# Applied Multivariate Techniques

Daniele Zago

February 11, 2022

# Contents

# Lecture 1: Principal component analysis

2022-01-13

Consider a sample of $n$ observations of $p$ variables, then we usually define the observed data matrix as the following quantity,

$$X_{n\times p} = \begin{pmatrix} x_{11} & \dots & x_{ip} \\ \\ x_{n1} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix}.$$

In the following sections, we briefly review a set of matrix definitions, identities, and properties that we will find useful throughout the course.

## 1.1 Matrix algebra review

**Def. (Orthogonal matrix)**
A square matrix $Q$ is called ***orthogonal*** if $Q^\top Q = I$.

**Properties**

- $Q^{-1} = Q^\top$
- $QQ^\top = I$
- $|Q| = \pm 1$
- $A, B$ orthogonal matrices, then $A^\top B = C$ is still an orthogonal matrix.

  *Proof.*
  $C^\top C = B^\top A A^\top B = B^\top B = I$.

  □

**Def. (Semi-orthogonal matrix)**
A matrix $Q_{n\times p}$ is called ***semi-orthogonal*** if either

$$Q^\top Q = I_p \quad \text{and} \quad QQ^\top \neq I_n$$

or

$$Q^\top Q \neq I_p \quad \text{and} \quad QQ^\top = I_n$$

**Def. (Eigenvalue)**
Let $A$ be a $p \times p$ square matrix, then the roots $\{\lambda_1, \dots, \lambda_p\} \in \mathbb{C}$ of the characteristic equation

$$\det(A - \lambda I) = 0$$

are called ***eigenvalues*** .

**Def. (Eigenvector)**

Counting multiplicity, for each $\lambda_i$ eigenvalue there exists a unique ***eigenvector*** $\gamma_i$ associated to $\lambda_i$ such that

$$A\gamma_i = \lambda_i \gamma_i.$$

The uniqueness is constrained to $\gamma_i^\top \gamma_i = 1$.

**Remark** If $A$ is symmetric, then $\lambda_i \in \mathbb{R}$ for all $i$.

**Remark** If all $\lambda_i > 0$ we have that $x^\top A x > 0$ for any $x \in \mathbb{R}^p$ and $A$ is called ***positive-definite***. If all $\lambda_i \geq 0$, then $x^\top A x \geq 0$ and $A$ is ***positive-semidefined***.

**Def. (Rank)**

The ***rank*** of $A$ is defined as $\operatorname{rank} A = \#(\lambda_i > 0)$.

**Properties** If $\lambda_i$ are eigenvalues for $A$, then

1. *Addition*: $|A + \alpha I - (\lambda + \alpha)I| = 0 \implies \alpha + \lambda_i$ are eigenvalues of $A + \alpha I$.
2. *Multiplication*: $|\alpha A - \alpha \lambda I| = 0 \implies \alpha \cdot \lambda_i$ are eigenvalues of $\alpha A$.

**Theorem 1 (Spectral decomposition)**

*A symmetric matrix $A$ can be written in terms of its eigenvalues and eigenvectors as*

$$A_{p \times p} = \Gamma \Lambda \Gamma^\top = \sum_{i=1}^{p} \lambda_i \gamma_i \gamma_i^\top,$$

*where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ and $\Gamma = (\gamma_1 \ \gamma_2 \ \ldots \ \gamma_p)$ is orthonormal, where eigenvalues and eigenvectors are counted with multiplicity and $\lambda_1 > \lambda_2 > \ldots > \lambda_p$.*

**Power** With the above decomposition, we have a fast way of computing $A^q$,

$$A^q = \Gamma \Lambda^q \Lambda^\top.$$

If $A \succ 0$ then $q \in \mathbb{Q} \setminus 0$, else if $A \succeq 0$, then $q \in \mathbb{Q}^+$.

**Principal components** We have that $\gamma_1$ is the solution to the following maximization problem

$$\gamma_1 = \operatorname*{argmax}_x x^\top A x \implies \gamma_1^\top A \gamma_1 = \gamma_1^\top \Gamma \Lambda \Gamma^\top \gamma_1 = \lambda_1.$$

The second eigenvalue maximizes

$$\gamma_2 = \operatorname*{argmax}_{\substack{x^\top x = 1 \\ x^\top \gamma_1 = 0}} x^\top A x$$

## 1.2 Singular value decomposition

**Theorem 2 (Singular value decomposition)**
*Let $X_{n\times p}$ be a general matrix, then $X$ can be written as*

$$X_{n\times p} = U_{n\times n} D_{n\times p} V'_{p\times p} = \sum_{i=1}^{\min\{n,p\}} d_i u_j v_i^\top,$$

*where $UU^\top = I_n$, $VV^\top = I_p$, and*

$$D = \begin{pmatrix} d_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & d_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & d_3 & 0 & \dots & 0 \\ 0 & 0 & 0 & d_4 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & d_{\min\{n,p\}} \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & & & & & \end{pmatrix}$$

**Remark** The matrix $D$ has lots of zeros, therefore if we set $d_i = 0$ for $i \geq h$, effectively truncating the approximation to the first $h$ components, we obtained a compressed representation of $X$.

**Example (Linear model)**
Consider a linear model $y = X\beta + \varepsilon$, and let $P$ be the projection matrix on the parameter model estimates, i.e.

$$Py = X(X^\top X)^{-1} X^\top y,$$

and consider now the singular value decomposition of $X = UDV^\top$:

$$\begin{aligned} P &= UDV^\top (VDU^\top UDV^\top)^{-1} VDU^\top \\ &= UDV^\top (VD^2V)^{-1} VDU^\top \\ &= UDV^\top VD^{-2} VVDU^\top \\ &= UDD^{-2}DU^\top \\ &= UU^\top. \end{aligned}$$

We have that $U_{n\times p}$ is a semi-orthogonal matrix, therefore $U^\top U = I_p$ but $UU^\top \neq I_n$.

