High-dimensional autocovariance matrices, Linear Process Bootstrap and optimal linear prediction

Dimitris N. Politis, UCSD
[joint work with Tim McMurry and Carsten Jentsch]
Setup

\(X_1, \ldots, X_n\) data from a stationary process with weak dependence

Mean \(\mu = EX_t\) is estimated by the sample mean \(\bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t\)

Lag-\(h\) autocovariance \(\gamma_h = \text{Cov}(X_t, X_{t+h})\) is estimated by the sample autocovariance \(\hat{\gamma}_h = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X})(X_{t+h} - \bar{X})\) for \(h \geq 0\)

**GOAL:** Estimate \(\Gamma_n\), the autocovariance matrix of \(X_1, \ldots, X_n\).

\[
\Gamma_n = \begin{bmatrix}
\gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\
\gamma_1 & \gamma_0 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \gamma_1 \\
\gamma_{n-1} & \cdots & \gamma_1 & \gamma_0
\end{bmatrix}
\]

But \(\hat{\Gamma}_n = [\hat{\gamma}_{|i-j|}]_{i,j=1}^{n}\) is not consistent for \(\Gamma_n\) (in operator norm).
Setup

\( X_1, \ldots, X_n \) data from a stationary process with weak dependence

Mean \( \mu = EX_t \) is estimated by the sample mean \( \bar{X} = \frac{1}{n} \sum_{t=1}^{n} X_t \)

Lag-\( h \) autocovariance \( \gamma_h = \text{Cov}(X_t, X_{t+h}) \) is estimated by the sample autocovariance \( \hat{\gamma}_h = \frac{1}{n} \sum_{t=1}^{n-h} (X_t - \bar{X})(X_{t+h} - \bar{X}) \) for \( h \geq 0 \)

**GOAL:** Estimate \( \Gamma_n \), the autocovariance **matrix** of \( X_1, \ldots, X_n \).

\[
\Gamma_n = \begin{bmatrix}
\gamma_0 & \gamma_1 & \cdots & \gamma_{n-1} \\
\gamma_1 & \gamma_0 & \ddots & \vdots \\
\vdots & \ddots & \ddots & \gamma_1 \\
\gamma_{n-1} & \cdots & \gamma_1 & \gamma_0 \\
\end{bmatrix}
\]

But \( \hat{\Gamma}_n = [\hat{\gamma}_{|i-j|}]_{i,j=1}^{n} \) is not consistent for \( \Gamma_n \) (in operator norm).
Banded estimates

Wu and Pourahmadi (2009) proposed an $l$-banded estimator

$$\hat{\Gamma}_{WP} = \begin{bmatrix}
\hat{\gamma}_0 & \ldots & \hat{\gamma}_l & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & 0 \\
\hat{\gamma}_l & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & \hat{\gamma}_l & \ddots & \ddots & \hat{\gamma}_0 \\
\end{bmatrix}$$
Banded and tapered estimates

Weight \( \hat{\gamma}_{|i-j|} \) by a \( \kappa_l(i-j) \), where \( l \) is a banding parameter, and \( \kappa \) is a flat-top lag-window.

\[
\kappa(x) = \begin{cases} 
1 & \text{if } |x| \leq 1 \\
g(|x|) & \text{if } 1 < |x| \leq c_\kappa \\
0 & \text{if } |x| > c_\kappa,
\end{cases}
\]

and

\[
\kappa_l(x) = \kappa(x/l).
\]
Banded and tapered estimates

Weight $\hat{\gamma}_{|i-j|}$ by a $\kappa_l(i - j)$, where $l$ is a banding parameter, and $\kappa$ is a flat-top lag-window.

$$\kappa(x) = \begin{cases} 1 & \text{if } |x| \leq 1 \\ g(|x|) & \text{if } 1 < |x| \leq c_\kappa \\ 0 & \text{if } |x| > c_\kappa, \end{cases}$$

and

$$\kappa_l(x) = \kappa(x/l).$$

Tapered estimator:

$$\hat{\Gamma}_{\kappa, l} = \left[ \kappa_l(i - j) \hat{\gamma}_{|i-j|} \right]_{i,j=1}^n$$
Convergence rates

Let $\rho(A) = \sqrt{\lambda_{\text{max}}(A^*A)}$ denote operator norm and $||Y||_p = E[|Y|^p]^{1/p}$.

