a general covariance matrix, such that also measurement error with a correlation structure beyond the uncorrelated case can be explored. It has been shown that the impact of uncorrelated errors on component scores will mainly be in terms of an increased variability. We have quantified the impact of the additive measurement error based on a small error assumption. In practice, what we need for the theory to work is that σ^2 is small relative to the eigenvalues. As shown in the example, this will often be the case, even if there is substantial measurement error relative to the variation in the data themselves. In the setting of microarray data, where the first eigenvalues can be substantially larger than the error variance, the relative impact of the error variability will be negligible. This suggests that the additive measurement error might be unproblematic in microarrays, when dealing only with the components corresponding to the largest eigenvalues, for instance, in the case of data visualization. However, the measurement error will also cause an increased variability in the loadings, which can be large relative to the loading values and thereby undermine their interpretation.

For the specific application of microarray data, the effects of multiplicative error should also be investigated, as Rocke & Durbin (2001) and Karakach & Wentzell (2007) suggest that the appropriate measurement error model for microarrays is a combination of additive and multiplicative errors.

Because our aim is to understand the direct impact of measurement error, we condition on the data **X**, fixing the model error. However, recent results raise issues regarding the consistent estimation of the population structure by PCA in the high-dimensional setting. Johnstone & Lu (2009), among others have shown that eigenvalue and eigenvector estimates are not asymptotically consistent when $p \gg n$, and they have introduced the asymptotically consistent sparse PCA methodology. Therefore, it remains an open question if the inconsistency may be a more severe problem than measurement error.

Acknowledgements

The authors wish to thank the reviewers, whose suggestions were very helpful in improving the presentation of this work. This work was supported by the Norwegian Cancer Society, (Project ID 744088).

References

Buonaccorsi, J. P. (2009). *Measurement error: models, methods, and applications*, Chapman and Hall, London.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*, CRC Press, London.

Faber, N. M., Buydens, L. M. C. & Kateman, G. (1993). Standard errors in the eigenvalues of a cross-product matrix: Theory and applications. *J. Chemom.* **7**, 495–526.

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

Faber, N. M., Meinders, M. J., Geladi, P., Sjöström, M., Buydens, L. M. C. & Kateman, G. (1995). Random error bias in principal component analysis. Part I. Derivation of theoretical predictions. *Anal. Chim. Acta* **304**, 257–271.

Fan, J., Lv, J. & Qi, L. (2011). Sparse high dimensional models in economics. *Annu. Rev. Econ.* **3**, 291–317.

Ferré, L. (1995). Selection of components in principal component analysis: a comparison of methods. *Comput. Stat. Data Anal.* **19**, 669–682.

Hein, A. M. K., Richardson, S., Causton, H. C., Ambler, G. K. & Green, P. J. (2005). Bgx: a fully bayesian integrated approach to the analysis of affymetrix genechip data. *Biostatistics* **6**, 349–373.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**, 295–327.

Johnstone, I. M. & Lu, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Am. Stat. Assoc.* **104**, 682–693.

Jolliffe, I. T. (2002). *Principal component analysis*, Springer, New York.

Kadane, J. B. (1970). Testing overidentifying restrictions when the disturbances are small. *J. Am. Stat. Assoc.* **65**, 182–185.

Karakach, T. K. & Wentzell, P. D. (2007). Methods for estimating and mitigating errors in spotted, dual-color dna microarrays. *Omics: J. Integr. Biol.* **11**, 186–199.

Kritchman, S. & Nadler, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemom. Intell. Lab. Syst.* **94**, (1), 19–32.

Li, X. & Xu, R. (2008). *High-dimensional data analysis in cancer research*, Springer, New York.

Nadler, B. (2008). Finite sample approximation results for principal component analysis: A matrix perturbation approach. *Ann. Stat.* **36**, 2791–2817.

Rao, N. R., Mingo, J. A., Speicher, R. & Edelman, A. (2008). Statistical eigen-inference from large wishart matrices. *Ann. Stat.* **36**, 2850–2885.

Rocke, D. M. & Durbin, B. (2001). A model for measurement error for gene expression arrays. *J. Comput. Biol.* **8**, 557–569.

Sanguinetti, G., Milo, M., Rattray, M. & Lawrence, N. D. (2005). Accounting for probe-level noise in principal component analysis of microarray data. *Bioinformatics* **21**, 3748–3754.

Stewart, G. W. (1990). Stochastic perturbation theory. *SIAM Rev.* **32**, 579–610.

Stewart, G. W. & Sun, J. (1990). *Matrix perturbation theory*, Academic press, New York.

Turro, E., Bochkina, N., Hein, A. M. K. & Richardson, S. (2007). Bgx: a bioconductor package for the bayesian integrated analysis of affymetrix genechips. *BMC Bioinformatics* **8**, 439–449.

Wentzell, P. D. & Hou, S. (2012). Exploratory data analysis with noisy measurements. *J. Chemom.* **26**, 264–281.

Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K. & Kowalski, B. R. (1997). Maximum likelihood principal component analysis. *J. Chemom.* **11**, 339–366.

Wilkinson, J. H. (1965). *The algebraic eigenvalue problem*, Clarendon Press, Oxford.

Received January 2013, in final form January 2014

Kristoffer H. Hellton, Department of Biostatistics, Institute of Basic Medical Sciences, University of Oslo,
PO Box 1122 Blindern, 0317 Oslo, Norway.
E-mail: k.h.hellton@medisin.uio.no

Supporting information

Additional supporting information for this article is available online, including the following:

Appendix A. Proof of Lemma 1 and Theorem 1.

Appendix B. Proof of Theorems 2 and 3.

© 2014 Board of the Foundation of the Scandinavian Journal of Statistics.

Supporting Information to the paper

The impact of measurement error on principal component analysis

Kristoffer Hellton, Magne Thorsen

A Appendix

A.1 Proof of Lemma 1

Proof. From the representation of the empirical covariance matrix in (1), it follows from standard results in perturbation theory of linear operators (Kato, 1995) that $\lambda_{W,i}$ and $\mathbf{v}_{W,i}$ are analytic in σ for eigenvalues of multiplicity one. Therefore, as $\sigma \to 0$, we can expand the eigenvalue $\lambda_{W,i}$ around λ_i and the eigenvector $\mathbf{v}_{W,i}$, up to a scaling, around $\mathbf{v}_i$

$$\begin{aligned}\lambda_{W,i} &= \lambda_i + \sigma\,\lambda_{1,i} + \sigma^2\,\lambda_{2,i} + O(\sigma^3),\\ \mathbf{v}_{W,i} &= \mathbf{v}_i + \sigma\,\mathbf{v}_{1,i} + \sigma^2\,\mathbf{v}_{2,i} + O(\sigma^3).\end{aligned}$$

The first- and second-order coefficient in the Taylor expansion of the eigenvectors are denoted by $\mathbf{v}_{1,i}$ and $\mathbf{v}_{2,i}$. By inserting these two expressions and $\mathbf{S}_W$ in expression (1) into the eigenequation $\mathbf{S}_W\mathbf{v}_W = \lambda_W\mathbf{v}_W$ and collecting the terms with the same power in σ, the following system of equations is specified

$$\begin{aligned}\mathbf{S}_X\mathbf{v}_i &= \lambda_i\mathbf{v}_i,\\ \mathbf{S}_X\mathbf{v}_{1,i} + \Delta\mathbf{S}_1\mathbf{v}_i &= \lambda_i\mathbf{v}_{1,i} + \lambda_{1,i}\mathbf{v}_i, &&(1)\\ \mathbf{S}_X\mathbf{v}_{2,i} + \Delta\mathbf{S}_1\mathbf{v}_{1,i} + \Delta\mathbf{S}_2\mathbf{v}_i &= \lambda_i\mathbf{v}_{2,i} + \lambda_{1,i}\mathbf{v}_{1,i} + \lambda_{2,i}\mathbf{v}_i. &&(2)\end{aligned}$$

The terms in the Taylor expansion of $\mathbf{v}_{W,i}$ must be orthogonal to the eigenvector $\mathbf{v}_i$; $\mathbf{v}_{1,i}^T\mathbf{v}_i = 0$ and $\mathbf{v}_{2,i}^T\mathbf{v}_i = 0$. Thus the terms can be written as linear combinations of all the other eigenvectors, such that $\mathbf{v}_{1,i} = \sum_{j\neq i}\alpha_{1,j}\mathbf{v}_j$ and $\mathbf{v}_{2,i} = \sum_{j\neq i}\alpha_{2,j}\mathbf{v}_j$. First, $\lambda_{1,i}$ is found by premultiplying equation (1) by $\mathbf{v}_i^T$, using the fact that $\mathbf{v}_i^T\mathbf{v}_{1,i} = 0$ and $\mathbf{v}_i^T\mathbf{v}_i = 1$

$$\begin{aligned}\mathbf{v}_i^T\mathbf{S}_X\mathbf{v}_{1,i} + \mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_i &= \mathbf{v}_i^T\lambda_i\mathbf{v}_{1,i} + \mathbf{v}_i^T\lambda_{1,i}\mathbf{v}_i,\\ \lambda_i\mathbf{v}_i^T\mathbf{v}_{1,i} + \mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_i &= \mathbf{v}_i^T\lambda_{1,i}\mathbf{v}_i,\\ \lambda_{1,i} &= \mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_i.\end{aligned}$$

Then $\mathbf{v}_{1,i}$ is found by inserting the linear combination $\mathbf{v}_{1,i} = \sum_{j\neq i}\alpha_{1,j}\mathbf{v}_j$ and premultiplying equation (1) by $\mathbf{v}_j^T$

$$\begin{aligned}\mathbf{v}_j^T\mathbf{S}_X\mathbf{v}_{1,i} + \mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i &= \mathbf{v}_j^T\lambda_i\mathbf{v}_{1,i} + \mathbf{v}_j^T\lambda_{1,i}\mathbf{v}_i,\\ \lambda_j\alpha_{1,j} + \mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i &= \lambda_{1,i}\alpha_{1,j},\\ \alpha_{1,j} &= \frac{\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i}{\lambda_i - \lambda_j}.\end{aligned}$$

