Instrumental Factor Models for High Dimensional Functional Data

Jihyun Kim[1] and Young Kim[2]

ESEM, Barcelona 2023

[1]Sungkyunkwan University
[2]Toulouse School of Economics

Table of Contents

Table of Contents

Introduction

- With the advance of data technology, **high dimensional data (HD)** and **functional data (FD)** became much more accessible. ▸ Example
- For HD, a factor model is a framework that assumes a few latent factors can explain the entire observed time series.
- The **principal component analysis (PCA)** is the most commonly used method to estimate factor models.
- In PCA, the factors are estimated by eigendecomposition of covariance matrix which corresponds to the solution of LS problem.

Limitations of the Model and Estimation Method

Limitation of PCA: PCA estimation requires both cross-sectional dimensionality (N) and time horizon (T) to be large.

- In some applications, only short panel available (T is short).
- Structural instability over long time span (λ time varying).

Limitation of Conventional Model: The conventional model is based on scalar-valued observation for each i and t.

- Often y_{it} is a vector/matrix/function rather than a scalar.
  - y_{it} =(Return, Volatility)$'$ of financial asset i
  - $y_{it}(r)$ = city i's temperature at time r, day t
  - $y_{it}(r)$ = asset i's volatility at time r, day t
  - $y_{it}(r)$ = distribution of i at t.

We introduce an **instrumental factor model (IFM) for HDFD** to overcome such limitations.

What We Have Done

Theoretical Studies:

- Identification and estimation methodology.
- Consistency as long as $N \to \infty$, regardless of T being finite or not.
- Eigenvalue ratio estimator for the number of factors.

Model Applications: Within/Between-period factor model

- Propose a new type of factor model for high-frequency data.
- The number of estimated factors often depends on the choice of data frequency.
- The proposed model provides a unified framework to explain the phenomena.

Empirical Application: Climate change and economic outcome

- Long-run relationship between global temperature and economic outcomes.

Table of Contents

Conventional Factor Model

A factor model has the following representation:

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{tk} + \varepsilon_{it}, \;\; i = 1, ..., N, \;\; t = 1, ..., T.$$

- y_{it} : observed data.
- f_{tk} : common factor.
- λ_{ik} : factor loading.
- ε_{it} : idiosyncratic error.

In the conventional model, the variables $(y, f, \lambda, \varepsilon)$ are all **real-valued**.

IFM for Functional Data

An instrumental factor model for HDFD:

$$y_{it}(r) = \sum_{k=1}^{K} \lambda_{ik}(r) f_{tk} + \varepsilon_{it}(r), \quad r \in [0,1].$$

- $y_{it}, \lambda_{ik}, \varepsilon_{it}$ belongs to a Hilbert space $\mathcal{H}$, and $f_{tk} \in \mathbb{R}$.
- Examples of $\mathcal{H}$: square integrable function; vector space; square integrable random variable.

In addition, loading coefficient is modeled as

$$\lambda_{ik}(r) = g_k(X_i, r) + \gamma_{ik}(r).$$

- $X_i = (X_{i1}, \ldots, X_{iH})'$ be observed characteristics.
- $g_k(X, r)$ be an unknown function $\mathbb{R}^H \times [0,1] \to \mathbb{R}$.
- $\gamma_{ik} \in \mathcal{H}$ be the remaining component unexplained by X_i.
- X_i is independent of $\gamma_{ik}, \varepsilon_{it}$.

IFM encompasses various forms of factor model exist in the literature.

IFM $\rightarrow$ Conventional Factor Model for Scalar Data

The conventional factor model becomes a special case of IFM.

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{tk} + \varepsilon_{it}.$$

- $y_{it}, \lambda_{ik}, \varepsilon_{it} \in \mathcal{H}$.
- A case of interest is $\mathcal{H} = \mathbb{R}$.

Loading coefficient is modeled as

$$\lambda_{ik} = \gamma_{ik}.$$

- X_i has no explaining power, i.e. $g_k(X_i) = 0$.
- γ_{ik} itself becomes the loading.

IFM $\rightarrow$ Semiparametric Factor Model for Scalar Data

The semiparametric factor model is another special case of IFM for FD.

$$y_{it} = \sum_{k=1}^{K} \lambda_{ik} f_{tk} + \varepsilon_{it}.$$

- $y_{it}, \lambda_{ik}, \varepsilon_{it} \in \mathcal{H}$.
- We assume $\mathcal{H} = \mathbb{R}$.

Loading coefficient is modeled as

$$\lambda_{ik} = g_k(X_i) + \gamma_{ik}.$$

The proposed model is estimated by a Projected-PCA method introduced by Fan et al. (2016).

IFM $\rightarrow$ Factor Model for Functional Data

A factor model can be also developed for the FD, which is a special case of IFM for FD. This model considered by Tavakoli et al. (2021).

$$y_{it}(r) = \sum_{k=1}^{K} \lambda_{ik}(r) f_{tk} + \varepsilon_{it}(r).$$

- $y_{it}, \lambda_{ik}, \varepsilon_{it} \in \mathcal{H}$, and $f_{tk} \in \mathbb{R}$.
- We assume $\mathcal{H} = L^2([0,1], \mathbb{R})$.

For each (i,k), loading coefficient is modeled as

$$\lambda_{ik}(r) = \gamma_{ik}(r).$$

- X_i has no explaining power, i.e. $g_k(X_i) = 0$.
- $\gamma_{ik} \in \mathcal{H}$.

IFM Representation

IFM representation provides a unified modeling approach for various factor models that exist in the literature.

An instrumental factor model for HDFD:

$$y_{it}(r) = \sum_{k=1}^{K} \lambda_{ik}(r) f_{tk} + \varepsilon_{it}(r), \quad r \in [0,1].$$

- We assume $y_{it}, \lambda_{ik}, \varepsilon_{it} \in L^2([0,1], \mathbb{R})$, and $f_{tk} \in \mathbb{R}$.

In addition, loading coefficient is modeled as

$$\lambda_{ik}(r) = g_k(X_i, r) + \gamma_{ik}(r).$$

- $X_i = (X_{i1}, \ldots, X_{iH})'$ be observed characteristics.
- $g_k(X, r)$ be an unknown function $\mathbb{R}^H \times [0,1] \to \mathbb{R}$.
- $\gamma_{ik} \in L^2$ be the remaining component unexplained by X_i.
- X_i is independent of $\gamma_{ik}, \varepsilon_{it}$.

Table of Contents

Principal Component Analysis

Suppose we have a factor model in the conventional setup:

$$Y = \Lambda F' + \varepsilon.$$

- Y : $N \times T$ matrix of y_{it}
- Λ : $N \times K$ matrix of λ_{ik}
- F : $T \times K$ matrix of f_{tk}
- ε : $N \times T$ matrix of ε_{it}

Under regularity conditions, Λ and F can be estimated as follows:

1. **(PCA)** Apply eigendecomposition to the covariance matrix $Y'Y$.
   - Extract matrix U consists of K number of eigenvectors.
   - The factor estimator is defined as $\widehat{F} = \sqrt{T}U$.
2. **(Regression)** Regress Y on $\widehat{F}$.
   - The loading estimator is $\widehat{\Lambda}' = (\widehat{F}'\widehat{F})^{-1}\widehat{F}'Y'$.

Projected Principal Component Analysis

Suppose we have a factor model as follows:

$$Y = \Lambda F' + \varepsilon, \quad F'F/T = I, \quad \Lambda'\Lambda/N = \text{diagonal}.$$

For an expositional purpose, let $\Lambda = X$, i.e., the loadings are observable.

$$Y = XF' + \varepsilon.$$

We can estimate F by simply running regression of Y on X for each t when N is large enough.

Equivalently, we may consider the projection matrix $P = X(X'X)^{-1}X'$ and the projected data $\widehat{Y} = PY$,

$$\widehat{Y} = XF' + P\varepsilon \approx XF'$$

as long as X and ε are asymptotically orthogonal ($N \to \infty$). Here the large T assumption is not necessary.

So, F can be estimated by applying the PCA to $\widehat{Y}'\widehat{Y}$ as long as N is large enough.

Here the eigenvector of $\widehat{Y}'\widehat{Y}$ becomes the estimator for F which is equivalent to the one obtained by the regression.

Projected Principal Component Analysis

This idea can be generalized for $\Lambda = G(X)$ by using a proper projection $P_J = G_J(X)(G_J(X)'G_J(X))^{-1}G_J(X)$, where $G_J(X)$ is the sieve approximation of unknown $G(X)$ with J basis functions.

This approach is proposed by Fan et al. (2016) for a scalar variable, and is called **projected-PCA (PPCA)**.

Clearly, factors and loadings are accurately estimable as long as $N \to \infty$ (and $J \to \infty$), regardless of T being fixed or not.

In this project, we generalize the PPCA method to functional data.

PCA vs Projected-PCA

Suppose we have a factor model:

$$Y = \Lambda F' + \varepsilon,$$

and assume that a projection P satisfies $P\Lambda = \Lambda$, and $P\varepsilon \approx 0$.

