# Gaussian Approximation for High Dimensional Vector Under Physical Dependence

Xianyang Zhang
Joint work with Guang Cheng

# High dimensional time series

- Modern time series datasets often defy traditional statistical assumptions.
- Key features:
  1. **high dimensional**
  2. **non-normally-distributed**
  3. **non-linear**
  4. **nonstationary**
- Application areas:
  1. Macroeconomics and finance
  2. Neuroscience
  3. Climate studies

# Statistical problems for high dimensional time series

- Factor modeling, time series PCA and clustering
- (Auto)covariance structure estimation, graphical modeling and causality
- Sparse modeling and regularized estimation
- Change-point detection and estimation
- Predictive inference and forecasting
- **Statistical inference and uncertainty quantification**
- ......

## CLT for low dimensional time series

- Consider $n$ observations $\{x_i\}_{i=1}^n$ from a $p$-dimensional time series with $p \ll n$.

- Central Limit Theorem (CLT):

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu_i) \to^d N(0, \Sigma),$$

$$\mu_i = \mathbb{E}[x_i], \quad \Sigma = \lim_{n \to +\infty} \frac{1}{n} \sum_{i,j=1}^{n} \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)'].$$

See Rosenblatt (1956), Ibragimov and Linnik (1971), Wu (2005) among others.

## Inference for low dimensional time series

- Continuous mapping theorem:

$$h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i-\mu)\right)\to^d h\left(N(0,\Sigma)\right),$$

where $h:\mathbb{R}^p\to\mathbb{R}$ is continuous.

- Special cases:

$$h(z)=\max_{1\le i\le p} z_i,$$
$$h(z)=z'Az,$$

where $z=(z_1,\ldots,z_p)'$ and $A\in\mathbb{R}^{p\times p}$.

## CLT fails in high dimension

- Portnoy (1986) showed that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu_i)$$

no longer converges to the Gaussian limit when $\sqrt{n} = o(p)$.

- For a specific $h$, does

$$h\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(x_i - \mu_i)\right) \to^d h\left(N(0, \Sigma)\right), \qquad (1)$$

still hold when $p \approx n$ or even $p \gg n$?

- For **independent** data, (1) holds when

$$h(z) = \max_{1 \le i \le p} z_i \quad \text{and} \quad h(z) = z' A z.$$

See Bai and Saranadasa (1996) and Chernozhukov et al. (2013).

## Our main contribution

- Develop a Gaussian approximation result for **high-dimensional**, **non-stationary**, **non-linear**, **non-Gaussian** time series when $h(z) = \max_{1 \le i \le p} z_i$.

- Let $y_i$ be a Gaussian sequence which preserves the autocovariance structure of $x_i$. Suppose $\mathbb{E}[x_i] = \mathbb{E}[y_i] = 0$.

- **Main result**:

$$\rho_n := \sup_{t \ge 0} \left| P\left( \max_{1 \le i \le p} X_{n,i} \le t \right) - P\left( \max_{1 \le i \le p} Y_{n,i} \le t \right) \right| \to 0,$$

where

$$X_n = (X_{n,1}, \ldots, X_{n,p})' = n^{-1/2} \sum_{i=1}^{n} x_i,$$

$$Y_n = (Y_{n,1}, \ldots, Y_{n,p})' = n^{-1/2} \sum_{i=1}^{n} y_i.$$

# Applications

- Multiplicity adjustment in large-scale inference
- Simultaneous inference for mean and covariance structure, white noise testing [Zhang and Cheng (2014); Zhang and Wu (2016); Chang et al. (2017)]
- Change-point detection [Dette and Gömann (2017)]

## Smooth approximation

- Note that

$$P\left(\max_{1\le i\le p} X_{n,i} \le t\right) = \mathbb{E}\left[\mathbf{1}\left\{\max_{1\le i\le p} X_{n,i} \le t\right\}\right].$$

  Both the maximum function and the indicator function $\mathbf{1}\{\cdot \le t\}$ are non-smooth.

