

# Essential Regression

Latent factor regression with interpretable factors

Marten Wegkamp

Department of Mathematics  
Department of Statistics and Data Science

Cornell University

Leiden, 17 June 2019

Van harte gefeliciteerd!

Meeting in honor of the 60th birthday of Aad van der Vaart



Joint work with

Xin Bing

Florentina Bunea

Yang Ning

Seth Strimas-Mackey

DEPARTMENT OF STATISTICS AND DATA SCIENCE  
CORNELL UNIVERSITY

Talk based on

- Xin Bing, Florentina Bunea, Yang Ning and W.  
*Sparse Latent Factor Models with Pure Variables for Overlapping Clustering.*  
arXiv:1704.06977
- Xin Bing, Florentina Bunea, W. and Seth Strimas-Mackey.  
*Essential Regression.*  
arXiv: 1905.12696
- Xin Bing, Florentina Bunea and W.  
*A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics.*  
arXiv:1805.06837

# Outline

## Introduction

## Essential Regression model

- Model
- Essential regression framework
- Identifiability
- Example
- Constructive recovery
- Identifiability results

## Estimation of $A$

- Estimation of pure variable partition
- Estimation of  $A_J$

## Estimation of $\beta$

- Estimator
- Rates
- Asymptotic normality
- Regression with cluster covariates
- Simulation
- Discussion

## Prediction of $Y$

- Predictor
- Rates
- Regression with cluster covariates
- Comparison with Factor Regression and PCR

## Summary

# Introduction

# Low dimensional regression models

Observed:  $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$  with  $p$  large ( $p > n$ ).

- **Sparse linear regression:**

$Y = \theta'X + e$ ;  $\theta \in \mathbb{R}^p$  supported on  $\mathbb{R}^K$  (sparse),  $K \ll p$ .

Strong conditions on  $X$  (incoherence, etc.)

- **Factor regression:**

There exists  $Z \in \mathbb{R}^K$  such that  $Y = \beta'Z + \varepsilon$  and  $X = AZ + E$  with  $\beta \in \mathbb{R}^K$ ,  $A \in \mathbb{R}^{p \times K}$ ;  $K \ll p$ ,  $Z$  not observable.

$E = 0$  yields highly correlated  $X$ .

# Brief History

- Factor models  $X = AZ + E$  extensively studied.

Anderson and Rubin (1956). Many sufficient conditions under which  $A$  and  $\text{Cov}(Z)$  are uniquely defined.

- Mathematically convenient, not practically interpretable.

(1)  $A'A$  diagonal,  $\text{Cov}(Z) = Id$ .

- Practically interpretable, but rigid.

Existence of **known** simple structure on  $A$  = pure variables.

- 50 years later: Practically interpretable.

(2) Existence of **unknown** simple structure on  $A$ .

All positive: Donoho and Stodden, 2008; Arora et al (2012, 2013); Bittorf et al (2015)

Arbitrary signs: Bing, Bunea, Ning and W.

# A Factor Regression Model for Inference

- Classical Factor Regression Models

- **Focus = Prediction**

- Non-Bayesian: Factors  $Z$  mimic principal components of  $X$ .

Stock and Watson, 2002

- Bayesian: Computationally very involved.

- Quasi-Bayesian (Variational Inference, Mean Field):  
Very sensitive to initializations; Multiple Minima, Uncorrelated with target; etc.

Blei, 2018; Ghorbani, Javadi, Montanari (2018), Mukherjee et al (2018).

- Essential Regression

- **Focus = Inference**

## Essential Regression Model

# An interpretable factor regression model

## Essential Regression

- A **new** type of factor regression:

$$\begin{aligned}Y &= \beta'Z + \varepsilon \\X &= AZ + E.\end{aligned}$$

- $Z, E, \varepsilon$  independent, mean zero subGaussian.
- **New:** Regression on **interpretable**, but non-observable,  $Z$ .
  - Interpretable  $Z$  = structure on  $A$  that captures the *essence* of  $X$ .
  - Conditions on  $A$  and  $\Sigma^Z$
  - $K$  is **unknown**.

# Essential Regression

## Conditions on $A$

### **Mixing/Scaling:**

$\|A_{j\cdot}\|_1 \leq 1$  for all  $j \in [p]$ .

### **Pure variables:**

For every  $k \in [K]$ , there exists at least two  $j \neq \ell \in [p]$ , such that  $|A_{jk}| = |A_{\ell k}| = 1$  and  $A_{jk'} = A_{\ell k'} = 0$  for any  $k' \neq k$ .

## Conditions on $\Sigma^Z$

### **Distinct latent factors:**

$\Sigma^Z > 0$  and  $\Delta(\Sigma^Z) := \min_{1 \leq a < b \leq K} \Sigma_{aa}^Z \wedge \Sigma_{bb}^Z - |\Sigma_{ab}^Z| > 0$ .