PCA : As $N, T \to \infty$

$$\psi_k\left(\frac{1}{NT}YY'\right) \approx \underbrace{\psi_k\left(\frac{1}{NT}F\Lambda'\Lambda F'\right)}_{O_p(1)} + \underbrace{\psi_k\left(\frac{1}{NT}\varepsilon'\varepsilon\right)}_{O_p(1/\min\{N,T\})}.$$

PPCA : As long as $N \to \infty$

$$\psi_k\left(\frac{1}{NT}Y'PY\right) \approx \underbrace{\psi_k\left(\frac{1}{NT}F\Lambda'\Lambda F'\right)}_{O_p(1)} + \underbrace{\psi_k\left(\frac{1}{NT}\varepsilon'P\varepsilon\right)}_{O_p(1/N)}.$$

PCA for Functional Data

Let $Y = (y_1, \ldots, y_T), \quad y_t = (y_{1t}, \ldots, y_{Nt})', \quad y_{it} \in L^2([0,1])$.

To apply PCA, a symmetric p.d. matrix Y^*Y can be constructed as follows:

$$Y^*Y \equiv \begin{pmatrix} \langle y_1 , y_1 \rangle & \cdots & \langle y_1 , y_T \rangle \\ \vdots & \ddots & \vdots \\ \langle y_T , y_1 \rangle & \cdots & \langle y_T , y_T \rangle \end{pmatrix} \in \mathbb{R}^{T \times T},$$

$$\langle y_t , y_s \rangle \equiv \sum_{i=1}^{N} \langle y_{it} , y_{is} \rangle_{L^2} = \sum_{i=1}^{N} \int_0^1 y_{it}(r) y_{is}(r)\, dr.$$

We call the matrix Y^*Y **the integrated covariance matrix** since

$$E\langle y_t , y_s \rangle = \sum_{i=1}^{N} \int_0^1 E(y_{it}(r) y_{is}(r)) dr = \sum_{i=1}^{N} \int_0^1 cov(y_{it}(r), y_{is}(r)) dr.$$

It can be shown that

$$Y^*Y \approx F \left(\int_0^1 \Lambda'(r) \Lambda(r) dr \right) F'$$

So, the eigenvectors of Y^*Y become a valid estimator for F.

▸ Remark 1 ▸ Remark 2

FPPCA for Functional Data

Our FPPCA method consists of three steps.

First step (projection)

For each $r \in [0,1]$, we project $Y(r)$ on X, i.e., $\widehat{Y}(r) \equiv PY(r)$.

Second step (pca $\rightarrow$ eigenvector = factor)

Apply the PCA to the ICM $\widehat{Q}$ of $\widehat{Y}$,

$$\widehat{Y}^* \widehat{Y} \equiv \begin{pmatrix} \langle \hat{y}_1 , \hat{y}_1 \rangle & \cdots & \langle \hat{y}_1 , \hat{y}_T \rangle \\ \vdots & & \vdots \\ \langle \hat{y}_T , \hat{y}_1 \rangle & \cdots & \langle \hat{y}_T , \hat{y}_T \rangle \end{pmatrix} \in \mathbb{R}^{T \times T},$$

The eigenvector of $\widehat{Y}^* \widehat{Y}$ becomes the estimator for F.

Third step (regression on $\widehat{F}$ $\rightarrow$ loadings $\Lambda(r) = G(X,r) + \Gamma(r)$)

$\Lambda(r)$ is estimated by regressing $Y(r)$ on $\widehat{F}$ for each r.

$G(X,r)$ is estimated by regressing $\widehat{Y}(r)$ on $\widehat{F}$ for each r.

$\widehat{\Gamma}(r) = \widehat{\Lambda}(r) - \widehat{G}(X,r)$ for each r.

Table of Contents

Within/Between-Period Factor Model

In some applications, the number of estimated factors depend on the choice of data frequency (see, e.g., Kong et al. (2021, JASA)).

The Instrumental FM for FD provides

- a unified framework to explain the phenomena
- a useful tool to separately identify and analyze within/between-period factors.

We consider the continuous-time factor model

- time index at high frequency: r
- time index at low frequency: t

  $\Rightarrow y_{it}(r)$ is an intraday observation indexed by r for a given day t.

  $\Rightarrow$ Goal is to separately identify and estimate the factors driving

  - the intraday variation at the frequency r

  - the day-to-day variation at the frequency t

Factor Model in Continuous Time

Let $t = 1, \cdots, T$ be the index of days.

Let r be the index of normalized intraday observations, i.e., $r = 1/R, \cdots, 1$ with the number of intraday observation R.

Let's assume that the price of financial asset follows an Ito semi-martingale

$$dP_{it}(r) = \underbrace{\{b_i dG(r)\}F_t}_{\text{common component}} + \underbrace{dU_{it}(r)}_{\text{idiosyncratic component}} \quad \text{(ignoring drift for simplicity)}$$

Note that $P_{it}(0) = P_{i,t-1}(1)$.

A special case with $F_t = I$ has been extensively considered in the literature on high frequency and high dimensional data analysis in econometrics/statistics/finance (see, e.g., Aït-Sahalia and Xiu (2017)).

In this case,

$$dP_i(r) = \underbrace{\{b_i dG(r)\}}_{\text{common component}} + \underbrace{dU_i(r)}_{\text{idiosyncratic component}} \quad \text{(ignoring drift for simplicity)}$$

Within/Between-Period Factors in Continuous Time

Now let $F_t \neq I$.

Assume that the discrete samples $(P_{it}(r), r = 1/R, \cdots, 1)$ are observed.

Then, the discrete samples satisfy (when R is large)

$$\Rightarrow \quad P_{it}(r+\delta) - P_{it}(r) \approx \{b_i(G(r+\delta) - G(r))\}F_t + U_{it}(r+\delta) - U_{it}(r)$$

$$\Rightarrow \quad y_{it}(r) \approx b_i \underbrace{(G(r+\delta) - G(r))}_{g(r)} F_t + \underbrace{U_{it}(r+\delta) - U_{it}(r)}_{\varepsilon_{it}(r)}$$

Therefore, the model admits a factor representation for functional data

$$y_{it}(r) = \underbrace{\Lambda_i(r)}_{1\times K_1}\underbrace{F_t}_{K_1\times 1} + \varepsilon_{it}(r), \quad \underbrace{\Lambda_i(r)}_{1\times K_1} = \underbrace{b_i}_{1\times K_2}\underbrace{g(r)}_{K_2\times K_1}, \quad K_2 \geq K_1$$

We may interpret

- F_t: between-period factor (K_1 vector)
  $\Rightarrow$ day-to-day variation
- $G(r) = \sum_{s\leq r} g(s)$: within-period factor ($K_2$ vector)
  $\Rightarrow$ intraday variation

Sampling Frequency Matters

Recall the within/between-period factor representation.

$$y_{it}(r) = \underbrace{\Lambda_i(r)}_{1\times K_1}\underbrace{F_t}_{K_1\times 1} + \varepsilon_{it}(r), \quad \underbrace{\Lambda_i(r)}_{1\times K_1} = \underbrace{b_i}_{1\times K_2}\underbrace{g(r)}_{K_2\times K_1}, \quad K_2 \geq K_1$$

The model can be rewritten as

$$y_{it}(r) = \underbrace{b_i}_{1\times K_2}\underbrace{g(r)}_{K_2\times K_1}\underbrace{F_t}_{K_1\times 1} + \varepsilon_{it}(r).$$

If one applies PCA or PPCA at daily frequency (t), then the number of factors (F_t) becomes K_1.

However, if one applies the same PCA or PPCA using intraday data (r), then the number of factors ($g(r)F_t$) becomes K_2.

Estimation of Within/Between-Period Factors

$$y_{it}(r) = \underbrace{\Lambda_i(r)}_{1\times K_1}\underbrace{F_t}_{K_1\times 1} + \varepsilon_{it}(r), \quad \underbrace{\Lambda_i(r)}_{1\times K_1} = \underbrace{b_i}_{1\times K_2}\underbrace{g(r)}_{K_2\times K_1}$$

(i) F_t and $\Lambda_i(r)$ can be estimated by PPCA for IFM.

(ii) We then apply PPCA to the estimated functional loading

$$\hat{\Lambda}_i(r) = b_i g(r) + \text{estimation error}_i(r)$$

PPCA provides consistent estimates for b_i and $g(r)$ in the second step.

Estimation of Within/Between-Period Factors

$$\text{Between Factor Rep.:} \quad y_{it}(r) = \underbrace{\Lambda_i(r)}_{1\times K_1}\underbrace{F_t}_{K_1\times 1} + \varepsilon_{it}(r), \quad \underbrace{\Lambda_i(r)}_{1\times K_1} = \underbrace{b_i}_{1\times K_2}\underbrace{g(r)}_{K_2\times K_1}$$

$$\text{Gross Factor Rep.:} \quad y_{it}(r) = \underbrace{b_i}_{1\times K_2}\underbrace{H_t(r)}_{K_2\times 1} + \varepsilon_{it}(r), \quad \underbrace{H_t(r)}_{K_2\times 1} = \underbrace{g(r)}_{K_2\times K_1}\underbrace{F_t}_{K_1\times 1}$$

(i) F_t and $\Lambda_i(r)$ can be estimated from BFR by PCA (for FM) or PPCA for (IFM).

