The impact of measurement error on principal component analysis

Kristoffer Hellton, Magne Thoresen

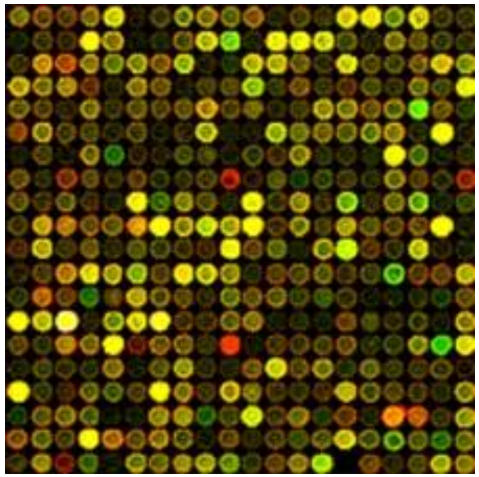
Dep. of Biostatistics, University of Oslo

August 10, 2012

Outline

Background

With genetic microarray data, which is highly affected by measurement error, PCA is a widely technique. In regression setting the measurement error cause bias and lack of power in the naive approach.



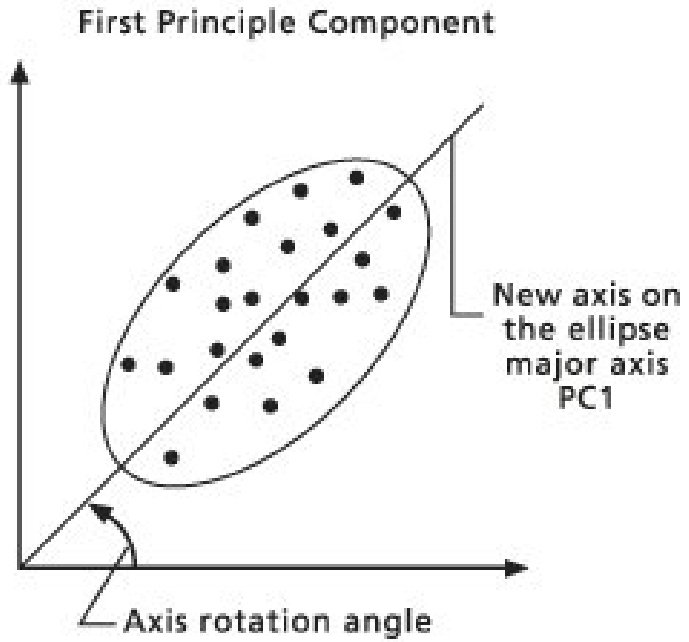
Our aim: To understand the effects of measurement error on principal component analysis.

Principal component analysis

A dimension reduction technique, which transform data, based on the eigendecomposition of the covariance matrix:

$$C_X = V^T \Lambda V$$

- New data: component scores, $Z = V^T X$.
- Weights: loadings, v_i.



Aim to answer the question: How does the error structure affect loadings and scores?

Our approach: Find the bias and variance of the eigenvalues and -vectors induced by measurement error.

Method: perturbation

Assume an additive error model

$$W = X + U,$$

where W is observed under the error U, with some distribution. Then the covariance matrix of W is

$$C_W = C_X + \Delta C.$$

We decompose the eigenvector and -values as

$$\lambda_W = \lambda_X + \Delta\lambda, \quad \mathbf{v}_W = \mathbf{v}_X + \Delta\mathbf{v}.$$

Perturbation

Since the perturbations must satisfy the eigen-equation the first-order approximation relate ΔC to the eigenvalues and -vectors (Stewart and Sun 1990):

$$\Delta\lambda_i = \mathbf{v}_i^T \Delta C \mathbf{v}_i, \qquad \Delta\mathbf{v}_i = \sum_{j \neq i} \frac{\mathbf{v}_j^T \Delta C \mathbf{v}_i}{\lambda_i - \lambda_j} \mathbf{v}_j.$$

To look directly at the impact of measurement error, we condition on X.