# Identifiable sparse latent variable models

## The pure variable assumption

A pure variable  $X_j$  associates with **only one** latent factor  $Z_k$ .

## Pure variables are crucial in building overlapping clusters

- Clusters  $G_k := \{j \in \{1, \dots, p\} : A_{ja} \neq 0\}$  are defined by **unobserved**  $Z_k$  ( $Z_k =$  (biological) function).
- A pure variable  $X_j$  is an observable proxy for a  $Z_k$  (**Observable**  $X_j$  performs function  $Z_k$ . It anchors  $G_k$ ).

# Identifiability

## Definition

**Pure variable set**  $I$  is the index set of pure variables.

$$\begin{aligned} I &= I_1 \cup \dots \cup I_K \quad (\text{partition}) \\ I_k &= \{i \in [p] : |A_{ik}| = 1, A_{i\ell} = 0, \text{ for any } \ell \neq k\} \end{aligned}$$

Challenge:

- To find  $K$ :  **$K$  is unknown!!**
- To distinguish between the set  $I$  and its complement  $J$ .

# Identifiability

## Identifiability of model parameters (Bing, Bunea, Ning, W.)

The set of pure variables  $I$ ,  
the covariance matrix  $\Sigma^Z$   
and allocation matrix  $A$ ,  
whence  $\beta$ ,  
are uniquely defined, up to a **signed permutation matrix**.

# Essential Regression: a benchmark example

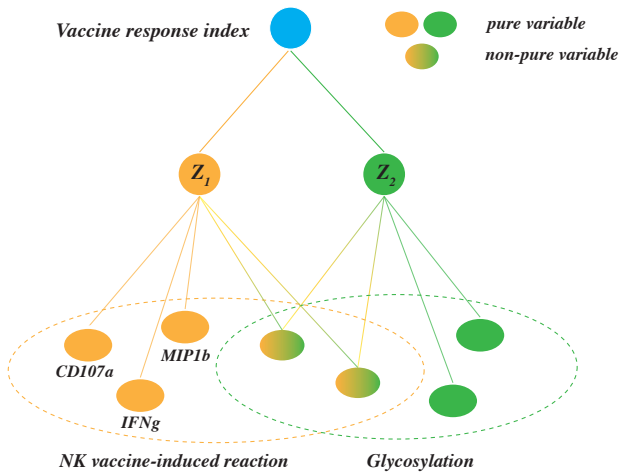
- Study on vaccine-induced humoral immune responses in monkeys conducted at BioEng MIT + NIH.

Goal: Determine antibody-centric **mechanisms** of vaccine induced protection against SIV.

- $X$  = antibody functional and biophysical properties,  $p = 191$ .  
(Fc effector functions, glycosylation profiles, etc)
- $Y$  = vaccine protection level index.
- $n = 60$  monkeys.
- $Z$  = **Mechanisms**

Each  $Z$  has pure variables.

(Ackerman, Das et al, Nature Medicine 2018)



# Recovery

Let  $\Sigma^Z = \text{diag}(1, 2, 3)$  and

$$A = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/2 & -1/2 & 0 \\ 2/3 & 1/6 & -1/6 \end{pmatrix}$$

$$A\Sigma^z A^T =$$

$$\begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & 1 & -1 & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\ 1 & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ 0 & 0 & 0 & 2 & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & 0 & -2 & 2 & 0 & 0 & -1 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 0 & 3 & 3 & 0 & -1/2 \\ 1/2 & -1/2 & 1/2 & 1 & -1 & 0 & 0 & 3/4 & 1/6 \\ 2/3 & -2/3 & 2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & 7/12 \end{pmatrix}$$

Look at diagonal!

$$A\Sigma^Z A^T + \Sigma^E =$$

$$\begin{pmatrix} * & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & * & -1 & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\ 1 & -1 & * & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ 0 & 0 & 0 & * & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & 0 & -2 & * & 0 & 0 & -1 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & * & 3 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 0 & 3 & * & 0 & -1/2 \\ 1/2 & -1/2 & 1/2 & 1 & -1 & 0 & 0 & * & 1/6 \\ 2/3 & -2/3 & 2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & * \end{pmatrix}$$

Oops! \* are arbitrary positive numbers

# Recovery

**Option 1:** Look for sparsity pattern  
(works for non-negative  $A$  and diagonal  $\Sigma^Z$ )

# Pure versus impure

Option 2: Look for maxima

## Proposition (Pure variable test)

For each row  $i$ , define

$$M_i := \max_{j \in [p] \setminus \{i\}} |\Sigma_{ij}|$$

$$S_i := \{j \in [p] \setminus \{i\} : |\Sigma_{ij}| = M_i\}.$$

For given  $A$  and its induced pure variable set  $I$ , we have

$$i \in I \iff M_i = \max_{k \in [p] \setminus \{i\}} |\Sigma_{ki}| \text{ for all } j \in S_i.$$