(ii) b_i and $H_t(r)$ can be estimated from GFR by applying PCA or PPCA to the realized covariance matrix P of $p_{it}(r)$

$$P = \sum_{t=1}^{T} \begin{pmatrix} \langle y_{1t}\,, y_{1t}\rangle & \cdots & \langle y_{1t}\,, y_{Nt}\rangle \\ \vdots & & \vdots \\ \langle y_{Nt}\,, y_{1t}\rangle & \cdots & \langle y_{Nt}\,, y_{Nt}\rangle \end{pmatrix}, \quad \langle y_{it}\,, y_{jt}\rangle = \sum_{r=1}^{R} y_{it}(r/R)y_{jt}(r/R)$$

(iii) $g(r)$ can be estimated by

- regressing $\hat{H}_t(r)$ on $\hat{F}_t$ for each r, or
- regressing $\hat{\Lambda}_i(r)$ on $\hat{b}_i$ for each r.

Number of Within/Between Factors

$$\text{Between Factor Rep.:} \quad y_{it}(r) = \underbrace{\Lambda_i(r)}_{1\times K_1}\underbrace{F_t}_{K_1\times 1} + \varepsilon_{it}(r), \quad \underbrace{\Lambda_i(r)}_{1\times K_1} = \underbrace{b_i}_{1\times K_2}\underbrace{g(r)}_{K_2\times K_1}$$

$$\text{Gross Factor Rep.:} \quad y_{it}(r) = \underbrace{b_i}_{1\times K_2}\underbrace{H_t(r)}_{K_2\times 1} + \varepsilon_{it}(r), \quad \underbrace{H_t(r)}_{K_2\times 1} = \underbrace{g(r)}_{K_2\times K_1}\underbrace{F_t}_{K_1\times 1}$$

Our approach with PPCA (PCA) provides consistent estimations of the between/within/gross factors and the between/gross loadings as $N, R \to \infty$ ($N, R, T \to \infty$).

Moreover, the proposed model can explain the phenomenon that the number of estimated factors varies over the choice of data frequency.

- the high frequency data $\Rightarrow$ Gross Factor (K_2)
- the low frequency data $\Rightarrow$ Between Factor (K_1)

In our model, the assumption $K_2 \geq K_1$ is a seemingly artificial and technical condition for the separate identification of $g(r)$ and F_t.

However, the assumption corresponds to (and is supported by) the recent empirical findings in Kong et al. (2021).

Table of Contents

Empirical Application

Global annually averaged air temperature has risen by around $1^{\circ}C$ since the mid-20th century, and this trend is expected to continue.

Key Question: Climate change and economic outcomes?

- **GDP growth:** Dell, Jones and Olken (2010), (2012).
- **Labor productivity:** Heyes and Saberian (2022).
- **Agriculture:** Burke and Emerick (2016), Chen and Gong (2021).

We analyze the effect of temperature rise on the European cereal markets; barley, maize, and wheat.

Specifically, we use factor-augmented VAR with a time series of ($Temp$, $Precip$, $Production$, $Price$) from $1962 - 2020$.

Cereal production shares by continent

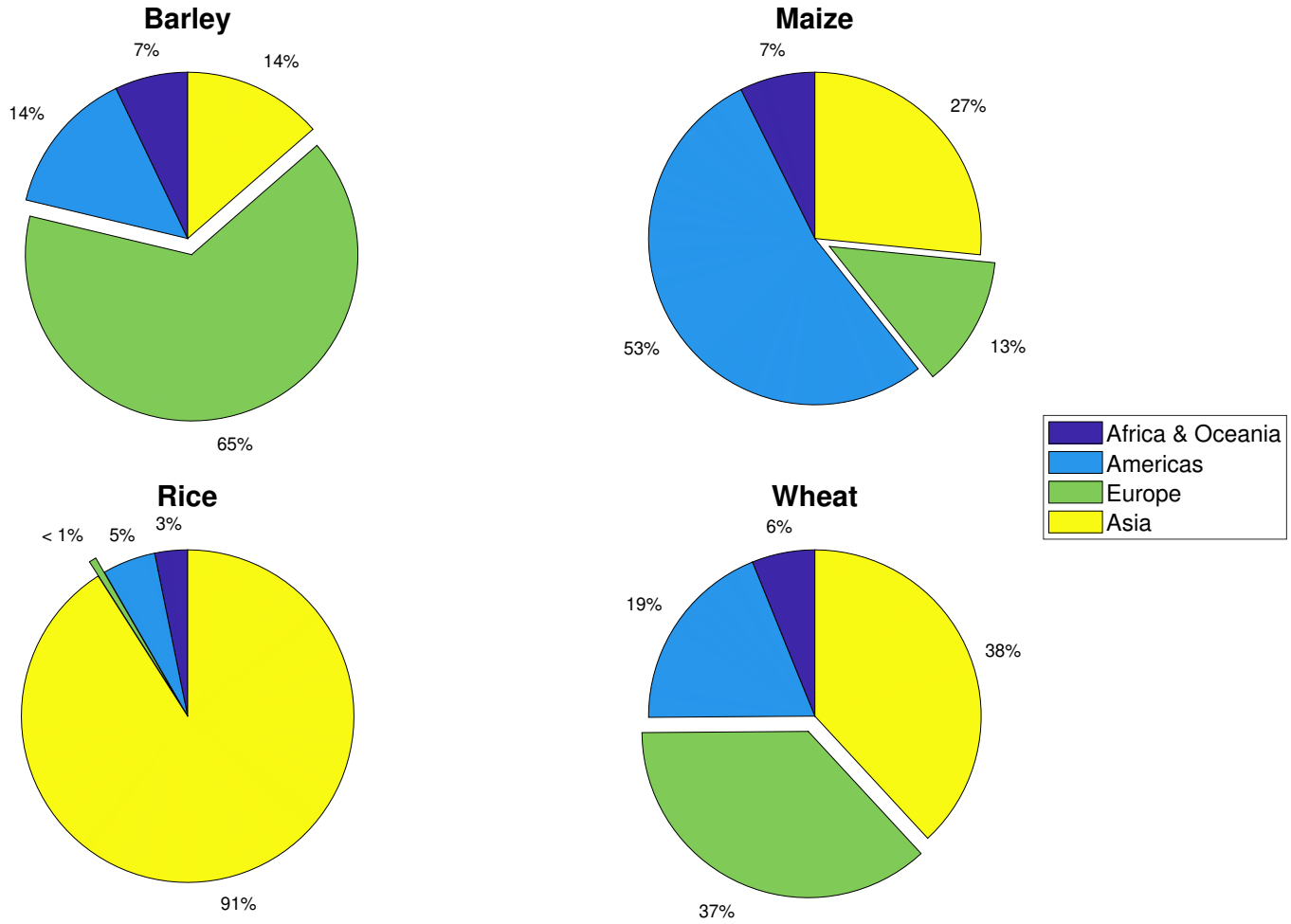


Figure: *Annual average production share by region 1961 - 2020*

Empirical Application: Global Temperature

Annual average temperature does not fully capture the increasing trend in global temperatures (IPCC 2014).

To improve the temperature measure, we use globally gridded data and extract factors using our instrumental factor model framework.

Data: NCEP/NCAR Reanalysis dataset

- Daily air temperature from 1948 to 2020
- A globally gridded data set consists of 2629 stations.

Characteristics of each station:

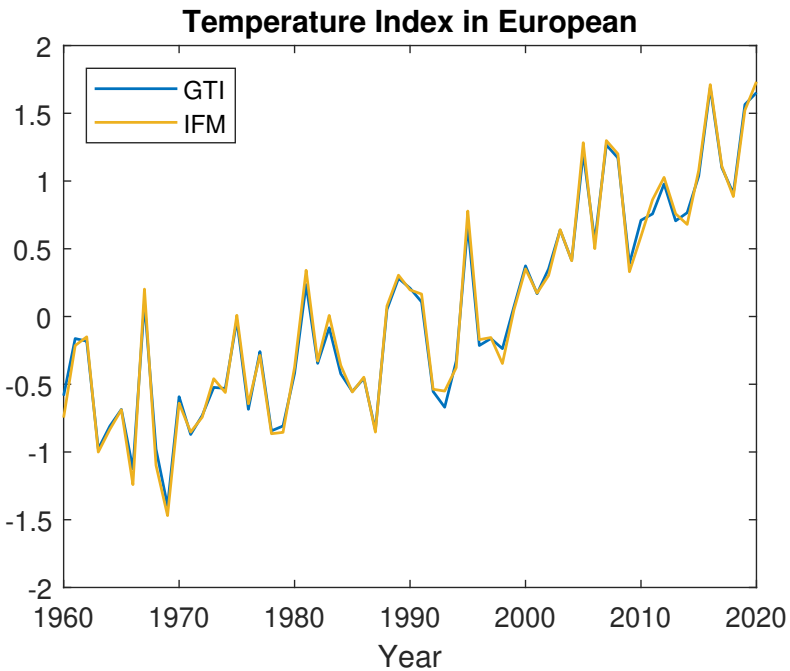
- Latitude and longitude.
- Köppen climate classification (1961).