Secondly, $\lambda_{2,i}$ is found by premultiplying equation (2) by $\mathbf{v}_i^T$:

$$\begin{aligned}
\mathbf{v}_i^T\mathbf{S}_X\mathbf{v}_{2,i} + \mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_{1,i} + \mathbf{v}_i^T\Delta\mathbf{S}_2\mathbf{v}_i &= \mathbf{v}_i^T\lambda_i\mathbf{v}_{2,i} + \mathbf{v}_i^T\lambda_{1,i}\mathbf{v}_{1,i} + \mathbf{v}_i^T\lambda_{2,i}\mathbf{v}_i,\\
\lambda_i\mathbf{v}_i^T\mathbf{v}_{2,i} + \mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_{1,i} + \mathbf{v}_i^T\Delta\mathbf{S}_2\mathbf{v}_i &= \mathbf{v}_i^T\lambda_{2,i}\mathbf{v}_i,\\
\lambda_{2,i} &= \mathbf{v}_i^T\Delta\mathbf{S}_2\mathbf{v}_i + \sum_{j\neq i}\frac{\mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_j\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i}{\lambda_i-\lambda_j}.
\end{aligned}$$

Lastly, the second-order Taylor expansion coefficient for the eigenvector is found by premultiplying equation (2) by $\mathbf{v}_j^T$:

$$\begin{aligned}
\mathbf{v}_j^T\mathbf{S}_X\mathbf{v}_{2,i} + \mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_{1,i} + \mathbf{v}_j^T\Delta\mathbf{S}_2\mathbf{v}_i &= \mathbf{v}_j^T\lambda_i\mathbf{v}_{2,i} + \mathbf{v}_j^T\lambda_{1,i}\mathbf{v}_{1,i} + \mathbf{v}_j^T\lambda_{2,i}\mathbf{v}_i,\\
\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_{1,i} + \mathbf{v}_j^T\Delta\mathbf{S}_2\mathbf{v}_i - \lambda_{1,i}\mathbf{v}_j^T\mathbf{v}_{1,i} &= \lambda_i\mathbf{v}_j^T\mathbf{v}_{2,i} - \lambda_j\mathbf{v}_j^T\mathbf{v}_{2,i},\\
\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_{1,i} + \mathbf{v}_j^T\Delta\mathbf{S}_2\mathbf{v}_i - \lambda_{1,i}\alpha_{1,j} &= (\lambda_i-\lambda_j)\alpha_{2,j},
\end{aligned}$$

which, when inserting the expressions for $\lambda_{1,i}$ and $\mathbf{v}_{1,i}$, results in

$$\alpha_{2,j} = \frac{\mathbf{v}_j^T\Delta\mathbf{S}_2\mathbf{v}_i}{\lambda_i-\lambda_j} + \sum_{k\neq i}\frac{\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_k}{(\lambda_i-\lambda_j)(\lambda_i-\lambda_k)} - \frac{\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i\mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_i}{(\lambda_i-\lambda_j)^2}.$$

□

A.2 Proof of Theorem 1

Proof. Since the scores are given by $\mathbf{Z}_{W,i} = \mathbf{v}_{W,i}^T\mathbf{W}$, the Taylor expansion of $\mathbf{Z}_{W,i}$ around $\mathbf{Z}_i = \mathbf{v}_i^T\mathbf{X}$ is found by combining the Taylor expansion of the eigenvector in expression (3) together with $\mathbf{W} = \mathbf{X} + \sigma\,\mathbf{U}$. By using the notation of proof A.1, we obtain

$$\begin{aligned}
\mathbf{Z}_W &= \left(\mathbf{v}_i + \sigma\,\mathbf{v}_{1,i} + \sigma^2\,\mathbf{v}_{2,i} + O(\sigma^3)\right)^T(\mathbf{X}+\sigma\,\mathbf{U}),\\
&= \mathbf{v}_i^T\mathbf{X} + \sigma\left(\mathbf{v}_{1,i}^T\mathbf{X} + \mathbf{v}_i^T\mathbf{U}\right) + \sigma^2\left(\mathbf{v}_{1,i}^T\mathbf{U} + \mathbf{v}_{2,i}^T\mathbf{X}\right) + O(\sigma^3).
\end{aligned}$$

□

B Appendix

B.1 Second-order moments of random matrices

For the second-order moments of random matrices, we have the following lemma:

Lemma 1. *(Ghazal & Neudecker, 2000, p.81) For a $p \times n$ matrix $\mathbf{U} = [\mathbf{U}_1, \ldots, \mathbf{U}_n]$ where $\mathbf{U}_r \sim N(0, \Sigma)$ for $r = 1, \ldots, n$, the following is given*

$$\mathbb{E}(\mathbf{U}\mathbf{A}\mathbf{U}^T) = \mathrm{tr}(\mathbf{A})\Sigma, \tag{3}$$

$$\mathbb{E}(\mathbf{U}^T\mathbf{A}\mathbf{U}) = \mathrm{tr}(\mathbf{A}\Sigma)I_n, \tag{4}$$

$$\mathbb{E}(\mathbf{U}^T\mathbf{A}\mathbf{U}^T) = \mathbf{A}^T\Sigma. \tag{5}$$

Second-order moments of $\Delta\mathbf{S}_1 = \frac{1}{n}\mathbf{X}\mathbf{U}^T + \frac{1}{n}\mathbf{U}\mathbf{X}^T$ are found by using Lemma 1.

$$\begin{aligned}\mathbb{E}([\mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i]^2 \mid \mathbf{X}) = \frac{1}{n^2}\,\mathbb{E}\left(\mathbf{v}_i^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i(\mathbf{v}_i^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i)^T\right) + \frac{1}{n^2}\,\mathbb{E}\left(\mathbf{v}_i^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i(\mathbf{v}_i^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i)^T \mid \mathbf{X}\right)\\ + \frac{2}{n^2}\,\mathbb{E}\left(\mathbf{v}_i^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i(\mathbf{v}_i^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i)^T \mid \mathbf{X}\right). \qquad (6)\end{aligned}$$

By using formula (3), the first term yields

$$\begin{aligned}\mathbb{E}\left(\mathbf{v}_i^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i(\mathbf{v}_i^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i)^T \mid \mathbf{X}\right) &= \mathbf{v}_i^T\,\mathbb{E}\left(\mathbf{U}\left[\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}\right]\mathbf{U}^T\right)\mathbf{v}_i = \mathbf{v}_i^T\,\mathrm{tr}\left(\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}\right)\Sigma_U\mathbf{v}_i,\\ &= n\mathbf{v}_i^T\,\mathrm{tr}\left(\mathbf{v}_i^T\frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v}_i\right)\Sigma_U\mathbf{v}_i = n\lambda_i\mathbf{v}_i^T\Sigma_U\mathbf{v}_i,\end{aligned}$$

by formula (4), the second term yields

$$\begin{aligned}\mathbb{E}(\mathbf{v}_i^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i(\mathbf{v}_i^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i)^T \mid \mathbf{X}) &= \mathbf{v}_i^T\mathbf{X}\,\mathbb{E}\left(\mathbf{U}^T\left[\mathbf{v}_i\mathbf{v}_i^T\right]\mathbf{U}\right)\mathbf{X}^T\mathbf{v}_i = \mathbf{v}_i^T\mathbf{X}\,\mathrm{tr}\left(\mathbf{v}_i\mathbf{v}_i^T\Sigma_U\right)I_n\mathbf{X}^T\mathbf{v}_i,\\ &= n\,\mathrm{tr}\left(\mathbf{v}_i^T\Sigma_U\mathbf{v}_i\right)\mathbf{v}_i^T\frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v}_i = n\lambda_i\mathbf{v}_i^T\Sigma_U\mathbf{v}_i,\end{aligned}$$

and by formula (5), the third term yields

$$\begin{aligned}\mathbb{E}(\mathbf{v}_i^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i(\mathbf{v}_i^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i)^T \mid \mathbf{X}) &= \mathbf{v}_i^T\mathbf{X}\,\mathbb{E}\left(\mathbf{U}^T\left[\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^T\right]\mathbf{U}^T\right)\mathbf{v}_i = \mathbf{v}_i^T\mathbf{X}\Sigma_U\left[\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}^T\right]^T\mathbf{v}_i,\\ &= n\mathbf{v}_i^T\frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v}_i\mathbf{v}^T\Sigma_U\mathbf{v}_i = n\lambda_i\mathbf{v}_i^T\Sigma_U\mathbf{v}_i,\end{aligned}$$

such that

$$\mathbb{E}([\mathbf{v}_i^T\Delta\mathbf{S}_1\mathbf{v}_i]^2 \mid \mathbf{X}) = \frac{4\lambda_i}{n}\,\mathbf{v}^T\Sigma_U\mathbf{v}. \qquad (7)$$

When the eigenvector indexes are different, the second-order moment is given by

$$\begin{aligned}\mathbb{E}([\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i]^2 \mid \mathbf{X}) = \frac{1}{n^2}\,\mathbb{E}\left(\mathbf{v}_j^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i(\mathbf{v}_j^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i)^T \mid \mathbf{X}\right) + \frac{1}{n^2}\,\mathbb{E}\left(\mathbf{v}_j^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i(\mathbf{v}_j^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i)^T \mid \mathbf{X}\right)\\ + \frac{2}{n^2}\,\mathbb{E}\left(\mathbf{v}_j^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i(\mathbf{v}_j^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i)^T \mid \mathbf{X}\right).\end{aligned}$$