Look for maxima in  $(\Sigma_{ij})_{i \neq j}$

$$\begin{pmatrix} * & -1 & 1 & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ -1 & * & -1 & 0 & 0 & 0 & 0 & -1/2 & -2/3 \\ 1 & -1 & * & 0 & 0 & 0 & 0 & 1/2 & 2/3 \\ 0 & 0 & 0 & * & -2 & 0 & 0 & 1 & -1/3 \\ 0 & 0 & 0 & -2 & * & 0 & 0 & -1 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & * & 3 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 0 & 3 & * & 0 & -1/2 \\ 1/2 & -1/2 & 1/2 & 1 & -1 & 0 & 0 & * & 1/6 \\ 2/3 & -2/3 & 2/3 & -1/3 & 1/3 & -1/2 & -1/2 & 1/6 & * \end{pmatrix}$$

We find  $\mathcal{I} = \{\{1, 2, 3\}, \{4, 5\}, \{6, 7\}\}$  and  $J = \{8, 9\}$ .

# Identifiability

## Theorem (Bing, Bunea, Ning, W.)

We can construct the pure variable set  $I$  and its partition  $\mathcal{I} =: \{I_1, \dots, I_K\}$  uniquely from  $\Sigma = \text{Cov}(\mathbf{X})$ , up to label permutations.

## Theorem (Bing, Bunea, Ning, W.)

- There exists a unique matrix  $A$ , up to a **signed permutation**, such that  $\mathbf{X} = \mathbf{AZ} + \mathbf{E}$ .
- The associated overlapping clusters  $G_1, \dots, G_K$  are identifiable, up to label switching.

## Estimation of $A$

# Estimation

## Outline

- Estimation of  $I$  and its partition
- Estimation of  $A_I$ .
- Estimation of  $A_J$ .

# Estimation of $I$

## Algorithm idea

- Use the constructive characterization of  $I$  at the population level.
- Replace  $\Sigma$  by the sample covariance  $\hat{\Sigma}$ .
- Allow for tolerance  $\delta =: \|\hat{\Sigma} - \Sigma\|_{\infty}$  when comparing maxima.
- Algorithm returns partition  $\hat{\mathcal{I}}$  (and  $\hat{K}$  and  $\hat{I}$ ).

# Estimation of the partition of pure variables

## Conditions

- $X$  is subGaussian with subGaussian constant  $\sigma^2$ .  
(This implies  $\delta = O(\sigma^2 \sqrt{(\log p)/n})$ .)
- $\Delta(\Sigma^Z) := \nu > 2 \max \left( 2\delta, \sqrt{2\delta \|\Sigma^Z\|_\infty} \right)$

## Definition (nearly pure variable set)

$$\begin{aligned} J_1 &= \{i \in J : \text{there exists } k \text{ such that } |A_{ik}| \geq 1 - 4\delta/\nu\} \\ J_1^k &= \{i \in J_1 : |A_{ik}| \geq 1 - 4\delta/\nu\}. \end{aligned}$$

Hence  $\{J_1^1, \dots, J_1^K\}$  forms a partition of  $J_1$ .

# Estimation of the partition of pure variables

## Theorem (Bing, Bunea, Ning, W.)

Under the above conditions, with high probability,

- (a)  $\hat{K} = K$ .
- (b)  $I \subseteq \hat{I} \subseteq I \cup J_1$ .

Moreover, there exists a label permutation  $\pi$ , such that

- (c)  $I_{\pi(k)} \subseteq \hat{I}_k \subseteq I_{\pi(k)} \cup J_1^{\pi(k)}$  for each  $k \in [K]$ .

# Estimation of the partition of pure variables

Minimal recovery mistakes: no conditions on  $A$

|            |                                        |                     |
|------------|----------------------------------------|---------------------|
| Pure       | $(1, 0, 0, 0, 0, 0)$                   | In, correct.        |
| Quasi Pure | $(0.99, 0.01, 0, 0, 0, 0)$             | In, slight mistake. |
| Impure     | $(0.25, 0.25, 0.001, 0.099, 0.2, 0.2)$ | Out, correct.       |

# Estimation of the allocation matrix $A$

Rearrange  $\Sigma$  and  $A$  as

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix} \text{ and } A = \begin{bmatrix} A_I \\ A_J \end{bmatrix}.$$

so

$$\Sigma = \begin{bmatrix} \Sigma_{II} & \Sigma_{IJ} \\ \Sigma_{JI} & \Sigma_{JJ} \end{bmatrix} = \begin{bmatrix} A_I \Sigma^Z A_I^T & A_I \Sigma^Z A_J^T \\ A_J \Sigma^Z A_I^T & A_J \Sigma^Z A_J^T \end{bmatrix} + \begin{bmatrix} \Sigma_{II}^E & 0 \\ 0 & \Sigma_{JJ}^E \end{bmatrix}.$$