Eigenvectors

Let the error be distributed $\mathbb{E}\, U = 0$ and $\mathrm{Var}\, U = \Sigma$, then we obtain for the eigenvectors

$$\mathbb{E}(\Delta \mathbf{v}_i | X) = \sum_{j \neq i} \frac{\mathbf{v}_i^T \Sigma_U \mathbf{v}_j}{\lambda_i - \lambda_j} \mathbf{v}_j,$$

$$\mathrm{Var}(\Delta \mathbf{v}_{ik} | X) = \sum_{j \neq i} K_{ij} \mathbf{v}_{jk}^2,$$

where

$$K_{ij} = \frac{\lambda_j \, \mathbf{v}_i^T \Sigma_U \mathbf{v}_i + \lambda_i \, \mathbf{v}_j^T \Sigma_U \mathbf{v}_j + 2(\mathbf{v}_i^T \Sigma_U \mathbf{v}_j)^2}{n \, (\lambda_i - \lambda_j)^2}.$$

Eigenvalues

For the eigenvalues we obtain

$$\mathbb{E}(\Delta\lambda_i|\mathbf{X}) = \mathbf{v}_i^T\Sigma_U\mathbf{v}_i,$$
$$\text{Var}(\Delta\lambda_i|\mathbf{X}) = \frac{4\lambda_i}{n}\,\mathbf{v}_i^T\Sigma_U\mathbf{v}_i + \frac{2}{n}\left(\mathbf{v}_i^T\Sigma_U\mathbf{v}_i\right)^2.$$

The projected error $\mathbf{v}_i^T\Sigma_U\mathbf{v}_i$ is crucial for all expression.

Homogeneous error

For homogeneous, uncorrelated errors, where $\Sigma_U = \sigma^2 I_p$, the projected error is $\mathbf{v}_i^T \Sigma \mathbf{v}_i = \sigma^2$.

The resulting impact is

- no bias in eigenvectors/loadings
- a loading variance:

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

- equal bias in all eigenvalues $\mathbb{E}(\Delta \lambda_i | \mathbf{X}) = \sigma^2$

Heterogeneous error

For heterogeneous, uncorrelated error, where $\Sigma_U = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$, then the projected error is a weighted mean of the variance.

The resulting impact is

- bias in loadings, approximately

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

- complicated variance expression, but same magnitude as for the homogeneous case
- different bias in eigenvalues, $\mathbb{E}(\Delta \lambda_i | \mathbf{X}) = \sum \sigma_k^2 \mathbf{v}_{ik}^2$