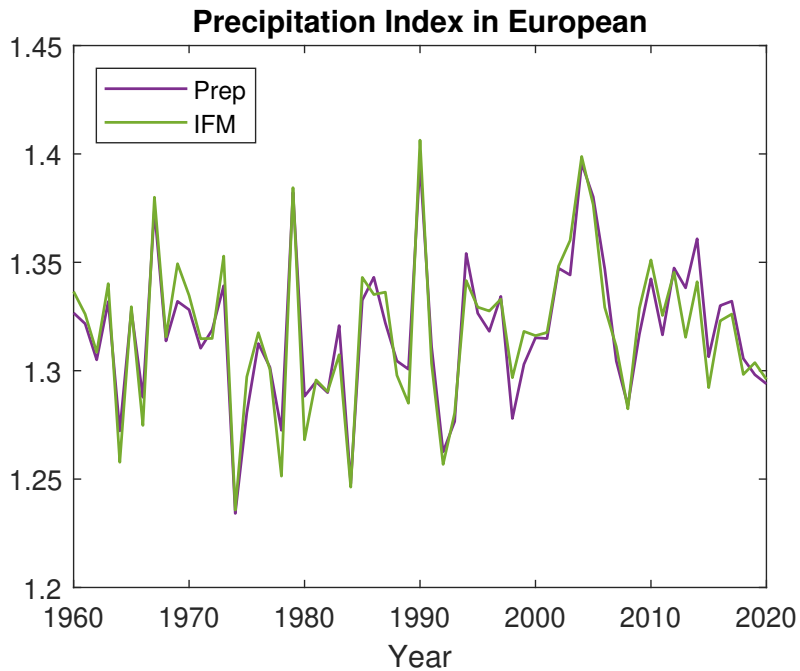
The variables are defined as:

$y_{it}(r) =$ Air temperature at station (i) for month (t) on day (r).

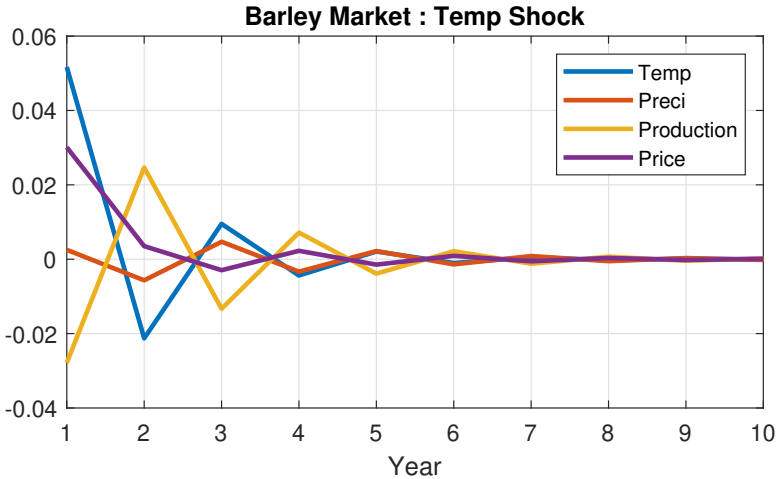
$X_i =$ [Latitude, Longitude, Köppen's classification] at station (i).

Estimation Result : Temperature and Precipitation





Temperature Shock : Barely Market



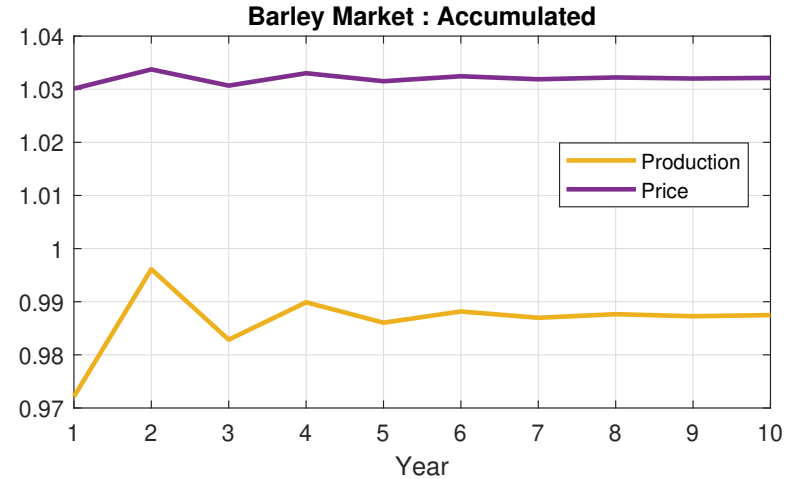
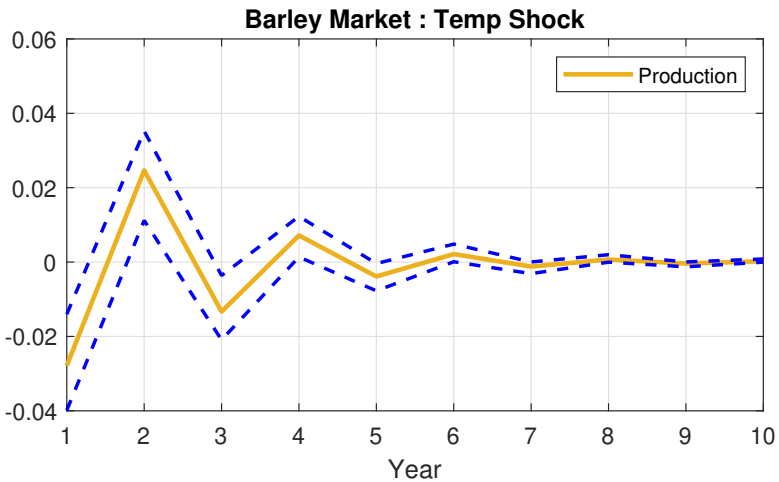


Figure: *Impulse response functions temperature shock on barely market*



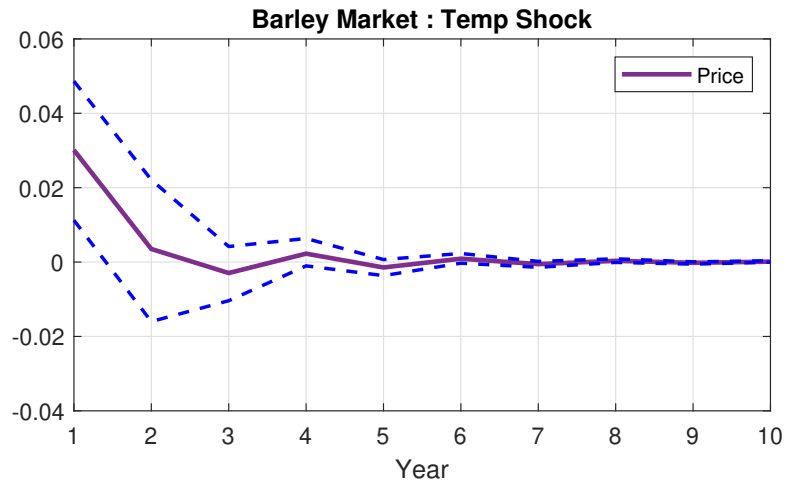
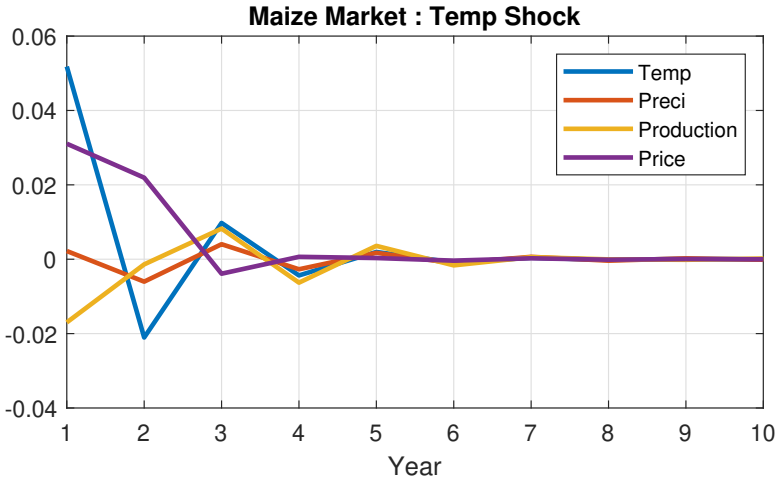


Figure: *68% confidence intervals of IRF on barely market*

Temperature Shock : Maize market



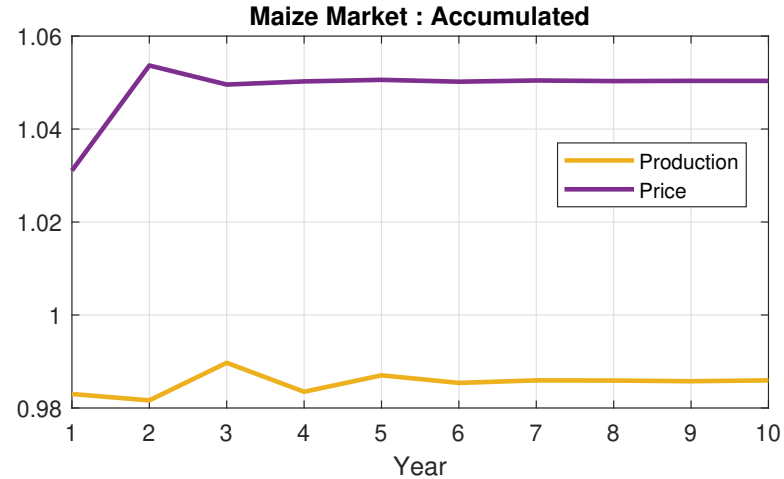
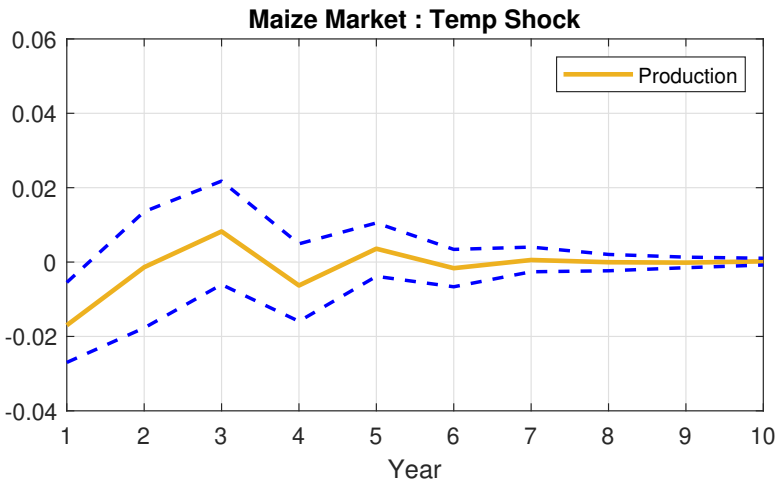