# Estimation of the allocation submatrix $A_J$

- For each row, compute  $\alpha_j := A_j$ .
- $\theta_j = \left[ |I_1|^{-1} \sum_{i \in I_1} A_{i1} \Sigma_{ij}, \dots, |I_K|^{-1} \sum_{i \in I_K} A_{iK} \Sigma_{ij} \right]^T$
- $\theta_j = \Sigma^Z \alpha_j$  and  $\|\alpha_j\|_1 \leq 1$
- $\tilde{\alpha}_j = \hat{\Omega} \hat{\theta}_j = \alpha_j + \hat{\Omega}(\hat{\theta}_j - \theta_j) + (\hat{\Omega} \Sigma^Z - I) \alpha_j$
- $\|\tilde{\alpha}_j - \alpha_j\|_\infty \leq \|\hat{\Omega}\|_{\infty,1} \|\hat{\theta}_j - \theta_j\|_\infty + \|\hat{\Omega} \Sigma^Z - I\|_\infty \|\alpha_j\|_1$
- **Linear Program:**  $\min_{t>0, \Omega \in \mathbb{R}^{\hat{K} \times \hat{K}}} t$  such that  
 $\Omega = \Omega^T, \|\Omega \hat{\Sigma}^Z - I\|_\infty \leq \lambda t, \|\Omega\|_{\infty,1} \leq t$
- **Projection:**  $\hat{\alpha}_j = \arg \min \|\alpha\|_1$  s.t.  $\|\alpha - \tilde{\alpha}_j\|_\infty \leq \mu$ .

Set  $\delta' = (8\|C\|_\infty/\nu - 3) \|\hat{\Sigma} - \Sigma\|_\infty$ .

Theorem (Bing, Bunea, Ning, W.)

Provided **SubGaussian errors, distinguishable factors**  $\Delta(\Sigma^Z)$

For all  $1 \leq q \leq \infty$  and  $1 \leq i \leq p$ ,

$$\min_{P \in \mathcal{H}_K} \|\hat{A}_i - (AP)_i\|_q \leq 10(\|A_i\|_0)^{1/q} \|\Sigma^Z\|_\infty^{-1} \delta'$$

with probability larger than  $1 - c_1(n \vee p)^{-c_2}$ .

## Estimation of $\beta$

# Estimation in Essential Regression

$$\begin{aligned}\beta &= [\text{Cov}(Z)]^{-1} \text{Cov}(Z, Y) \\ &= (\Sigma^Z)^{-1} (A^T A)^{-1} A^T \text{Cov}(X, Y) \\ &= (\Sigma^Z)^{-1} (A_I^T A_I)^{-1} A_I^T \text{Cov}(X_I, Y).\end{aligned}$$

- $K = \dim(Z)$ , **unknown**
- $\Sigma^Z = \text{Cov}(Z)$ ,  $Z$  **not observable**
- Pure variable set  $I$ , **unknown**

# Estimation in Essential Regression

(1) We have estimated:

Number of factors  $\widehat{K}$ .

All non-overlapping groups  $\widehat{l}_k$  of pure variables.

All pure variables  $\widehat{l}$  and therefore  $\widehat{A}_{\widehat{l}}$ .

(2) Estimate  $\Sigma^z$  by  $\widehat{\Sigma}^z$ .

$$\widehat{\Sigma}_{aa}^z = \frac{1}{|\widehat{l}_a|(|\widehat{l}_a| - 1)} \sum_{i \neq j \in \widehat{l}_a} |\widehat{\Sigma}_{ij}|, \quad \widehat{\Sigma}_{ab}^z = \frac{1}{|\widehat{l}_a||\widehat{l}_b|} \sum_{i \in \widehat{l}_a, j \in \widehat{l}_b} \widehat{A}_{ia} \widehat{A}_{jb} \widehat{\Sigma}_{ij}.$$

(3) Plug-in estimator.  $K < n$ .

$$\widehat{\beta} = (\widehat{\Sigma}^z)^{-1} (\widehat{A}_{\widehat{l}}^T \widehat{A}_{\widehat{l}})^{-1} \widehat{A}_{\widehat{l}}^T (n^{-1} X_{\widehat{l}}^T Y).$$

# Minimax optimal lower bounds in $\ell_2$ norm

Target (minimax) rate for estimators of  $\beta$  (Bing, Bunea, W, Strimas-Mackey)

$$\inf_{\hat{\beta}} \sup_{(\beta, \Sigma^Z, A) \in \mathcal{S}(R, m)} \mathbb{P} \left\{ \|\hat{\beta} - \beta\| \geq c' \left( 1 \vee \frac{R}{\sqrt{m}} \right) \cdot \sqrt{\frac{K}{n}} \right\} \geq c''$$

- Factor  $\left( 1 \vee \frac{R}{\sqrt{m}} \right)$  = price to pay for not observing  $Z$ .

$$\mathcal{S}(R, m) = \left\{ (\beta, \Sigma^Z, A) : \|\beta\| \leq R, C_{\min} \leq \lambda_{\min}(\Sigma^Z) \leq \lambda_{\max}(\Sigma^Z) < C_{\max}, \right. \\ \left. A \text{ satisfies convex and pure with } \min_k |I_k| = m \right\}.$$

- Errors-in-variables models, Belloni et al. (2017).

