MPhil Advanced Econometrics

# Principal component analysis

Martin Weidner

University of Oxford

2022-23, Hilary Term

# Principal component analysis (PCA)

- PCA is useful for
  (1) data compression
  (2) data representation
  (3) noise reduction
  (4) (it is also closely related to "matrix completion")

- The original data need to be in matrix form (i.e. a two-dimensional array)

- Mathematically, PCA is very closely related to the singular value decomposition (SVD) of a matrix, which is why we will discuss the concepts of matrix rank and SVD first.

# Rank of a matrix I

- Notation: For a matrix $C$ we denote its transpose by $C'$.
- Rank: For an $n \times m$ matrix $A$ the rank of $A$ is the the smallest non-negative integer $r$ such that there exists an $n \times r$ matrix $B$ and an $m \times r$ matrix $C$ which satisfy

$$A = BC'$$

  We then write $r = \text{rank}(A)$.
- We have $0 \leq \text{rank}(A) \leq \min(n, m)$.
- Examples:
  - $\text{rank}(A) = 0 \quad \Leftrightarrow \quad A = 0_{n \times m}$ (a matrix with all entries zeroes)
  - $\text{rank}(A) = 1 \quad \Leftrightarrow \quad A = vw'$ for some vectors $v$ and $w$.

# Rank of a matrix II

- Equivalently a matrix $A = (A_{ij})$ with $\text{rank}(A) = r$ can be written as

$$A_{ij} = \sum_{q=1}^{r} B_{iq} C_{jq} = \underbrace{B_{i1} C_{j1} + B_{i2} C_{j2} + \ldots + B_{ir} C_{jr}}_{\text{sum of } r \text{ matrices of rank one}}$$

- A concrete numerical example with $\text{rank}(A) = 2$:

$$\begin{pmatrix} -1 & 3 & 0 \\ -5 & 6 & 1 \\ 2 & 3 & -1 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} \begin{pmatrix} -1 \\ 3 \\ 0 \end{pmatrix}' + \begin{pmatrix} 0 \\ -1 \\ 1 \end{pmatrix} \begin{pmatrix} 3 \\ 0 \\ -1 \end{pmatrix}'$$

- Dimensional reduction idea: The number of parameters in the $n \times m$ matrix $A$ equals $n \cdot m$. But, if $\text{rank}(A) \ll \min(n, m)$, then we can represent the matrix in terms of only $(n + m) \cdot \text{rank}(A)$ parameters, which may be much smaller than $n \cdot m$.

# Singular value decomposition (SVD)

- Notation: We denote by $\mathbb{I}_q$ the $q \times q$ identity matrix.
- SVD: Every $n \times m$ matrix $A$ with real entries can be written as

$$A = U\,S\,V'$$

where

  - $U$ is an $n \times \mathrm{rank}(A)$ matrix such that $U'U = \mathbb{I}_{\mathrm{rank}(A)}$
  - $S$ is an $\mathrm{rank}(A) \times \mathrm{rank}(A)$ diagonal matrix with positive diagonal entries.
  - $V$ is an $m \times \mathrm{rank}(A)$ matrix such that $V'V = \mathbb{I}_{\mathrm{rank}(A)}$

This is called the singular value decomposition of $A$.

- The columns of $U$ and $V$ are called the (left and right) singular vectors. The diagonal entries of

$$S = \begin{pmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & \dots & s_{\mathrm{rk(A)}} \end{pmatrix}$$

are called the singular values, $s_q > 0$.

# Singular value decomposition (cont.)

- Equivalently, the SVD of a matrix $A = (A_{ij})$ with $\text{rank}(A) = r$ can be written as

$$A = \sum_{q=1}^{r} s_q\, u_q\, v_q', \qquad s_q > 0,\;\; \|u_q\| = 1,\;\; \|v_q\| = 1.$$

  where $s_q \in \mathbb{R}$ are the singular values and $u_q \in \mathbb{R}^n$, $v_q \in \mathbb{R}^m$ are the singular vectors, whose Euclidian norm $\|\cdot\|$ equals one, and who are mutually orthogonal, e.g. $u_1' u_2 = 0$, $v_3' v_5 = 0$.

