Two Scatter Matrices for Dimension Reduction

K. Nordhausen, E. Liski, and H. Oja

[1] Tampere School of Public Health, University of Finland, 33014 University of Tampere, Finland

Abstract

Dimension reduction plays an increasingly important role in high dimensional data analysis. In this talk we will revisit PCA, ICA and SIR and show that they all can be seen within the framework of invariant coordinate selection (ICS) and supervised invariant coordinate selection (SICS).

Let $\mathbf{x}$ be a q-variate random vector and assume for simplicity that $E(\mathbf{x}) = \mathbf{0}$. Then PCA finds a transformation matrix $\mathbf{A}$ (and a transformation $\mathbf{x} \to \mathbf{z} = \mathbf{Ax}$) such that $\mathbf{AI}_q\mathbf{A}' = \mathbf{AA}' = \mathbf{I}_q$ and $\mathbf{A}E(\mathbf{xx}')\mathbf{A}' = \mathbf{L}$ where $\mathbf{L}$ is a diagonal matrix with positive diagonal elements in a decreasing order, i.e. $l_1 \geq ... \geq l_q > 0$. The transformation matrix $\mathbf{A}$ is orthogonal and the components in the new coordinate system are ordered according to their variances l_i, $i = 1, ..., q$. If *"high variation means high information content"*, one should use then components with the highest variances in the future analysis of data.

In ICA the so-called fourth-order blind identification (FOBI) procedure finds another $\mathbf{A}$ (and another transformation $\mathbf{x} \to \mathbf{z} = \mathbf{Ax}$) such that now $\mathbf{A}E(\mathbf{xx}')\mathbf{A}' = \mathbf{I}_q$ and $\mathbf{A}E(\mathbf{xx}'\mathbf{xx}')\mathbf{A}' = \mathbf{L}$ with another diagonal $\mathbf{L}$ with diagonal elements $l_1 \geq ... \geq l_q > 0$. In ICA, the new components are ordered according to their kurtosis values. If *"high/low kurtosis means high information content"*, then components with extreme kurtosis should be kept for the future analysis of the data.

The variables in the new coordinate system are however often used to predict the value of a p-variate response variable $\mathbf{y}$. The joint distribution of $\mathbf{x}$ and $\mathbf{y}$ should in this case be used in dimension reduction for $\mathbf{x}$. SIR is again based on the comparison of the values of two scatter matrices; now one finds a transformation matrix $\mathbf{A}$ (and a transformation $\mathbf{x} \to \mathbf{z} = \mathbf{Ax}$) such that $\mathbf{A}E(\mathbf{xx}')\mathbf{A}' = \mathbf{I}_q$ and $\mathbf{A}E\left(E(\mathbf{x}|\mathbf{y})E(\mathbf{x}|\mathbf{y})'\right)\mathbf{A}' = \mathbf{L}$ where $\mathbf{L}$ is again a diagonal matrix with diagonal elements $l_1 \geq ... \geq l_q \geq 0$. The coordinates with zero eigenvalues l_i do not carry any information about the dependence between $\mathbf{x}$ and $\mathbf{y}$.

Basically all the classical approaches above (PCA, ICA, and SIR) are special cases of ICS or SICS. Our approaches uses two different location vector functionals and two different scatter matrix functionals. Let $\mathbf{T}_1$ and $\mathbf{T}_2$, and $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, be the values of these four functionals at the distribution of $\mathbf{x}$. We then find a transformation $\mathbf{x} \to \mathbf{z} = \mathbf{Ax} + \mathbf{b}$ such that

$$\mathbf{AT}_1 + \mathbf{b} = \mathbf{0}, \quad \mathbf{A}(\mathbf{T}_2 - \mathbf{T}_1) \geq \mathbf{0}, \quad \mathbf{AS}_1\mathbf{A}' = \mathbf{I}_q, \quad \text{and} \quad \mathbf{AS}_2\mathbf{A}' = \mathbf{L}.$$

Here $\mathbf{L}$ is again a diagonal matrix with diagonal elements $l_1 \geq ... \geq l_q \geq 0$. Affine transformation $\mathbf{x} \to \mathbf{z} = \mathbf{Ax} + \mathbf{b}$ transforms the variable to an invariant coordinate system. If the two scatter matrices have the so-called independence property, then the invariant coordinate system is a solution of the independent component analysis (ICA) problem. The invariant coordinate system is called supervised (SICS) if the second scatter matrix is supervised as in SIR (depends on the joint distribution of $\mathbf{x}$ and $\mathbf{y}$).

References

E. Liski, K. Nordhausen, and H. Oja (2010). Supervised invariant coordinate selection. Manuscript.

K. Nordhausen, H. Oja, and E. Ollila (2010). Multivariate Models and the First Four Moments. Festschrift in Honour of Tom Hettmansperger, to appear.

D. E. Tyler, F. Critchley, L. Dümbgen, and H. Oja (2009). Invariant coordinate selection. Journal of the Royal Statistical Society, Series B, 71, 549-592.