Figure: *Impulse response functions temperature shock on maize market*



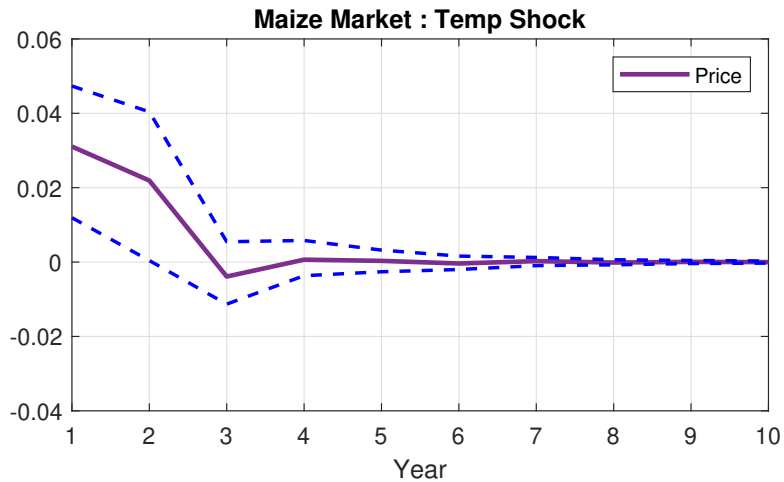
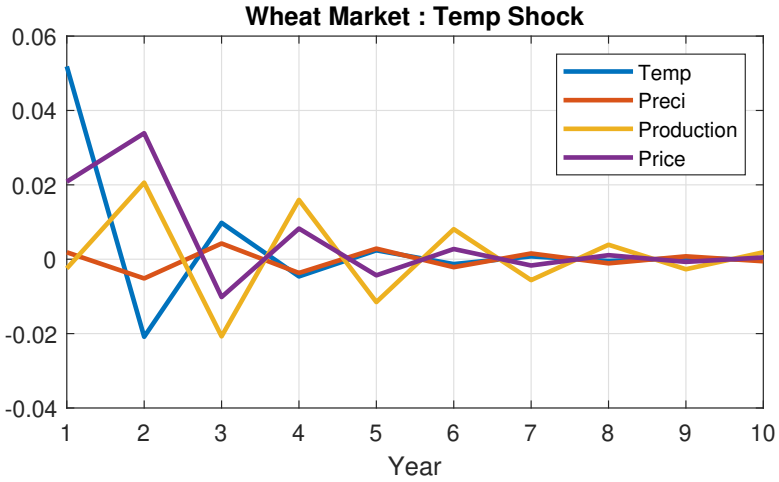


Figure: *68% confidence intervals of IRF on maize market*

Temperature Shock : Wheat market



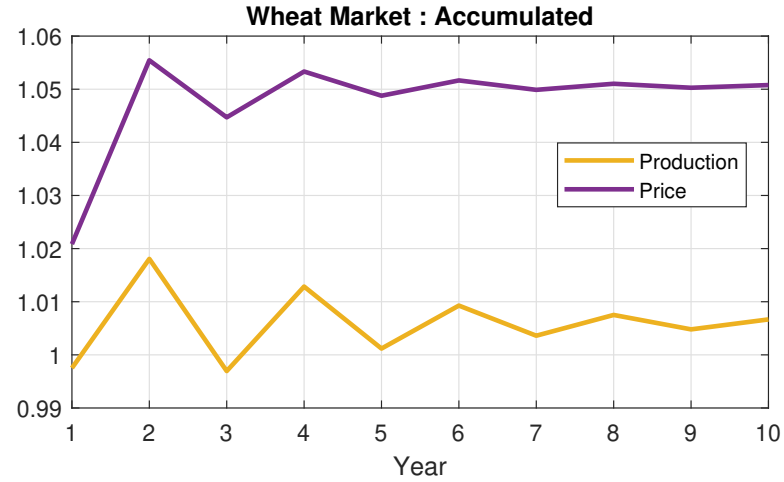
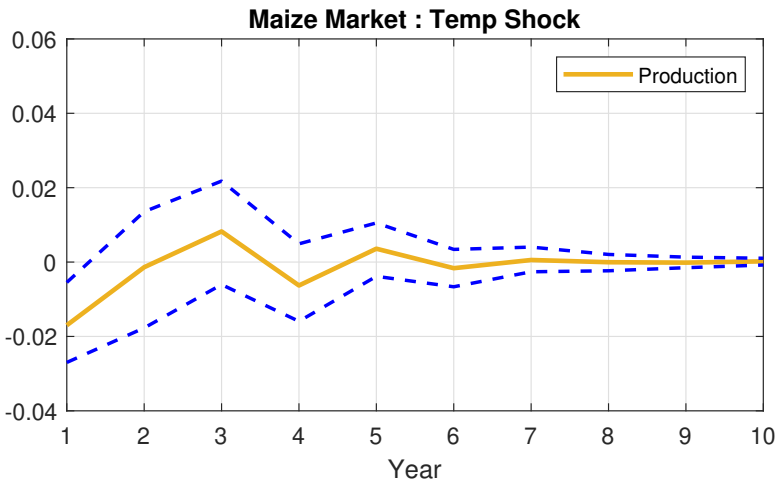


Figure: *Impulse response functions temperature shock on wheat market*



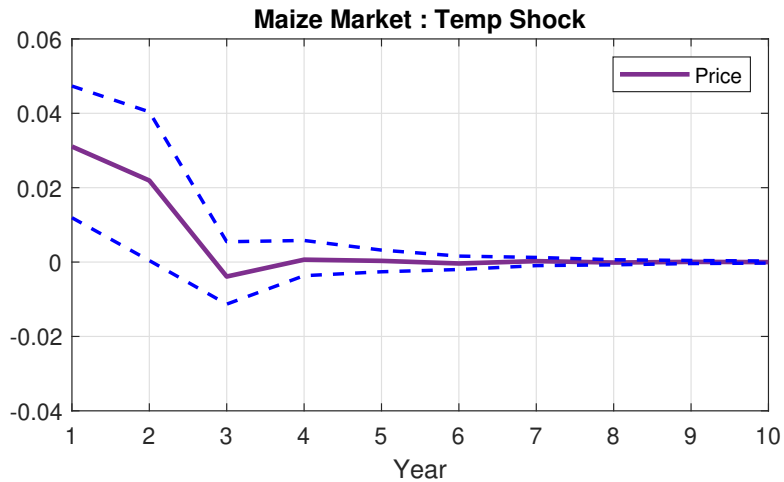


Figure: *68% confidence intervals of IRF on wheat market*

Table of Contents

Assumptions

Assumption 1 (Random functions)

$y_{it}, \lambda_{ik}, \varepsilon_{it} \in L^2([0,1], \mathbb{R})$, and $f_{tk} \in \mathbb{R}$.

Assumption 2 (Identification)

1. Almost surely,

$$\frac{F'F}{T} = I_K, \qquad \frac{G(X)^* G(X)}{N} = D,$$

   where $D \in \mathbb{R}^{K \times K}$ is a diagonal matrix with distinctive elements.

2. There exists two positive constants c_{min} and c_{max} such that with probability approaching one, as $N \to \infty$

$$c_{min} \leq \psi_{min}\left(\frac{G(X)^* G(X)}{N}\right) \leq \psi_{max}\left(\frac{G(X)^* G(X)}{N}\right) \leq c_{max}.$$

Assumptions

Assumption 3 (Basis functions)

1. As $N \to \infty$, with probability approaching one

$$d_{min} \leq \psi_{min}\left(\frac{\Phi(X)'\Phi(X)}{N}\right) \leq \psi_{max}\left(\frac{\Phi(X)'\Phi(X)}{N}\right) \leq d_{max},$$

   where d_{min} and d_{max} denote two positive constants.

2. $\max_{j \leq J, i \leq N, h \leq H} E[\phi_j(X_{ih})^2] < \infty$.

Assumption 4.1 (Data generating process)

1. A mean zero functional process $\{\varepsilon_t\}_{t \leq T}$ is independent of $\{X_i, f_t\}_{i \leq N, t \leq T}$.
2. $\{f_t, \varepsilon_t\}_{t \leq T}$ is strictly stationary.