- In components:

$$A_{ij} = \sum_{q=1}^{r} s_q\, u_{iq}\, v_{jq} = \underbrace{s_1\, u_{i1}\, v_{j1} + s_2\, u_{i2}\, v_{j2} + \ldots + s_r\, u_{ir}\, v_{jr}}_{\text{sum of } r \text{ matrices of rank one}}$$

- It is customary (and we will always assume this) to sort the singular values in decreasing order:

$$s_1 \geq s_2 \geq \ldots \geq s_{\text{rank}A}$$

# Singular value decomposition (cont.)

- The singular values $s_q$ are uniquely determined from $A$.

- If all the singular values $s_q$ are mutually different, then the singular vectors are also unique, apart from the trivial transformation,

$$u_q \mapsto -u_q, \qquad v_q \mapsto -v_q,$$

for each $q \in \{1, \ldots, \operatorname{rank}(A)\}$.

- If multiple singular values are equal, e.g. $s_q = s_{q+1}$, then there is some freedom to transform the corresponding singular vectors into each other. If $A$ is an observational data matrix, then this usually doesn't happen. For our purposes we can consider the singular value decomposition to be unique.

# Principal components

- For a matrix $A$ with SVD

$$A = \sum_{q=1}^{\mathrm{rank}(A)} s_q\, u_q\, v_q'$$

we denote the leading few terms $s_q\, u_q\, v_q'$ as the leading principal components.

- The magnitude of the principal components is given by $s_q$.
- By choosing an integer $R < \mathrm{rank}(A)$ we can approximate $A$ by its leading $R$ principal components as

$$A \approx A2 = \sum_{q=1}^{R} s_q\, u_q\, v_q'$$

- (This is just our first definition of principal components, more statistical definitions are given below.)

# Grayscale Image Example

orginal image

- This grayscale image can be interpreted as a matrix $A$ of dimension $750 \times 1125$.

# Grayscale Image Example

- Given the matrix $A$ we can extract the $R \in \{1, 2, 3, \ldots\}$ leading principal components and then recombine them back into a new matrix $A2$ of the same dimensions as A.

- matlab code:

```
[U,S,V]  =  svd(A);
s = diag(S);
A2 = U(:,1:R) * diag(s(1:R)) * V(:,1:R)';
```

- In matlab the singular value decomposition command svd applied to an $n \times m$ matrix $A = USV'$ returns an $n \times n$ matrix $U$, an $n \times m$ matrix $S$, and an $m \times m$ matrix $V$. Thus, for $\mathrm{rank}(A) < \min(n, m)$ some of the singular values in $S$ are zero.

- The following slides show $A2$ for $R = 50$, 20, 5 and 1.

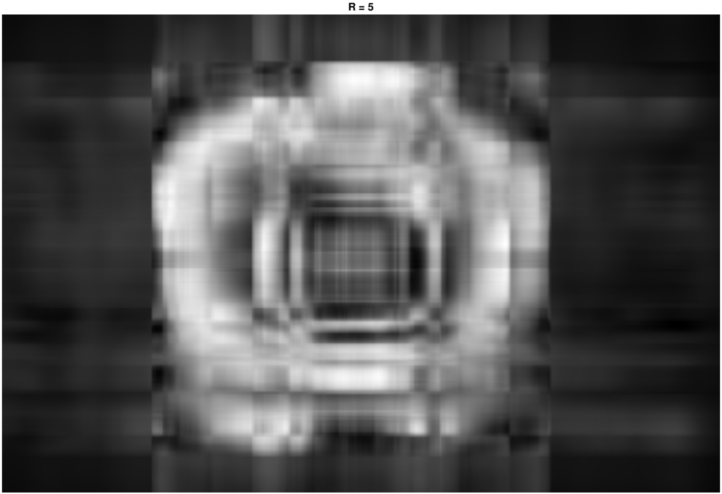
# Grayscale Image Example (cont.)


- Using only 50 principal components to reconstruct the image.