- Approximate $\max_{1\le i\le p} z_i$ by the "soft maximum"

$$F_\beta(z) := \beta^{-1}\log\left(\sum_{j=1}^{p}\exp(\beta z_j)\right), \quad \text{where } z = (z_1,\dots,z_p)'.$$

  We have

$$0 \le F_\beta(z) - \max_{1\le i\le p} z_i \le \beta^{-1}\log p.$$

- Approximate $\mathbf{1}\{\cdot \le t\}$ by a sufficiently smooth function say $g(\cdot)$.

## Moment match

- By the smooth approximation,

$$\left| P\left(\max_{1\le i\le p} X_{n,i} \le t\right) - P\left(\max_{1\le i\le p} Y_{n,i} \le t\right) \right|$$
$$\approx \left| \mathbb{E} g \circ F_\beta(X_n) - \mathbb{E} g \circ F_\beta(Y_n) \right| .$$

  From now on, we write $g \circ F_\beta(\cdot)$ as $m(\cdot)$.

- How can we compare $\mathbb{E}[m(X_n)]$ with $\mathbb{E}[m(Y_n)]$?

- Two classical methods

  1. Slepian-Stein smart path interpolation: **second moment match**.
  2. Lindeberg exchange method: **third or higher moment match**.

## Slepian-Stein interpolation

- Smart interpolation:

$$Z_n(t) = \sqrt{t}X_n + \sqrt{1-t}\,Y_n = \sum_{i=1}^{n}(z_{i,1}(t), \ldots, z_{i,p}(t))',$$

  where $\mathrm{var}(Z_n(t)) = \mathrm{var}(X_n) = \mathrm{var}(Y_n)$.

-

$$\begin{aligned}
\mathbb{E}[m(X_n)] - \mathbb{E}[m(Y_n)] =& \mathbb{E}[m(Z_n(1))] - \mathbb{E}[m(Z_n(0))] \\
=& \int_0^1 \frac{\partial \mathbb{E}[m(Z_n(t))]}{\partial t} dt \\
=& \sum_{i=1}^{n}\sum_{j=1}^{p}\int_0^1 \mathbb{E}[\partial_j m(Z_n(t))]\frac{\partial z_{i,j}(t)}{\partial t} dt.
\end{aligned}$$

- We develop a new argument to analyze the RHS when $x_i$ is a M-dependent time series.

## Physical dependence

- Consider a $p$-dimensional random vector with the following causal representation:

$$x_i := \mathcal{G}_i(\ldots, \epsilon_{i-1}, \epsilon_i),$$

  where $\mathcal{G}_i = (\mathcal{G}_{i,1}, \ldots, \mathcal{G}_{i,p})'$ and $\{\epsilon_i\}_{i\in\mathbb{Z}}$ are i.i.d elements.

- Define

$$\theta_{k,j,q} = \sup_i (\mathbb{E}|\mathcal{G}_{i,j}(\mathcal{F}_i) - \mathcal{G}_{i,j}(\mathcal{F}_{i,i-k})|^q)^{1/q}, \quad \Theta_{k,j,q} = \sum_{l=k}^{+\infty} \theta_{l,j,q},$$

  where

$$\begin{aligned} \mathcal{F}_i &= (\ldots, \epsilon_{i-1}, \epsilon_i), \\ \mathcal{F}_{i,i-k} &= (\ldots, \epsilon_{k-1}, \epsilon'_{i-k}, \epsilon_{i-k+1}, \ldots, \epsilon_{i-1}, \epsilon_i). \end{aligned}$$

## M-dependent approximation

- Construct a M-dependent time series:

$$x_i^{(M)} = E[x_i | \epsilon_{i-M}, \epsilon_{i-M+1}, \ldots, \epsilon_i].$$

- Derive a finite sample upper bound for

$$\left| \mathbb{E}[m(X_n^{(M)})] - \mathbb{E}[m(Y_n^{(M)})] \right|,$$

  where $X_n^{(M)} = n^{-1/2} \sum_{i=1}^{n} x_i^{(M)}$.