Assumptions

Assumption 4.2 (Data generating process)

Let M_1 be a positive constant. Then

$$\max_{i \leq N} \sum_{q=1}^{N} \int_0^1 |E[\varepsilon_{it}(r_1)\varepsilon_{qt}(r_2)]| dr_1 dr_2 < M_1$$

$$\frac{1}{NT} \sum_{i,q=1}^{N} \sum_{t,s=1}^{T} \int_0^1 |E[\varepsilon_{it}(r_1)\varepsilon_{qs}(r_2)]| dr_1 dr_2 < M_1$$

$$\max_{i \leq N} \frac{1}{NT} \sum_{i,q=1}^{N} \sum_{t,s=1}^{T} \int_0^1 |cov[\varepsilon_{it}(r_1)\varepsilon_{qt}(r_2), \varepsilon_{is}(r_1)\varepsilon_{ms}(r_2)]| dr_1 dr_2 < M_1$$

Assumptions

Assumption 5 (Unexplained loading components)

1. $\{\gamma_{ik}\}_{i\leq N,k\leq K}$ is independent of $\{X_i\}_{i\leq N}$, and $E[\gamma_{ik}(r)] = 0$.
2. Let $\rho_N = \sup_{r\in[0,1],\, k\leq K} \frac{1}{N}\sum_{i=N}^{N} E[\gamma_{ik}^2(r)] < \infty$. Then we have

$$\sup_{r_1,r_2\in[0,1],i\leq N,k\leq K} \sum_{q=1}^{N} |E[\gamma_{ik}(r_1)\gamma_{qk}(r_2)]| = O(\rho_N).$$

3. $\sup_{r\in[0,1],i\leq N,k\leq K} E[g_k^2(X_i, r)] < \infty$.

Assumptions

Assumption 6 (Sieve approximation)

1. For all $h \leq H$, $k \leq K$, the loading component $g_{kh}(\cdot)$ belongs to a Hölder space $\mathcal{G}(\omega, \beta, L)$ defined as

$$\mathcal{G}(\omega, \beta, L) = \{g : |D^{\omega} g(v_1) - D^{\omega} g(v_2)| \leq L\|v_1 - v_2\|^{\beta}\}$$

   for some $L > 0$, $v_1, v_2 \in \mathbb{R} \times [0, 1]$.

2. Suppose $\kappa = (\omega + \beta) \geq 2$. As $J \to \infty$, the sieve coefficients $\{b_{k,jh}\}_{j \leq J}$ satisfy, for all $h \leq H$, $k \leq K$,

$$\sup_{r \in [0,1],\, x \in \mathcal{X}_h} |g_{kh}(x, r) - \sum_{j=1}^{J} b_{k,jh}(r)\phi_j(x)|^2 = O(J^{-\kappa}),$$

   where $\mathcal{X}_h$ denotes the support of X_{ih}.

3. $$\sup_{r \in [0,1],\, k \leq K, j \leq J, h \leq H} b_{k,jh}^2(r) < \infty.$$

Asymptotics for F and $G(X)$

Consider IFM:

$$Y = \Lambda F' + \varepsilon, \quad \Lambda = G(X) + \Gamma.$$

Theorem 1

Suppose $J = o(\sqrt{N})$. Under the assumptions 1-6, as $N, J \to \infty$ (T may stay constant or simultaneously grow with N and J),

$$\frac{1}{T}\|\widehat{F} - F\|^2 = O_p\Big(\frac{1}{N} + \frac{1}{J^\kappa}\Big),$$
$$\frac{1}{N}\|\widehat{G}(X) - G(X)\|^2 = O_p\Big(\frac{J}{N^2} + \frac{J}{NT} + \frac{J}{J^\kappa} + \frac{J\rho_N}{N}\Big).$$

Only large N assumption is needed for the consistency of F and $G(X)$.

Under the correct specification with $\Gamma = 0$, Λ is consistently estimable for a fixed T via our FPPCA.

Asymptotics for Γ Under Misspecification

Now let the model be misspecified and $\Lambda = G(X) + \Gamma$ with $\Gamma \neq 0$.

Corollary 1

Under assumptions of Theorem 1, as $T \to \infty$ simultaneously with N and J,

$$\frac{1}{N}\|\widehat{\Gamma} - \Gamma\|^2 = O_p\Big(\frac{J}{N^2} + \frac{1}{T} + \frac{1}{J^\kappa} + \frac{J\rho_N}{N}\Big).$$

Large N assumption is not sufficient for the consistency of Λ.

We additionally require the large T for the consistent estimation of Λ.

Estimating the Number of Factors

Assumption 7 (Error structure)

The error matrix $\varepsilon(r)$ can be decomposed as

$$\varepsilon(r) = A_N^{1/2} U(r) Z_T^{1/2},$$

1. $A_N \in \mathbb{R}^{N \times N}$ and $Z_T \in \mathbb{R}^{T \times T}$ are non-stochastic positive definite matrices where eigenvalues are bounded away from zero and infinity.
2. $U(r)$ is the $N \times T$ matrix of $u_{it}(r)$, where u_{it} is mean-zero and independent over i and t. In addition, $u_t = (u_{1t}, \ldots, u_{Nt})'$ is *iid* sub-Gaussian, that is, there exists $M_2 > 0$ such that

$$E\left[\exp\{\tau \langle u_t, v \rangle\}\right] \leq \exp\{\tau^2 M_2 \|v\|^2\},$$

for all $\tau > 0$, $v \in \mathcal{H}_N$.

Estimating the Number of Factors

The number of factors can be estimated in various ways:

- Eigenvalue Ratio (Ahn and Horenstein (2013))
- Eigenvalue Difference (Onatski (2010))
- AIC/BIC (Bai and Ng (2002), Alessi et al. (2010)).

Among the methods, we consider the ER.

Theorem 2 (Number of factors)

The number of factors estimator is defined as

$$\widehat{K} = \underset{1 \leq \ell < \ell_{max}}{\text{argmax}} \frac{\psi_\ell\left(\widehat{Y}^* \widehat{Y}\right)}{\psi_{\ell+1}\left(\widehat{Y}^* \widehat{Y}\right)}.$$

Under assumptions 1-7, and $1 \leq K < JH/2$, as $N \to \infty$, and $J = o\left(\min\{N, T\}\right)$, we have

$$P(\widehat{K} = K) \to 1.$$

Table of Contents

Estimation of Factors, Loadings, and Common Component

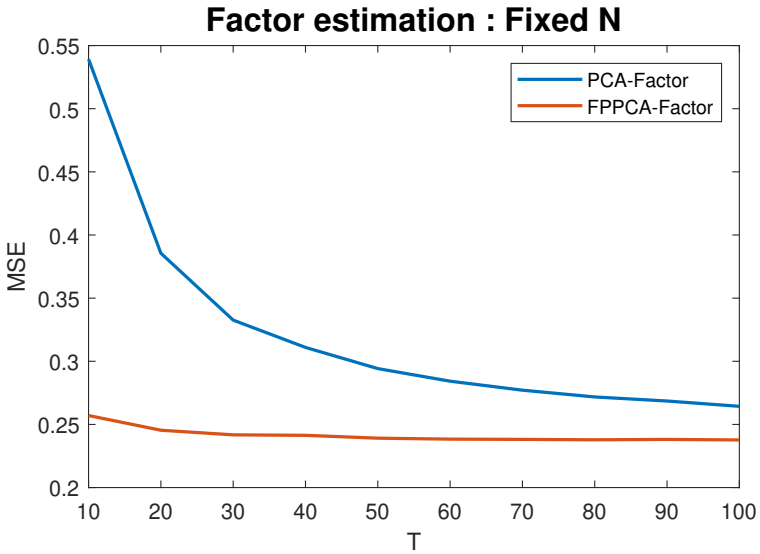
The purpose of the simulation is to compare the performance of FPPCA and PCA estimators.