Estimator:

$$\hat{\beta} = \left(\hat{\Sigma}^z\right)^{-1} \left(\hat{A}_I^T \hat{A}_I\right)^{-1} \hat{A}_I^T \left(\frac{1}{n} X_I^T Y\right)$$

Minimax adaptive estimation of  $\beta$  (Bing, Bunea, W, Strimas-Mackey)

$$\text{W.h.p. } \inf_P \left\| \hat{\beta} - P\beta \right\| \lesssim \left(1 \vee \frac{\|\beta\|}{\sqrt{m}}\right) \sqrt{\frac{K \log(p \vee n)}{n}}$$

- If  $K < n$ ,  $(\hat{\Sigma}^z)^{-1}$  exists, w.h.p.
- When  $(\hat{\Sigma}^z)^{-1}$  does not exist, use Dantzig-type estimation:

$$\hat{\beta}_d = \arg \min_{\beta \in \mathbb{R}^K} \{ \|\beta\|_1 : \|\hat{\Sigma}^z \beta - \hat{h}\|_\infty \leq \lambda + \mu \|\beta\|_1 \}$$

- Rate slower than that of  $\hat{\beta}$  by a factor in  $[\sqrt{m}, \sqrt{mK}]$ :

$$\inf_P \left\| \hat{\beta}_d - P\beta \right\| \lesssim (1 \vee \|\beta\|_1) \sqrt{\frac{K \log(p \vee n)}{n}},$$

$m$  is the size of the smallest pure variable group,  $m = \min_{1 \leq k \leq K} |I_k|$ .

# Inference with Essential Regression

## Component-wise AN (Bing, Bunea, W, Strimas-Mackey)

$$\sqrt{n}Q_{kk}^{-1/2}(\hat{\beta}_k - (P\beta)_k) \xrightarrow{d} N(0, 1), \quad \text{as } n \rightarrow \infty.$$

Set  $\Omega =: (\Sigma^Z)^{-1}$ ,

$$Q_{kk} = \left( \sigma^2 + \frac{\tau^2}{m} \|\beta\|^2 \right) \left( \Omega_{kk} + \frac{\tau^2}{m} \|\Omega_{k\cdot}\|^2 \right) + \frac{\tau^4}{m^2(m-1)} \sum_{a=1}^K \beta_a^2 \Omega_{ka}^2.$$

$Q_{kk}$  can be estimated consistently.

Each  $\hat{\beta}_k$  is asymptotically "adaptive".

Asymptotic variance of  $\hat{\beta}_k =$  Cramér-Rao bound, if  $\tau^2$ ,  $\sigma^2$ ,  $A$ ,  $I$  and  $\Sigma^Z$  were all known.

Uniform confidence intervals in  $\beta_k$ .

# Assumptions for valid **inference** in E-Regression

- Sub-Gaussian Errors (Can be relaxed).
- $C_{\min} < \lambda_{\min}(\Sigma^Z) \leq \lambda_{\max}(\Sigma^Z) < C_{\max}$ .
- Distinguishable factors:  $\Delta(\Sigma^Z) > \nu > 0$ .
- Not too many quasi-pure variables: condition on  $\rho := \max_k |J_1^k| / (|I_k| + |J_1^k|)$  implying  $|I^q| = o\left(\sqrt{m / \log(p \vee n)}\right)$ .  
Not needed for prediction.
- $K \log(p \vee n) = o(\sqrt{n})$ . For prediction, relax to  $K \log(p \vee n) = O(n)$ .

# Regression with clustered covariates

- $X = AZ + E$  under Essential Regression framework defines  $A$  uniquely.
- Clusters  $G_k =: \{j : |A_{jk}| > 0\}$  are uniquely defined.  
Overlap allowed !
- Clusters of  $X$  are defined "around"  $Z$ .
- $Y = \beta^T Z + \varepsilon$ : these clusters are informative for  $Y$ , via  $Z$ .
- Inference: for  $\beta$ , relative to cluster "centers"  $Z$ .

Don't replace  $Z$  by cluster averages !