Theorem (McMurry and Politis, 2011)

Assume $||X_1||_4 < \infty$, weak dependence conditions, and $0 < l \leq n$. Then

$$\rho\left(\hat{\Gamma}_{\kappa,l} - \Gamma_n^{-1}\right) = O_p(r_n), \text{ where } r_n = \ln^{-1/2} + \sum_{i=l}^{\infty} |\gamma_j|.$$
Optimal banding parameter

**Corollary**

1. $|\gamma_i| = O(i^{-d})$ for some $d > 1$
   - **Optimal l**: $l \propto n^{1/(2d)}$
   - **Convergence rate bound**: $O \left( n^{-(d-1)/(2d)} \right)$

2. $|\gamma_i| = O(\theta^i)$ for some $\theta$ with $|\theta| < 1$
   - **Optimal l**: $l = \lfloor a \log n \rfloor$ for a large enough $a$
   - **Convergence rate bound**: $O \left( n^{-1/2} \log n \right)$

3. There exists $B$ such that $\gamma_i = 0$ for all $i > B$
   - **Optimal l**: $l = B$
   - **Convergence rate bound**: $O \left( n^{-1/2} \right)$
Convergence rates for the inverse

**Theorem (McMurry and Politis, 2011)**

Assume the conditions of previous theorem, and that \( l \) grows at a rate that ensures \( r_n \to 0 \).

Also assume that the spectral density \( f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-ih\omega} \) satisfies

\[
0 < c_1 \leq f(\omega) \leq c_2 < \infty.
\]

Then \( \hat{\Gamma}_{\kappa,l} \) is positive definite with probability tending to 1, and

\[
\rho \left( \hat{\Gamma}_{\kappa,l}^{-1} - \Gamma_n^{-1} \right) = O_p(r_n), \text{ where } r_n = \ln^{-1/2} + \sum_{i=l}^{\infty} |\gamma_j|.
\]
Finite-sample positive definiteness

\[ \hat{\Gamma}_{\kappa, l} = TDT^t \] where \( D = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \) and \( T \) is orthogonal.

Define the thresholded estimate

\[ \hat{\Gamma}^\epsilon_{\kappa, l} = T D^\epsilon T^t, \]

where \( D^\epsilon = \text{diag}(\lambda^\epsilon_1, \ldots, \lambda^\epsilon_n) \) and \( \lambda^\epsilon_i = \max(\hat{\lambda}_i, \epsilon \hat{\gamma}_0 / n^\beta) \).
Finite-sample positive definiteness

\( \hat{\Gamma}_{\kappa,l} = TDT^t \) where \( D = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \) and \( T \) is orthogonal.

Define the thresholded estimate

\[ \hat{\Gamma}^\varepsilon_{\kappa,l} = TD^\varepsilon T^t, \]

where \( D^\varepsilon = \text{diag}(\lambda_1^\varepsilon, \ldots, \lambda_n^\varepsilon) \) and \( \lambda_i^\varepsilon = \max(\hat{\lambda}_i, \varepsilon \hat{\gamma}_0 / n^\beta) \)

**Theorem**

*Under the conditions of the previous theorem, with \( \varepsilon \geq 0 \) and \( \beta > 1/2 \), we have*

\[ \rho \left( (\hat{\Gamma}^\varepsilon_{\kappa,l})^{-1} - \Gamma_n^{-1} \right) = O_p(r_n). \]
Connection to Spectral estimation

The $j$th eigenvalue of $\Gamma_n$ is approximately given by $2\pi f(\omega_j)$ where $f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-ih\omega}$ and $\omega_j = 2\pi j/n$ for $j = 0, 1, \ldots, n-1$. 

$\hat{f}(\omega)$ is a higher-order (actually: infinite order) estimate and thus it is not guaranteed to be $\geq 0$. 
Connection to Spectral estimation

The \( j \)th eigenvalue of \( \Gamma_n \) is approximately given by \( 2\pi f(\omega_j) \) where \( f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-ih\omega} \) and \( \omega_j = 2\pi j/n \) for \( j = 0, 1, \ldots, n - 1 \).

The \( j \)th eigenvalue of \( \hat{\Gamma}_{\kappa,l} \) is approximately given by \( 2\pi \hat{f}(\omega_j) \) where \( \hat{f}(\omega) = (2\pi)^{-1} \sum_h \kappa(h/l) \hat{\gamma}_h e^{-ih\omega} \) is the flat-top spectral density estimate.
Connection to Spectral estimation

The $j$th eigenvalue of $\Gamma_n$ is approximately given by $2\pi f(\omega_j)$ where $f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-ih\omega}$ and $\omega_j = 2\pi j/n$ for $j = 0, 1, \ldots, n-1$.