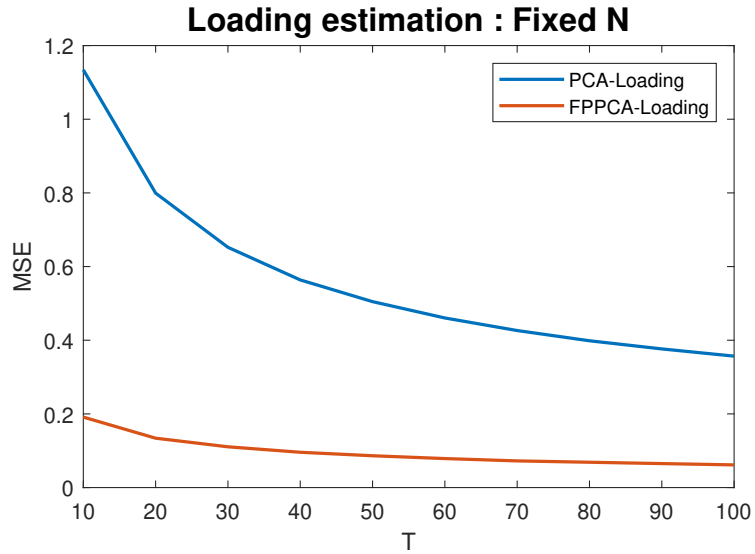
Consider a model with two factors and three characteristics:

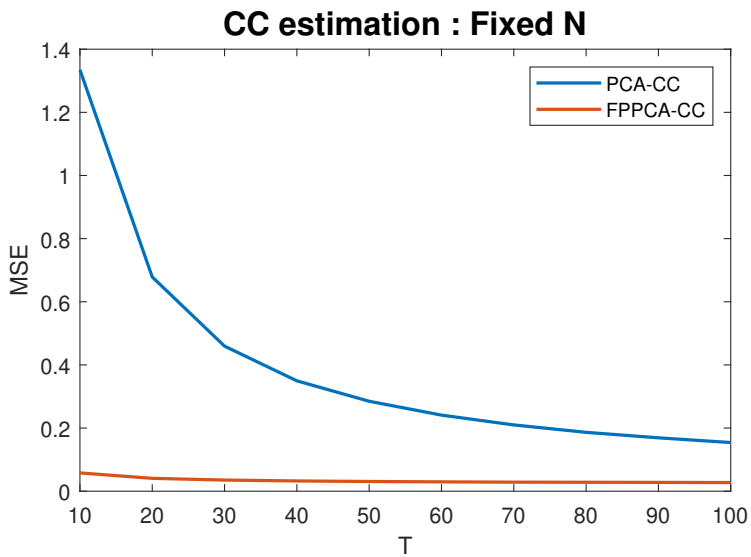
$$y_{it}(r) = g_1(X_i, r)f_{t1} + g_2(X_i, r)f_{t2} + \varepsilon_{it}(r).$$

- Basis functions : $\{\phi_1(r), \ldots \phi_5(r)\}$.
- Factor loading : $g_k(X_i, r) = (X_{i1}\beta_{k1} + X_{i2}\beta_{k2} + X_{i3}\beta_{k3})\phi_k(r)$.
  - $X_{ih} \overset{iid}{\sim} N(0, 1)$.
- Factors : $f_{tk} \overset{iid}{\sim} N(0, 1)$.
- Errors : $\varepsilon_{it}(r) = a_{1,it}\phi_1(r) + \cdots + a_{5,it}\phi_5(r), \quad a_{\ell,it} \overset{iid}{\sim} N(0, 1)$
- Sample size :
  - Fixed N sample : $N = 100,\ T = 10, 20 \ldots 100.$
  - Fixed T sample : $T = 100,\ N = 10, 20 \ldots 100.$

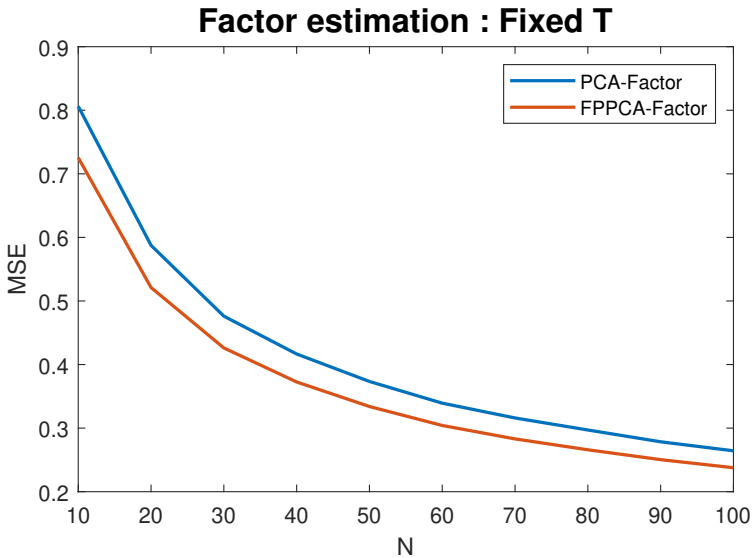
Estimation Results : Fixed N

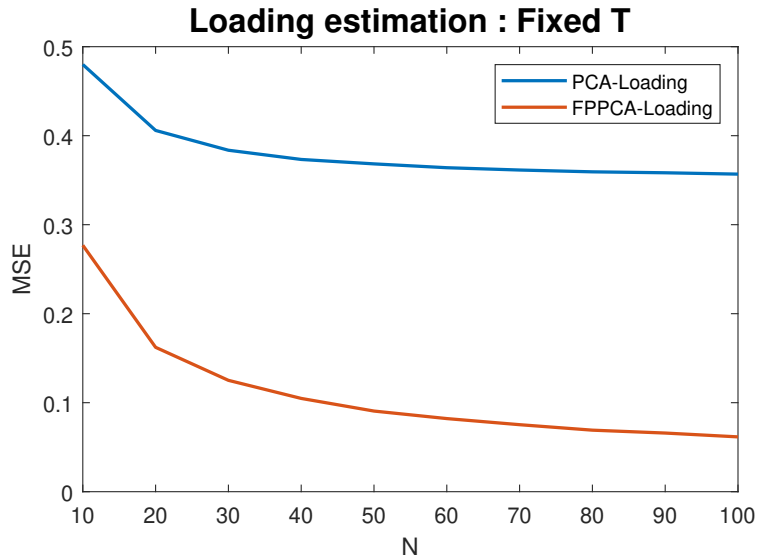


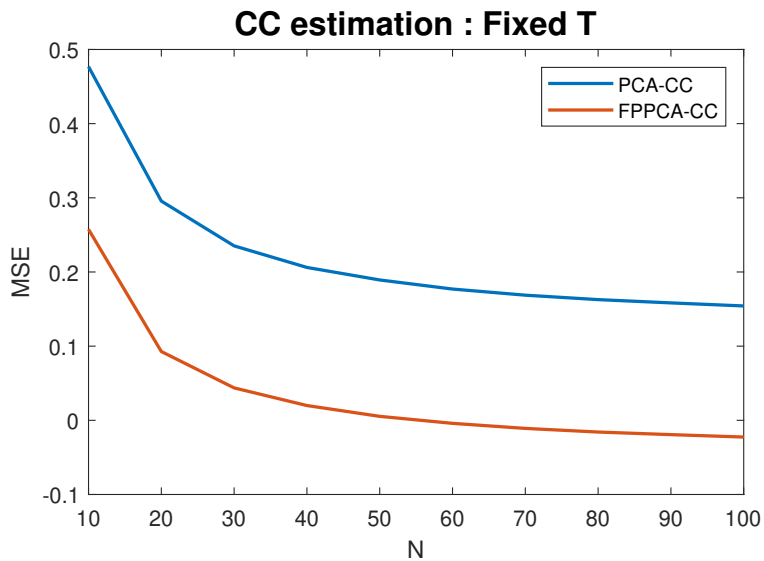




Estimation Results : Fixed T







Estimation Results

$$MSE\text{-}Ratio := \frac{MSE(FPPCA)}{MSE(PCA)}.$$

Table: MSE-ratio for fixed $N = 100$ (in %).

Size of T	10	20	30	40	50	60
(*Factors*)	47.7	63.6	72.7	77.6	81.3	83.9
(*Loading*)	16.8	16.8	16.9	17.0	17.1	17.1
(*CC*)	4.3	6.0	7.7	9.3	10.8	12.3

Table: MSE-ratio for fixed $T = 100$ (in %).

Size of N	10	20	30	40	50	60
(*Factors*)	90.0	88.7	88.5	88.4	88.4	87.1
(*Loading*)	57.7	40.0	32.6	28.1	24.6	22.6
(*CC*)	64.5	48.3	39.8	33.9	29.2	25.9

