# Factor Models and Variable Selection in High-dimensional Regression Analysis

PASCAL SARDA

Mathematical Institute of Toulouse

Group of Statistics and Probability

University Paul Sabatier

118, route de Narbonne,

31062 Toulouse Cedex, France

sarda@cict.fr

*Working group STAPH*
*http://www.math.univ-toulouse.fr/staph/*

joint work with Alois KNEIP

# High Dimensional Regression

Starting Model:

$$(1) \qquad Y_i = \boldsymbol{\beta}^\tau \mathbf{X}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

- $Y_i \in \mathbb{R}$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\tau \in \mathbb{R}^p, i = 1, \ldots, n$ are independent r.v.,
- $\boldsymbol{\beta}$ is a vector of parameters in $\mathbb{R}^p$
- $(\epsilon_i)_{i=1,\ldots,n}$ are centered i.i.d. r.r.v. independent with $\mathbf{X}_i$ with $Var(\epsilon_i) = \sigma^2$.

# High Dimensional Regression

Starting Model:

$$(1) \qquad Y_i = \boldsymbol{\beta}^\tau \mathbf{X}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

- $Y_i \in \mathbb{R}$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\tau \in \mathbb{R}^p, i = 1, \ldots, n$ are independent r.v.,
- $\boldsymbol{\beta}$ is a vector of parameters in $\mathbb{R}^p$
- $(\epsilon_i)_{i=1,\ldots,n}$ are centered i.i.d. r.r.v. independent with $\mathbf{X}_i$ with $Var(\epsilon_i) = \sigma^2$.

The dimension p is much larger than the sample size n

# High Dimensional Regression

Starting Model:

$$(1) \qquad Y_i = \boldsymbol{\beta}^\tau \mathbf{X}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

- $Y_i \in \mathbb{R}$, $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})^\tau \in \mathbb{R}^p, i = 1, \ldots, n$ are independent r.v.,
- $\boldsymbol{\beta}$ is a vector of parameters in $\mathbb{R}^p$
- $(\epsilon_i)_{i=1,\ldots,n}$ are centered i.i.d. r.r.v. independent with $\mathbf{X}_i$ with $Var(\epsilon_i) = \sigma^2$.

The dimension p is much larger than the sample size n

Two different situations:

- $\mathbf{X}_i$: high dimensional vector of different predictor variables
- functional data: $X_{ij}$, $j = 1, \ldots, p$ are discretization points of a same curve: $X_{ij} = X_i(t_j)$

In the following both situations are analyzed in a same way

# General Outline

- Two main approaches for high dimensional regression in the literature:
  - Variable selection: select only a (small) set of variables with influence on the response
  - Functional (linear) regression: model (1) is a discrete version of an underlying FLR model. No variable has a particular influence on the response but *all together* explain a part of variability of the response. (Nonparametric models are also considered)

# General Outline

- Two main approaches for high dimensional regression in the literature:
  - Variable selection: select only a (small) set of variables with influence on the response
  - Functional (linear) regression: model (1) is a discrete version of an underlying FLR model. No variable has a particular influence on the response but *all together* explain a part of variability of the response. (Nonparametric models are also considered)
- **Objective**. Combine the two approaches with the aim of considering:
  - possible high correlations between the predictors, Factor models. Roughly speaking, the predictors are decomposed in two components which respectively represent *common* and *specific* variabilities
  - variable selection in an augmented model which extend model (1) and includes principal components which may posses an additional power for predicting the response

## General ideas

Studies of the “High dimensional model” rest on conditions on the coefficient vector $\beta$ and/or the predictors $X_{ij}$.

# General ideas

Studies of the "High dimensional model" rest on conditions on the coefficient vector $\beta$ and/or the predictors $X_{ij}$.

- **Variable selection**:
  - $\beta$ has coefficients that are mostly 0: sparseness;
  - To retrieve non null coefficients, correlations between $X_{ij}$ and $X_{il}$, $j \neq l$, are sufficiently "weak": almost uncorrelated e.g. in Candes and Tao (2007), more general Restricted Eigenvalue condition in Bickel, Ritov, Tsybakov (2009).

# General ideas

Studies of the "High dimensional model" rest on conditions on the coefficient vector $\boldsymbol{\beta}$ and/or the predictors $X_{ij}$.

- **Variable selection**:
  - $\boldsymbol{\beta}$ has coefficients that are mostly 0: sparseness;
  - To retrieve non null coefficients, correlations between $X_{ij}$ and $X_{il}$, $j \neq l$, are sufficiently "weak": almost uncorrelated e.g. in Candes and Tao (2007), more general Restricted Eigenvalue condition in Bickel, Ritov, Tsybakov (2009).
- **Functional regression**:
  - $\beta_j = \frac{\beta(t_j)}{p}$, $\beta \in L^2([0,1])$, $t_j = \frac{1}{p}$, continuous slope function, and as $p \to \infty$, $\sum_j \beta_j X_{ij} \to \int_0^1 \beta(t) X_i(t) dt$;
  - the predictors are heavily correlated. As $p \to \infty$, $corr(X_i(t_j), X_i(t_{j+m})) \to 1$ for any fixed $m$