The $j$th eigenvalue of $\hat{\Gamma}_{\kappa,l}$ is approximately given by $2\pi \hat{f}(\omega_j)$ where $\hat{f}(\omega) = (2\pi)^{-1} \sum_h \kappa(h/l) \hat{\gamma}_h e^{-ih\omega}$ is the flat-top spectral density estimate.

$\hat{f}(\omega)$ is a higher-order (actually: infinite order) estimate and thus it is not guaranteed to be $\geq 0$. 
Positive estimators

Positive spectral density estimators \( \hat{f}_{2o}(\omega) \) are available based on 2nd order kernels, e.g., Parzen’s.
Positive estimators

Positive spectral density estimators $\hat{f}_{2o}(\omega)$ are available based on 2nd order kernels, e.g., Parzen’s.

But flat-top kernel estimators are more accurate!
Positive estimators

Positive spectral density estimators $\hat{f}_{2o}(\omega)$ are available based on 2nd order kernels, e.g., Parzen’s.

But flat-top kernel estimators are more accurate!

$$\hat{f}_{2o}(\omega) = f(\omega) + O_P(\frac{1}{n^{2/5}})$$

$$\hat{f}(\omega) = f(\omega) + O_P(\frac{1}{n^{d/(2d+1)}}) \text{ with } d \text{ = number of derivatives of } f(\omega)$$
Positive estimators

Positive spectral density estimators \( \hat{f}_{2o}(\omega) \) are available based on 2nd order kernels, e.g., Parzen’s.

But flat-top kernel estimators are more accurate!

\[
\hat{f}_{2o}(\omega) = f(\omega) + O_P\left(\frac{1}{n^{2/5}}\right)
\]

\[
\hat{f}(\omega) = f(\omega) + O_P\left(\frac{1}{n^{d/(2d+1)}}\right) \text{ with } d = \text{number of derivatives of } f(\omega)
\]

If \( f \) is very smooth, then \( \hat{f}(\omega) \) is almost \( \sqrt{n} \)—convergent.
Shrinkage correction

Same accuracy holds for $\hat{f}^+(\omega) = \max(0, \hat{f}(\omega))$, i.e., $\hat{f}^+(\omega)$ has the same fast rate of convergence as $\hat{f}(\omega)$. 
Shrinkage correction

Same accuracy holds for $\hat{f}^+(\omega) = \max(0, \hat{f}(\omega))$, i.e., $\hat{f}^+(\omega)$ has the same fast rate of convergence as $\hat{f}(\omega)$.

But: finite-sample singularity of the autocovariance matrix ensues when $\hat{f}^+(\omega) = 0$.
Shrinkage correction

Same accuracy holds for $\hat{f}^+(\omega) = \max(0, \hat{f}(\omega))$, i.e., $\hat{f}^+(\omega)$ has the same fast rate of convergence as $\hat{f}(\omega)$.

But: finite-sample singularity of the autocovariance matrix ensues when $\hat{f}^+(\omega) = 0$

Previous threshold correction is tantamount to using the spectral estimator $\hat{f}^\epsilon(\omega) = \max(\epsilon \hat{\gamma}_0 n^{-\beta}, \hat{f}(\omega))$
Shrinkage correction

Same accuracy holds for $\hat{f}^+(\omega) = \max(0, \hat{f}(\omega))$, i.e., $\hat{f}^+(\omega)$ has the same fast rate of convergence as $\hat{f}(\omega)$.

But: finite-sample singularity of the autocovariance matrix ensues when $\hat{f}^+(\omega) = 0$

Previous threshold correction is tantamount to using the spectral estimator $\hat{f}^\epsilon(\omega) = \max(\epsilon \hat{\gamma}_0 n^{-\beta}, \hat{f}(\omega))$

**Shrinkage idea:** For $\omega$ where $\hat{f}^+(\omega) = 0$ (or very small) use the convex combination:

$$\hat{f}^*(\omega) = c\hat{f}_{2o}(\omega) + (1-c)\hat{f}^+(\omega) = c\hat{f}_{2o}(\omega)$$

i.e., a shrunk form of $\hat{f}_{2o}(\omega)$ since $c \in [0, 1]$. 
\[ \hat{f}^*(\omega) = c\hat{f}_{2o}(\omega) + (1 - c)\hat{f}^+(\omega) = c\hat{f}_{2o}(\omega) \]
Shrinkage correction to finite-sample positive definiteness

\[ \hat{\Gamma}_{\kappa,l} = TDT^t \] where \( D = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \) and \( T \) is orthogonal.
Shrinkage correction to finite-sample positive definiteness

\[ \hat{\Gamma}_{\kappa,l} = TDT^t \] where \( D = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \) and \( T \) is orthogonal.