The first term yields

$$\begin{aligned}\mathbb{E}\left(\mathbf{v}_j^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i(\mathbf{v}_j^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i)^T \mid \mathbf{X}\right) &= \mathbf{v}_j^T\,\mathbb{E}\left(\mathbf{U}\left[\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}\right]\mathbf{U}^T\right)\mathbf{v}_j = \mathbf{v}_j^T\,\mathrm{tr}\left(\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{X}\right)\Sigma_U\mathbf{v}_j,\\ &= n\,\mathrm{tr}\left(\mathbf{v}_i^T\frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v}_i\right)\mathbf{v}_j^T\Sigma_U\mathbf{v}_j = n\lambda_i\mathbf{v}_j^T\Sigma_U\mathbf{v}_j,\end{aligned}$$

the second term yields

$$\begin{aligned}\mathbb{E}\left(\mathbf{v}_j^T\mathbf{X}\mathbf{U}^T\mathbf{v}_i(\mathbf{v}_j^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i)^T \mid \mathbf{X}\right) &= \mathbf{v}_j^T\mathbf{X}\,\mathbb{E}\left(\mathbf{U}^T\left[\mathbf{v}_i\mathbf{v}_i^T\right]\mathbf{U}\right)\mathbf{X}^T\mathbf{v}_j = \mathbf{v}_j^T\mathbf{X}\,\mathrm{tr}\left(\mathbf{v}_i\mathbf{v}_i^T\Sigma_U\right)I_n\mathbf{X}^T\mathbf{v}_j,\\ &= n\,\mathrm{tr}\left(\mathbf{v}_i^T\Sigma_U\mathbf{v}_i\right)\mathbf{v}_j^T\frac{1}{n}\mathbf{X}\mathbf{X}^T\mathbf{v}_j = n\lambda_j\mathbf{v}_i^T\Sigma_U\mathbf{v}_i,\end{aligned}$$

and the third term yields

$$\begin{aligned}\mathbb{E}(\mathbf{v}_j^T\mathbf{U}\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{U}\mathbf{X}^T\mathbf{v}_j \mid \mathbf{X}) &= \mathbf{v}_j^T\,\mathbb{E}\left(\mathbf{U}\left[\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\mathbf{U}\right]\right)\mathbf{X}^T\mathbf{v}_j,\\ &= \mathbf{v}_j^T\Sigma_U\left[\mathbf{X}^T\mathbf{v}_i\mathbf{v}_i^T\right]^T\mathbf{X}^T\mathbf{v}_j, = \mathbf{v}_j^T\Sigma_U\mathbf{v}_i\;\mathbf{v}_i^T\mathbf{X}\mathbf{X}^T\mathbf{v}_j = 0,\end{aligned}$$

due to the fact that the ith and jth score vectors are uncorrelated by definition, $\mathrm{Cov}(\mathbf{v}_i^T\mathbf{X}, \mathbf{v}_j^T\mathbf{X}) = 0$. Thus, the second-order moment is given by

$$\mathbb{E}([\mathbf{v}_j^T\Delta\mathbf{S}_1\mathbf{v}_i]^2 \mid \mathbf{X}) = \frac{\lambda_j}{n}\,\mathbf{v}_i^T\Sigma_U\mathbf{v}_i + \frac{\lambda_i}{n}\,\mathbf{v}_j^T\Sigma_U\mathbf{v}_j. \qquad (8)$$

For $\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_k)$ for $k \neq j$, we obtain

$$\mathbb{E}\left(\mathbf{v}_j^T \mathbf{U}\mathbf{X}^T \mathbf{v}_i (\mathbf{v}_j^T \mathbf{U}\mathbf{X}^T \mathbf{v}_k)^T \mid \mathbf{X}\right) = \frac{\lambda_j}{n} \mathbf{v}_k^T \Sigma_U \mathbf{v}_i,$$

while all other combinations yield zero, such that

$$\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_k \mid \mathbf{X}) = \frac{\lambda_j}{n} \mathbf{v}_k^T \Sigma_U \mathbf{v}_i, \quad k \neq j. \tag{9}$$

To calculate the moment $\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i \mid \mathbf{X})$, we first obtain

$$\begin{aligned}
\mathbb{E}(\mathbf{v}_j^T \mathbf{U}\mathbf{X}^T \mathbf{v}_i \mathbf{v}_i^T \mathbf{X}\mathbf{U}^T \mathbf{v}_i) &= n\lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_i, \\
\mathbb{E}(\mathbf{v}_j^T \mathbf{U}\mathbf{X}^T \mathbf{v}_i \mathbf{v}_i^T \mathbf{U}\mathbf{X}^T \mathbf{v}_i) &= n\lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_i,
\end{aligned}$$

while the two other combinations yield zero, such that

$$\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i \mid \mathbf{X}) = \frac{2\lambda_i}{n} \mathbf{v}_j^T \Sigma_U \mathbf{v}_i. \tag{10}$$

In addition, the formulas in Lemma 1 give

$$\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \mathbf{U} \mid \mathbf{X}) = \mathbf{v}_j^T \Sigma_U \mathbf{v}_i \mathbf{v}_j^T \mathbf{X} + \mathbf{v}_j^T \Sigma_U \mathbf{v}_j \mathbf{v}_i^T \mathbf{X}, \tag{11}$$

$$\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \mathbf{U} \mid \mathbf{X}) = \mathbf{v}_j^T \Sigma_U \mathbf{v}_i \mathbf{v}_i^T \mathbf{X} + \mathbf{v}_i^T \Sigma_U \mathbf{v}_i \mathbf{v}_j^T \mathbf{X}. \tag{12}$$

B.2 Proof of Theorem 2

Proof. The bias in the ith eigenvalue is found by moving λ_i in expression (2) to the other side and take the expectation

$$\mathbb{E}(\Delta\lambda_i \mid \mathbf{X}) = \sigma\, \mathbf{v}_i^T\, \mathbb{E}(\Delta\mathbf{S}_1)\mathbf{v}_i + \sigma^2\, \mathbf{v}_i^T\, \mathbb{E}(\Delta\mathbf{S}_2)\mathbf{v}_i + \sigma^2 \sum_{j\neq i} \frac{\mathbb{E}(\mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_j \mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i)}{\lambda_i - \lambda_j} + O(\sigma^3).$$

When the error $\mathbf{U}_r$ is *iid*, $\mathbf{U}_r \sim N(0, \Sigma_U)$, it follows that $\mathbb{E}(\Delta\mathbf{S}_1) = \mathbf{0}$ and $\mathbb{E}(\Delta\mathbf{S}_2) = \Sigma_U$. By inserting these expectations together with expression (8), derived in B.1, we obtain the result.

The bias in the ith eigenvector is found by moving $\mathbf{v}_i$ in expression (3) to the other side and take the expectation

$$\begin{aligned}
\mathbb{E}(\Delta\mathbf{v}_i \mid \mathbf{X}) = {} & \sigma \sum_{j\neq i} \frac{\mathbf{v}_j^T\, \mathbb{E}(\Delta\mathbf{S}_1)\mathbf{v}_i}{\lambda_i - \lambda_j}\, \mathbf{v}_j + \sigma^2 \sum_{j\neq i} \frac{\mathbf{v}_j^T\, \mathbb{E}(\Delta\mathbf{S}_2)\mathbf{v}_i}{\lambda_i - \lambda_j}\, \mathbf{v}_j \\
& + \sigma^2 \sum_{j\neq i}\sum_{k\neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_k)}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)}\, \mathbf{v}_j - \sigma^2 \sum_{j\neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta\mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i)}{(\lambda_i - \lambda_j)^2}\, \mathbf{v}_j + O(\sigma^3).
\end{aligned}$$

We insert $\mathbb{E}(\Delta\mathbf{S}_1) = \mathbf{0}$ and $\mathbb{E}(\Delta\mathbf{S}_2) = \Sigma_U$ and the expressions (9) and (10) derived in B.1 and get the result by rearranging the last two terms.

The variance of the change in the ith eigenvalue is given by

$$\operatorname{Var}(\Delta\lambda_i \mid \mathbf{X}) = \sigma^2 \operatorname{Var}(\mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i) + O(\sigma^3) = \sigma^2\, \mathbb{E}([\mathbf{v}_i^T \Delta\mathbf{S}_1 \mathbf{v}_i]^2) - \sigma^2(\mathbf{v}_i^T\, \mathbb{E}\, \Delta\mathbf{S}_1 \mathbf{v}_i)^2 + O(\sigma^3).$$

As $\mathbb{E}(\Delta\mathbf{S}_1) = \mathbf{0}$, the variance is equal to the second-order moment in equation (7), such that

$$\operatorname{Var}(\Delta\lambda_i \mid \mathbf{X}) = \frac{4\sigma^2\lambda_i}{n}\, \mathbf{v}_i^T \Sigma_U \mathbf{v}_i + O(\sigma^3).$$

The variance of the change in the kth coordinate of the ith eigenvector is given by

$$\begin{aligned}\operatorname{Var}(\Delta \mathbf{v}_{ik} \mid \mathbf{X}) &= \sigma^2 \operatorname{Var}\left(\sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_{jk}\right) + O(\sigma^3),\\ &= \sigma^2 \sum_{j\neq i} \frac{\operatorname{Var}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i)}{(\lambda_i - \lambda_j)^2} \mathbf{v}_{jk}^2 + 2\,\sigma^2 \sum_{j,l\neq i, j<l} \frac{\operatorname{Cov}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i, \mathbf{v}_l^T \Delta \mathbf{S}_1 \mathbf{v}_i)}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_l)} \mathbf{v}_{jk}\mathbf{v}_{lk} + O(\sigma^3).\end{aligned}$$