Suppose now that we apply a linear transformation on $X$ before computing the estimates, i.e.

$$Z = XC, \quad C \text{ orthogonal and } \operatorname{rank} C = p,$$

then $Z = UD(VC)^\top = UDC^\top V^\top$ and the projection matrix $P_Z$ can be calculated as

$$\begin{aligned} P_Z &= UDC^\top V^\top (VCDU^\top UDC^\top V^\top)^{-1} VCDU^\top \\ &= UDC^\top V^\top VCD^{-2}C^\top V^\top VCDU^\top \\ &= UU^\top \end{aligned}$$

**Remark** The above result is slightly more complicated but still holds if $C$ is not orthogonal but has rank $p$.

**Exercise:** Let $P$ be the projection matrix of rank $p$, then prove that the eigenvalues are all 1 and that the ***residual maker matrix*** $I - P$ has rank $n - p$ and $\lambda_1, \dots, \lambda_{n-p} = 1$.
Use the properties of the eigenvalues (addition/multiplication).

**Def. (Centering matrix)**
Consider the matrix $H = \frac{1}{n}\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix} = \mathbb{1}(\mathbb{1}^\top \mathbb{1})^{-1}\mathbb{1}^\top$, which is the linear model when only the intercept term is available. Then, $I - H$ is the ***centering matrix*** and

$$X_C := (I - H)X,$$

which has column-wise zero mean.

**Notation** Starting from now, we will not use a centering matrix anymore, and we assume that $X$ has been already centered.

**Def. (Variance-covariance matrix)**
For a centered matrix $X$ we define $\frac{1}{n}X^\top X$ as the ***variance-covariance*** matrix of $X$.

We want to look at the first principal component, which is defined as the direction $g$ such that

$$\begin{aligned} \widehat{g} &= \underset{g}{\operatorname{argmax}}\, \mathbb{V}[Xg] \\ &= \underset{g}{\operatorname{argmax}}\, g^\top X^\top X g, \end{aligned}$$

and this problem has the solution given by the singular value decomposition of $X^\top X$. Let

$$X^\top X \overset{\text{s.d.}}{=} \Gamma \Lambda \Gamma^\top,$$

we know that the maximum is attained in the first eigenvector, $\widehat{g} = \gamma_1$.

**Def. (Principal components)**
The $j^{\text{th}}$ principal component is the $j^{\text{th}}$ direction of maximum variance constrained to being uncorrelated with the previous $j-1$ directions of maximum variance, and

$$g_j = \gamma_j,$$

where $\gamma_j$ are the eigenvectors of the SVD of $X^\top X$.

The variance of the $i^{\text{th}}$ principal components are given by

$$X\gamma_i = \frac{1}{n}\lambda_i.$$

The fraction of explained variance by the $i^{\text{th}}$ principal component is

$$\%\mathbb{V}_i = \frac{\lambda_i/n}{\operatorname{tr}(\Lambda)/n} = \frac{\lambda_i}{\sum_{i=1}^p \lambda_i}.$$

We now describe the connection between the SVD and $X$ and the SVD of $X^\top X$:

$$X^\top X \xrightarrow{\text{sp. dec.}} \Gamma\Lambda\Gamma^\top$$

$$X \xrightarrow{\text{sing. val.}} UDV^\top$$

Then, we have that

$$X^\top X = VDU^\top UDV^\top = VD^2V,$$

therefore the singular value decomposition of $X$ is such that $V = \Gamma$ and $D = \Lambda^2$.
In general, the principal components of $X$ are defined by $UD$, therefore we can obtain them by applying the transformation

$$UD = X\Gamma.$$

**Example (Problems with SVD)**
Suppose we have a biometric test where we have different unit of scale: if we change the unit of measurement, we get different results in terms of principal components.

To do so, we usually apply the SVD to the standardized variables whenever we do not have variables on the same scale.

**Exercise** Prove that $P = X(X^\top X)^{-1}X^\top$ has rank $p$, then prove that $(I-P)$ has rank $n-p$ and find the possible eigenvalues of $P$ and $I-P$.

*Proof.*
Since $P$ is the projection matrix on $\langle x_1, x_2, \dots, x_p\rangle$, we have that

$$\operatorname{rank} P = \operatorname{rank} X = p.$$

Moreover, we have that $I - P$ is the projection matrix on the orthogonal subspace $\langle x_1, x_2, \dots, x_p \rangle^\perp$, which is a linear subspace of dimension $n - p$, and therefore $\operatorname{rank} I - P = n - p$.

Since the projection matrix $P$ is such that $P = P^2$, we have that if $\lambda$ is an eigenvalue of $P$ relative to an eigenvector $v$, then

$$\lambda^2 v = P^2 v = Pv = \lambda v,$$

hence $\lambda^2 = \lambda$, and this can only happen if either $\lambda = 0$ or $\lambda = 1$. The same applies for $(I - P)$, since $(I - P)^2 = (I - P)$.

□

**Exercise** Let $X_{n\times p}$ and $P_X = X(X^\top X)^{-1}X^\top$ be the projection matrix, let now $R$ be a rotation matrix such that $R^\top R = I$ and $RR^\top = I$. Define $Y = XR$, prove that $P_Y = Y(Y^\top Y)^{-1}Y^\top = P_X$.

*Proof.*

$$\begin{aligned}
P_Y &= Y(Y^\top Y)^{-1}Y^\top \\
&= XR(R^\top X^\top XR)^{-1}R^\top X^\top \\
&= XRR^\top(X^\top X)^{-1}RR^\top X^\top && \text{(since } R^{-1} = R^\top\text{)} \\
&= X(X^\top X)^{-1}X^\top.
\end{aligned}$$

□

# Lecture 2: Multidimensional Scaling

2022-01-20

MultiDimensional Scaling (MDS) is a technique which starts from an observed matrix $D$ of pairwise distances and aims to reconstruct an approximate low-dimensional ***configuration*** of points which could have produced $D$. This in turn is very useful for obtaining a low-dimensional representation of the data in order to visualize clusters and extract relevant information.

Suppose that we have $x_1, \dots, x_n$ observations in a general space $\mathbb{R}^p$, and we know the distances between each pair of elements $d_{ij} = d(x_i, x_j)$. This distance can be any arbitrary distance function, as long as it satisfies the three following properties

1. $d_{ij} \geq 0$ and $d_{ij} = 0 \iff i = j$.
2. $d_{ij} = d_{ji}$
3. $d_{ij} + d_{jk} \geq d_{ik}$

**Example (Euclidean distance in $\mathbb{R}^p$)**
If $x_i, \dots, x_n \in \mathbb{R}^p$, then

$$d_{ij} = \sqrt{(x_i - x_j)^\top (x_i - x_j)} = \|x_i - x_j\|_2.$$

**Note** Since we can start from an arbitrary distance matrix $D$, it's possible to apply the multidimensional scaling even without knowing *a*) the original data which produces $D$ and *b*) the true dimension of the underlying space.