Let \( \hat{\Gamma}_n^{2o} \) be the autocovariance matrix obtained via tapering with a 2nd order kernel, e.g., Parzen’s.
Shrinkage correction to finite-sample positive definiteness

\[ \hat{\Gamma}_{\kappa,l} = T D T^t \]

where \( D = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \) and \( T \) is orthogonal.

Let \( \hat{\Gamma}_{2o}^n \) be the autocovariance matrix obtained via tapering with a 2nd order kernel, e.g., Parzen’s.

\( \hat{\Gamma}_{2o}^n \) and \( \hat{\Gamma}_{\kappa,l} \) are both Toeplitz so approximately diagonalizable using the same orthogonal matrix \( T \), i.e.,
Shrinkage correction to finite-sample positive definiteness

\[ \hat{\Gamma}_{\kappa,l} = T D T^t \] where \( D = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \) and \( T \) is orthogonal.

Let \( \hat{\Gamma}^{2o}_n \) be the autocovariance matrix obtained via tapering with a 2nd order kernel, e.g., Parzen’s.

\( \hat{\Gamma}^{2o}_n \) and \( \hat{\Gamma}_{\kappa,l} \) are both Toeplitz so approximately diagonalizable using the same orthogonal matrix \( T \), i.e.,

\[ \hat{\Gamma}^{2o}_n \sim T D^{2o} T^t \] where \( D^{2o} = \text{diag}(\hat{\lambda}_1^{2o}, \ldots, \hat{\lambda}_n^{2o}) \)

Define the shrinkage estimate

\[ \hat{\Gamma}^*_{\kappa,l} = T D^* T^t, \]

where \( D^* = \text{diag}(\lambda^*_1, \ldots, \lambda^*_n) \) and \( \lambda^*_i = \hat{\lambda}_i \) if \( |\hat{\lambda}_i| > \) some threshold; else let \( \lambda^*_i = c \hat{\lambda}_i \), a shrunk form of the positive eigenvalue.
Shrinkage correction to finite-sample positive definiteness

\[ \hat{\Gamma}_{\kappa,l} = TDT^t \] where \( D = \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_n) \) and \( T \) is orthogonal.

Let \( \hat{\Gamma}_n^{2o} \) be the autocovariance matrix obtained via tapering with a 2nd order kernel, e.g., Parzen’s.

\( \hat{\Gamma}_n^{2o} \) and \( \hat{\Gamma}_{\kappa,l} \) are both Toeplitz so approximately diagonalizable using the same orthogonal matrix \( T \), i.e.,

\[ \hat{\Gamma}_n^{2o} \approx TD^{2o}T^t \] where \( D^{2o} = \text{diag}(\hat{\lambda}_1^{2o}, \ldots, \hat{\lambda}_n^{2o}) \)

Define the \textit{shrinkage} estimate

\[ \hat{\Gamma}^*_{\kappa,l} = TD^*T^t, \]

where \( D^* = \text{diag}(\lambda_1^*, \ldots, \lambda_n^*) \) and \( \lambda_i^* = \hat{\lambda}_i \) if \( |\hat{\lambda}_i| > \text{some threshold} \); else let \( \lambda_i^* = c\hat{\lambda}_i^{2o} \), a \textit{shrunk} form of the positive eigenvalue.
Empirical rule of picking $l$: (Politis, 2003) Define $\hat{\gamma}(k) = \hat{\gamma}_k / \hat{\gamma}_0$. Let $\hat{l}$ be the smallest positive integer such that

$$|\hat{\gamma}(\hat{l} + k)| < c \sqrt{\frac{\log n}{n}}$$

for $k = 1, \ldots, K_N$, where $K_N$ is very slowly increasing.
Let $\hat{l}$ be the smallest positive integer such that

$$|\hat{\varrho}(\hat{l} + k)| < c\sqrt{\frac{\log_{10} n}{n}}$$

for $k = 1, \ldots, K_N$; here $c = 2$ and $K_N = 5$. 
Asymptotics of Bandwidth Choice

Theorem

Under modest conditions,

1. If $\gamma_i = C i^{-d}$ for $i > i_0$, then

   $\hat{l} \overset{P}{\sim} \frac{A_1 n^{1/2d}}{(\log n)^{1/2d}}$

2. If $\gamma_i = C \theta^i$ for $i > i_0$ and $|\theta| < 1$, then

   $\hat{l} \overset{P}{\sim} A_2 \log n$

   where $A_2 = -1/\log |\theta|$.