# Functional regression: basis expansion

- Model is rewritten in term of a "sparse" basis expansion of the predictor functions $X_i$

  Best possible basis, minimizing the $L^2$-error, for a $k$-dimensional approximation of random functions $X_i$: eigenfunctions corresponding to the $k$ largest eigenvalues of the covariance operator of $X_i$

$$\mathbb{E}(X_i \otimes X_i)$$

  i.e. leading elements of the Karhunen-Loève decomposition

# Functional regression: basis expansion

- Model is rewritten in term of a "sparse" basis expansion of the predictor functions $X_i$
  Best possible basis, minimizing the $L^2$-error, for a $k$-dimensional approximation of random functions $X_i$: eigenfunctions corresponding to the $k$ largest eigenvalues of the covariance operator of $X_i$

$$\mathbb{E}(X_i \otimes X_i)$$

  i.e. leading elements of the Karhunen-Loève decomposition

- Important feature of the covariance operator of $X_i$: compact, nuclear
  $\longrightarrow$ The (infinite) set of eigenvalues decrease rapidly to zero: actually, the sum is finite

# Functional regression: estimation

- **Discretized case (model (1))**: this amounts to consider the eigenvalues $l_1 \geq l_2 \geq \ldots$ and corresponding eigenvectors $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ of the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X}_i \mathbf{X}_i^T)$
  $\longrightarrow$ even if $p > n$, $\boldsymbol{\psi}_r$ $1 \leq r \leq k$, can be well estimated by the eigenvectors (principal components) $\widehat{\boldsymbol{\psi}}_r$ of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i^T$.

# Functional regression: estimation

- **Discretized case (model (1))**: this amounts to consider the eigenvalues $l_1 \geq l_2 \geq \ldots$ and corresponding eigenvectors $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ of the covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X}_i \mathbf{X}_i^T)$
  $\longrightarrow$ even if $p > n$, $\boldsymbol{\psi}_r$ $1 \leq r \leq k$, can be well estimated by the eigenvectors (principal components) $\widehat{\boldsymbol{\psi}}_r$ of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i^T$.
- **approximate model**:

$$Y_i \approx \sum_{l=1}^{k} \alpha_l \widehat{\xi}_{il} + \epsilon_i$$

  $\widehat{\xi}_{il} = \sum_{j=1}^p X_i(t_j) \widehat{\psi}_{lj}$
  ($k$ serves as smoothing parameter).
  Coefficients $\alpha_j$ are estimated by least squares, then $\widehat{\beta}_j = \sum_{l=1}^k \widehat{\alpha}_l \widehat{\psi}_{jl}$
  Hall and Horowitz (2008)

# Variable selection: L1 penalized estimators

- **Sparseness**: $S := \#\{j|\beta_j \neq 0\} \ll p$

# Variable selection: L1 penalized estimators

- **Sparseness**: $S := \#\{j|\beta_j \neq 0\} \ll p$
- **Lasso** (Tibshirani, 1996, Bickel, Ritov and Tsybakov, 2009):

$$\widehat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + 2\rho\sum_{i=1}^{n}|\beta_j| \right\},$$

# Variable selection: L1 penalized estimators

- **Sparseness**: $S := \#\{j|\beta_j \neq 0\} \ll p$
- **Lasso** (Tibshirani, 1996, Bickel, Ritov and Tsybakov, 2009):

$$\widehat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + 2\rho\sum_{i=1}^{n}|\beta_j| \right\},$$

- **Dantzig selector** (Candes and Tao, 2007):

$$\widehat{\boldsymbol{\beta}} \;=\; \arg\min \left\{ \|\widetilde{\boldsymbol{\beta}}\|_1 : \left\| \frac{1}{n}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \right\|_\infty \;\leq\; \rho \right\},$$

where $\mathbf{X}$ is the $n \times p$-dimensional matrix with entries $X_{ij}$

# Variable selection: L1 penalized estimators

- **Sparseness**: $S := \#\{j|\beta_j \neq 0\} \ll p$
- **Lasso** (Tibshirani, 1996, Bickel, Ritov and Tsybakov, 2009):

$$\widehat{\boldsymbol{\beta}} \;=\; \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i\boldsymbol{\beta})^2 + 2\rho\sum_{i=1}^{n}|\beta_j| \right\},$$

- **Dantzig selector** (Candes and Tao, 2007):

$$\widehat{\boldsymbol{\beta}} \;=\; \arg\min \left\{ \|\widetilde{\boldsymbol{\beta}}\|_1 : \left\| \frac{1}{n}\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \right\|_\infty \;\leq\; \rho \right\},$$

  where $\mathbf{X}$ is the $n \times p$-dimensional matrix with entries $X_{ij}$

- Unlike L2 penalized estimators (such as Ridge Regression), Lasso and Dantzig selector will find coefficients that are exactly 0

# Variable selection: General conditions

- The diagonal elements of $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\tau\mathbf{X}$ are equal to 1