Application: Microarray data

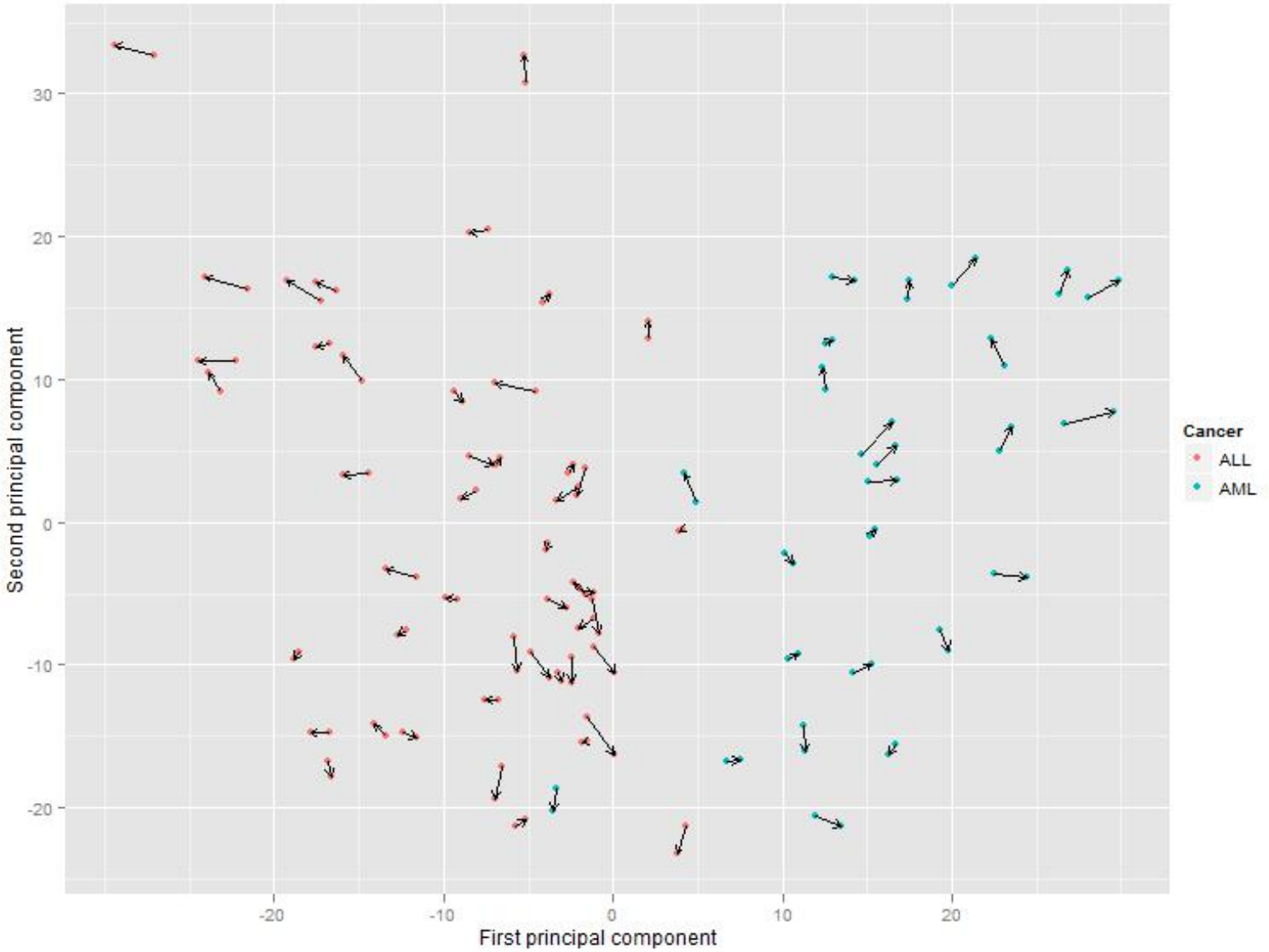
For the high-dimensional data, such as microarray data,

- the variance in loadings becomes larger than the mean loading
- most variables are correlated resulting in high eigenvalues compared to the bias from uncorrelated homogeneous or heterogeneous error.

Example:
The Golub data set with 3050 genes, 72 patients, mean genetic variance $\bar{\sigma}^2 = 0.35$ and largest eigenvalue $\lambda_1 = 183$. We simulate an additive independent and heterogeneous normally distributed error.

Application: Microarray data

Small relative changes in transformed data (arrows)



Summary

- Uncorrelated errors in high-dimensional genetic data have little impact on component scores.
- The variance for each loading will be large relative to the loading value.
- Results are easily extended to correlated errors.
- Further work: characterising the effect of multiplicative error.
- This approach does not take into account the problem of sample and population structure in high-dimensional PCA.

The end

Thank you!

I.T. Jollille, *Principal component analysis*, Wiley Online Library (2002)
G.W. Stewart and J. Sun, *Matrix perturbation theory*,
Academic press New York (1990)
G. Sanguinetti et.al. *Accounting for probe-level noise in principal component analysis of microarray data*, Bioinformatics (2005)
NM Faber et. al. *Random error bias in principal component analysis*,
Analytica chimica acta (1995)