3. If $\gamma_i = 0$ for all $k > B$, but $\gamma_B \neq 0$, then

   $\hat{l} = B + o_P(1)$. 
Simulations—autocovariance matrix estimation

Losses in max absolute row sum norm with $n = 250$;
[similar results in operator norm]

<table>
<thead>
<tr>
<th>$X_t = \epsilon_t + \phi \epsilon_{t-1}$</th>
<th>$\phi$</th>
<th>0.5</th>
<th>0.00 (0.00)</th>
<th>0.15 (0.05)</th>
<th>1.7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>0.95 (0.22)</td>
<td>0.27 (0.24)</td>
<td>2.7</td>
<td></td>
</tr>
</tbody>
</table>

| $X_t = \phi |X_{t-1}| + \epsilon_t$ | $\phi$ | 0.5 | 0.01 (0.10) | 0.67 (0.07) | 2.2 |
|------------------------|-------|-----|-------------|-------------|-----|
|                        | 0.9   | 4.43 (3.56) | 12.13 (6.31) | 14.3 |

<table>
<thead>
<tr>
<th>$X_t = \phi X_{t-1} + \epsilon_t$</th>
<th>$\phi$</th>
<th>0.5</th>
<th>1.22 (0.46)</th>
<th>1.18 (0.29)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.9</td>
<td>10.01 (5.18)</td>
<td>9.63 (3.71)</td>
<td>14.3</td>
<td></td>
</tr>
</tbody>
</table>

Inferior performance of WP mainly due to suboptimal bandwidth choice; $\infty$-Rect is only slightly worse than $\infty$-Trap when both bandwidths are chosen using the BC Rule.
Simulations—autocovariance matrix estimation

Losses in max absolute row sum norm with $n = 250$; [similar results in operator norm]

<table>
<thead>
<tr>
<th>$X_t = \epsilon_t + \phi \epsilon_{t-1}$</th>
<th>$\phi$</th>
<th>$l$</th>
<th>$\infty$-Trap</th>
<th>$\infty$-WP</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>0.95 (0.22)</td>
<td>0.27 (0.24)</td>
<td>2.7</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.00 (0.00)</td>
<td>0.15 (0.05)</td>
<td>1.7</td>
<td></td>
</tr>
<tr>
<td>$X_t = \phi</td>
<td>X_{t-1}</td>
<td>+ \epsilon_t$</td>
<td>0.5</td>
<td>0.01 (0.10)</td>
</tr>
<tr>
<td>0.9</td>
<td>4.43 (3.56)</td>
<td>12.13 (6.31)</td>
<td>14.3</td>
<td></td>
</tr>
<tr>
<td>$X_t = \phi X_{t-1} + \epsilon_t$</td>
<td>0.5</td>
<td>1.22 (0.46)</td>
<td>1.18 (0.29)</td>
<td></td>
</tr>
<tr>
<td>0.9</td>
<td>10.01 (5.18)</td>
<td>9.63 (3.71)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Inferior performance of WP mainly due to suboptimal bandwidth choice; $\infty$-Rect is only slightly worse than $\infty$-Trap when both bandwidths are chosen using the BC Rule.
Application I: Linear process bootstrap

Popular time series bootstraps include:

- Block bootstrap – resample blocks of $b$ consecutive observations
- AR-sieve – fit an AR($p$) model, bootstrap residuals in conjunction with the fitted model for bootstrap data
Why not an MA-sieve?

MA coefficients are estimated by

- Numerical optimization
- Innovations algorithm
Why not an MA-sieve?

MA coefficients are estimated by

- Numerical optimization
- Innovations algorithm

No closed-form estimator plus lack of efficiency.
Linear process bootstrap

- Center the data: \( \tilde{Y} = (Y_1, \ldots, Y_n) = (X_1, \ldots, X_n) - \bar{X} \)
- Standardize: \( \tilde{W} = (\hat{\Gamma}_c^{\epsilon, \kappa,l})^{-1/2} \tilde{Y} \)
- \( \tilde{Z} = (\tilde{W} - \bar{W})/\hat{\sigma}_W \)
- Generate \( Z_1^*, \ldots, Z_n^* \) by and i.i.d. bootstrap of \( Z_1, \ldots, Z_n \)
- \( \tilde{Y}^* = (\hat{\Gamma}_c^{\epsilon, \kappa,l})^{1/2} \tilde{Z}^* \)
Linear process bootstrap