# How to use the cluster structure for inference

- $\beta = \arg \min_{\eta} \mathbb{E} [Y - \eta^T Z]^2$ , but  $Z$  not observed.
- Possible cluster summaries for replacing/predicting  $Z$ :
  1. Weighted averages of all variables  $\bar{X} := (A^T A)^{-1} A^T X \in \mathbb{R}^K$ .
  2. Averages of pure variables  $\bar{X}_I := (A_I^T A_I)^{-1} A_I^T X_I \in \mathbb{R}^K$ .
  3. The BLP of  $Z$  from  $\bar{X}$ :  $\tilde{Z} \neq \bar{X}$ .
  4. The BLP of  $Z$  from  $\bar{X}_I$ :  $\tilde{Z}_I \neq \bar{X}_I$ .
- Morale: Do not replace  $Z$  by (weighted) averages!

$$\beta = \arg \min_{\eta} \mathbb{E} [Y - \eta^T \tilde{Z}_I]^2 = \arg \min_{\eta} \mathbb{E} [Y - \eta^T \tilde{Z}]^2$$

$$\neq \arg \min_{\eta} \mathbb{E} [Y - \eta^T \bar{X}_I]^2 \neq \arg \min_{\eta} \mathbb{E} [Y - \eta^T \bar{X}]^2.$$

# Inference in E-Regression: the clustering interpretation

- $Y = \beta^T Z + \varepsilon \quad X = AZ + W.$
- Cluster averages, sample level:  $\hat{\bar{X}}$
- BLP of  $Z$  from  $\bar{X}_I$ , sample level:  $\hat{\mathbf{Z}}_I$

$$\hat{\beta} = \left( \hat{\mathbf{Z}}_I^T \hat{\mathbf{Z}}_I \right)^{-1} \hat{\mathbf{Z}}_I^T \mathbf{y} \neq \left( \hat{\bar{X}}^T \hat{\bar{X}} \right)^{-1} \hat{\bar{X}}^T \mathbf{y} = \hat{\beta}_{ave}.$$

- $\hat{\beta}$  already analyzed: consistent; AN with small variance.
- $\hat{\beta}_{ave}$  provably not consistent for  $\beta$ .

# Simulation

We compare  $\hat{\beta}$  with

- $\hat{\beta}_{full} = \frac{1}{n} [\hat{\Sigma}^Z]^{-1} [\hat{A}^T \hat{A}]^{-1} \hat{A}^T X^T Y.$
- $\tilde{\beta} := \arg \min_{\beta} \|Y - \hat{Z}\beta\|^2$  based on plug-in estimator  $\hat{Z}$  of

$$\begin{aligned} \tilde{Z} &= \text{Cov}(Z, \bar{X}) [\text{Cov}(\bar{X})]^{-1} \bar{X} \\ &= \Sigma^Z \left( \Sigma^Z + \Pi^T \Sigma^E \Pi \right)^{-1} \bar{X} \end{aligned}$$

with  $\Pi = A[A^T A]^{-1}.$

- $\hat{\beta}_{oracle} = (Z^T Z)^{-1} Z^T Y$ , the oracle LSE using the true  $Z$ .

$\ell_2$  error of different estimators and the coverage of the 95% CI of  $\beta_1$ .

|                                          | $\hat{\beta}$ | $\tilde{\beta}$ | $\hat{\beta}_{full}$ | $\hat{\beta}_{oracle}$ | coverage(%) |
|------------------------------------------|---------------|-----------------|----------------------|------------------------|-------------|
| Vary $n$ with $p = 500, m = K = 10$      |               |                 |                      |                        |             |
| $n = 200$                                | 0.52          | 0.76            | 1.02                 | 0.16                   | 98          |
| $n = 400$                                | 0.32          | 0.51            | 0.75                 | 0.11                   | 99          |
| $n = 600$                                | 0.26          | 0.45            | 0.63                 | 0.08                   | 100         |
| $n = 800$                                | 0.22          | 0.41            | 0.51                 | 0.08                   | 100         |
| $n = 1000$                               | 0.20          | 0.36            | 0.49                 | 0.07                   | 98          |
| Vary $m$ with $n = 300, p = 500, K = 10$ |               |                 |                      |                        |             |
| $m = 5$                                  | 0.57          | 0.71            | 1.10                 | 0.12                   | 96          |
| $m = 10$                                 | 0.35          | 0.56            | 0.83                 | 0.12                   | 99          |
| $m = 20$                                 | 0.25          | 0.37            | 0.43                 | 0.12                   | 100         |
| $m = 30$                                 | 0.21          | 0.27            | 0.29                 | 0.12                   | 98          |
| $m = 40$                                 | 0.20          | 0.22            | 0.22                 | 0.12                   | 98          |

# Inference for $\beta$ in other factor models: under-explored

- Inference for  $\beta$  not of direct interest; not studied.
- Factor Regression (FR): developed for high dim prediction.
- AN-type results in FR: vehicles for deriving prediction-related results.
- $\sqrt{n}(\hat{\beta} - Q\beta) \rightarrow N(0, D)$ , when  $p, n \rightarrow \infty$ , but  $K$  **constant**,  
for some **unknown invertible matrix**  $Q$ . (Bai and Ng, 2006)

Enough for prediction intervals, but not for inference on  $\beta$  itself.