# Grayscale Image Example (cont.)

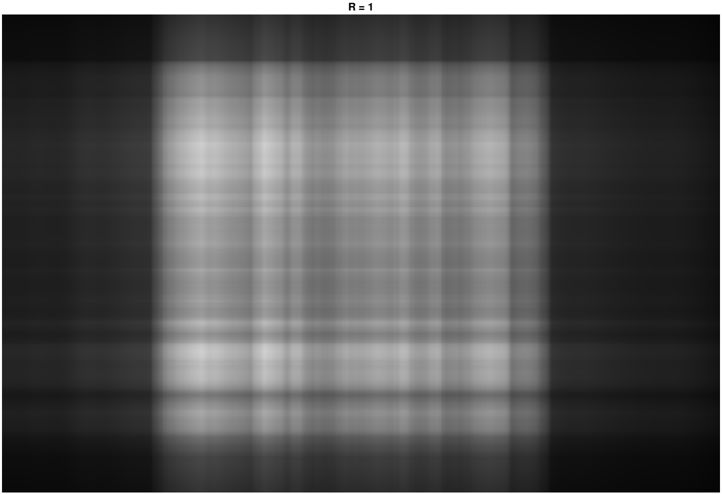

- Using only 20 principal components to reconstruct the image.

# Grayscale Image Example (cont.)



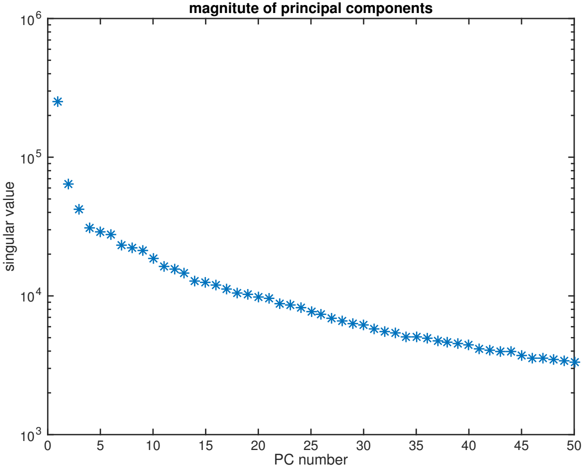

- Using only 5 principal components to reconstruct the image.

# Grayscale Image Example (cont.)



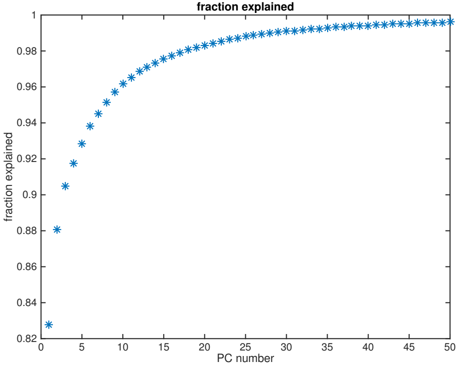

- Using only 1 principal component to reconstruct the image.

# Grayscale Image Example (cont.)




- The magnitude of the principal components is quickly decreasing with $R$.

# Grayscale Image Example (cont.)




- The leading few principal components can explain the vast majority of the total variation in the image matrix.
- fraction explained $= \dfrac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} A2_{ij}^2}{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2} = 1 - \dfrac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} (A_{ij} - A2_{ij})^2}{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}^2}$

# Factor Model / Interactive Fixed Effects

- Panel data: $i = 1, \ldots, n$ cross-sectional units; $t = 1, \ldots, T$ time periods (or $t = 1, \ldots, T$ variables observable for every $i$).
- A factor model with $R \in \{1, 2, 3, \ldots\}$ factors for the observed outcomes $y_{it} \in \mathbb{R}$ is given by

$$y_{it} = \sum_{r=1}^{R} \lambda_{ir} \, f_{tr} + e_{it},$$

  where $\lambda_{ir} \in \mathbb{R}$ are unobserved "factor loading" ($R$ individual specific effects), $f_{tr} \in \mathbb{R}$ are unobserved "factors" ($R$ time specific effects), and $e_{it} \in \mathbb{R}$ are unobserved "idiosyncratic errors" (noise, modeled as mean zero random variables, either independent or only weakly dependent across $i$ and over $t$).
- In matrix notation we can write this as

$$\underset{n \times T}{y} = \underset{n \times R}{\lambda} \underset{(T \times R)'}{f'} + \underset{n \times T}{e}$$

# Least Squares Estimator

- One could write down a stochastic model for $\lambda_{ir}$ and $f_{tr}$ ("random effects"), but in the following we treat $\lambda_{ir}$ and $f_{tr}$ as parameters to be estimated ("fixed effects").