- Center the data $\vec{Y} = (Y_1, \ldots, Y_n) = (X_1, \ldots, X_n) - \bar{X}$
- Standardize $\vec{W} = (\hat{\Gamma}_{\kappa,l}^{\epsilon})^{-1/2} \vec{Y}$
- $\vec{Z} = (\vec{W} - \bar{W})/\hat{\sigma}_W$
- Generate $Z_1^*, \ldots, Z_n^*$ by an i.i.d. bootstrap of $Z_1, \ldots, Z_n$
- $\vec{Y}^* = (\hat{\Gamma}_{\kappa,l}^{\epsilon})^{1/2} \vec{Z}^*$
Linear process bootstrap

- Center the data \( \bar{Y} = (Y_1, \ldots, Y_n) = (X_1, \ldots, X_n) - \bar{X} \)
- Standardize \( \bar{W} = (\hat{\Gamma}_{\kappa, l})^{-1/2} \bar{Y} \)
- \( \bar{Z} = (\bar{W} - \bar{W})/\hat{\sigma}_W \)
- Generate \( Z_1^*, \ldots, Z_n^* \) by and i.i.d. bootstrap of \( Z_1, \ldots, Z_n \)
- \( \bar{Y}^* = (\hat{\Gamma}_{\kappa, l})^{1/2} \bar{Z}^* \)
Linear process bootstrap

- Center the data \( \bar{Y} = (Y_1, \ldots, Y_n) = (X_1, \ldots, X_n) - \bar{X} \)
- Standardize \( \bar{W} = (\hat{\Gamma}_{\kappa, l})^{-1/2} \bar{Y} \)
- \( \bar{Z} = (\bar{W} - \bar{W})/\hat{\sigma}_W \)
- Generate \( Z_1^*, \ldots, Z_n^* \) by and i.i.d. bootstrap of \( Z_1, \ldots, Z_n \)
- \( \bar{Y}^* = (\hat{\Gamma}_{\kappa, l})^{1/2} \bar{Z}^* \)
Linear process bootstrap

- Center the data \( \tilde{Y} = (Y_1, \ldots, Y_n) = (X_1, \ldots, X_n) - \bar{X} \)
- Standardize \( \tilde{W} = (\hat{\Gamma}_{\epsilon, l}^{\kappa, l})^{-1/2} \tilde{Y} \)
- \( \tilde{Z} = (\tilde{W} - \bar{W})/\hat{\sigma}_W \)
- Generate \( Z_1^*, \ldots, Z_n^* \) by and i.i.d. bootstrap of \( Z_1, \ldots, Z_n \)
- \( \tilde{Y}^* = (\hat{\Gamma}_{\epsilon, l}^{\kappa, l})^{1/2} \tilde{Z}^* \)
Bootstrap for the mean

Theorem (McMurry and Politis, 2011)

Let $E[X_i] = \mu$. Under the conditions of the previous theorems

$$\sup_x \left| P \left[ n^{1/2} (\bar{X} - \mu) \leq x \right] - P^* \left[ n^{1/2} \bar{Y}^* \leq x \right] \right| \rightarrow_P 0.$$
Bootstrap for the mean

**Theorem (McMurry and Politis, 2011)**

Let $E[X_i] = \mu$. Under the conditions of the previous theorems

\[
\sup_x \left| P \left[ n^{1/2} (\bar{X} - \mu) \leq x \right] - P^* \left[ n^{1/2} \bar{Y}^* \leq x \right] \right| \rightarrow_P 0.
\]

The \{X_t\} process can be **nonlinear**!
Bootstrap for the mean

Theorem (McMurry and Politis, 2011)

Let $E[X_i] = \mu$. Under the conditions of the previous theorems

$$\sup_x \left| P \left[ n^{1/2}(\bar{X} - \mu) \leq x \right] - P^* \left[ n^{1/2} \bar{Y}^* \leq x \right] \right| \to_P 0.$$ 

The $\{X_t\}$ process can be nonlinear!

