High-dimensional dimension reduction

Joni Virta
Department of Mathematics and Statistics

15th December 2022

The 1st Science Seminar of the Faculty of Science

Premise

- A common denominator behind all (natural) sciences:

Premise

- A common denominator behind all (natural) sciences:

$\boxed{\text{Data from } n \text{ observational units and } p \text{ variables.}}$

Premise

- A common denominator behind all (natural) sciences:

Data from n observational units and p variables.

- This talk discusses a specific **shift/trend** in the way statistical methodology is viewing data.

Classical statistics

- Classical statistical procedures assume that the sample size $n \to \infty$.

$$\downarrow \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

Classical statistics

- Classical statistical procedures assume that the sample size $n \to \infty$.

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

- Increasing amount of information smooths out randomness and lets us make precise probabilistic statements.
- Simple interpretation: We recruit more subjects.

High-dimensional statistics

- High-dimensional methodology assumes $n \to \infty$ AND $p \to \infty$!

$$
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1p} & \cdots \\
x_{21} & x_{22} & \cdots & x_{2p} & \cdots \\
\vdots & \vdots & \ddots & \vdots & \cdots \\
x_{n1} & x_{n2} & \cdots & x_{np} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix}
$$

High-dimensional statistics

- High-dimensional methodology assumes $n \to \infty$ AND $p \to \infty$!

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} & \cdots \\ x_{21} & x_{22} & \cdots & x_{2p} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

- Current literature on statistical methodology has a strong emphasis on such *high-dimensional regimes*.

Practical considerations

- Targeted to (and motivated by) datasets where
  1. The sample size n is large, and
  2. The number of variables p is large **relative to** n.

Practical considerations

- Targeted to (and motivated by) datasets where
  1. The sample size n is large, and
  2. The number of variables p is large **relative to** n.
- For example, *microarray gene expression data* can have $n \sim 100$ and $p \sim 10000$.

Practical considerations

- Targeted to (and motivated by) datasets where
  1. The sample size n is large, and
  2. The number of variables p is large **relative to** n.
- For example, *microarray gene expression data* can have $n \sim 100$ and $p \sim 10000$.
- High-dimensional (HD) versions of classical methods, e.g., correlation matrix estimation, are being developed.

Practical considerations

- Targeted to (and motivated by) datasets where
    1. The sample size n is large, and
    2. The number of variables p is large **relative to** n.
- For example, *microarray gene expression data* can have $n \sim 100$ and $p \sim 10000$.
- High-dimensional (HD) versions of classical methods, e.g., correlation matrix estimation, are being developed.
- HD methods typically yield significantly more useful results on HD data sets, compared to classical methods.

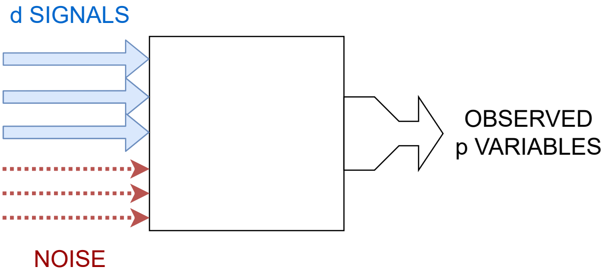
Theoretical considerations

1. Unexpected phenomena:
   - Basic estimators (means) might no longer converge.
   - Standard techniques, such as the central limit theorem, no longer work.

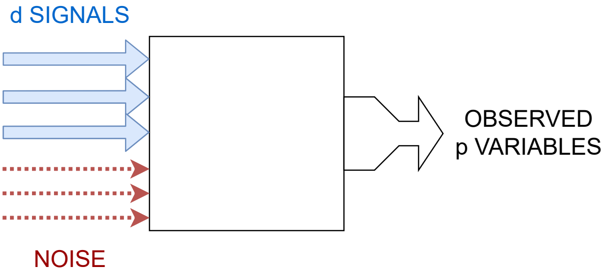
Theoretical considerations

1. Unexpected phenomena:
   - Basic estimators (means) might no longer converge.
   - Standard techniques, such as the central limit theorem, no longer work.
2. Random matrix theory allows constructing new tools.

Dimension reduction

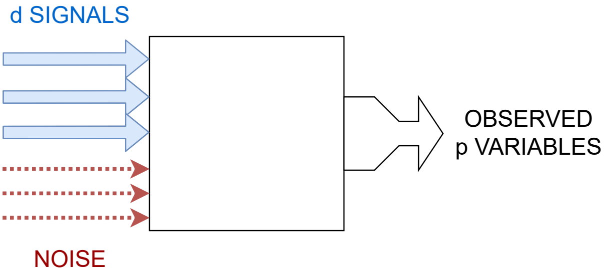


Dimension reduction



- Dimension reduction (DR) attempts to recover a low-dimensional signal behind the data, without any loss of important information.

Dimension reduction



- Dimension reduction (DR) attempts to recover a low-dimensional signal behind the data, without any loss of important information.
- This makes DR especially valuable in high-dimensional data analysis.

Principal component analysis

- *Principal component analysis* (PCA) is a classical dimension reduction method.

Principal component analysis

- *Principal component analysis* (PCA) is a classical dimension reduction method.
- A key question in PCA is how to choose the number of signals.

Principal component analysis

- *Principal component analysis* (PCA) is a classical dimension reduction method.
- A key question in PCA is how to choose the number of signals.
- A modification of a certain classical method can be used to construct the high-dimensional estimator $\hat{d}$ for their number (Schott, 2006).

A result

Theorem (Schott, 2006; Virta, 2021)

Assume that the ratio

$$\frac{p}{n} \to c,$$

where either $c = 0$, $c \in (0, \infty)$ or $c = \infty$.

Under certain regularity conditions, the limiting behavior of the estimator $\hat{d}$ is identical in all three cases.

Schott, J. R. (2006). A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *Journal of Multivariate Analysis*, 97(4):827–843.

Virta, J. (2021). Testing for subsphericity when *n* and *p* are of different asymptotic order. *Statistics & Probability Letters*, 179.