- For given $R$, consider the (non-linear) least squares estimator

$$\left\{\widehat{\lambda}, \widehat{f}\right\} \in \underset{\{\lambda \in \mathbb{R}^{n \times R}, f \in \mathbb{R}^{T \times R}\}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \sum_{t=1}^{T} \left( y_{it} - \sum_{r=1}^{R} \lambda_{ir} \, f_{tr} \right)^2}_{= \|y - \lambda f'\|_F^2}$$

(Here, $\|A\|_F = \sqrt{\sum_i \sum_t A_{it}^2}$ is the Frobenius norm of matrix $A$.)

# Least Squares Estimator: Normalization

- Here, the solution for the $n \times T$ matrix $\widehat{\lambda}\widehat{f'}$ is unique, but the individual components $\widehat{\lambda}$ and $\widehat{f}$ are not unique (under standard regularity conditions), because for any invertible $R \times R$ matrix $A$ we can reparameterize

$$\lambda \mapsto \lambda A \qquad\qquad f \mapsto f(A^{-1})'$$

  without changing $\lambda f'$.

- A very common normalization is to impose that

$$\frac{1}{T}\, f' f = \mathbb{I}_R \qquad\qquad \frac{1}{n}\, \lambda' \lambda = \text{diagonal matrix}$$

  Imposing those extra conditions gives unique solutions $\widehat{\lambda}$ and $\widehat{f}$.

- However, for many purposes (e.g. prediction) the normalization does not matter.

# Principal Components = Least Squares Estimator

- The FOC of the least squares problem read $y\,\widehat{f} = \widehat{\lambda}\,\widehat{f}'\,\widehat{f}$ and $y'\,\widehat{\lambda} = \widehat{f}\,\widehat{\lambda}'\,\widehat{\lambda}$. Plugging one of those into the other gives

$$(y'y)\,\widehat{f} = \widehat{f}\,\widehat{B} \qquad\qquad (yy')\,\widehat{\lambda} = \widehat{\lambda}\,\widehat{B}',$$

  where $\widehat{B} = (\widehat{\lambda}'\,\widehat{\lambda})(\widehat{f}'\,\widehat{f})$ is an $R \times R$ matrix.

- The last display shows that $\widehat{f}$ is a collection of $R$ eigenvectors of the $T \times T$ matrix $y'y$, and analogously $\widehat{\lambda}$ is a collection of $R$ eigenvectors of the $n \times n$ matrix $yy'$.

- A more careful analysis (involving SOC) shows that $\widehat{f}$ and $\widehat{\lambda}$ are in fact eigenvectors corresponding to the largest $R$ eigenvalues of $y'y$ and $yy'$. Those "principal eigenvectors" are often called principal components of $y$ (or of $y'y$ and $yy'$).

## Computation (for balanced panel case)

- Minimizing the (non-convex) objective function $\|y - \lambda f'\|_F^2$ over $\lambda \in \mathbb{R}^{n\times R}$ and $f \in \mathbb{R}^{T\times R}$ is practically infeasible, except for very small $n$ and $T$.
- However, computing eigenvalues and eigenvectors is very quick on modern computers. Therefore if $T \leq n$ we would
  (1) Calculate $\widetilde{f} \in \mathbb{R}^{T\times R}$ as the eigenvectors corresponding to the $R$ largest eigenvalues of the $T \times T$ matrix $y'y$.
  (2) Impose the normalization $\frac{1}{T}\widehat{f}'\widehat{f} = \mathbb{I}_R$ by defining
  $$\widehat{f} = \widetilde{f}\left(\frac{1}{T}\widetilde{f}'\widetilde{f}\right)^{-1/2}$$
  (3) Use the FOC $y\,\widehat{f} = \widehat{\lambda}\,\widehat{f}'\,\widehat{f}$ to calculate
  $$\widehat{\lambda} = \frac{1}{T}\,y\,\widehat{f}.$$

  (if $n < T$ we turn things around, that is, we first calculate $\widehat{\lambda}$ as eigenvectors of the $n \times n$ matrix $yy'$.)