# Variable selection: General conditions

- The diagonal elements of $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\mathbf{X}^\tau\mathbf{X}$ are equal to 1
- **Restricted eigenvalue assumption RE(S,$c_0$)** (Bickel et al., 2009)

  $$C(S, c_0) = \{\boldsymbol{\delta} \in \mathbb{R}^p, \exists J_0 \subset \{1, \ldots, p\}, |J_0| \leq S, \|\boldsymbol{\delta}_{J_0^c}\|_1 \leq c_0\|\boldsymbol{\delta}_{J_0}\|_1\}$$

  $\longrightarrow$ with high probability $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \in C(S, c_0)$, with $\widehat{\boldsymbol{\beta}}$ Lasso ($c_0 = 3$) or Dantzig ($c_0 = 1$) estimator and $J_0 = J(\boldsymbol{\beta})$ is the set of non null coefficients of $\boldsymbol{\beta}$
- $$\kappa(S, c_0) := \min_{\boldsymbol{\delta} \in C(S,c_0)\setminus\{0\}} \frac{(\boldsymbol{\delta}^T\widehat{\mathbf{\Sigma}}\boldsymbol{\delta})^{1/2}}{\|\boldsymbol{\delta}_{J_0}\|_2} > 0$$
- $RE(S, c_0)$ means that there is a kind of "restricted" positive definiteness which is valid only for vectors in $C(S, c_0)$

# Variable selection: Results

- bounds on prediction loss and $L^1$ loss are obtained under $RE(S, c_0)$
- The bounds depends on the value of $\kappa(S, c_0)$: lower bounds are obtained for great values of $\kappa(S, c_0)$

# Variable selection: Results

- bounds on prediction loss and $L^1$ loss are obtained under $RE(S, c_0)$
- The bounds depends on the value of $\kappa(S, c_0)$: lower bounds are obtained for great values of $\kappa(S, c_0)$
- For "purely" functional predictors, $\kappa(S, c_0)$ tends to zero as $p$ tends to infinity.

  In any case, variable selection such as penalized L1 procedures will not be efficient for this kind of data (at least when they are applied directly, solutions exit: work in progress in that direction)

  When predictors are too heavily correlated, usual variable selection procedures will be not efficient to select a small set of variables that have influence on the response

# Variable selection and Factor Models

- Structure of predictors: **factor model**

$$\mathbf{X}_i \;=\; \mathbf{W}_i + \mathbf{Z}_i,$$

where $\mathbf{W}_i$ and $\mathbf{Z}_i$ are two uncorrelated r. v. in $\mathbb{R}^p$
$Z_{i1}, \ldots, Z_{ip}$ independent with $Var(Z_{ij}) = \sigma_j^2$

# Variable selection and Factor Models

- Structure of predictors: **factor model**

$$\mathbf{X}_i \;=\; \mathbf{W}_i + \mathbf{Z}_i,$$

where $\mathbf{W}_i$ and $\mathbf{Z}_i$ are two uncorrelated r. v. in $\mathbb{R}^p$
$Z_{i1}, \ldots, Z_{ip}$ independent with $Var(Z_{ij}) = \sigma_j^2$

- $W_{ij}$ describes *common* variability while $Z_{ij}$ induces *specific* variability

# Variable selection and Factor Models

- Structure of predictors: **factor model**

$$\mathbf{X}_i \;=\; \mathbf{W}_i + \mathbf{Z}_i,$$

where $\mathbf{W}_i$ and $\mathbf{Z}_i$ are two uncorrelated r. v. in $\mathbb{R}^p$
$Z_{i1}, \ldots, Z_{ip}$ independent with $Var(Z_{ij}) = \sigma_j^2$

- $W_{ij}$ describes *common* variability while $Z_{ij}$ induces *specific* variability
- $\mathbf{\Sigma}$ covariance matrix of $\mathbf{X}_i$; with $\mathbf{\Gamma} = \mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$ covariance matrix of $\mathbf{W}_i$

$$\mathbf{\Sigma} \;=\; \mathbf{\Gamma} + \mathbf{\Psi}$$

- $\mathbf{\Psi} \;=\; Diag(\sigma_1^2 \ldots \sigma_p^2)$.

# Variable selection and Factor Models

- Structure of predictors: **factor model**

$$\mathbf{X}_i \;=\; \mathbf{W}_i + \mathbf{Z}_i,$$

where $\mathbf{W}_i$ and $\mathbf{Z}_i$ are two uncorrelated r. v. in $\mathbb{R}^p$
$Z_{i1}, \ldots, Z_{ip}$ independent with $Var(Z_{ij}) = \sigma_j^2$

- $W_{ij}$ describes *common* variability while $Z_{ij}$ induces *specific* variability
- $\mathbf{\Sigma}$ covariance matrix of $\mathbf{X}_i$; with $\mathbf{\Gamma} = \mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$ covariance matrix of $\mathbf{W}_i$

$$\mathbf{\Sigma} \;=\; \mathbf{\Gamma} + \mathbf{\Psi}$$

- $\mathbf{\Psi} \;=\; Diag(\sigma_1^2 \ldots \sigma_p^2)$.
- A small number of eigenvectors of $\mathbf{\Gamma}$ suffices to approximate $\mathbf{W}_i$ with high accuracy (in spirit: $\mathbf{W}_i$ is of "functional nature")