As $\mathbb{E}(\Delta \mathbf{S}_1) = 0$, the variance and covariance expressions are given by the second-order moments

$$\begin{aligned}\operatorname{Var}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i) &= \mathbb{E}([\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i]^2)\\ \operatorname{Cov}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i, \mathbf{v}_l^T \Delta \mathbf{S}_1 \mathbf{v}_i) &= \mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i \mathbf{v}_l^T \Delta \mathbf{S}_1 \mathbf{v}_i)\end{aligned}$$

which are given by expressions (8) and (9). □

B.3 Proof of Theorem 3

Proof. The bias in the scores is found by moving $\mathbf{Z}_i$ in expression (4) to the other side and take the expectation

$$\begin{aligned}\mathbb{E}(\Delta \mathbf{Z}_i \mid \mathbf{X}) = \sigma \sum_{j\neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i)}{\lambda_i - \lambda_j}\, \mathbf{v}_j^T \mathbf{X} + \sigma\, \mathbf{v}_i^T\, \mathbb{E}\,\mathbf{U} + \sigma^2 \sum_{j\neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \mathbf{U})}{\lambda_i - \lambda_j} + \sigma^2 \sum_{j\neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_2 \mathbf{v}_i)}{\lambda_i - \lambda_j}\, \mathbf{v}_j^T \mathbf{X}\\ + \sigma^2 \sum_{j\neq i}\sum_{k\neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i \mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_k)}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_k)}\, \mathbf{v}_j^T \mathbf{X} - \sigma^2 \sum_{j\neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \Delta \mathbf{S}_1 \mathbf{v}_i)}{(\lambda_i - \lambda_j)^2}\, \mathbf{v}_j^T \mathbf{X} + O(\sigma^3).\end{aligned}$$

Apart from $\mathbb{E}(\Delta \mathbf{S}_1) = \mathbf{0}$, $\mathbb{E}(\Delta \mathbf{S}_2) = \Sigma_U$ and $\mathbb{E}(\mathbf{U}) = \mathbf{0}$, the expectations are given by the expressions in (9), (10) and (11). The result is obtained by inserting all expectations.

The variance of the change in the kth score of the ith component is given by

$$\begin{aligned}\operatorname{Var}(\Delta \mathbf{Z}_{ik} \mid \mathbf{X}) &= \sigma^2 \operatorname{Var}\left(\sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j^T \mathbf{X}_k + \mathbf{v}_i^T \mathbf{U}_k \mid \mathbf{X}\right) + O(\sigma^3)\\ &= \sigma^2 \operatorname{Var}\left(\sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j^T \mathbf{X}_k \mid \mathbf{X}\right) + \sigma^2 \operatorname{Var}(\mathbf{v}_i^T \mathbf{U}_k \mid \mathbf{X})\\ &\quad + 2\,\sigma^2 \operatorname{Cov}\left(\sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j^T \mathbf{X}_k, \mathbf{v}_i^T \mathbf{U}_k \mid \mathbf{X}\right) + O(\sigma^3).\end{aligned}$$

The first term is directly given by the variance of the eigenvector

$$\begin{aligned}\operatorname{Var}\left(\sum_{j\neq i} \frac{\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j^T \mathbf{X}_k \mid \mathbf{X}\right) &= \sum_{j\neq i} \frac{\operatorname{Var}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i)}{(\lambda_i - \lambda_j)^2} (\mathbf{v}_j^T \mathbf{X}_k)^2 + 2 \sum_{j,l\neq i, j<l} \frac{\operatorname{Cov}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i, \mathbf{v}_l^T \Delta \mathbf{S}_1 \mathbf{v}_i)}{(\lambda_i - \lambda_j)(\lambda_i - \lambda_l)} \mathbf{v}_j^T \mathbf{X}_k \mathbf{v}_l^T \mathbf{X}_k,\\ &= \sum_{j\neq i} \frac{\lambda_j \mathbf{v}_i^T \Sigma_U \mathbf{v}_i + \lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_j}{n(\lambda_i - \lambda_j)^2} (\mathbf{v}_j^T \mathbf{X}_k)^2 + \sum_{j,l\neq i, j<l} \frac{2\lambda_i \mathbf{v}_j^T \Sigma_U \mathbf{v}_l}{n(\lambda_i - \lambda_j)(\lambda_i - \lambda_l)} \mathbf{v}_j^T \mathbf{X}_k \mathbf{v}_l^T \mathbf{X}_k.\end{aligned}$$

The second term, the direct contribution from the error, is given by

$$\operatorname{Var}(\mathbf{v}_i^T \mathbf{U}_k \mid \mathbf{X}) = \mathbf{v}_i^T \Sigma_U \mathbf{v}_i.$$

The last term is given by the interaction between the error and the eigenvector, as $\mathbb{E}(\mathbf{U}) = \mathbf{0}$

$$\mathrm{Cov}\left(\sum_{j \neq i} \frac{\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j^T \mathbf{X}_k, \mathbf{v}_i^T \mathbf{U}_k \mid \mathbf{X} \right) = \sum_{j \neq i} \frac{\mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \mathbf{U}_k)}{\lambda_i - \lambda_j} \mathbf{v}_j \mathbf{X}_k,$$

where the expectation is given by the kth coefficient in expression (12):

$$\mathbb{E}(\mathbf{v}_j^T \Delta \mathbf{S}_1 \mathbf{v}_i \mathbf{v}_i^T \mathbf{U}_k \mid \mathbf{X}) = \frac{1}{n} \mathbf{v}_j^T \Sigma_U \mathbf{v}_i \mathbf{v}_i^T \mathbf{X}_k + \frac{1}{n} \mathbf{v}_i^T \Sigma_U \mathbf{v}_i \mathbf{v}_j^T \mathbf{X}_k.$$

The result is given by combining the three terms. □

References

Ghazal, G. A. & Neudecker, H. (2000). On second-order and fourth-order moments of jointly distributed random matrices: a survey. *Linear Algebra Appl.* **321**, 61–93.

Kato, T. (1995). *Perturbation theory for linear operators.* Springer, Berlin.

Integrative clustering of high-dimensional data with joint and individual clusters, with an application to the Metabric study

Kristoffer Hellton, Magne Thoresen
Department of Biostatistics, University of Oslo,
P.O.Box 1122 Blindern N-0317, Oslo, Norway
k.h.hellton@medisin.uio.no

September 26, 2014

Abstract

When measuring a range of different genomic, epigenomic, transcriptomic and other variables, an integrative approach to analysis can strengthen inference and give new insights. This is also the case when clustering patient samples, and several integrative cluster procedures have been proposed. Common for these methodologies is the restriction of a joint cluster structure, which is equal for all data layers. We instead present Joint and Individual Clustering (JIC), which estimates both joint and data type-specific clusters simultaneously, as an extension of the JIVE algorithm (Lock et al., 2013). The method is compared to iCluster, another integrative clustering method, and simulations show that JIC is clearly advantageous when both individual and joint clusters are present. The method is used to cluster patients in the Metabric study, integrating gene expression data and copy number aberrations (CNA). The analysis suggests a division into three joint clusters common for both data types and seven independent clusters specific for CNA. Both the joint and CNA-specific clusters are significantly different with respect to survival, also when adjusting for age and treatment.

Keywords: Breast cancer; Clustering; Integrative genomics; Latent variable estimation; Singular value decomposition.

1 Introduction

The rapid development in genomic technologies has enabled the analysis of an increasing range of data layers or data types. This increases the need for integrative procedures that can handle several data types. When studying diseases that build on several molecular processes, we need to consider the interplay between the genomic layers to fully understand the phenotypic traits. We should therefore attempt to integrate different data types in a single joint analysis, and this is the core principle of integrative genomics. As the information content is higher in an integrative framework compared to individual analyses, it is possible to gain statistical power to detect relevant signals. This is especially relevant for genetically driven diseases such as cancer in general or breast cancer, as studied in this paper.

An integrative approach is especially relevant in the exploratory field of unsupervised clustering, and such procedures have been suggested earlier (Shen et al., 2009, 2013; Lock and Dunson, 2013). The aim of clustering is to discover novel disease subtypes, which can aid the understanding of survival and mortality risk differences or enable personalized treatments. Earlier integrative clustering approaches include the iCluster methodology (Shen et al., 2009, 2013) and the Bayesian consensus clustering (Lock and Dunson, 2013). The iCluster method clusters observations based on joint latent variables, utilizing the connection between k-means clustering and latent factor modeling. In Bayesian consensus clustering, observations are clustered for each data type separately with a last step of combining the different groupings into a consensus solution.

However, when several highly heterogeneous genomic data types are integrated, some cluster structures are typically not shared between all the data layers. If there are clear clusters present in some of the data types, but not in others, these can confound or obscure the joint clusters shared by all data types. Data type-specific cluster structures can be caused by biological confounders, such as ethnicity, or technical and measurement-related differences, such as samples processed at different labs or changes in techniques over time, affecting only a single data type. But more importantly from a biomedical point of view, there could exist disease-related patient clusters that are independent of the joint subtypes, but still relevant and interesting for treatment and disease-understanding.

Our aim is to take into account the presence of data type-specific clusters together with joint clusters in an integrative framework. We will therefore present a clustering extension of the JIVE algorithm (Lock et al., 2013), which decomposes several data sets into joint and individual latent structures in an iterative procedure. In our extension, termed Joint and Individual Clustering (JIC), the joint cluster structure is estimated simultaneously with the individual or data type-specific clustering. JIC will be compared to the iCluster methodology in different simulation settings and will be used to find joint and data type-specific clusters of patients in the Metabric study (Curtis et al., 2012).

2 Integrative clustering

The iCluster method (Shen et al., 2009, 2013) has become an established method for integrative clustering of multiple genomic data types. We extend the JIVE methodology (Lock et al., 2013) to accommodate clustering of observations, as done by iCluster. Both approaches are based on estimating latent variables as continuous representations of the cluster assignment vectors. An important difference between JIC or JIVE and iCluster is the assumed noise structure in the latent variable model. iCluster allows the factor residuals to have different variances for each variable, while JIC, assuming equal variance, allows for additional latent variables specific for each data type. Both approaches can incorporate sparsity in the loadings matrices.