# Asymptotic Theory for $\widehat{f}$ and $\widehat{\lambda}$

- For $n, T \to \infty$, with $T/n^2 \to 0$ and $n/T^2 \to 0$, Bai (2003) shows that

$$\sqrt{n}\left(\widehat{f}_t - H' f_t^0\right) \Rightarrow \mathcal{N}(0, V_f),$$
$$\sqrt{T}\left(\widehat{\lambda}_i - H^{-1} \lambda_i^0\right) \Rightarrow \mathcal{N}(0, V_\lambda),$$

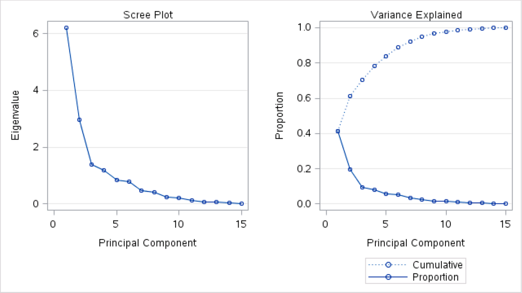
  where $H$ is an $R \times R$ matrix that depends the normalization of $\widehat{\lambda}$ and $\widehat{f}$.

# Determining the Number of Factors $R$

- There are many Statistics and Econometrics papers that suggest methods to estimate $R$ from observing $y$.
- See e.g. Bai and Ng (2002), Onatski (2010), Ahn and Horenstein (2013).

# Using PCA for dimensional reduction

- The principal components methods allows to approximate the $n \times T$ matrix $y$ by $\widehat{\lambda}\,\widehat{f}'$. Together, $\widehat{\lambda}$ and $\widehat{f}'$ correspond to $(n+T)R$ parameters. (once we account for the normalization it is $(n+T-R)R$ parameters).
- In most applications just a few principal components will explain most of the observable variation in $y$.
- Example from Megyesiova and Lieskovska (2018), where $i \in \{35$ OECD countries$\}$ and $t \in \{15$ economic and public health indicators$\}$ in the year 2000.



# Examples of possible Applications

Example 1: Reducing the number of control variables in a regression.

- Consider the same problem as for "double variable selection" before:

$$y_i = d_i\,\alpha + x_i\,\beta + u_i,$$

where $\alpha \in \mathbb{R}$ is the parameter of interest, and $\beta \in \mathbb{R}^K$ is high-dimensional.

- Apply principal components analysis to the $n \times K$ matrix $X = [x_i \,:\, i = 1, \ldots, n]$ to find

$$X \approx \widehat{\lambda}\,\widehat{f}',$$

where $\widehat{\lambda}$ is an $n \times R$ matrix and $\widehat{f}$ is an $K \times R$ matrix, $R < K$.

- Estimate $\alpha$ by applying OLS to

$$y_i = d_i\,\alpha + \widehat{\lambda}_i'\,\gamma + u_i,$$

that is, we replace the many $(K)$ controls $x_i$ by the few $(R)$ controls $\widehat{\lambda}_i$, which capture the major part of the variation in $x_i$.

# Examples of possible Applications

Example 2: Diffusion Index Forecasting: Stock and Watson (2002)

- Want to predict future values of one variable $y_t \in \mathbb{R}$ (e.g. GDP growth) in terms of many predictor variables $x_t \in \mathbb{R}^n$ (CPI, industrial production and sales in various sectors, ...).
- Consider a factor model for those predictor variables $x_{it}$:

  $$x_{it} = \lambda_i' f_t + e_{it},$$

  Estimate $\lambda_i \in \mathbb{R}^R$ and $f_t \in \mathbb{R}^R$ by principal components. (actually Stock and Watson (2002) use a "dynamic factor model", but both is possible)
- A forecast model for $y_{t+1}$ reads

  $$y_{t+1} = \beta(L) f_t + \gamma(L) y_t + \epsilon_{i,t+1},$$

  where $\beta(L)$ and $\gamma(L)$ are polynomials in the "lag-operator" $L$. Estimate those parameters (e.g. OLS) and forecast:

  $$\widehat{y}_{t+1} = \widehat{\beta}(L) \widehat{f}_t + \widehat{\gamma}(L) y_t$$

# Examples of possible Applications

Example 3: Imputation / Matrix Completion

- Assume that we only observe $y_{it}$ for a subset $\mathcal{O} \subset \{1, \ldots, n\} \times \{1, \ldots, T\}$ of all possible observations, and we want to impute $y_{it}$ for $(i,t) \notin \mathcal{O}$.