# Variable selection and Factor Models

- Structure of predictors: **factor model**

$$\mathbf{X}_i \;=\; \mathbf{W}_i + \mathbf{Z}_i,$$

where $\mathbf{W}_i$ and $\mathbf{Z}_i$ are two uncorrelated r. v. in $\mathbb{R}^p$
$Z_{i1}, \ldots, Z_{ip}$ independent with $Var(Z_{ij}) = \sigma_j^2$

- $W_{ij}$ describes *common* variability while $Z_{ij}$ induces *specific* variability
- $\mathbf{\Sigma}$ covariance matrix of $\mathbf{X}_i$; with $\mathbf{\Gamma} = \mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$ covariance matrix of $\mathbf{W}_i$

$$\mathbf{\Sigma} \;=\; \mathbf{\Gamma} + \mathbf{\Psi}$$

- $\mathbf{\Psi} \;=\; Diag(\sigma_1^2 \ldots \sigma_p^2)$.
- A small number of eigenvectors of $\mathbf{\Gamma}$ suffices to approximate $\mathbf{W}_i$ with high accuracy (in spirit: $\mathbf{W}_i$ is of "functional nature")
- Both $\mathbf{W}_i$ and $\mathbf{Z}_i$ are not observed

# Sparse model for Factor Models

- For factor models, the Dantzig selector or the Lasso will retrieve the coefficients of a sparse model provided that the *specific* component $\mathbf{Z}_i$ contributes in a determining way in the variability of $\mathbf{X}_i$. One of the central hypothesis is:
  $Var(Z_{ij}) = \sigma_j^2$ such that for some positive constants $D_1$ and $D_2$

$$(A.1) \qquad 0 < D_1 < \sigma_j^2 < D_2.$$

# Sparse model for Factor Models

- For factor models, the Dantzig selector or the Lasso will retrieve the coefficients of a sparse model provided that the *specific* component $\mathbf{Z}_i$ contributes in a determining way in the variability of $\mathbf{X}_i$. One of the central hypothesis is:
  $Var(Z_{ij}) = \sigma_j^2$ such that for some positive constants $D_1$ and $D_2$

  $$(A.1) \qquad 0 < D_1 < \sigma_j^2 < D_2.$$

- The initial model (1) is normalized as
  $Y_i = \sum_{j=1}^{p} \beta_j^* X_{ij}^* + \epsilon_i,\ i = 1, \ldots, n.$ with $X_{ij}^* = \frac{X_{ij}}{\left(\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right)^{1/2}}$
  and $\beta_j^* = \beta_j \left(\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right)^{1/2}$,

# Sparse model for Factor Models

- For factor models, the Dantzig selector or the Lasso will retrieve the coefficients of a sparse model provided that the *specific* component $\mathbf{Z}_i$ contributes in a determining way in the variability of $\mathbf{X}_i$. One of the central hypothesis is:
  $Var(Z_{ij}) = \sigma_j^2$ such that for some positive constants $D_1$ and $D_2$

  $$(A.1) \qquad 0 < D_1 < \sigma_j^2 < D_2.$$

- The initial model (1) is normalized as
  $Y_i = \sum_{j=1}^{p} \beta_j^* X_{ij}^* + \epsilon_i,\ i = 1, \ldots, n.$ with $X_{ij}^* = \frac{X_{ij}}{\left(\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right)^{1/2}}$
  and $\beta_j^* = \beta_j \left(\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right)^{1/2}$,
- Sparseness. $\sharp\{\beta_j^* | \beta_j^* \neq 0\} \leq S, S \ll p$

# Sparse model for Factor Models

- For factor models, the Dantzig selector or the Lasso will retrieve the coefficients of a sparse model provided that the *specific* component $\mathbf{Z}_i$ contributes in a determining way in the variability of $\mathbf{X}_i$. One of the central hypothesis is:
  $Var(Z_{ij}) = \sigma_j^2$ such that for some positive constants $D_1$ and $D_2$

  $$(A.1) \qquad 0 < D_1 < \sigma_j^2 < D_2.$$

- The initial model (1) is normalized as
  $Y_i = \sum_{j=1}^{p} \beta_j^* X_{ij}^* + \epsilon_i,\ i = 1, \ldots, n.$ with $X_{ij}^* = \frac{X_{ij}}{\left(\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right)^{1/2}}$
  and $\beta_j^* = \beta_j \left(\frac{1}{n}\sum_{i=1}^{n} X_{ij}^2\right)^{1/2},$
- Sparseness. $\sharp\{\beta_j^* | \beta_j^* \neq 0\} \leq S, S \ll p$
- The parameters $\beta_j^*$ (and then $\beta_j$) are estimated either with Lasso or the Dantzig selector