Compare to AR-sieve result by Kreiss, Paparoditis and Politis (AoS, 2011)
Bootstrap confidence intervals for the mean

\( n = 250, \ R = 1000, \) each repeated 1000 times

<table>
<thead>
<tr>
<th>Model</th>
<th>( \phi )</th>
<th>Coverage</th>
<th>Avg. Length</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LPB</td>
<td>Blk</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \phi</td>
<td>X_{t-1}</td>
<td>+ \epsilon_t )</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \phi X_{t-1}</td>
<td>+ \epsilon_t )</td>
<td>0.5</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \epsilon_t</td>
<td>+ \phi \epsilon_{t-1} )</td>
<td>0.9</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \epsilon_t</td>
<td>+ \phi \epsilon_{t-1} )</td>
<td>0.1</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \epsilon_t</td>
<td>+ \phi \epsilon_{t-1} )</td>
<td>0.5</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \epsilon_t</td>
<td>+ \phi \epsilon_{t-1} )</td>
<td>0.9</td>
<td>0.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \epsilon_t</td>
<td>+ \phi \epsilon_{t-1} )</td>
<td>0.1</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \epsilon_t</td>
<td>+ \phi \epsilon_{t-1} )</td>
<td>0.5</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( X_t = \epsilon_t</td>
<td>+ \phi \epsilon_{t-1} )</td>
<td>0.9</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Bootstrap for the spectral density

Recall spectral density: $f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-i\omega}$
Bootstrap for the spectral density

Recall spectral density: \( f(\omega) = (2\pi)^{-1} \sum h \gamma_h e^{-ih\omega} \)

**Flat-top** spectral density estimate:
\( \hat{f}(\omega) = (2\pi)^{-1} \sum h \kappa(h/m) \hat{\gamma}_h e^{-ih\omega} \)
Bootstrap for the spectral density

Recall spectral density: \( f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-ih\omega} \)

**Flat-top** spectral density estimate:
\[ \hat{f}(\omega) = (2\pi)^{-1} \sum_h \kappa(h/m) \hat{\gamma}_h e^{-ih\omega} \]

**2nd-order** spectral density estimate:
\[ \hat{f}^{2o}(\omega) = (2\pi)^{-1} \sum_h \kappa^{2o}(h/m) \hat{\gamma}_h e^{-ih\omega}; \text{ optimal } m \sim n^{1/5}. \]
Bootstrap for the spectral density

Recall spectral density: \( f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-ih\omega} \)

**Flat-top** spectral density estimate:
\( \hat{f}(\omega) = (2\pi)^{-1} \sum_h \kappa(h/m) \hat{\gamma}_h e^{-ih\omega} \)

**2nd-order** spectral density estimate:
\( \hat{f}^{2o}(\omega) = (2\pi)^{-1} \sum_h \kappa^{2o}(h/m) \hat{\gamma}_h e^{-ih\omega}; \text{ optimal } m \sim n^{1/5}. \)

**Theorem (Jentsch and Politis, 2013)**

Let \( \hat{f}^*(\omega) \) and \( \hat{f}^{2o,*}(\omega) \) be computed from an LPB resample. Then,

\[
\sup_x \left| \mathbb{P} \left[ n^{2/5} (\hat{f}^{2o} - f) \leq x \right] - \mathbb{P}^* \left[ n^{2/5} (\hat{f}^{2o,*} - \hat{f}) \leq x \right] \right| \to_P 0
\]

for any (fixed) \( \omega \).
Bootstrap for the spectral density

Recall spectral density: \( f(\omega) = (2\pi)^{-1} \sum_h \gamma_h e^{-ih\omega} \)

**Flat-top** spectral density estimate:
\( \hat{f}(\omega) = (2\pi)^{-1} \sum_h \kappa(h/m) \hat{\gamma}_h e^{-ih\omega} \)

**2nd-order** spectral density estimate:
\( \hat{f}^{2o}(\omega) = (2\pi)^{-1} \sum_h \kappa^{2o}(h/m) \hat{\gamma}_h e^{-ih\omega}; \text{ optimal } m \sim n^{1/5} \).

**Theorem (Jentsch and Politis, 2013)**

Let \( \hat{f}^*(\omega) \) and \( \hat{f}^{2o,*}(\omega) \) be computed from an LPB resample. Then,

\[
\sup_x \left| P \left[ n^{2/5} (\hat{f}^{2o} - f) \leq x \right] - P^* \left[ n^{2/5} (\hat{f}^{2o,*} - \hat{f}) \leq x \right] \right| \rightarrow_P 0
\]

for any (fixed) \( \omega \). **Centering by flat-top \( \hat{f}(\omega) \) takes care of the bias.**
Multivariate and high-dimensional LPB

Consider the \( \{X_t\} \) process being \( d \)-dimensional.
Multivariate and high-dimensional LPB

Consider the $\{X_t\}$ process being $d$-dimensional.

LPB idea is straightforward but the $\Gamma_n$ matrix is very complicated!
Multivariate and high-dimensional LPB

Consider the \( \{ X_t \} \) process being d-dimensional.