Integrative clustering aims to cluster observations simultaneously in different data types. Let $X_1, \dots, X_M$ be M different genome-scale data types (typically expression, copy number variation, methylation) or genome-related data types (such as miRNA, proteins, transcription factors) that are all measured on the same n patients, indexed $j = 1, \dots, n$. Then each X_m is a $p_m \times n$ data matrix for $m = 1, \dots, M$ with p_m variables, indexed by $i = 1, \dots, p_m$. The data types can be highly heterogeneous with respect to scale, unit or variation.

The M data matrices can be combined into a single concatenated matrix

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_M \end{bmatrix},$$

of dimension $p \times n$ where $p = p_1 + \cdots + p_M$. A scaled version of the concatenated matrix can be constructed by first scaling each data matrix X_i by some norm $\|X_i\|$. Then each data type will contribute equally to the integrative solution.

2.1 Clustering and dimension reduction

Both iCluster and JIC are closely linked to k-means clustering, where clusters are defined by minimizing the distance between each observation and the cluster centroid. To simplify the procedure of k-means clustering, one can use principal component analysis (PCA) as an initial step to reduce the dimension of the data matrix. This two-step procedure, called “tandem clustering” (Arabie and Hubert, 1996; Terada, 2014), clusters the reduced subset of PC scores, but have been criticized in the statistics literature.

However, in machine learning, Zha et al. (2001); Ding and He (2004) have shown that principal components are the continuous solution to the k-means optimization problem, such that the PC scores correspond to a continuous version of the discrete cluster indicators. Specifically, if the k-means clustering solution is denoted $Z^T = [z_1, \ldots, z_{K-1}]$, a matrix of $K-1$ indicator vectors

$$z_k^T = n_k^{-1/2}[0, \ldots, 0, \underbrace{1, \ldots, 1}_{n_k}, 0, \ldots, 0],$$

where n_k is the number of observations in each cluster, the $K-1$ first principal component scores will minimize the k-means objective function. Therefore, k-means clustering (into K groups) can be solved in two steps: first find the $K-1$ (standardized) principal component scores, and then reconstruct the discrete cluster assignments from the continuous scores, for instance with k-means clustering. In a high-dimensional setting, this is highly efficient as the data matrix is reduced from $p \times n$ to $(K-1) \times n$.

The estimation of the continuous matrix Z can also be done through Gaussian latent variable modeling, where the data matrix X_m is modeled as

$$X_m = W_m^T Z + \varepsilon_m, \quad \varepsilon_m \sim N(0, \Sigma),$$

where W_m is a loading coefficient matrix and ε_m is a set of independently distributed errors. Tipping and Bishop (1999) connected the latent factor model and PCA, showing that under homogeneous and normally distributed errors, $\Sigma = \sigma^2 I_{p_m}$, the maximum likelihood estimates of W_m yield the same solution as classical principal component analysis. The use of latent variable modeling as a part of the k-means clustering is motivated by the natural extension of the latent variables to multiple data types.

2.2 iCluster

The iCluster method extends k-means clustering to an integrative clustering procedure, following the same approach as Deun et al. (2009, 2011). The latent variables Z, representing the clusters, are assumed to be common for all the data types. iCluster assumes the following model for M data types:

$$\begin{aligned} X_1 &= W_1^T Z + \varepsilon_1, \\ &\vdots \\ X_M &= W_M^T Z + \varepsilon_M, \end{aligned}$$

where the noise terms are heterogeneous, $\varepsilon_m \sim N(0, \Psi_m)$, $\Psi_m = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{p_m}^2)$. The parameter estimates are obtained by maximum likelihood estimation using the EM algorithm. If ε_m was homogeneous, the solution is analytically given by the singular value decomposition. In iCluster, one can also enforce sparsity on the loading matrices by penalizing the data log-likelihood. After convergence of the EM algorithm, the rows of Z are clustered by the k-means algorithm to obtain the group membership of each observation. In this way, the latent variable Z corresponds to a cluster indicator matrix shared between all data sets.

2.3 Joint and Individual Clustering (JIC)

Clustering based on estimated latent variables can also include other noise structures. We present a novel clustering extension of JIVE, the Joint and Individual Clustering (JIC), where clustering is carried out on both joint and data type-specific latent variables. The JIVE scheme proposed by Lock et al. (2013) decomposes multiple data matrices into joint and individual structures. Both the shared and the data type-specific latent variables can be used to obtain a clustering of patients in a finale reduced k-means step.

In JIC, the data types are assumed to be realizations of a combination of common and data type-specific latent variables

$$\begin{aligned} X_1 &= W_1^T Z + V_1^T Z_1 + \varepsilon_1, \\ &\vdots \\ X_M &= W_M^T Z + V_M^T Z_M + \varepsilon_M, \end{aligned}$$

where $\varepsilon_m \sim N(0, \sigma_m^2 I), m = 1, \ldots, M$ and the joint loading matrices form a concatenated matrix

$$W = \begin{bmatrix} W_1^T \\ \vdots \\ W_M^T \end{bmatrix}.$$

When each individual latent clustering matrix Z_m, is orthogonal to the joint latent matrix, such that $ZZ_m^T = 0_{(K-1)\times(K_m-1)}$, there exists a unique decomposition of X (Lock et al., 2013, Supplementary material). The decomposition can be found by minimizing the reconstruction error

$$\|R\|^2 = \sum_{m=1}^{M} \|R_m\|^2 = \sum_{m=1}^{M} \|X_m - W_m^T Z - V_m^T Z_m\|^2.$$

If the rank of $W^T Z$, r, and the rank of $V_m^T Z_m$, r_m, for $m = 1, \ldots M$ are fixed, the decomposition can be found by iteratively estimating the joint and individual structures: First fix $W^T Z$ and estimate each $V_m^T Z_m$ by minimizing $\|R_m\|$. Then fix $V_1^T Z_1, \ldots, V_M^T Z_M$ and estimate $W_m^T Z$ by minimizing $\|R\|$. This procedure is repeated until a suitable convergence criterion is reached.

When errors are assumed homogenous across variables (of same type), the solution minimizing the reconstruction error is given by the singular value decomposition and the latent variables corresponds to the left singular vectors or standardized principal component scores estimated as follows:

- Calculate $W^T Z$ by the r rank singular value decomposition of X, and subtract $W^T Z$ from X,
- Calculate $V_m^T Z_m$ by the r_m rank singular value decomposition of the sub-matrix $X_m - W_m^T Z$, for $m = 1, \ldots M$
- Form the concatenated matrix of $X_m^{(l+1)} = X_m^{(l)} - V_m^T Z_m$ for $m = 1, \ldots M$ and repeat all steps until convergence.

At convergence, the rows of Z^T are clustered into $r + 1$ groups and the rows of Z_m^T are clustered into $r_m + 1$ groups for $m = 1, \ldots M$, respectively, using k-means clustering.

2.4 Procedure for selection number of clusters

To choose the number of clusters is a difficult task, and in general there is no optimal procedure. However, the selection procedure can be tailored to the method and relevant data, and we will use a procedure enlightening the subjective choices always present in such analyses.

Firstly, we exploit the subspace structure in JIC. The number of dimensions present in the clustering step is directly given by the number of clusters we aim to find; for K clusters, we use $K - 1$ component scores. As these are given by the singular value decomposition, the variables are by construction uncorrelated with each other, $ZZ^T = I_{K-1}$, such that each dimension contains independent information regarding the clustering. As shown by Ding and He (2004), a new cluster should be separated out in each dimension specified by a component. We exploit this property, and check if a new cluster is present in each added dimension. When no new cluster separates out, the total number of relevant dimensions is found. We use the following procedure:

1. For the ith component, check if the k-means clustering into two clusters is better than one cluster by a chosen procedure.
2. If two clusters are better, proceed to the next component. If instead only one cluster is supported, stop and set the number of clusters to the current component number.

Instead of checking K clusters in a $K-1$ dimensional space, we will check two clusters in a one-dimensional space, until we find the first component where no new cluster is present.

How to check the presences of a new cluster should depend on the application and data characteristics. Some possible choices of procedures are:

- *Prediction strength (Tibshirani and Walther, 2005; Shen et al., 2013):* evaluates clusters based on reproducibility between random splits of the data into discovery and validation sets. A predicted and validation clustering are evaluated by a similarity index, and the K with the highest index is chosen. However, in the $p \gg n$ setting, the component scores are very stable (Lee et al., 2014; Hellton and Thoresen, 2014), such that the sub-sampling induces little variability. Therefore component scores representing noise can exhibit very good cluster reproducibility, a property which is not desirable.

- *Cluster separation*: clusters can be evaluated by a separation criterion, such as the Calinski-Harabasz, the Dunn criterion or within group sum-of-squares. This requires the index value for a single cluster, which can be difficult to assess. The approach seems to work best in low-dimensional settings with well-separated clusters (Milligan and Cooper, 1987).

- *Approach of G-means (Hamerly and Elkan, 2003):* evaluates the normality of the continuous scores. When no clusters are present, the component scores should behave as noise and follow a normal distribution, instead of a mixing distribution. We can evaluate this normality by qq-plots or normality tests. If the scores deviate significantly from normality, they do not resemble pure noise and clusters are present in the data. If the test is not significant, there is no evidence of clusters beyond the normally distributed noise. This approach seems to work well when clusters are not well-separated, and instead resemble a continuum.

2.5 Cluster procedure for JIC

As genetic data usually do not exhibit well-separated clusters, we will utilize the idea behind the G-means method together with the notion of the independent subspaces. We use qq-plots, complemented by the Anderson-Darling test, to evaluate the normality of each component.