- Quantify the M-dependent approximation error:

$$P(|X_n^{(M)} - X_n|_\infty > t)$$

  where $|\cdot|_\infty$ is the $l_\infty$ norm.

## Proof roadmap

$$\max_{1 \le i \le p} X_{n,i}$$

$$\Big\downarrow \textbf{ M-dependent approximation}$$

$$\max_{1 \le i \le p} X_{n,i}^{(M)}$$

$$\Big\downarrow \textbf{ Modified Stein's method}$$

$$\max_{1 \le i \le p} Y_{n,i}^{(M)}$$

$$\Big\downarrow \textbf{ Gaussian-to-Gaussian approximation}$$

$$\max_{1 \le i \le p} Y_{n,i}$$

## Key result

Assume that

- **High dimensionality**:

$$p \lesssim \exp(n^b) \quad \text{for} \quad 0 \leq b < 1/11.$$

- **Weak dependence**:

$$\max_{1 \leq j \leq p} \Theta_{k,j,q} \lesssim \varrho^k \quad \text{for} \quad \varrho < 1, q \geq 2.$$

- **Moment condition**: one of the following two conditions holds

$$\max_{1 \leq i \leq n} \mathbb{E}(\max_{1 \leq j \leq p} |x_{ij}|/\mathfrak{D}_n)^4 \leq 1, \quad \mathfrak{D}_n \lesssim n^{(3-25b)/32},$$

$$\max_{1 \leq i \leq n} \max_{1 \leq j \leq p} \mathbb{E} \exp(|x_{ij}|/\mathfrak{D}_n) \leq 1, \quad \mathfrak{D}_n \lesssim n^{(3-17b)/8}.$$

Then

$$\rho_n \lesssim n^{-(1-11b)/8}.$$

## Key result (con't)

Dependence adjusted norm [Zhang and Wu (2016)]:

$$\omega_{j,q} = \max_i \| \, \|\mathcal{G}_i(\mathcal{F}_i) - \mathcal{G}_i(\mathcal{F}_{i,i-j})\|_\infty \|_q, \quad \Omega_{M,q} = \sum_{j=M}^{+\infty} \omega_{j,q}.$$

Assume that

- **High dimensionality**:

$$p \lesssim \exp(n^b) \quad \text{for} \quad 0 \leq b < 1/11.$$

- **Weak dependence** $+$ **Moment condition**:

$$\Omega_{M+1,q} \asymp M^{-\alpha} \quad \text{for} \quad \alpha > (1+b)/(1-7b).$$

Then

$$\rho_n \lesssim n^{-c}, \quad c > 0.$$

# Nonstationary linear model

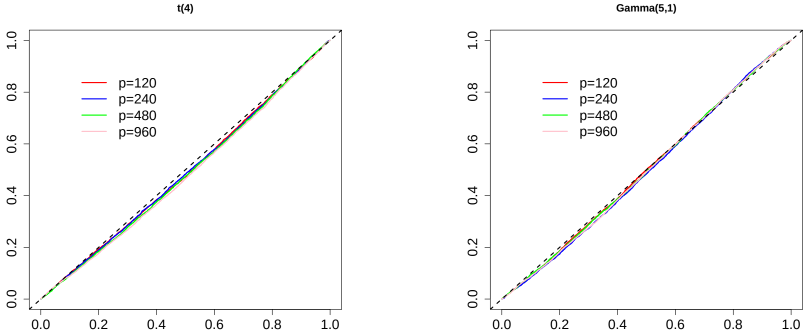
- Nonstationary linear model:

$$x_i = \sum_{l=0}^{+\infty} \mathbf{A}^{i,l} \epsilon_{i-l}.$$

- Our assumptions are satisfied if
  1. $\sup_i \max_{1 \le j \le p} ||\mathbf{A}^{i,l}_{j,\cdot}||_2 \lesssim \varrho^l$, for some $\varrho < 1$.
  2. The components of $\epsilon_i$ are sub-exponential.