**Def. (Multidimensional scaling)**
Consider the observed symmetric square matrix of distances $D_{n\times n} = (d_{ij})_{i,j=1,\dots,n}$, the ***multidimensional scaling*** (MDS) procedure aims to obtain a low-dimensional representation $z_1, \dots, z_n \in \mathbb{R}^k$ such that

$$z_1, \dots, z_n = \operatorname*{argmin}_{v_1,\dots,v_n} \sum_{i,j} (d_{ij} - \|v_i - v_j\|_2)^2. \tag{1}$$

**Interpretation** With the above minimization, we obtain a low-dimensional representation of the higher-dimensional observed data. The obtained configuration is thus as similar as possible in terms of distance structure to the original points $x_1, \dots, x_n$.

**Notation**

- We denote by $D_2 = (d_{ij}^2)_{i,j}$ the matrix of squared distances, and note that $D_2 \neq D^2$.
- We also define the residualizing matrix by $H = I - \frac{1}{n}\mathbb{1}\mathbb{1}^\top = I - \frac{1}{n}J$

- Finally, we define $B = -\frac{1}{2}HD_2H$, which is a ***double-centering*** of $D_2$. The resulting row-wise and column-wise sums are both zeros:

$$\mathbb{1}^\top HD_2H = \mathbf{0}$$
$$HD_2H\mathbb{1} = \mathbf{0}^\top$$

There is a very strong connection between the principal component analysis and the multidimensional scaling.

**Def. (Euclidean matrix)**
We say that the matrix $D = (d_{ij})_{i,j}$ is ***euclidean*** if there exists a configuration $z_1, \ldots, z_n \in \mathbb{R}^p$ such that $d_{ij} = \|z_i - z_j\|_2$

**Note** In the following, we denote by $Z$ the matrix of the corresponding configuration of $n$ vectors,

$$Z = \begin{pmatrix} z_1^\top \\ z_2^\top \\ \vdots \\ z_n^\top \end{pmatrix}. \tag{2}$$

**Theorem 3 (Euclidean matrix and $B$ matrix)**
*Let $D$ be a matrix and define $B = -\frac{1}{2}HDH$, then $D$ is euclidean $\iff$ $B$ is positive semidefinite. We call the matrix $B$ the* ***inner product matrix****.*

*Proof.*
Since $-2B = HD_2H$ and $H = I - \frac{1}{n}J$, then

$$\begin{aligned} -2B &= D_2H - \frac{1}{n}JD_2H \\ &= D_2 - \frac{1}{nJ} - \frac{1}{n}JD_2 + \frac{1}{n}JD_2J. \end{aligned}$$

For each element of $-2B$, we have

$$(-2B)_{ij} = d_{ij}^2 - \frac{1}{n}\sum_h d_{ih}^2 - \frac{1}{n}\sum_k d_{kj}^2 + \sum_h \sum_k \frac{1}{n^2} d_{hk}, \tag{3}$$

now since $D$ is euclidean, we can express $d_{ij}$ in terms of a distance between each element $z_i$ and $z_j$, $d_{ij}^2 = (z_i - z_j)^\top (z_i - z_j) = z_i^2 - 2z_i z_j + z_j^2$, hence

$$\frac{1}{n}\sum_h d_{ih}^2 = \frac{1}{n} n z_i^2 + \sum_h \frac{z_h^2}{n} - 2z_i \frac{1}{n}\sum_h z_h$$

$$= z_i^2 + \sum_{i=1}^{h} \frac{z_h^2}{n} - 2z_i \bar{z}$$

$$\frac{1}{n}\sum_{h,k} d_{hk} = \frac{\sum_h z_h^2}{n} + \frac{\sum_h z_h^2}{n} - 2\bar{z}^2$$

If we substitute the above terms in Equation (3), then we obtain (exercise)

$$(-2B)_{ij} = -2(z_i - \bar{z})^\top (z_j - \bar{z}).$$

□

**Note** We have that $B = (b_{ij})_{i,j} = (z_i^\top z_j)_{i,j}$, hence the name *inner product matrix*,

$$B = HZ(HZ)^\top.$$

## 2.1 Relationship with PCA

The following theorem states the link between the metrix MDS and the principal component, and gives an algorithm for immediately obtaining the solution to the MDS problem (1).

**Theorem 4 (MDS and principal components)**
*Let $D$ be euclidean, then if we define $Z$ as (2), we have that if $B \geq 0$, then there exists $Z = US$ such that*

$$B = US^2U^\top,$$

*where $UU^\top = I$ and $S^2 = \operatorname{diag}(s_1^2, s_2^2, \ldots, s_k^2)$*

**Remark** From the above theorem, if we compute the singular value decomposition on a positive-semidefined $B = -\frac{1}{2}HD_2H$, then we obtain a representation $Z = US$ which minimizes the multidimensional scaling problem.

**Low-dimension** If we choose a lower-dimensional representation, say $z_1, \ldots, z_n \in \mathbb{R}^k$ with $k < p$, then we obtain the *optimal* configuration with minimal discrepancy from the observed matrix $D$.

*Proof.*
Define $B = US^2U^\top = ZZ^\top$, where $Z = US$. Then, we know that if we write

$$(z_i - z_j)^\top (z_i - z_j) = z_i^2 + z_j^2 - 2z_i z_j$$

$$= b_{ii} + b_{jj} - 2b_{ij},$$

but then we can write each $b_{ij}$ in terms of the distances, since $B = -\frac{1}{2}HD_2H$. Check that

$$\begin{aligned} b_{ij} &= d_{ij}^2 - \frac{1}{n}\sum_h d_{ih}^2 - \frac{1}{n}\sum_k d_{kj}^2 + \frac{1}{n^2}\sum_{h,k} d_{hk}^2 \\ &= -\frac{1}{2}(-2d_{ij}^2) \\ &= d_{ij}^2. \end{aligned}$$

□

## 2.2 Non-metric MDS

*References* Chen and Buja (2013)

The above discussion states the optimality of metric MDS, i.e. when $D$ is euclidean, and its equivalence to principal component analysis. However, most of the times $D$ is not euclidean and the resulting matrix $B = -\frac{1}{2}HD_2H$ is not guaranteed to have non-negative eigenvalues. Therefore, a lot of research has developed non-metric variants of the multidimensional scaling procedure, which extend its analysis to more general ***dissimilarity metrics***.