To identify the number of joint and individual clusters, we use the fact that the total rank of the cluster structure in the concatenated matrix, X, is given by

$$E = r + r_1 + \cdots + r_M,$$

	iCluster	JIC: joint	X_1	X_2	X_3
Setting I: Precision	0.998	0.985	-	-	-
Correctly estimated K		97%	96%	95%	98%
Setting II: Precision	0.415	0.933	0.950	0.791	0.874
Correctly estimated K		89%	90%	88%	88%

Table 1: Mean precision of estimated cluster assignment (over 100 simulations), when the numbers of clusters are known. Percentage of times the numbers of clusters are correctly estimated.

and the rank of the cluster structure in the original data X_m is $E_m = r + r_m$ for $m = 1, \ldots, M$. As the number of clusters is given by $r+1$ and r_m+1 respectively, we can determine E and $E_1, \ldots, E_M$ in the data and use them to calculate K and $K_1, \ldots, K_M$. We follow the two step procedure:

1. Estimate the number of relevant subspaces E in X, when the ranks of the individual structures are fixed to zero: test the normality of the ith joint component scores for increasing i, until the last non-normally distributed component is found and set E to the component number.

2. Estimate the number of relevant subspaces E_m in X_m: For each $m = 1, \ldots, M$, test the normality of the ith component scores for increasing i, until the last non-normally distributed component is found and set E_m to the component number.

Now, the number of joint clusters is given as

$$K = \frac{E_1 + \cdots + E_M - E}{M - 1} + 1, \tag{1}$$

while the number of individual clusters is given as $K_m = E_m - K + 2$ for $m = 1, \ldots, M$.

3 Simulations

We compare JIC to the iCluster procedure by simulating two different settings; only joint clusters and both joint and data type-specific clusters. In both settings, three different data types are integrated, $M = 3$, and the number of clusters is first assumed known, then estimated by the procedure described in Section 2.5.

3.1 Setting I: Joint cluster structure

First, we simulate 5 joint clusters, present in all three data sets. Specifically, $n = 150$, where $j = 1, \ldots, 30$ belongs to the first cluster, $j = 31, \ldots, 60$ belongs to the second cluster and so on, giving 30 observations in each cluster. The joint latent variable Z_J^T, with the indicator vectors as columns, is an $n \times 4$ matrix

$$Z_J^T = \begin{bmatrix} 1 & 0 & \cdots \\ \vdots & \vdots & \\ 0 & 1 & \cdots \\ \vdots & \vdots & \\ 0 & 0 & \cdots \end{bmatrix}.$$

Each row contains a single '1' indicating the assignment of the observation to the cluster corresponding to the column number. The last cluster is, however, specified by only zeros. The loading matrices W_1, W_2 and W_3 are of the same dimension 200×4 ($p_1 = p_2 = p_3 = 200$). We generate the loadings according to a standard normal distribution and normalize the matrices, such that $W_m^T W_m = I$ for $m = 1, 2, 3$. Within each W_i, the columns are also made orthogonal to each other. The three data sets are generated by

$$\begin{aligned} X_1 &= cW_1^T Z_J + \varepsilon_1, \\ X_2 &= cW_2^T Z_J + \varepsilon_2, \\ X_3 &= cW_3^T Z_J + \varepsilon_3, \end{aligned}$$

with standard normally distributed errors, $\varepsilon_m \sim N(0, I)$, and $c = 80$.

In the simulation, we first assume $K = 5$ known and compare the estimated cluster assignments to the true clusters in terms of the precision. Secondly, we assume K unknown and estimate it by the procedure in Section 2.5. Under Setting I in Table 1, the precision of JIC compared to the iCluster methodology is shown. We see that iCluster and JIC perform equally well in the situation with only joint clusters. In the case of unknown number of clusters, K was correctly estimated in 97% of the simulated cases, as seen in Table 1.

3.2 Setting II: Joint and individual clusters

In the second setting, two data type-specific clusters are added in each of the three data sets. The observations are randomly assigned to one of two clusters, such that the data type-specific latent variables Z_1, Z_2 and Z_3 are vectors with random ones and zeros. For the loadings matrices V_1, V_2 and V_3 of dimension 200×1, the loadings are randomly generated according to a standard normal distribution and normalized, such that $V_m^T V_m = 1$ for $m = 1, 2, 3$.

To obtain an identifiable decomposition, each Z_m is made orthogonal to the columns of Z_J. The three data sets are generated by the model

$$
\begin{aligned}
X_1 &= cW_1^T Z_J + c_1 V_1^T Z_1 + \varepsilon_1,\\
X_2 &= cW_2^T Z_J + c_2 V_2^T Z_2 + \varepsilon_2,\\
X_3 &= cW_3^T Z_J + c_3 V_3^T Z_3 + \varepsilon_3,
\end{aligned}
$$

with standard normally distributed noise, $\varepsilon_m \sim N(0, I)$, $c = 80$ and $c_1 = c_2 = c_3 = 30$. First, the correct numbers of clusters, $K = 5$ and $K_1 = K_2 = K_3 = 2$, are assumed known and the joint and individual clustering are compared to the true cluster memberships. The precisions are shown in Table 1 under Setting II. For iCluster, only the precision of the joint clustering is displayed.

We see that JIC is highly superior to the iCluster method in recovering the joint cluster as the individual clusters clearly obscure the joint signal. We also see that JIC performs well with a high precision for both the joint and individual clusters. Table 1 shows that when K, K_1, K_2 and K_3 are assumed unknown, they can be correctly estimated by the procedure in Section 2.5.

4 Example: the Metabric study

To illustrate JIC, we will analyze the data from the Metabric study (Curtis et al., 2012) with a discovery set consisting of the gene expression and somatic copy number aberrations (CNAs) of 997 breast cancer tumor samples. For the analysis, we select the 1000 genes and CNA locations with the largest variability. The CNAs are considered gene locations with tumor-specific differences in copy number compared to a healthy control, and are recorded as the count of gene copies, transformed to a log2 scale. Also recorded is disease-specific survival, together with the clinical variables: age, estrogen status, treatment and PAM50 classification. The outline of the analysis is as follows: First, the number of joint and individual clusters is chosen. Then, both clusterings are tested for differences in survival time and explored with regard to the available clinical variables.

We determine the number of joint, expression-specific and CNA-specific clusters, K, K_1, K_2 according to the procedure described in Section 2.5. Figure 1 displays the qq-plots of the first 9 joint component scores, not allowing for individual structures. Generally, it is seen that the component scores are closer to being normally distributed as the component number increases. The first, second, third and fourth joint components are clearly not normally distributed, while the 5th and 6th are borderline cases. Then, again the 7th and 8th component scores clearly deviate from normality, while the 9th component does not seem to deviate significantly. This is confirmed by the

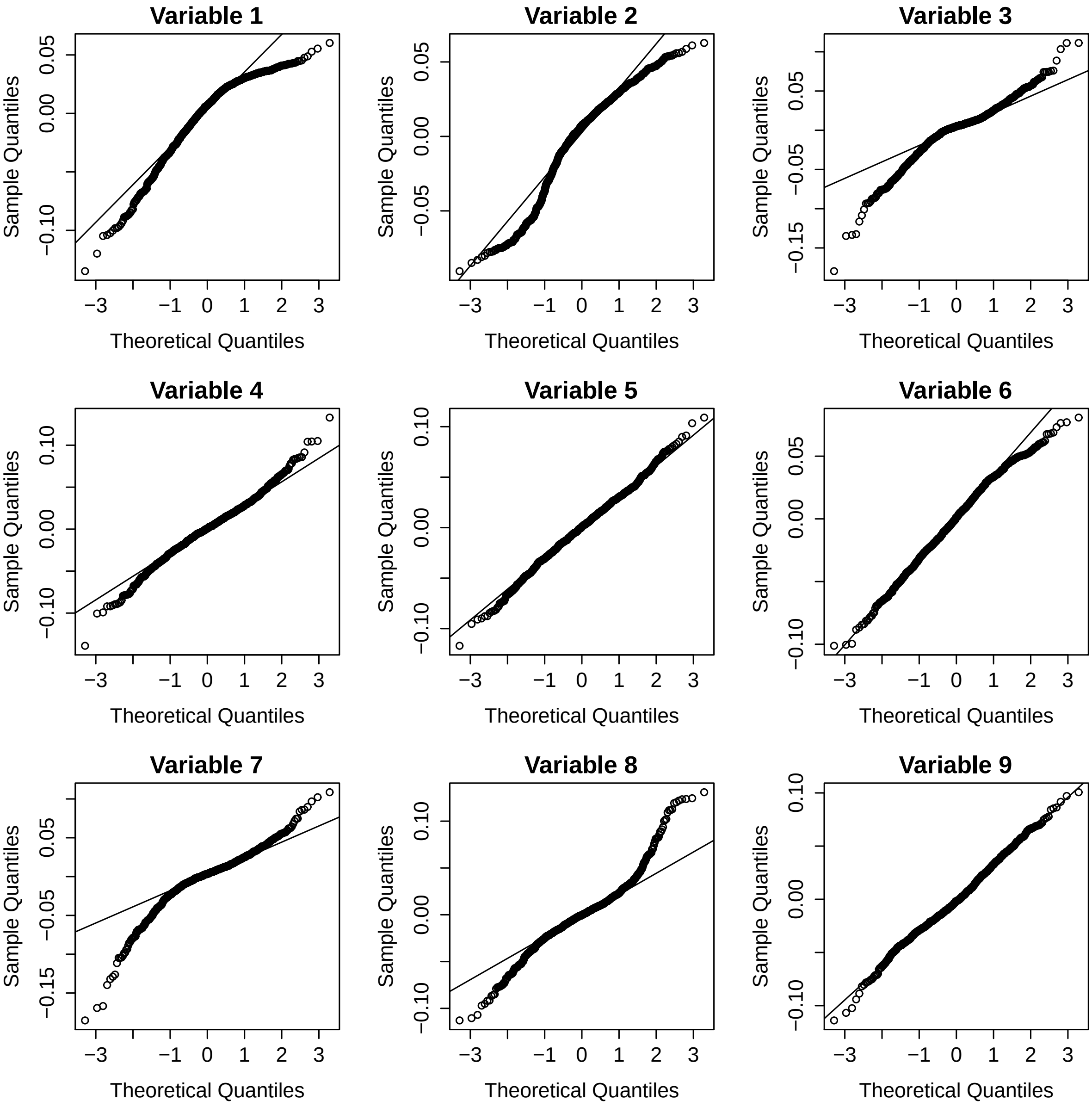

Figure 1: Normal quantile-quantile plots for the first 9 joint component scores. The 5th and 9th do not exhibit clear deviations from normality.