## Prediction of $Y$

# Prediction with Essential Regression

## Population level

$$\bar{X} = \Pi^T X := (A^T A)^{-1} A^T X = Z + \bar{W}.$$

BLP of  $Z$  from  $\bar{X}$ :

$$\begin{aligned}\tilde{Z} &= \text{Cov}(Z, \bar{X}) \text{Cov}(\bar{X})^{-1} \bar{X} \\ &= \Sigma^Z \left( \Sigma^Z + \Pi^T \Sigma^E \Pi \right)^{-1} \bar{X}.\end{aligned}$$

# Prediction with Essential Regression

## Population level

Since

$$\beta = \arg \min_{\alpha} \mathbb{E}(Y - Z^T \alpha)^2 = \arg \min_{\alpha} \mathbb{E}(Y - \tilde{Z}^T \alpha)^2,$$

we predict  $Y$  by

$$\tilde{Y} = \tilde{Z}^T \beta = \dots = X^T A \text{Cov}(A^T X)^{-1} \text{Cov}(A^T X, Y).$$

# Prediction with Essential Regression

## Prediction in E-Regression

Input:  $\mathbf{X} \in \mathbb{R}^{n \times p}$ ,  $\mathbf{y} \in \mathbb{R}^n$ ; new observation  $X^{\text{new}} \in \mathbb{R}^p$ .

Estimate  $A$  by  $\hat{A}$  from BBYW.

Predict  $Y$  by:

$$\hat{Y}_{ER} = (X_{\text{new}})^T \hat{\theta} = (X_{\text{new}})^T \hat{A} \left( \hat{A}^T \mathbf{X}^T \mathbf{X} \hat{A} \right)^+ \hat{A}^T \mathbf{X}^T \mathbf{y}$$

# Prediction error in Essential Regression

Theorem (Bing, Bunea, W, Strimas-Mackey)

$$\mathbb{E} \left[ (\beta^T Z^{new} - \hat{Y}_{ER})^2 \wedge T \right] \lesssim \frac{K}{n} \sigma^2 + \frac{\|\beta\|^2}{\lambda_{\min}(A^T A)} \left\{ 1 + \frac{\|A_J\|_0 \log(p \vee n)}{n} \right\}.$$

- Price to pay for not observing  $Z$ , at the population level.

$$\mathbb{E} \left[ \beta^T Z - \beta^T \tilde{Z} \right]^2 \lesssim \frac{\|\beta\|^2}{\lambda_{\min}(A^T A)}.$$

- $\lambda_{\min}(A^T A) \geq \lambda_{\min}(A_I^T A_I) + \lambda_{\min}(A_J^T A_J) \geq m + \lambda_{\min}(A_J^T A_J)$
- Estimation error of  $A$ , shown to be minimax optimal.

$$\mathbb{E}[\|\hat{A} - A\|^2] \lesssim \|A_J\|_0 \frac{\log(p \vee n)}{n}.$$

- No rates available in the literature for other FR models; only (few) convergence results, when  $K$  is fixed.

# Regression with clustered covariates

- Clusters of  $X$  are defined "around"  $Z$ , via  $X = AZ + E$ .
- $Y = \beta^T Z + \varepsilon$ : these clusters are informative for  $Y$ , via  $Z$ .
- Prediction: use cluster structure to predict  $Y$ .

OK to replace  $Z$  by cluster averages !

## Prediction: the clustering interpretation

Safe to use cluster (weighted) averages  $\bar{X} = (A^T A)^{-1} A^T X$  for predicting  $Y$  !

If  $\tilde{Z}$  is the BLP of  $Z$  from  $\bar{X}$  then:

$$\min_{\eta} \mathbb{E} \left[ Y - \eta^T \tilde{Z} \right]^2 = \min_{\eta} \mathbb{E} \left[ Y - \eta^T \bar{X} \right]^2$$

Better than using  $X_{(l)}$  or  $\tilde{Z}_{(l)}$

Prediction: same as before, already analyzed

$$\hat{Y} = (X_{new})^T \hat{\theta} = (X_{new})^T \hat{A} \left( \hat{A}^T \mathbf{X}^T \mathbf{X} \hat{A} \right)^+ \hat{A}^T \mathbf{X}^T \mathbf{y}$$

# Prediction in E-Regression: non-overlapping clusters

The familiar particular case of non-overlapping clusters

If clusters do not overlap, are balanced, under ER setting:

$$\begin{aligned}\mathbb{E} \left[ \left( \beta^T Z^{new} - \hat{Y} \right)^2 \wedge T \right] &\lesssim \frac{K}{n} \sigma^2 + \frac{\|\beta\|^2 K}{p} \\ &\lesssim \frac{K}{n} \sigma^2 + \frac{K^2}{p}.\end{aligned}$$