# Sparse model for Factor Model: theoretical results - 1

- In the following $\mathbb{E}(X_{ij}) = 0$ and

$$\sup_j \mathbb{E}(X_{ij}^2) \leq D_0 < \infty.$$

# Sparse model for Factor Model: theoretical results - 1

- In the following $\mathbb{E}(X_{ij}) = 0$ and

$$\sup_j \mathbb{E}(X_{ij}^2) \leq D_0 < \infty.$$

- $Var(Z_{ij}) = \sigma_j^2$ such that for some positive constants $D_1$ and $D_2$

$$(A.1) \qquad 0 < D_1 < \sigma_j^2 < D_2.$$

# Sparse model for Factor Model: theoretical results - 1

- In the following $\mathbb{E}(X_{ij}) = 0$ and

$$\sup_j \mathbb{E}(X_{ij}^2) \leq D_0 < \infty.$$

- $Var(Z_{ij}) = \sigma_j^2$ such that for some positive constants $D_1$ and $D_2$

$$(A.1) \qquad 0 \ < \ D_1 \ < \ \sigma_j^2 \ < \ D_2.$$

- $(A.2)$ There exists a $C_0 < \infty$ such that
  $\sup_{1\leq j,l\leq p} |\frac{1}{n}\sum_{i=1}^n W_{ij}W_{il} - cov(W_{ij}, W_{il})| \leq C_0\sqrt{\log p/n}$
  $\sup_{1\leq j,l\leq p} |\frac{1}{n}\sum_{i=1}^n Z_{ij}Z_{il} - cov(Z_{ij}, Z_{il})| \leq C_0\sqrt{\log p/n}$
  $\sup_{1\leq j,l\leq p} |\frac{1}{n}\sum_{i=1}^n Z_{ij}W_{il}| \leq C_0\sqrt{\log p/n}$
  $\sup_{1\leq j,l\leq p} |\frac{1}{n}\sum_{i=1}^n X_{ij}X_{il} - cov(X_{ij}, X_{il})| \leq C_0\sqrt{\log p/n}$
  hold simultaneously with probability $A(n,p) > 0$, where $A(n,p) \to 1$ as $n, p \to \infty$, $\frac{\log p}{n} \to 0$.
  $\longrightarrow$ Condition satisfied for instance for normally distributed random vectors

# Sparse model for Factor Model: theoretical results - 2

- **RE condition**. *Let $c_0 = 1, 3$ and assume (A.1), (A.2) as well as $D_1 - 3C_0 n^{-1/2}\sqrt{\log p} > 0$. Then for $S \leq p/2$ the following inequality holds with probability $A(n,p)$*

$$\kappa(S, c_0) := \min_{\boldsymbol{\delta} \in C(S, c_0 \setminus \{0\}} \frac{[\boldsymbol{\Delta}^T \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^* \mathbf{X}_i^{*T} \boldsymbol{\delta}]^{1/2}}{\|\boldsymbol{\delta}_{J_0}\|_2}$$

$$\geq \left( \frac{D_1}{D_0 + C_0 n^{-1/2}\sqrt{\log p}} - \frac{8 S c_0 C_0 n^{-1/2}\sqrt{\log p}}{D_1 - 3C_0 n^{-1/2}\sqrt{\log p}} \right)_+^{1/2}.$$

# Sparse model for Factor Model: theoretical results - 2

- **RE condition**. *Let* $c_0 = 1, 3$ *and assume (A.1), (A.2) as well as* $D_1 - 3C_0 n^{-1/2}\sqrt{\log p} > 0$. *Then for* $S \leq p/2$ *the following inequality holds with probability* $A(n,p)$

$$\kappa(S, c_0) := \min_{\boldsymbol{\delta} \in C(S, c_0 \setminus \{0\}} \frac{[\boldsymbol{\Delta}^T \frac{1}{n} \sum_{i=1}^n \mathbf{X}_i^* \mathbf{X}_i^{*T} \boldsymbol{\delta}]^{1/2}}{\|\boldsymbol{\delta}_{J_0}\|_2}$$

$$\geq \left( \frac{D_1}{D_0 + C_0 n^{-1/2}\sqrt{\log p}} - \frac{8 S c_0 C_0 n^{-1/2}\sqrt{\log p}}{D_1 - 3C_0 n^{-1/2}\sqrt{\log p}} \right)_+^{1/2}.$$

- For $n$ and $p$ large enough $\kappa(S, c_0) > 0$ holds with high probability and thus the RE condition is satisfied. Then, results of Bickel et al. (2009) imply that bounds on prediction loss and L1 loss can be derived.
- In our Factor Model setup the Lasso or the Dantzig selector will retrieve the coefficients of a sparse model

# Sparse model for Factor Model: some remarks

- The assumption (A.1) plays a crucial role: bounds depend on the smallest value of $\sigma_j^2$ the variances of the $Z_{ij}$. When this value is too small, the estimation procedure will not be efficient.