**Exercise** On Moodle, try to analyze the uploaded dataset using the MDS approach.

## 2.3 Stress function for nonlinear MDS

1. Classical scaling indirectly approximate the distance through inner products using eigendecompositions.
2. Distance scaling tries to approximate the target distance using high-dimensional approximation.

$$\boldsymbol{X} = \underset{\boldsymbol{x}}{\operatorname{argmin}} \sum_{i,j} \|D - d_{ij}\|_2$$

We can consider the following stress function in terms of the found solutions $d_{ij}$'s and the observed matrix $D = (D_{ij})$,

$$\begin{aligned} S(d|D) &= \sum_{i,j}(d_{ij} - D_{ij})^2, \\ &= \sum_{i,j} \underbrace{d_{ij}^2}_{\text{attractive energy}} - 2\underbrace{D_{ij}d_{ij}}_{\text{repulsing energy}}. \end{aligned}$$

Since this function can be interpreted in terms of attractive and repulsive energies between nodes, it can be optimized using techniques from the graph-drawing literature. There is no universally better stress function, therefore some solutions have been proposed in the literature:

- Embed stress functions in a parametric family, avoiding *ad hoc* choices.
- Measure goodness of stress choice using meta-criteria.

### 2.3.1 Parametric stress functions

We can use the Box-Cox transformation to define a family of stress functions parametrized by $\alpha \in \mathbb{R}$,

$$\mathrm{BC}_\alpha(d) = \begin{cases} \frac{d^\alpha - 1}{\alpha} & \alpha \neq 0 \\ \log d & \alpha = 0 \end{cases}$$

which includes the following stress functions:

1. Power laws and logarithmic laws
2. Power law for up- or down-weighting of small/large distances
3. Regularization parameter for incomplete distance data.

$$S(d|D) = \sum_{i,j} D_{ij}^{\nu} \left( BC_{\mu+\lambda}(d_{ij}) - D_{ij}^{\lambda} \mathrm{BC}_\mu(d_{ij}) \right),$$

where $\mu$ is a repulsive strength, $\lambda$ is the relative strength btw attracting and repulsive force, and $\nu$ is the weight parameter.

**Prop. 1 (Edgewise unbiasedness)**
*All BC stress functions are minimized by the embeddings that produces exactly $D$.*

The parameters produce different type of compromises.

The BC stress functions can be extended to incomplete data by imputing missing information using an infinitesimally-small weight,

$$S(d|D) = \sum_{i,j \in E} D_{ij}^{\nu} \left( BC_{\mu+\lambda}(d_{ij}) - D_{ij}^{\lambda} \mathrm{BC}_\mu(d_{ij}) \right) - t^{\nu-\lambda} \sum_{i,j \notin E} \mathrm{BC}_\mu(d_i, j),$$

where $t$ is a balancing parameter.

The choice of parameters can be guided by meta-criteria based on the KNN embedding. The idea is to define two neighborhoods for each point $i$, $\mathcal{N}_D(i)$ and $\mathcal{N}_d(i)$ based on $D_{ik}$ and $d_{ij}$ respectively, and to compare the observed overlap

$$M_d(i) = \frac{|\mathcal{N}_D(i) \cap \mathcal{N}_d(i)|}{|\mathcal{N}_D(i)|},$$

which is adjusted using a hypergeometric distribution as a baseline expected value under completely random overlap of points.

## Lecture 3: Canonical correlation analysis

2022-01-27

Canonical correlation analysis (CCA) is a rather old technique which has seen a big resurgence of interest, especially in psychological and psychometric analysis. We consider the following problem: given $n$ observation of two sets of variables,

$$X = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ x_{21} & \dots & x_{2p} \\ \vdots & \dots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}, \quad Y = \begin{pmatrix} y_{11} & \dots & y_{1q} \\ y_{21} & \dots & y_{2q} \\ \vdots & \dots & \vdots \\ y_{n1} & \dots & y_{nq} \end{pmatrix}$$

the goal is to find a linear combination $C_x = Xa$ and a linear combination $C_y = Yb$ such that

$$(a_1, b_1) = \underset{a,b}{\operatorname{argmax}} \operatorname{Corr}(Xa, Yb). \tag{4}$$

**Notation** The quantities $C_x$ and $C_y$ are called ***scores***.

**Notation** We define the following matrices:

$$\begin{aligned} \mathbb{V}[X]: \quad & S_{11\, p\times p} = \frac{1}{n} X^\top H^\top H X = \frac{1}{n} X^\top H X \\ \mathbb{V}[Y]: \quad & S_{22\, q\times q} = \frac{1}{n} Y^\top H Y \\ \operatorname{Cov}(X,Y): \quad & S_{12\, p\times q} = \frac{1}{n} X^\top H Y \end{aligned}$$

The maximization problem in (4) thus becomes

$$(a_1, b_1) = \underset{a,b}{\operatorname{argmax}} \frac{a^\top S_{12} b}{\sqrt{a^\top S_{11} a \cdot b^\top S_{22} b}} = \frac{\operatorname{Cov}(C_x, C_y)}{\sqrt{\mathbb{V}[C_x] \cdot \mathbb{V}[C_y]}} \tag{5}$$

and if we define $C_X = HXa$, we have $S_{C_xC_x} = \frac{1}{n} a^\top X^\top H X a = a^\top S_{11} a$, and the same applies to $S_{C_yC_y} = b^\top S_{22} b$. Finally, $\operatorname{Cov}(C_x, C_y) = a^\top S_{12} b$, hence the final equality.