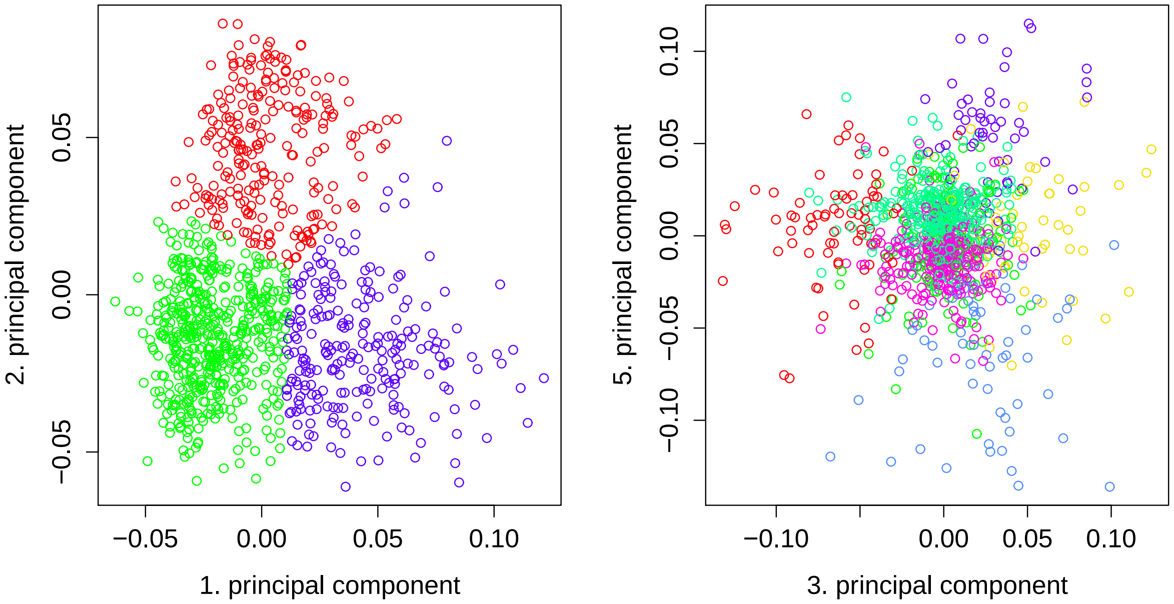

Figure 2: a) The 1. and 2. joint component scores with the three joint clusters in different coloring. b) The 3. and 5. CNA-specific component scores with the seven CNA clusters in different coloring.

Anderson-Darling test, and we therefore determine the rank of the complete joint and individual cluster structure to be $E = 8$. It would also be possible stop at the fifth component, but with an exploratory aim of the analysis and the clear signs of structure in the 7th and 8th component in mind, we choose to include more components.

We examine the qq-plots of the first three component scores of the original expression data. This shows that the first component is clearly non-normal, while the second component is a borderline case and the third component does not deviate significantly from normality. We therefore determine the number of relevant subspaces in the expression data to be $E_1 = 2$. We also examine the qq-plots of the first 8 component scores of the original CNA data. However, when analyzing the CNA data individually, the assumption of normally distributed noise is not properly fulfilled due to the discrete nature of the copy number counts . All of the qq-plots therefore show a clear deviation from normality, and as the total rank of the original data cannot exceed E, we set $E_2 = 8$.

With $E = 8, E_1 = 2, E_2 = 8$, we calculate the number of clusters using (1):

$$K = 3, \quad K_1 = 1, \quad K_2 = 7,$$

meaning we use three joint clusters, no expression-specific clusters and seven CNA-specific clusters.

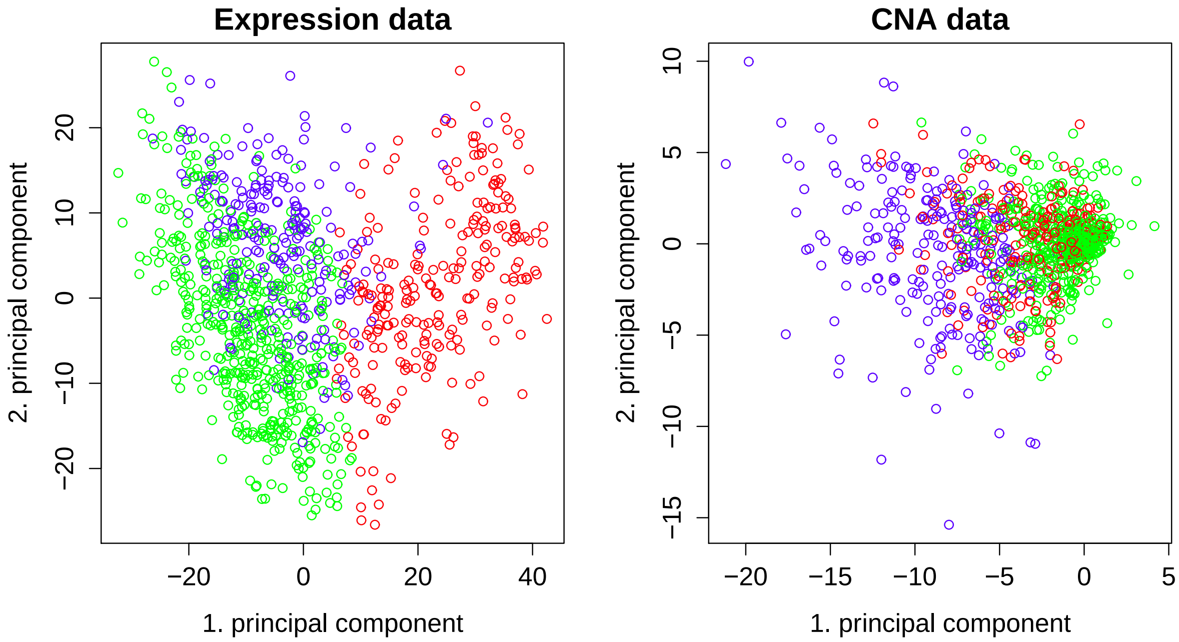

Figure 3: The first and second principal component of the original expression data and the copy number aberrations data, colored with the three joint clusters.

Figure 2a) displays the first and second joint component scores, and we see that the first component discriminates between the 'purple' and 'green' cluster, while the second component separates out the 'red' cluster. Comparing the clusters in terms of clinical covariates, reveals that the 'red' cluster coincide with the Estrogen Receptor (ER) status of the patients, as most ER-negative patient cases are present in the 'red' cluster. Within the PAM50 classification, ER-negative cases are mainly of Basal or HER2-type, meaning the 'red' cluster mainly consists of these two cancer subtypes, as observed in Table 2.

To investigate the relationship between the joint clusters and the original data, Figure 3 displays the first and second principal component scores of the original expression and CNA data with the coloring of the joint clusters. For the expression data, it is clear that the main differences are between the 'red' cluster and the two other clusters. In the CNA data, on the other hand, the observations in the 'red' cluster are randomly scattered, while the two other clusters are quite distinct.

To visualize the seven CNA-specific clusters, we look at the 3rd and 5th component scores, as seen in Figure 2b). For the Figure, it is seen that the 3rd component distinguish between the

Risk	Basal	Her2	LumA	LumB	Normal
High	115	63			37
Low			390	100	
Intermediate			63	152	
Total	118	86	456	268	58

Table 2: The distribution of patients from the PAM classification in the three joint clusters. For clarity, entries constituting less than 10% row-wise are not shown.

'yellow' and 'red' cluster, while the 5th shows the difference between the 'light blue' and 'purple' group. It is also observed that the remaining three clusters, especially the 'green' and 'lilac', are neutral groups situated at the origin.

4.1 Connections with survival, Metabric- and PAM50 classification

The joint and CNA-specific clusters are independently evaluated with regard to survival through Kaplan-Meier estimates. When comparing the three joint clusters against each other and the seven CNA-specific clusters against each other, both clusterings were shown to give significant differences by the logrank test ($p = 8.7 \cdot 10^{-7}$ and $p = 1.8 \cdot 10^{-7}$ for joint and CNA clusters, respectively). Also, when adjusting for age and treatment in a Cox proportional hazards model, both the joint and CNA-specific clusters are significant ($p = 0.02$ and $p = 0.0004$, respectively) by the likelihood ratio test.

Figure 4a) displays the Kaplan-Meier plot of the three joint clusters, revealing the 'red' cluster to be a high mortality risk group, the 'purple' cluster to be an intermediate risk group and the 'green' cluster to be a low risk group. Figure 4b) displays the Kaplan-Meier plot for the seven clusters only present in the CNA data. Interestingly, the two neutral 'dark green' and 'lilac' clusters, situated at the origin of Figure 2b), are low-risk mortality groups. These exhibit few somatic changes in the overall copy number patterns compared to healthy tissue. Conversely, the 'red','blue', 'purple' and 'yellow' groups with quite specific aberration patterns, all exhibit an increased risk of mortality. Especially, the copy number aberrations associated with a negative 3rd component in CNA structure results in highly increased risk, compared to the other groups.

The clusters found by JIC are related to the PAM50 classification (Perou et al., 2000) and the 10 breast cancer subgroups identified by the initial Metabric study Curtis et al. (2012). The Tables 2-5 display the distribution of patients according to the different clusterings.