# Sparse model for Factor Model: some remarks

- The assumption (A.1) plays a crucial role: bounds depend on the smallest value of $\sigma_j^2$ the variances of the $Z_{ij}$. When this value is too small, the estimation procedure will not be efficient.
- The traditional sparseness assumption is restrictive:

  The *common* variability of the predictors may also influence the response (each component $\mathbf{W}_i$ and $\mathbf{Z}_i$ may posses a significant influence).

# The augmented model - Introduction

- If $\mathbf{W}_i$ and $\mathbf{Z}_i$ were known, a possible improvement of the model would be

$$Y_i = \sum_{j=1}^{p} \beta_j^* W_{ij} + \sum_{j=1}^{p} \beta_j Z_{ij} + \epsilon_i, \; i = 1, \ldots, n$$

with different sets of parameters $\beta_j^*$ and $\beta_j$. Model can be rewritten as

$$Y_i = \sum_{j=1}^{p} (\beta_j^* - \beta_j) W_{ij} + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$$

# The augmented model - Introduction

- If $\mathbf{W}_i$ and $\mathbf{Z}_i$ were known, a possible improvement of the model would be

$$Y_i = \sum_{j=1}^{p} \beta_j^* W_{ij} + \sum_{j=1}^{p} \beta_j Z_{ij} + \epsilon_i, \; i = 1, \ldots, n$$

with different sets of parameters $\beta_j^*$ and $\beta_j$. Model can be rewritten as

$$Y_i = \sum_{j=1}^{p} (\beta_j^* - \beta_j) W_{ij} + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i$$

- $\mathbf{W}_i$ can be rewritten in terms of principal components (the $W_{ij}$ are heavily correlated). Denote $\lambda_1 \geq \lambda_2 \geq \ldots$ the eigenvalues of the standardized covariance matrix of $\mathbf{W}_i$, $\frac{1}{p}\mathbf{\Gamma} = \frac{1}{p}\mathbb{E}(\mathbf{W}_i \mathbf{W}_i^T)$ and $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ corresponding orthonormal eigenvectors. Then

$$\mathbf{W}_i = \sum_{r=1}^{p} (\boldsymbol{\psi}_r^T \mathbf{W}_i) \boldsymbol{\psi}_r$$

# The augmented model - Definition

- Assuming that a small number of leading PC suffice to describe the effects of $\mathbf{W}_i$ leads to the following augmented model

$$Y_i \;=\; \sum_{r=1}^{k} \alpha_r \xi_{ir} + \sum_{j=1}^{p} \beta_j X_{ij} \;+\; \epsilon_i, \quad i = 1, \ldots, n,$$

where $\xi_{ir} = \boldsymbol{\psi}_r^T \mathbf{W}_i / \sqrt{p\lambda_r}$

# The augmented model - Definition

- Assuming that a small number of leading PC suffice to describe the effects of $\mathbf{W}_i$ leads to the following augmented model

$$Y_i \;=\; \sum_{r=1}^{k} \alpha_r \xi_{ir} + \sum_{j=1}^{p} \beta_j X_{ij} \;+\; \epsilon_i, \quad i = 1, \ldots, n,$$

  where $\xi_{ir} = \boldsymbol{\psi}_r^T \mathbf{W}_i / \sqrt{p\lambda_r}$

- $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T \in \mathbb{R}^k$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ vectors of parameters.
- the dimension $k$ is fixed
- the vector $\boldsymbol{\beta}$ satisfies the sparseness condition for a fixed $S \ll p$.

# The augmented model - Estimation

- **Step 1. Estimation of $\boldsymbol{\xi_{ir}}$.** As the $W_{ij}$ are unknown, we use the eigenelements of standardized empirical covariance matrix $\frac{1}{np}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i$: $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots$ eigenvalues and $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ orthonormal eigenvectors.

  $\longrightarrow \xi_{ir}$ is estimated by $\widehat{\xi}_{ir} = \widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i/\sqrt{p\widehat{\lambda}_r}$

# The augmented model - Estimation

- **Step 1. Estimation of $\boldsymbol{\xi_{ir}}$.** As the $W_{ij}$ are unknown, we use the eigenelements of standardized empirical covariance matrix $\frac{1}{np}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i$: $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots$ eigenvalues and $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ orthonormal eigenvectors.

  $\longrightarrow \xi_{ir}$ is estimated by $\widehat{\xi}_{ir} = \widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i/\sqrt{p\widehat{\lambda}_r}$

- **Step 2. Decorrelation of the $\boldsymbol{X_{ij}}$.** In the second term, $X_{ij}$ is replaced by $(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j$, where $\widehat{\mathbf{P}}_k = \mathbf{I}_p - \sum_{r=1}^{k}\widehat{\boldsymbol{\psi}}_r\widehat{\boldsymbol{\psi}}_r^{\tau}$.

# The augmented model - Estimation

- **Step 1. Estimation of $\boldsymbol{\xi_{ir}}$.** As the $W_{ij}$ are unknown, we use the eigenelements of standardized empirical covariance matrix $\frac{1}{np}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i$: $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots$ eigenvalues and $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ orthonormal eigenvectors.