Since the solution is invariant under rescaling of vectors $a$ and $b$, we can find an infinite number of solutions unless we impose some constraints on the maximization procedure. In this case, we impose the following constraints to Equation (5), which guarantee that the solution is unique:

$$a^\top S_{11} a = 1$$

$$b^\top S_{22} b = 1$$

After finding the first solution, we can proceed similarly to principal component analysis in order to find the second pair of canonical vectors, such that

$$(a_2, b_2) = \underset{\substack{a,b: \\ a^\top S_{11} a=1 \\ b^\top S_{22} b=1 \\ a_1^\top S_{11} a=0 \\ b_1^\top S_{22} b=0}}{\operatorname{argmax}} \frac{a^\top S_{12} b}{\sqrt{a^\top S_{11} a \cdot b^\top S_{22} b}} = \frac{\operatorname{Cov}(C_x, C_y)}{\sqrt{\mathbb{V}[C_x] \cdot \mathbb{V}[C_y]}} \tag{6}$$

**Theorem 5 (Canonical correlation analysis)**
*The $k$ solutions to the canonical correlation problem can be found by defining the following matrix,*

$$S_{11}^{-1/2} S_{12} S_{22}^{-1/2} \overset{SVD}{=} UDV^\top.$$

*Then, the solution $A = (a_1 \ \cdots \ a_k)$ and $B = (b_1 \ \cdots \ b_k)$ is given by the first $k$ eigenvectors of $U$ and $V$, respectively.*

*Proof.*
Let us start by considering $a^\top S_{12} b$ under the constraint that $a^\top S_{11} a = 1$ and $b^\top S_{22} b = 1$. Apply the following change of coordinates,

$$u_0 = S_{11}^{1/2} a \implies a = S_{11}^{-1/2} u_0$$

$$v_0 = S_{22}^{1/2} b, \implies b = S_{22}^{-1/2} v_0$$

then the problem (5) becomes

$$\underset{u_0, v_0}{\operatorname{argmax}}\, u_0^\top S_{11}^{-1/2} S_{12} S_{22}^{-1/2} v_0,$$

under the constraints $u_0^\top u_0 = 1$ and $v_0^\top v_0 = 1$. Hence, the solution is given by the first eigenvectors of the $U$ and $V$ matrices from the SVD of the matrix

$$S_{11}^{-1/2} S_{12} S_{22}^{-1/2} = UDV^\top.$$

Repeating the argument yields the following solutions to the canonical correlations problem.

□

**Remark** Note that if $k = \operatorname{rank}\left(S_{11}^{-1/2} S_{12} S^{-1/2}\right)$, then we have that in most cases

$$k \approx \min\left\{\operatorname{rank} X, \operatorname{rank} Y\right\},$$

hence we can find at most $k$ canonical vectors

$$U = (a_1, a_2, \ldots, a_k), \quad V = (b_1, b_2, \ldots, b_k).$$

As always, this solution is unique up to a change in sign of the eigenvectors.

**Partial least squares** CCA has connection to the Partial Least Squares (PLS) estimator, which

Consider the SVD applied to the residualized matrices,

$$\begin{aligned}
HX &= U_X D_X V_X^\top \\
S_{11} &= V_X D_X^2 V_X^\top \\
HY &= U_Y D_Y V_Y^\top \\
S_{22} &= V_Y D_Y V_Y^\top \\
S_{12} &= V_X D_X U_X^\top U_Y D_Y V_Y^\top
\end{aligned}$$

then, if we write the matrix solution in terms of the above SVD, we have

$$\begin{aligned}
S_{11}^{-1/2} S_{12} S_{22}^{-1/2} &= V_X D_X^{-1} \cancel{V_X^\top V_X} D_X U_X^\top U_Y D_Y \cancel{V_Y^\top V_Y} D_Y^{-1} V_Y^\top \\
&= V_X U_X^\top U_Y V_Y^\top,
\end{aligned}$$

and we have that $U_Y V_Y^\top$ is the SVD of the normalized data, i.e. all variances are equal. Hence, we conclude that this solution is invariant under any linear transformation of the data (unlike the PLS).

## Lecture 4: Closed-testing framework

2022-02-02

In this lecture we will consider the problem of performing multiple tests while controlling the overall Type I error at the specified $\alpha$ level. We will do so by casting the usual multiple comparison adjustments into the closed-testing framework (Goeman and Solari, 2011). This framework offers a unified view of multiple testing and is the *de-facto* standard for hypothesis testing.

### 4.1 Multiple testing

Consider two groups $y_1$ and $y_2$, which we assume are drawn from two densities,

$$y_1 \sim P_1, \quad y_2 \sim P_2.$$

Our goal is to compare the two groups and see if the samples come from the same distribution. Consider for example when we assume a parametric form for $P_i$, for instance $P_1 = \mathcal{N}(\mu_i, \sigma^2)$, then the hypothesis would become

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

With the usual $t$-test, we consider the test statistic

$$t_{\text{obs}} = \frac{\bar{y}_1 - \bar{y}_2}{\widehat{\sigma}_{\bar{y}_1 - \bar{y}_2}} \sim t_{n-2},$$

and we define the ***p-value*** as the probability under the null hypothesis of observing a result as extreme as the observed statistic,

$$p = \mathbb{P}(|T| \geq t_{\text{obs}} | H_0), \quad T \sim t_{n-2}.$$

The ***statistical test*** is an object which yields a binary outcome, either 1 for a rejection and 0 for a non-rejection, depending on the limit $L$ that we choose,

$$\varphi = \begin{cases} 1 & \text{if } p \leq L \\ 0 & \text{if } p \geq L \end{cases} \tag{7}$$

We do have different types of errors, for instance

$$\begin{aligned} \text{Type-I error} \quad & \mathbb{P}(\varphi = 1 | H_0) = \mathbb{P}(p \leq L | H_0) \leq \alpha. \\ \text{Power} \quad & \mathbb{P}(\varphi = 1 | H_1) \geq \alpha \\ \text{Type-II error} \quad & 1 - \text{power} = \beta \end{aligned}$$

if $(1-\beta) \geq \alpha$, the test is called ***unbiased***, whereas if $1 - \beta \to 1$, the test is ***consistent***.
We have that the $p$-value of a continuous statistic $t$ is uniformly distributed in $[0, 1]$ under the null hypothesis (Murdoch et al., 2008), i.e.

$$P | H_0 \sim U(0, 1),$$

whereas if the test is consistent, then under $H_1$ the $p$-value is more skewed towards 0.