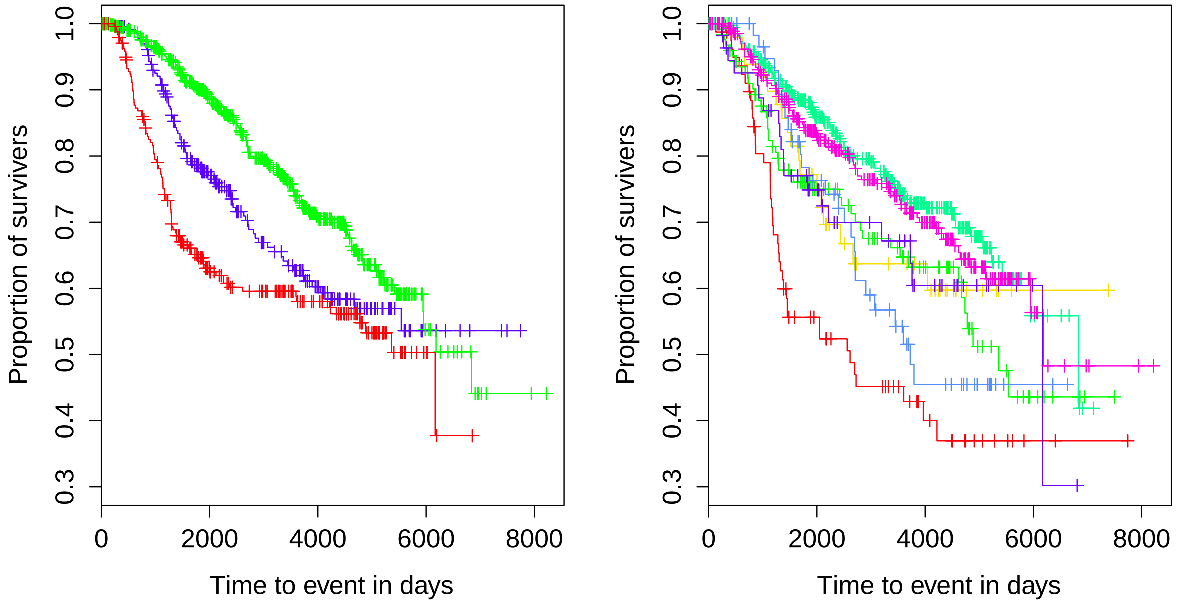

Figure 4: a) A Kaplan-Meier survival plot of the 3 joint clusters. b) A Kaplan-Meier survival plot of the 7 CNA clusters.

Risk	Basal	Her2	LumA	LumB	Normal
Very high (red)		31	13	29	
High (yellow)		6	7	35	
High (light blue)			31	25	
High (purple)			23	26	
High (lime)	19		56	37	
Low (green)	51		184	69	
Low (pink)	36		151	47	
Total	118	86	456	268	58

Table 3: The distribution of patients from the PAM classification in the seven individual clusters. For clarity, entries constituting less than 10% row-wise are not shown.

Risk	1	2	3	4	5	6	7	8	9	10
High				68	50					87
Low			150	95			68	127		
Intermediate	64					32	38		44	
Total	75	45	155	167	94	44	109	143	67	96

Table 4: The distribution of patients from the ten Metabric clusters (Curtis et al., 2012) in the three joint clusters. For clarity, entries constituting less than 10% row-wise are not shown.

Risk	1	2	3	4	5	6	7	8	9	10
Very high (red)					69					
High (yellow)	38									
High (light blue)		40								
High (purple)						34				
High (lime)							33	19	29	18
Low (green)			76	81				66		39
Low (pink)			61	64			37	49		31
Total	75	45	155	167	94	44	109	143	67	96

Table 5: The distribution of patients from the ten Metabric clusters (Curtis et al., 2012) in the seven individual clusters. For clarity, entries constituting less than 10% row-wise are not shown.

Table 2 displays the agreement between the three joint clusters and five subtypes in the PAM50 classification, and it is clear that the high risk cluster consists of Basal, Her2 and Normal-type tumors, while the low and intermediate are dominated by Luminal A and B. The low risk group has a majority of Luminal A cases, while the intermediate group has a majority of Luminal B cases. Table 3 displays the agreement between the seven CNA clusters and PAM50, but we observe no clear patterns here. An interesting observation is that the Basal and Her2 cases do not belong to the same cluster, indicating that the two classes differ in specific copy number alterations as also suggested by the Metabric study (Curtis et al., 2012). The Her2 group is mainly found in the very high risk 'red' group. The Luminal A and B cases are evenly distributed among all the clusters, but with a pivot in the two low risk groups.

Table 4 shows the distribution of patients between the 10 integrative Metabric clusters found by Curtis et al. (2012) and the three joint clusters. Here we observe that the high risk group mainly consists of the Metabric cluster 10, 4 and 5, where the 10th subgroup largely corresponds to the Basal subtype in the PAM50 classification. Further the low risk group consists mainly of Metabric clusters 3 and 8, together with 4 and 7. The intermediate risk group is less clear, but corresponds largely to Metabric clusters 1, 6 and 9.

Table 5 displays the distinct pattern of the correspondence between the ten Metabric clusters and the seven CNA-specific clusters found by JIC. The four groups with the highest risk profile corresponds uniquely to four Metabric clusters: The very high risk 'red' group corresponds to the 5th cluster, the high risk 'yellow' group to the 1st cluster, the high risk 'light blue' group to the 2nd cluster and the high risk 'purple' group to the 6th Metabric cluster. The 9th Metabric cluster is only found as a part of the high risk 'lime' group, while the remaining Metabric clusters 3,4,7,8 and 10 are evenly distributed between the high risk 'lime' group and the two low risk groups.

In conclusion, these observations suggest that there are two independent mechanisms influencing patient survival. From the PAM50 classification, there is a substantial mortality risk difference between the Basal and Her2 on one side and the Luminal A and B on the other. This seems to be the main driver of survival differences, but specific copy number alterations will in addition have an effect. This is seen from the highest risk CNA-specific cluster, which contains a large degree of Luminal A and B (Table 3), but only the 5th Metabric cluster (Table 5). There exist certain copy number aberrations, which override the overall group differences between the Basal/Her2 and the Luminal subtypes. The same reasoning also applies to the other high risk CNA-specific clusters.

5 Discussion

The Joint and Individual Clustering (JIC) contributes to the increased need for integrative procedures within genomics, by decomposing patient samples into joint and individual clusters simultaneously. This improves the understanding of cancer subtypes across genetic data types, as completely independent clusterings can both explain significant differences in survival. This suggests that in addition to clusters of cancer subtypes, found jointly in different data types, there exists, in for instance CNA data, independent groups related to other clinical variables, possibly age, smoking or other environmental influences. The results also agree with earlier analysis of the Metabric data by Curtis et al. (2012), where the iCluster method was used to identify 10 joint clusters. Specifically, four of the seven CNA-specific clusters correspond exactly to four of the joint clusters found by Curtis et al. (2012), suggesting that these are not joint clusters, but instead specific for the CNA data.

The crucial step of how to select the number of clusters proved to be difficult in our setting due to the high-dimensionality of the data. The use of cluster separation measures or cluster reproducibility by sub-sampling did not yield good results within JIC and therefore the more subjective normality-based approach was used. The selection of the number of clusters will always contain subjective aspects, and our selection procedure makes these choices particularly transparent.

Acknowledgment

The authors would like to thank J.S. Marron for pointing out certain of the inner workings of the JIVE algorithm. Funding for the project was provided by the Norwegian Cancer society under award 744088.

References

Arabie, P. and L. Hubert (1996). Advances in cluster analysis relevant to marketing research. In *From Data to Knowledge*, pp. 3–19. Springer.

Curtis, C., S. P. Shah, S.-F. S. Chin, G. Turashvili, O. M. Rueda, M. J. Dunning, D. Speed, A. G. A. G, S. Samarajiwa, Y. Yuan, et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature 486*(7403), 346–352.

Deun, K. V., A. K. Smilde, M. J. van der Werf, A. L. Kiers, and I. V. Mechelen (2009). A structured overview of simultaneous component based data integration. *BMC bioinformatics 10*(1), 246.

Deun, K. V., T. W. R. van den Berg, A. Antoniadis, and I. V. Mechelen (2011). A flexible framework for sparse simultaneous component based data integration. *BMC bioinformatics 12*(1), 448.

Ding, C. and X. He (2004). K-means clustering via principal component analysis. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 29. ACM.

Hamerly, G. and C. Elkan (2003). Learning the k in k-means. In *NIPS*, Volume 3, pp. 281–288.

Hellton, K. and M. Thoresen (2014). Asymptotic distribution of principal component scores for pervasive, high-dimensional eigenvectors. *arXiv preprint arXiv:1401.2781*.

Lee, S., F. Zou, and F. A. Wright (2014). Convergence of sample eigenvalues, eigenvectors, and principal component scores for ultra-high dimensional data. *Biometrika 101*(2), 484–490.

Lock, E. F. and D. B. Dunson (2013). Bayesian consensus clustering. *Bioinformatics 29*(20), 2610–2616.

Lock, E. F., K. A. Hoadley, J. S. Marron, and A. B. Nobel (2013). Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The annals of applied statistics 7*(1), 523–542.

Milligan, G. W. and M. C. Cooper (1987). Methodology review: Clustering methods. *Applied Psychological Measurement 11*(4), 329–354.

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

Shen, R., A. B. Olshen, and M. Ladanyi (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics 25*(22), 2906–2912.

Shen, R., S. Wang, and Q. Mo (2013). Sparse integrative clustering of multiple omics data sets. *The annals of Applied statistics 7*(1), 269–294.

Terada, Y. (2014). Strong consistency of reduced k-means clustering. *Scandinavian Journal of Statistics 10*(3), 515–534.

Tibshirani, R. and G. Walther (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics 14*(3), 511–528.

Tipping, M. E. and C. M. Bishop (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61*(3), 611–622.

Zha, H., X. He, C. Ding, M. Gu, and H. D. Simon (2001). Spectral relaxation for k-means clustering. In *NIPS*, Volume 1, pp. 1057–1064.