  $\longrightarrow \xi_{ir}$ is estimated by $\widehat{\xi}_{ir} = \widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i/\sqrt{p\widehat{\lambda}_r}$

- **Step 2. Decorrelation of the $\boldsymbol{X_{ij}}$.** In the second term, $X_{ij}$ is replaced by $(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j$, where $\widehat{\mathbf{P}}_k = \mathbf{I}_p - \sum_{r=1}^{k}\widehat{\boldsymbol{\psi}}_r\widehat{\boldsymbol{\psi}}_r^{\tau}$.

- After normalization, this finally leads to the approximated model

$$Y_i = \sum_{r=1}^{k}\widetilde{\alpha}_r\widehat{\xi}_{ir} + \sum_{j=1}^{p}\widetilde{\beta}_j\frac{(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j}{\left(\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2\right)^{1/2}} + \widetilde{\epsilon}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

# The augmented model - Estimation

- **Step 1. Estimation of $\boldsymbol{\xi_{ir}}$.** As the $W_{ij}$ are unknown, we use the eigenelements of standardized empirical covariance matrix $\frac{1}{np}\sum_{i=1}^{n}\mathbf{X}_i^T\mathbf{X}_i$: $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \ldots$ eigenvalues and $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ orthonormal eigenvectors.

  $\longrightarrow \xi_{ir}$ is estimated by $\widehat{\xi}_{ir} = \widehat{\boldsymbol{\psi}}_r^T\mathbf{X}_i/\sqrt{p\widehat{\lambda}_r}$

- **Step 2. Decorrelation of the $\boldsymbol{X_{ij}}$.** In the second term, $X_{ij}$ is replaced by $(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j$, where $\widehat{\mathbf{P}}_k = \mathbf{I}_p - \sum_{r=1}^{k}\widehat{\boldsymbol{\psi}}_r\widehat{\boldsymbol{\psi}}_r^{\tau}$.

- After normalization, this finally leads to the approximated model

$$Y_i = \sum_{r=1}^{k}\widetilde{\alpha}_r\widehat{\xi}_{ir} + \sum_{j=1}^{p}\widetilde{\beta}_j\frac{(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j}{\left(\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2\right)^{1/2}} + \widetilde{\epsilon}_i + \epsilon_i, \quad i = 1, \ldots, n,$$

- Lasso or Dantzig selector are used to estimate the vector of parameters $(\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_k, \widetilde{\beta}_1, \ldots, \widetilde{\beta}_k)^T \longrightarrow$ estimators of $\alpha_r$ and $\beta_j$ follow

# Efficiency of principal components -1

Question: for which setting the empirical eigenelements of the empirical covariance matrix of $\mathbf{X}_i$ approximate well the eigenelements of the covariance matrix of the unknown $\mathbf{W}_i$.

# Efficiency of principal components -1

Question: for which setting the empirical eigenelements of the empirical covariance matrix of $\mathbf{X}_i$ approximate well the eigenelements of the covariance matrix of the unknown $\mathbf{W}_i$.

- $\longrightarrow$ $\mathbf{W}_i$: $\lambda_1 \geq \lambda_2 \geq \ldots, \boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ eigenelements of $\frac{1}{p}\boldsymbol{\Gamma}$,
  $\longrightarrow$ $\mathbf{X}_i$: $\mu_1 \geq \mu_2 \geq \ldots, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots$ eigenelements of $\frac{\mathbf{1}}{\boldsymbol{p}}\boldsymbol{\Sigma}$
  $\longrightarrow \widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq, \widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ eigenelements of the standardized empirical covariance matrix $\frac{1}{p}\widehat{\boldsymbol{\Sigma}}$

# Efficiency of principal components -1

Question: for which setting the empirical eigenelements of the empirical covariance matrix of $\mathbf{X}_i$ approximate well the eigenelements of the covariance matrix of the unknown $\mathbf{W}_i$.

- $\longrightarrow$ $\mathbf{W}_i$: $\lambda_1 \geq \lambda_2 \geq \ldots$, $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ eigenelements of $\frac{1}{p}\boldsymbol{\Gamma}$,
  $\longrightarrow$ $\mathbf{X}_i$: $\mu_1 \geq \mu_2 \geq \ldots$, $\boldsymbol{\delta}_1, \boldsymbol{\delta}_2, \ldots$ eigenelements of $\frac{\mathbf{1}}{\boldsymbol{p}}\boldsymbol{\Sigma}$
  $\longrightarrow \widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq$, $\widehat{\boldsymbol{\psi}}_1, \widehat{\boldsymbol{\psi}}_2, \ldots$ eigenelements of the standardized empirical covariance matrix $\frac{1}{p}\widehat{\boldsymbol{\Sigma}}$

- (A.3) $\min_{j,l\leq k, j\neq l} |\lambda_j - \lambda_l| \geq v(k), \quad \min_{j\leq k} \lambda_j \geq v(k)$ for some $1 \geq v(k) > 0$.