## 4.2 Multivariate framework

Consider now a setting in which we perform a statistical test on a multiple variable, i.e.

$$y_1 \sim P_1, \quad y_2 \sim P_2, \quad P_i \in \mathbb{R}^n,$$

then the null hypothesis becomes

$$\begin{cases} H_1 : \mu_{11} = \mu_{21} \\ H_2 : \mu_{12} = \mu_{22} \\ \dots \\ H_n : \mu_{1n} = \mu_{2n} \end{cases} \implies H_0 : \bigcap_{i=1}^{n} H_i$$

We can solve the problem using Hotelling's $T$, i.e.

$$T^2 = (\bar{y}_1 - \bar{y}_2)^\top \Sigma^{-1} (\bar{y}_1 - \bar{y}_2),$$

which has a $\chi^2$ distribution if $\Sigma$ does not have to be estimated. Whenever $\Sigma$ has to be estimated by a $\widehat{\Sigma}$, the $T^2$ statistic has a Hotelling's $T$ distribution. If $p < L$ we conclude that there is a difference between the distributions, but we do not know *where* this difference lies.

The concept is that there is a true set $\tau \subseteq \{1, 2, \dots, n\}$ that collect the true variables which differ between he populations. Hence, the true null hypothesis is

$$H_0 : \bigcap_{i \in \tau} H_i.$$

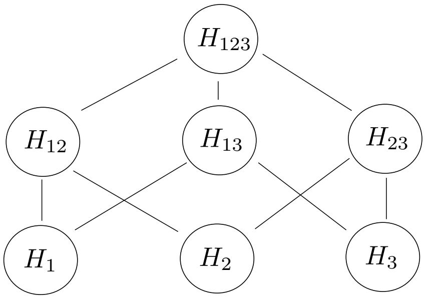

Figure 1: Graph of the hierarchical relationship between the null hypotheses.

We want a testing procedure such that all cases depicted in Figure 1 are considered and the rejection happens at the $\alpha$ level. This is an extension of the Type-I error given by the ***family-wise error rate***, which can be loosely defined as

$$\text{FWER} = \{\text{at least 1 error among all hypotheses}\}$$

We can apply a Hotelling's $T$ test for any of the above situations, however we do not know which of the $i = 1, \dots, 2^3$ null hypotheses is actually true.

A good solution to the above problem is provided by the ***closed testing*** procedure, which has been proven to be the only admissible procedure (Goeman and Solari, 2011), i.e. if there is another procedure which controls the FWER then it must be a closed testing procedure.

**Closed-testing procedure** Consider $p_{123}$ to be the $p$-value which tests $H_{123}$, $p_{12}$ the $p$-value which tests $H_{12}$, and so on. Suppose that we want to test individual hypotheses $H_1$ and $H_2$. We reject $H_1$ if we reject all hypotheses $H_{ij}$, $H_{ijk}$ which contain the subscript 1, and the same applies for $H_2$. Then,

$$H_1 \text{ rejected } \iff p_1, p_{12}, p_{13}, p_{123} \leq \alpha$$

$$H_2 \text{ rejected } \iff p_2, p_{12}, p_{23}, p_{123} \leq \alpha$$

In general, the adjusted test using the above procedure for a general subset of null hypotheses $S \subseteq \{1, 2, \dots, n\}$, denoted by $\tilde{\varphi}_S$, is

$$\tilde{\varphi} = \min_{\mathcal{S} \supseteq S} \varphi_{\mathcal{S}},$$

You can check using the definition (7) of statistical test that this indeed is the correct definition of the closed testing procedure. Hence if $\tilde{\varphi}_S = 1 \implies$ we reject $H_1$. This closed-testing procedure has been first described by Marcus et al. (1976) and the proof of the fact that the FWER is controlled by $\alpha$ is very simple.

*Proof.*
Consider $H_0 : \bigcap_{i \in \tau} H_i$ and the following sets,

$$A = \{\text{at least 1 false rejection}\}$$

$$B = \{\varphi_\tau = 1\}$$

and observe that $A \cap B = A$ by construction of the closed-testing procedure. We know that

$$\mathbb{P}(A) = \mathbb{P}(A \cap B) \leq \mathbb{P}(B) \leq \alpha,$$

since $B$ is a proper test. Hence, the probability of making *any* false rejection is bounded by $\alpha$.

□

## 4.3 Bonferroni correction

The most frequent approach to multiple testing is the Bonferroni procedure, which can be shown to be a special case of the closed-testing procedure. For $i \in \{1, \dots, m\}$, the statistical test for the $i$-th hypothesis is

$$\tilde{\varphi}_i = \mathbb{1}_{p_i \leq \frac{\alpha}{m}} = \mathbb{1}_{m \cdot p_i \leq \alpha},$$

hence we usually talk about ***adjusted p-values*** instead of adjusted limit.

*Proof.*
Assume that the set of true null hypotheses is $\tau$, then the FWER for the Bonferroni procedure is

$$\mathbb{P}\Big(\bigcup_{i \in \tau} p_i \leq \frac{\alpha}{m} | H_0\Big) \leq \sum_{i \in \tau} \mathbb{P}\Big(p_i \leq \frac{\alpha}{m} | H_0\Big) = |\tau| \cdot \frac{\alpha}{m} \leq m \cdot \frac{\alpha}{m} = \alpha.$$

□

**Remark** This is a very powerful result which does not assume any type of dependence between the $p$-values. However, when the dependence is very high we have an extremely conservative test which tends to be too strict.

## 4.4 Bonferroni-Holm

The Bonferroni-Holm procedure uses ordered $p$-values, and starts computing

$$\begin{aligned} p_{(1)} \cdot m \leq \alpha &\implies \text{reject } H_1\text{, otherwise stop} \\ p_{(2)} \cdot (m-1) \leq \alpha &\implies \text{reject } H_2\text{, otherwise stop} \\ &\vdots \\ p_{(m)} \cdot 1 \leq \alpha &\implies \text{reject } H_m\text{, otherwise stop} \end{aligned}$$

We will now see whether Bonferroni and Bonferroni-Holm procedures can be seen as special cases of the closed-testing procedure. Suppose that we want to test the global null hypothesis $H_{123}$, then using Bonferroni we would test

$$\begin{aligned} \text{Reject } H_{123} &\iff \min p_i \cdot 3 = p_{(1)} \cdot 3 \leq \alpha \\ \text{Reject } H_{12} &\iff \min\{p_1, p_2\} \cdot 2 = p_{(1)} \cdot 2 \leq \alpha \end{aligned}$$

hence, if we reject for $H_{123}$ we automatically reject all the connected null hypotheses. Consider now rejecting $H_2$, by the closed testing procedure we now only have to check for $H_{23}$ if $p_2 \cdot 2 \leq \alpha$, and we get a rejected $H_2$ for free. Finally, we only need to check for $H_3$, which can be done by only checking if $p_3 \leq \alpha$.