# Numerical results

**Figure:** P-P plots comparing the distributions of $|X_n|_\infty$ and $|Y_n|_\infty$, where the data are generated from the time-varying VAR(1) model.



## Estimating the covariance structure

- The Gaussian approximation theory says that

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu_i) \right|_\infty \approx^d |N(0, \Sigma_n)|_\infty,$$

where $\Sigma_n = \operatorname{var}\left(n^{-1/2} \sum_{i=1}^{n} x_i\right)$.

- Subsampling estimator for $\Sigma_n$:

$$\hat{\Sigma}_n = \frac{M}{n-M+1} \sum_{i=1}^{n-M+1} \left( \frac{1}{M} \sum_{j=i}^{i+M-1} x_j - \bar{x} \right) \left( \frac{1}{M} \sum_{j=i}^{i+M-1} x_j - \bar{x} \right)',$$

where $1/M + M/n \to 0$.

- Approximate the distribution of

$$\left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (x_i - \mu_i) \right|_\infty \quad \text{by that of } |N(0, \hat{\Sigma}_n)|_\infty.$$

## Testing second-order stationarity

- Consider the null hypothesis

$$H_0 : \mathbb{E}[x_{i+h}x_i'] = \Gamma(h) \quad \text{for } 0 \leq h \leq H \text{ and all } i.$$

- Define

$$\hat{\Gamma}^{(k)}(h) := (\hat{\gamma}_{i,j}^{(k)}(h))_{i,j=1}^{p} = \frac{1}{n}\sum_{i=1}^{n-h} \phi_k\left(\frac{i-1}{n}\right) x_{i+h}x_i',$$

where $\phi_k(\cdot)$ is a sequence of orthonormal basis on $[0,1]$ such that

$$\int_0^1 \phi_k(u)du = 0, \quad 1 \leq k \leq K.$$

- Our statistic:

$$\mathcal{G} = \sqrt{n} \max_{1\leq i,j\leq p} \max_{0\leq h\leq H} \max_{1\leq k\leq K} |\hat{\gamma}_{i,j}^{(k)}(h)|.$$

# Testing second-order stationarity (Con't)

| | | $p=20$ | | $p=30$ | | $p=40$ | |
|---|---|---|---|---|---|---|---|
| | $n$ | 10% | 5% | 10% | 5% | 10% | 5% |
| $H_0$ | 120 | 13.6 | 4.9 | 11.1 | 4.4 | 9.9 | 3.5 |
| | 240 | 11.7 | 5.1 | 9.4 | 3.4 | 7.0 | 3.1 |
| $H_a$ | 120 | 64.4 | 40.1 | 59.9 | 36.1 | 60.9 | 35.2 |
| | 240 | 100.0 | 99.7 | 100.0 | 99.9 | 100.0 | 99.9 |

**Table:** Rejection percentages for testing second-order stationarity. Under the null, the data are generated from a VAR(1) model. Under the alternative, the data are generated from a time varying VAR(1) model. The actual number of parameters is equal to $p^2\kappa H$ (i.e., 4800, 10800, and 19200 for $p = 20, 30, 40$ respectively).

# Conclusion

- Develop a Gaussian approximation theory for maxima of sums of dependent random vectors.
- A modified Stein's method for dependent data and M-dependent approximation.
- Future directions:
  1. Improve the rate on $p$ using Lindeberg exchange method [Deng and Zhang (2017)].
  2. Develop a rigorous bootstrap theory for locally stationary time series.
  3. Inference for high dimensional locally stationary time series.

Thank you!