- (A.4) $C_0(\log p/n)^{1/2} \geq \frac{D_0}{pv(k)}$ and $v(k) \geq 3C_0(\log p/n)^{1/2}$.

# Efficiency of principal components -2

- Under the above Assumptions (A.2)-(A.4) and under events with probability $A(n,p)$ we have for all $r \leq k$ and all $j = 1, \ldots, p$

$$|\lambda_r - \widehat{\lambda}_r| \leq \frac{D_2}{p} + C_0(\log p/n)^{1/2},$$

$$|\mu_r - \widehat{\lambda}_r| \leq C_0(\log p/n)^{1/2}$$

$$\|\boldsymbol{\psi}_r - \widehat{\boldsymbol{\psi}}_r\|_2 \leq 5\frac{\frac{D_2}{p} + C_0(\log p/n)^{1/2}}{v(k)},$$

$$\|\boldsymbol{\delta}_r - \widehat{\boldsymbol{\psi}}_r\|_2 \leq 3\frac{C_0(\log p/n)^{1/2}}{v(k)}$$

# Efficiency of principal components - 3

- Assume (A.1) and (A.2). There then exist constants $M_1$, $M_2 < \infty$, such that for all $n, p, k$ satisfying (A.3) and (A.4), all $j \in \{1, \ldots, p\}$,

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2 \geq \sigma_j^2 - M_1\frac{kn^{-1/2}\sqrt{\log p}}{v(k)^{1/2}},$$

$$|\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j^2 - \sigma_j^2| \leq \mathbb{E}\left((\mathbf{P}_k\mathbf{W_i})_j^2\right) + M_2\frac{kn^{-1/2}\sqrt{\log p}}{v(k)^{3/2}},$$

  hold with probability $A(n, p)$.

- If $\mathbf{X}_i$ satisfies a $k$-dimensional factor model, $\mathbf{P}_k\mathbf{W}_i = 0$, The results state that for $n$ and $p$ large $(\widehat{\mathbf{P}}_k\mathbf{X}_i)_j$ behaves "in average" similar to the specific variables $Z_{ij}$.

# Augmented model: properties

- the restricted eigenvalues conditions is satisfied with high probability. Define $\mathbf{\Phi}_i := (\widehat{\xi}_{i1}, \ldots, \widehat{\xi}_{ik}, \widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip})^T$, where $\widetilde{X}_{ij} = (\widehat{\mathbf{P}}_k \mathbf{P}_i)_j$.
- (A.5) $D_1/2 > M_1 \frac{kn^{-1/2}\sqrt{\log p}}{v(k)^{1/2}}$
- *Assume (A.1) and (A.2). There then exists a constant $M_3 < \infty$ such that for all $n, p, k, S$, $k + S \leq (k + p)/2$, satisfying (A.3)-(A.5), and $c_0 = 1, 3$*

$$\kappa_S(k+S, k+S, c_0) := \min_{\boldsymbol{\delta} \in C(k+S, c_0) \setminus \{0\}} \frac{[\boldsymbol{\delta}^T \frac{1}{n} \sum_{i=1}^n \mathbf{\Phi}_i \mathbf{\Phi}_i^T \boldsymbol{\delta}]^{1/2}}{\|\boldsymbol{\delta}_{J_0}\|_2}$$

$$\geq \left( \frac{D_1}{D_0 + C_0 n^{-1/2}\sqrt{\log p}} - \frac{8(k+S)c_0 M_3 k^2 n^{-1/2}\sqrt{\log p}}{v(k)D_1 - kv(k)^{1/2} n^{-1/2}\sqrt{\log p}} \right)_+^{1/2},$$

*holds with probability $A(n, p)$.*

# Bounds for the Dantzig selector

- $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- Compute the Dantzig selector with $\rho = A\sigma\sqrt{\frac{\log(k+p)}{n}} + \frac{kM_4 \sum_{r=1}^{k} |\alpha_r|}{v(k)^2}\sqrt{\frac{\log p}{n}}$, $A < \sqrt{2}$, $M_4$ is a positive constant.
- Assume (A.1)-(A.3)
- If $M_5$ is sufficiently large, then for all $n, p, k, k + S \leq (k + p)/2$, satisfying (A.4), (A.5) as well as $\kappa(k + S, c_0) > 0$ the following inequalities hold with probability at least $A(n, p) - (p + k)^{-A^2/2}$

$$\sum_{r=1}^{k} |\widehat{\alpha}_r - \alpha_r| \leq \frac{8(k+S)}{\kappa^2}\rho\left(1 + \frac{k(D_0 + C_0 n^{-1/2}\sqrt{\log p})^{1/2}}{(D_1 - M_1 \frac{kn^{-1/2}\sqrt{\log p}}{v(k)^{1/2}})^{1/2}}\right),$$

$$\sum_{j=1}^{p} |\widehat{\beta}_j - \beta_j| \leq \frac{8(k+S)}{\kappa^2(D_1 - M_1 \frac{kn^{-1/2}\sqrt{\log p}}{v(k)^{1/2}})^{1/2}}\rho,$$

where $\kappa = \kappa(k + S, 1)$.