- We can still estimate

$$\left\{ \widehat{\lambda}, \widehat{f} \right\} \in \underset{\{\lambda \in \mathbb{R}^{n \times R}, f \in \mathbb{R}^{T \times R}\}}{\operatorname{argmin}} \sum_{(i,t) \in \mathcal{O}} \left( y_{it} - \lambda_i' f_t \right)^2$$

  (actually this may be difficult to compute, see nuclear-norm minimization comments below)

- Imputation for $(i,t) \notin \mathcal{O}$:

$$y_{it} = \widehat{\lambda}_i' \, \widehat{f}_t$$

- See Recht, Fazel and Parrilo (2010) and Hastie, Tibshirani and Wainwright (2015) for surveys on “matrix completion”.

# Nuclear Norm Minimization (side comment)

- The problem

$$\min_{\lambda, f} \sum_{(i,t) \in \mathcal{O}} \left( y_{it} - \lambda_i' f_t \right)^2$$

can equivalently also be expressed as

$$\min_{\Gamma \in \mathbb{R}^{n \times T}} \sum_{(i,t) \in \mathcal{O}} (y_{it} - \Gamma_{it})^2 \quad \text{s.t.} \quad \text{rank}(\Gamma) \leq R,$$

where $\Gamma$ is an $n \times T$ matrix.

- Used here:

$$\Gamma = \lambda f' \quad \Leftrightarrow \quad \text{rank}(\Gamma) \leq R \quad \Leftrightarrow \quad \sum_{r=1}^{\min(n,T)} 1\left(s_r(\Gamma) > 0\right) \leq R,$$

where $s_1(\Gamma) \geq s_2(\Gamma) \geq \ldots \geq s_{\min(n,T)}(\Gamma) \geq 0$ are the singular values of $\Gamma$.

# Nuclear Norm Minimization (side comment)

- $\text{rank}(\Gamma) \leq R$ is a non-convex constraint.
- Convex relaxation of this constraint:

$$\underbrace{\sum_{r=1}^{\min(N,T)} s_r(\Gamma)}_{=:\|\Gamma\|_*} \leq \text{const.}$$

  where $\|\Gamma\|_*$ is the nuclear norm (or trace norm).
- An estimate for $\Gamma = \lambda f'$ is given by

$$\begin{aligned} \widehat{\Gamma} &= \underset{\Gamma \in \mathbb{R}^{n \times T}}{\text{argmin}} \sum_{(i,t) \in \mathcal{O}} (y_{it} - \Gamma_{it})^2 \quad \text{s.t.} \quad \|\Gamma\|_* \leq \text{const.} \\ &= \underset{\Gamma \in \mathbb{R}^{n \times T}}{\text{argmin}} \sum_{(i,t) \in \mathcal{O}} (y_{it} - \Gamma_{it})^2 + \psi \, \|\Gamma\|_*, \end{aligned}$$

  where $\psi > 0$ is a penalty parameter. This is a convex problem.
- Again, see Recht, Fazel and Parrilo (2010) and Hastie, Tibshirani and Wainwright (2015) for surveys on "matrix completion".

# Bibliography I

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica 71*(1), 135–171.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Megyesiova, S. and V. Lieskovska (2018). Analysis of the sustainable development indicators in the oecd countries. *Sustainability 10*(12), 4554.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92(4)*, 1004–1016.

# Bibliography II

Recht, B., M. Fazel, and P. A. Parrilo (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review 52*(3), 471–501.

Stock, J. H. and M. W. Watson (2002). Macroeconomic forecasting using diffusion indexes. *Journal of Business & Economic Statistics 20*(2), 147–162.