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 $\operatorname{Var} \mathbf{v}_1^T \mathbf{x} = \mathbf{v}_1^T \Sigma \mathbf{v}_1$, 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 $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ 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

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

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*

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

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

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

$$(3)$$

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*

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

$$(4)$$

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 $\Delta \mathbf{v}_i$:

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

$$(5)$$

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 $\mathbf{v}_{W,i}$.

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*

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

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

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

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

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

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

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*

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

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

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

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

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

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

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

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 $Var\left(\sigma\mathbf{U}_r\right) = \sigma^2 I$. In the case of uncorrelated, heterogeneous measurement error, the variance of $\mathbf{U}_r$ is different in each variable, such that $\Sigma_U = \mathrm{diag}(c_1, \ldots, c_p)$, where the constants c_k give the relative size of the variances. Then the covariance matrix of the error is $Var\left(\sigma\mathbf{U}_r\right) = \mathrm{diag}(\sigma^2 c_1, \ldots, \sigma^2 c_p)$, 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 $\Sigma_U = \mathrm{diag}(c_1, \ldots, c_p)$, 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 $\mathbf{v}_{ik}^2$ is 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.

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

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

(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

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

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, $\Sigma_U = \text{diag}(c_1, \ldots, c_p)$, 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:

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

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 $\text{Var}(\Delta \mathbf{v}_{ik} \,|\, \mathbf{X})$ in (9). If the error is uncorrelated and homogeneous, the variance is given by

$$\text{Var}(\Delta \mathbf{v}_{ik} \,|\, \mathbf{X}) = \frac{\sigma^2}{n} \sum_{j \neq i} \frac{\lambda_i + \lambda_j}{(\lambda_i - \lambda_j)^2} \mathbf{v}_{jk}^2 + O(\sigma^3).$$

(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

$$\text{Var}(\Delta \mathbf{v}_{ik} \,|\, \mathbf{X}) \simeq \frac{\sigma^2}{\lambda_i} \frac{1}{n}.$$

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

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,

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

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

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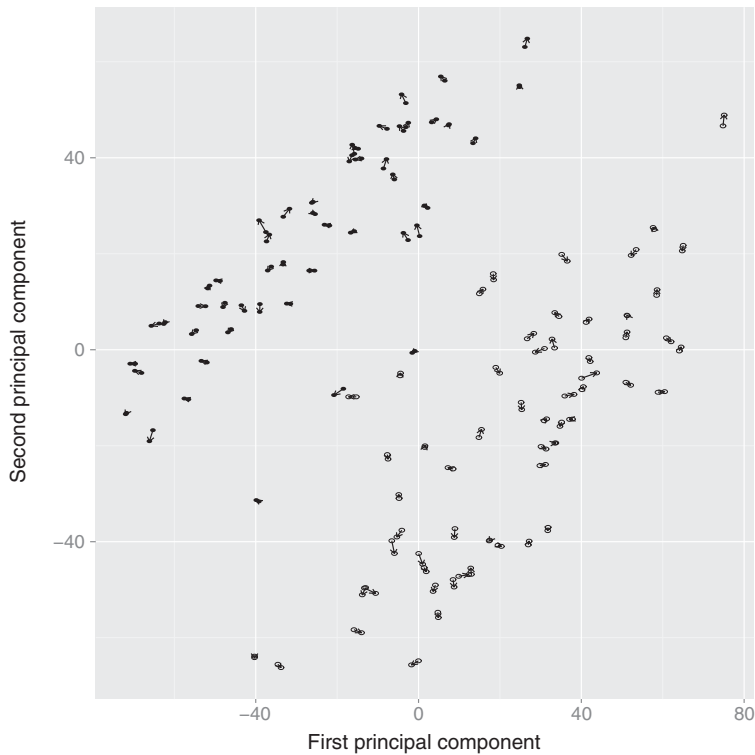
Fig. 1. Plot of first and second component scores from original data and data with simulated error based on the estimated error structure. An arrow indicates the change in scores from the original value. Black dots indicate adenocarcinoma tumour tissue, and open circles indicate non-tumour tissues.

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

absolute value. The attenuation factor gives the expected decrease as $\beta_{1,\mathbf{z}_{W,1}} = \psi\beta_{1,\mathbf{z}_1}$ with $\psi = \operatorname{Var} \mathbf{Z}_1/\operatorname{Var} \mathbf{Z}_{W,1}$, (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 $\mathbf{X}$, 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.

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.

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.

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.