Hence, by applying the closed-testing procedure using the minimum function we are employing the Bonferroni-Holm procedure.

In conclusion, the closed-testing procedure only needs the definition of

1. A hierarchical multiple testing setting.
2. Any kind of statistical testing procedure to put on each node (likelihood ratio, permutations, bootstrap, . . . ).

**Issues** Given $m$ tests, we have a total graph consisting of $2^m - 1$ nodes, hence we need to find shortcuts in order to compute the overall procedure. In the Bonferroni case, we only need to sort the $p$-values and we have a complexity of $\mathcal{O}(m)$.

Multiple testing procedures often tried to maximize the power in univariate leaf tests $H_1, H_2, \ldots, H_m$. However, it is often the case that we can reject $H_{12}$ under the closed testing procedure but neither $H_1$ nor $H_2$ can be rejected. As a consequence, we get some information in which combinations yield the difference between distributions.

Therefore, we can define a ***upper bound*** for the number of null hypotheses

$$\overline{m}_0(S = H_{123}) = \max_k \{|k| : \tilde{\varphi}_k = 0\}.$$

As a consequence, the ***lower bound*** on the number of alternative hypotheses

$$\underline{\mathrm{m}}_1(S) = \min_k \{|k| : \tilde{\varphi}_k = 1\} = |S| - \overline{m}_0.$$

For instance, rejecting $H_{123}$, $H_{12}$ and $H_{13}$ means that among $H_1, H_2, H_3$ we're not able to judge whether we have $H_1, H_2$ or $H_3$ alternative hypotheses, but we are able to tell that two of them are alternative.

**Conclusion**

*With the closed-testing procedure, we are calculating confidence intervals in the number of null hypotheses.*

## 4.5 Multiple testing presentation

Corrections and shortcuts for huge number of hypothesis tests. Define $R$ to be the number of total rejected null hypotheses, $V$ the number of false discoveries and $U$ the true discoveries

$$\mathrm{FWER} = \mathbb{P}(V \geq 1)$$

$$\mathrm{FDP} = \frac{V}{R},$$

$$\mathrm{FDR} = \mathbb{E}[\mathrm{FDP}]$$

FWER methods are used in confirmatory analysis, when the level $\alpha$ has to be made before seeing the data. the FDR method is used in exploratory research, since they are less stringent.

Closed testing allows mild, flexible and *post-hoc* inference.

1. Do not decide the hypotheses to be rejected.
2. Freely choose the collection of hypotheses
3. Confidence sets

**Pros** Simultaneous confidence sets over all sets and upper bounds

**Cons** Hypotheses have to be specified a priori, and low scalability in the number $m$ of hypotheses.

One can be confident into making at least one of two rejections, while it is not possible to decide whether to reject either $H_1$ or $H_2$.
Using the closed testing procedure with every test at level $\alpha$, the whole procedure contains the FWER at level $\alpha$ for all intersections. Upper confidence bounds given by (Goeman and Solari, 2011), both

for true discoveries and the FDP

$$q_\alpha(\mathcal{S}) = \frac{t_\alpha(S)}{|S|}, \quad \text{for } \pi(S).$$

As $m$ increases, the power per hypothesis vanishes, whereas if there is sufficient signal the FDR does not vanish. If all null hyp are true, FWER = FDR (**benjamini1995**), whereas in all other situations FWER > FDR and therefore the FDR is less stringent with more power.

**Weak control** Control the FWER when all hypotheses are true.

**Strong control** Control the FWER even when all hypotheses are not true.

**Def. (Consonant procedure)**
Consonant $\iff$ for all $I \in \mathcal{I}$, $H_I rejected$ means that at least one $H_i$ is rejected with $i \in I$.

**Remark** Only consonant procedures are admissible in hte FWER control.

**Remark** For controlling the FDP, non-consonant procedures should be taken into account. Non-consonant closed testing procedures have false discovery proportion confidence bounds.

Some common FWER procedures and FDR controling procedures are based on the ***Simes test***, which rejects an intersection hypothesis $H_i \iff$ there is $i \in [1, |I|]$ such that

$$\ldots$$

which is similar to the Bonferroni test.

In general the Simes test contain the Bonferroni test, and therefore is a more powerful test. In general, it holds for *independent p-values* and under certain forms of positive correlation.

Non-consonant procedures are starting to become relevant since the individual hypotheses gradually became less relevant. In this case we want to find sets of hypotheses in which the proportion of false rejection is low enough.

In particular, the Simes local test has non-vanishing power for the control of the false discovery rate. Goeman 2017 proved that the Simes local test rejects an hypothesis if

$$\min\left\{\frac{h_\alpha}{i} p_{(i:I)}\right\} \leq \alpha,$$

where $h_\alpha = \max\{i \in (0, \ldots, m) : i \cdot p_{m-i+j:B} > j \cdot \alpha\}$ is the size of the largest voxel set not rejected by the procedure. A shortcut for estimating $t_\alpha(S)$ can be defined as

$$d_\alpha(S) = \max_{1 \leq u \leq |S|} \{1 - u + |\{i \in S : \ldots\}|\}.$$

Meijer (2019) provided an algorithm to calculate $h_\alpha$ for all $\alpha$ simultaneously with order of time $\mathcal{O}(m \log m)$ and allows the computation of $d_\alpha(S)$ with order $\mathcal{O}(m \log m)$.
Note that $h_\alpha$ does not depend on $S$, hence if it has been calculated then $d_\alpha$ can be calculated for many $S$'s in linear time. Finally, an upper bound for the FDP can be calculated as whereas the true discovery proportion (TDP) has lower bound

$$\overline{\mathrm{FDP}}(S) = \frac{t_\alpha}{|S|}$$

$$\underline{\mathrm{TDP}}(S) = 1 - \overline{\mathrm{FDP}}(S),$$

and (Goeman and Solari, 2011) and goeman 2017 showed that

$$\mathbb{P}(\text{for all } S \in \mathcal{S}\colon \overline{\mathrm{TDP}}(S) \leq \mathrm{TDP}(S)) \geq 1 - \alpha,$$

hence we can come back and change $S$ after looking at the data since the procedure is valid for an arbitrary choice of $S$.