High dimensions ( $p$ ) are a blessing 😊

# Consistent prediction in FR

- Factor Regression:

$$Y = \beta^T Z + \varepsilon, \quad X = BZ + E$$

$$\Sigma = B\Sigma^Z B^T + \Sigma^E; \quad \text{SNR } \xi = \frac{\lambda_K(B\Sigma^Z B^T)}{\lambda_1(\Sigma^E)}.$$

- BLP of  $Y$  from  $B^T X$ :

$$Y_{FR}^* = X^T B [\text{Cov}(B^T X)]^{-1} \text{Cov}(B^T X, Y)$$

- Principal-component based prediction :

$$Y_{PCR}^* = X^T U_K [\text{Cov}(U_K^T X)]^{-1} \text{Cov}(U_K^T X, Y)$$

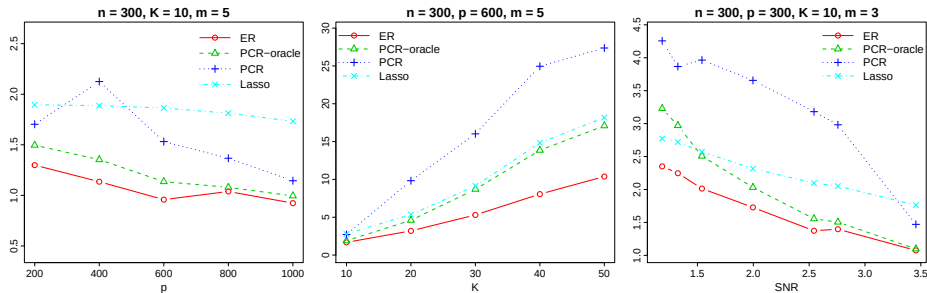
Replace  $B$  by matrix  $U_K$  of first  $K$  eigenvectors of  $\Sigma$ .

## Theorem (Bing, Bunea, W, Strimas-Mackey)

- 1 If  $\Sigma^E = \tau^2 I_p$ , then  $Y_{PCR}^* = Y_{FR}^*$  almost surely.
- 2 Under eigengap condition on  $B\Sigma^Z B^T$  and if  $\xi \rightarrow \infty$  at appropriate rate,

$$\mathbb{E}[Y_{PCR}^* - Y_{FR}^*]^2 \lesssim \beta^T \Sigma^Z \beta \left( \frac{\kappa K}{\xi} \right)^2.$$

# Prediction errors: a numerical comparison



Prediction risks of different predictors as  $p$ ,  $K$  and SNR vary separately

# Comparison

## • PCR

- $K$  is fixed,  $\|\beta\|_\infty \leq C$ ,  $\|\Sigma^E\|_{op} \leq C$ , all eigenvalues of  $p^{-1}B^TB$  and  $\Sigma^Z$  are of constant order. (Chamberlain and Rothschild, 1983; Connor and Korajczyk, 1986; Bai and Ng, 2008; Bai, 2003; Fan, Liao and Mincheva, 2011, 2013)
- $B^TB = pI_K$ ,  $\|\Sigma^E\|_{op} = o(1)$ ,  $\Sigma^Z$  diagonal. Stock and Watson (2002)
- Signal-to-noise ratio

$$\xi = \frac{\lambda_K(B\Sigma^Z B^T)}{\lambda_1(\Sigma^E)} \asymp p,$$

## • Essential regression:

- Pure variables,  $\Delta(\Sigma^Z)$ ,  $\Sigma^E$  diagonal, eigenvalues of  $\Sigma^Z$  constant order
- 

$$\xi = \frac{\lambda_K(A\Sigma^Z A^T)}{\lambda_1(\Sigma^E)} \asymp \lambda_K(A^T A).$$

- Consistent prediction under weaker SNR.

# Prediction with E-Regression: Summary

- Sufficient conditions for consistent prediction with low SNR.
- Explicit prediction error bounds, valid for all  $K, p, n$
- Low computational complexity.
- Other approaches, introduced for prediction with low SNR:

Fully Bayesian: computationally complex.

Variational Bayes / Mean Field Approx: Multiple maxima.

Potential problems for low SNR.

## Summary

# Summary

Essential Regression  $Y = \beta^T Z + \varepsilon, X = AZ + E$

- A different modeling perspective in high dimensions:  
A factor model with interpretable factors.
- Inference questions posed at a different resolution level.  
Computationally efficient  $\hat{\beta}$  is minimax-rate optimal  
The coordinate-wise asymptotic normality of  $\hat{\beta}$  holds uniformly over  $\beta$ .
- Simple prediction strategy; explicit prediction error.  
Addresses prediction in factor models with low SNR.
- Essential Regression provides a statistical framework for performing and analyzing post-clustering inference and prediction.



Thanks!