Estimation with Weak Instrument

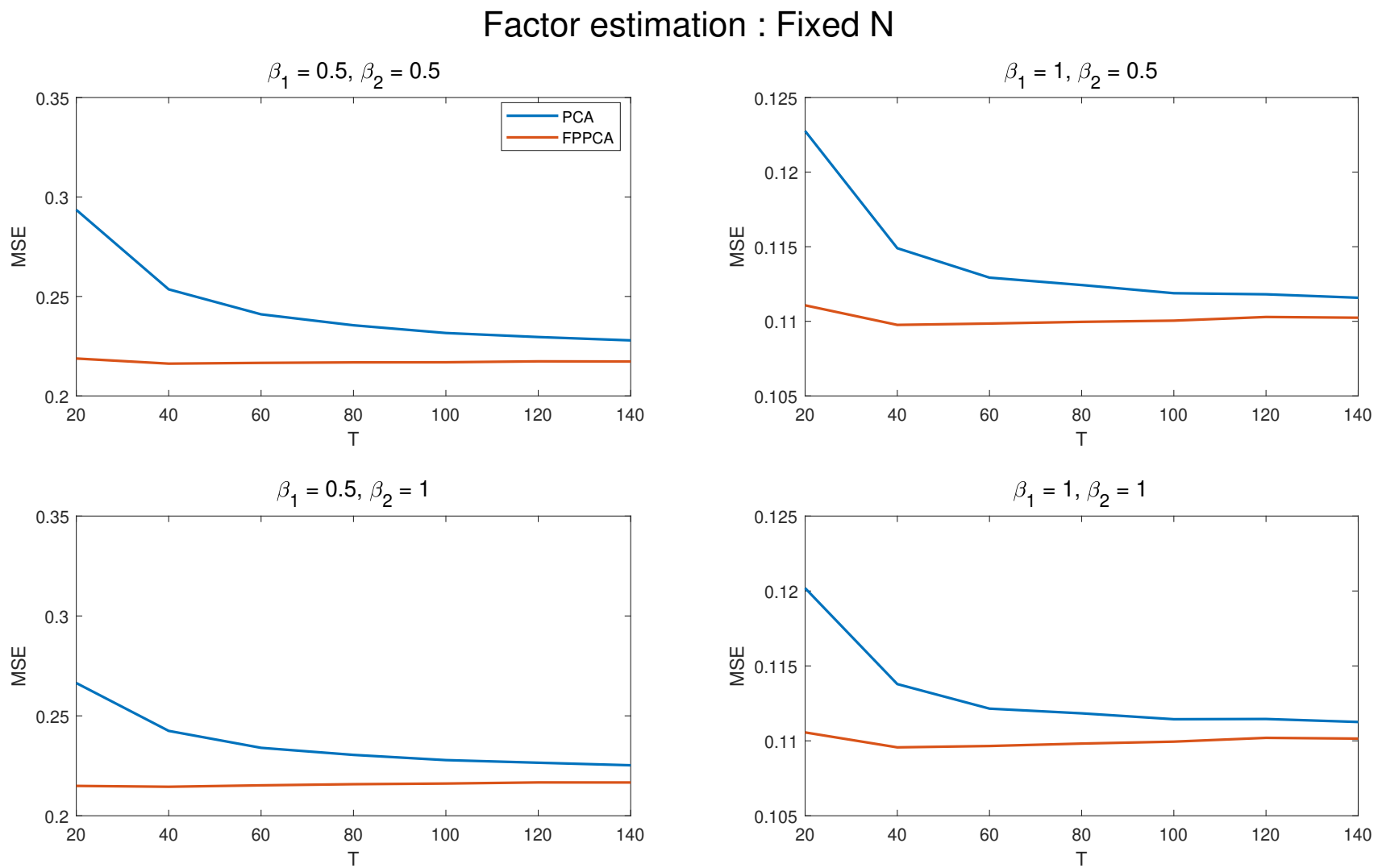
The objective is to examine the performance of the FPPCA when we observe an incomplete set of instruments that may be strong/weak.

Consider a model with one factor and two characteristics:

$$y_{it}(r) = \left[X_{i1}\beta_1\phi_1(r) + X_{i2}\left(\frac{\beta_2}{d_{NT}}\right)\phi_2(r) \right] f_t + \varepsilon_{it}(r), \qquad (1)$$

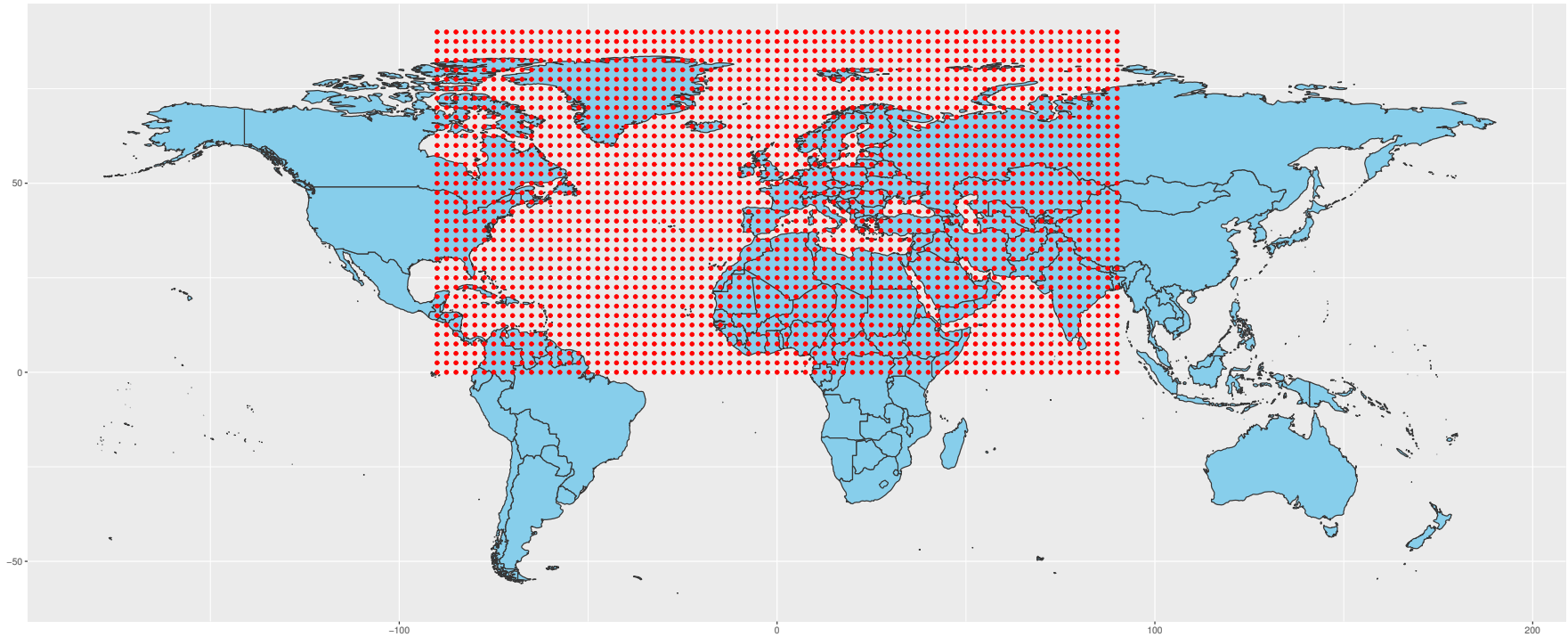
- Factor loading :
  - The parameter for X_2 is modeled as local-to-zero.
  - $\beta_1, \beta_2 \in \{0.5, 1\}$, $d_{NT} = T$.
  - $X_1, X_2 \sim N(0,1)$, and $cov(X_1, X_2) = 0.5$.
- Sample size :
  - Fixed N sample : $N = 100$, $T = 20, 40 \ldots 140$, $d_{NT} = T$.
- Instrument set :
  - X_1, strong instrument observed.

Estimation Results



Thank you for your attention!

Global Temperature : NCEP/NCAR Climate Dataset



- A globally gridded dataset of 2701 stations (i).
- Air temperature recorded daily (t) at time (r).
- One observation : $Temp_{it}(r)$.

Go back

PCA for Functional Data: Remark 1

Let $\mathcal{S}_N = \mathcal{H}_1 \oplus \mathcal{H}_2 \oplus \cdots \oplus \mathcal{H}_N$, where $\oplus$ denotes the direct some of spaces. The inner product of the space $\mathcal{S}_N$ is defined by

$$\langle v, w \rangle_{\mathcal{S}_N} = \sum_{i=1}^{N} \langle v_i, w_i \rangle_{\mathcal{H}_i}, \quad \forall v, w \in \mathcal{S}_N.$$

If $\mathcal{H}_i = \mathbb{R}$,

$$\langle v, w \rangle_{\mathcal{S}_N} = \sum_{i=1}^{N} \langle v_i, w_i \rangle_{\mathbb{R}} = v'w.$$

If $\mathcal{H}_i = \mathbb{R}^R$,

$$\langle v, w \rangle_{\mathcal{S}_N} = \sum_{i=1}^{N} \langle v_i, w_i \rangle_{\mathbb{R}^K} = \sum_{i=1}^{N} \sum_{r=1}^{R} v_{ir} w_{ir}.$$

If $\mathcal{H}_i = L^2([0,1])$,

$$\langle v, w \rangle_{\mathcal{S}_N} = \sum_{i=1}^{N} \langle v_i, w_i \rangle_{L^2} = \sum_{i=1}^{N} \int_0^1 v_i(r) w_i(r) dr$$

Go back

PCA for Functional Data: Remark 2

Let $Y = (y_1, \ldots, y_T)$, and $y_t = (y_{1t}, \ldots, y_{Nt}) \in \mathcal{S}_N$.

Then, we can define $Q = Y^* Y$, a real-valued $T \times T$ matrix such that

$$Q = \begin{pmatrix} \langle y_1 , y_1 \rangle_{\mathcal{S}_N} & \cdots & \langle y_1 , y_T \rangle_{\mathcal{S}_N} \\ \vdots & \ddots & \vdots \\ \langle y_T , y_1 \rangle_{\mathcal{S}_N} & \cdots & \langle y_T , y_T \rangle_{\mathcal{S}_N} \end{pmatrix}.$$

If $y_{it} \in \mathbb{R}$, we have $Q = Y' Y$ since

$$Q = \begin{pmatrix} y_1' y_1 & \cdots & y_1' y_T \\ \vdots & \ddots & \vdots \\ y_T' y_1 & \cdots & y_T' y_T \end{pmatrix}.$$

If $y_{it} \in L^2([0,1])$, we have $Q = \int_0^1 Y(r)' Y(r) dr$ since

$$Q = \begin{pmatrix} \int_0^1 y_1'(r) y_1(r) dr & \cdots & \int_0^1 y_1'(r) y_T(r) dr \\ \vdots & \ddots & \vdots \\ \int_0^1 y_T'(r) y_1(r) dr & \cdots & \int_0^1 y_T'(r) y_T(r) dr \end{pmatrix}.$$

Go back