# Lecture 5: Data splitting

2022-02-08

Data splitting is a way of solving the multiple comparison problem: the first half is used for selecting the hypotheses we want to test, whereas the second part is used only for running the tests.

**Algorithm 1** Data-splitting procedure

```
Divide the sample in two portions $(L, I)$
Choose the sample such that $s = \text{argmax}_k \overline{X}^L_k$
Perform the one-sided test of the corresponding second portion
    $$\varphi = \mathbb{1}\Big\{\overline{X}^I_s > \frac{\sigma}{\sqrt{n_I}} z_{1-\alpha}\Big\}.$$
```

**Advantages**

1. *Simplicity*: we are allowed to do any kind of selection on the first dataset.
2. *Correctness*: we do not have to worry about data snooping, since all inference is carried out on a “fresh” dataset, after having decided the hypotheses that we want to test in the first dataset.

**Disadvantages**

1. Cox has proved that there is an *effective power* of the procedure, which has also been proven to be lower than the Bonferroni procedure. Hence, in the i.i.d case the data splitting procedure is *always worse* than the Bonferroni correction.
2. The “$p$-value lottery” is the fact that we get different results based on the splitting. The randomness of the $p$-values depends on the split that we perform, and the variability is quite substantial.

## 5.1 High-dimensional inference

Consider the linear model when $p > n$,

$$y = X\beta^0 + \varepsilon, \quad \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n),$$

and denote the set of “active” predictors as $S_0 = \big\{j \in \{1, \dots, p\} : \beta^0_j \neq 0\big\}$ and $N_0 = S^c$. Suppose that we split the data into two parts of size $n/2$ (Wasserman and …2009) and run a penalized regression to reduce the number of variables under scrutiny. After having selected the relevant predictor, we use the second half of the data to perform hypothesis testing, using the classical least squares.

Let $\widehat{S}$ be the set of selected hypotheses during the screening part, then we hope that $S \subseteq \widehat{S}$. Indeed, for a model

$$y = X_{\widehat{S}}\beta_{\widehat{S}} + \varepsilon,$$

if we were to miss some of the active variables, then we incur in the problem of omitted variable bias, and our inferences would be grossly misled.

We construct the $p$-values on the second half of the dataset using

$$\tilde{p}_j = \begin{cases} p_j & \text{if } j \in \widehat{S} \\ 1 & \text{if } j \notin \widehat{S} \end{cases}$$

and then we adjust the $p$-values using standard adjustments for multiple comparisons, such as the Bonferroni procedure.

**Algorithm 2** Single-split procedure

```
Input: $y, X, \alpha \in (0,1)$ and a variable selection procedure $\widehat{S}$
Partition $\{1, \dots, n\}$ into portions $L, I$ of size $n_L$ and $n_I$
Using $L$ only, select $\widehat{S}^L \subseteq \{1, \dots, p\}$,
Apply OLS of $y^I$ on $X^I_{\widehat{S}^L}$, compute the $p$-values testing $H_{0j} : \beta^0_j = 0$ for $j \in \widehat{S}^L$
Adjust the $p$-values using some multiple testing adjustment
```

**Theorem 6 (Single-split controls the FWER)**
*Assume that*

1. *The linear model $y \sim X\beta_0 + \varepsilon$ is the true model*
2. *The variable selection procedure satisfies the screening property, i.e.*

$$\mathbb{P}(S_0 \subseteq \widehat{S}^L) \geq 1 - \delta,$$

*for some $\delta \in (0, 1)$.*

3. *The reduced design matrix for the second half of the sample satisfies*

$$\operatorname{rank}(X^I_{\widehat{S}^L}) = |\widehat{S}^L|.$$

*Then, the single-split procedure yields FWER control at level*

$$\mathbb{P}(\tilde{S} \cap N_0 \neq \emptyset) < \alpha + \delta.$$

*Proof.*
Complete proof on the paper, we use the fact that the probability of false rejection is either when we do not select the important variable or when we apply a false rejection.

□

Over the years, this procedure became much more attractive since the linear regression context makes the Bonferroni no longer the uniformly more powerful approach.

***p*-value lottery** There is still a problem of $p$-value uncertainty, which can be ameliorated by multiple sample-splitting, similarly to the cross-validation procedure. In general, we cannot simply average $p$-values since even the average of two independent $U(0, 1)$'s is not uniformly distributed.

**Algorithm 3** Multi sample-splitting procedure

```
for $b = 1, \dots, B$ do
    Apply the single-split procedure to obtain $\tilde{p}_j^{(b)}$ for $j = 1, \dots, p$.
end for
Aggregate the $p$-values using
    $$\tilde{p}_j = 2 \cdot \text{median}(\tilde{p}_j^{(1)}, \dots, \tilde{p}_j^{(B)}), \quad j = 1, \dots, p.$$
Selected predictors are $\tilde{S} = \{j \in \{1, \dots, p\} : \tilde{p}_j \leq \alpha\}$.
```

**Remark** The intuition is that the median of $U(0,1)$'s is approximately 0.5, hence by multiplying by 2 we obtain on average 1. Using Markov's inequality we obtain a bounded probability by $\alpha$, which is quite conservative.

# References

Chen, L. and Buja, A. (2013). «Stress Functions for Nonlinear Dimension Reduction, Proximity Analysis, and Graph Drawing». In: *Journal of Machine Learning Research* 14.Apr, 1145–1173.

Goeman, J. J. and Solari, A. (2011). «Multiple Testing for Exploratory Research». In: *Statistical Science* 26.4.

Marcus, R. et al. (1976). «On Closed Testing Procedures with Special Reference to Ordered Analysis of Variance». In: *Biometrika* 63.3, 655–660.

Murdoch, D. J. et al. (2008). «P-Values Are Random Variables». In: *The American Statistician* 62.3, 242–245.