LPB idea is straightforward but the \( \Gamma_n \) matrix is very complicated!

Sample mean result valid even if \( d \to \infty \) as \( n \to \infty \).
Multivariate and high-dimensional LPB

Consider the \( \{X_t\} \) process being \( d \)-dimensional.

LPB idea is straightforward but the \( \Gamma_n \) matrix is very complicated!

Sample mean result valid even if \( d \to \infty \) as \( n \to \infty \).

**Theorem (Jentsch and Politis, 2013)**

\[
L \text{et } b = b(d) \text{ be a } d\text{-dimensional vector of constants, i.e., } b(d) = (b_1(d), \ldots, b_d(d))^t.
\]
Multivariate and high-dimensional LPB

Consider the \( \{X_t\} \) process being \( d \)-dimensional.

LPB idea is straightforward but the \( \Gamma_n \) matrix is very complicated! Sample mean result valid even if \( d \to \infty \) as \( n \to \infty \).

**Theorem (Jentsch and Politis, 2013)**

Let \( b = b(d) \) be a \( d \)-dimensional vector of constants, i.e., \( b(d) = (b_1(d), \ldots, b_d(d))^t \). If \( \sup_{d \in \mathbb{N}} \sum_{j=1}^{d} b_j^2(d) < \infty \), then

\[
\sup_x \left| \mathbb{P} \left[ n^{1/2} b^t (\bar{X} - \mu) \leq x \right] - \mathbb{P}^* \left[ n^{1/2} b^t \bar{Y}^* \leq x \right] \right| \to_P 0
\]
Multivariate and high-dimensional LPB

Consider the \{X_t\} process being \(d\)-dimensional.

LPB idea is straightforward but the \(\Gamma_n\) matrix is very complicated!

Sample mean result valid even if \(d \to \infty\) as \(n \to \infty\).

**Theorem (Jentsch and Politis, 2013)**

Let \(b = b(d)\) be a \(d\)-dimensional vector of constants, i.e., \(b(d) = (b_1(d), \ldots, b_d(d))^t\). If \(\sup_{d \in \mathbb{N}} \sum_{j=1}^{d} b_j^2(d) < \infty\), then

\[
\sup_x \left| \mathbb{P} \left[ n^{1/2} b^t (\bar{X} - \mu) \leq x \right] - \mathbb{P}^* \left[ n^{1/2} b^t \bar{Y}^* \leq x \right] \right| \to_P 0
\]

Also need \(\log^2(dn)d^2l = o(\sqrt{n})\) and other regularity conditions.
Multivariate and high-dimensional LPB

Consider the \( \{X_t\} \) process being \( d \)-dimensional.

LPB idea is straightforward but the \( \Gamma_n \) matrix is very complicated!

Sample mean result valid even if \( d \to \infty \) as \( n \to \infty \).

**Theorem (Jentsch and Politis, 2013)**

Let \( b = b(d) \) be a \( d \)-dimensional vector of constants, i.e., \( b(d) = (b_1(d), \ldots, b_d(d))^t \). If \( \sup_{d \in \mathbb{N}} \sum_{j=1}^{d} b_j^2(d) < \infty \), then

\[
\sup_x |P \left[ n^{1/2} b^t (\bar{X} - \mu) \leq x \right] - P^* \left[ n^{1/2} b^t \bar{Y}^* \leq x \right] | \to_p 0
\]

Also need \( \log^2 (dn) d^2 l = o(\sqrt{n}) \) and other regularity conditions.

So if \( l \sim n^\beta \) for \( \beta \in (0, \frac{1}{2}) \), then we need \( d = o(n^{\frac{1}{2} \left( \frac{1}{2} - \beta \right) \log^{-1} n}) \).
Application II: Linear prediction using the complete history

**PROBLEM:** Predict \(X_{n+1}\) by a linear combination of \(X_1, \ldots, X_n\)

\[
\hat{X}_{n+1} = \phi_{n1} X_n + \phi_{n2} X_{n-1} + \cdots + \phi_{nn} X_1,
\]

where the coefficients \(\phi_{ni}\) are given by Yule-Walker equations

\[
\phi_n \equiv \begin{bmatrix} \phi_{n1} \\ \vdots \\ \phi_{nn} \end{bmatrix} = \Gamma_n^{-1} \gamma_n.
\]

\(\Gamma_n = [\gamma_n(i-j)]_{n \times n} \) is the covariance matrix of \(X_1, \ldots, X_n\), and \(\gamma_n = [\gamma_n(1), \ldots, \gamma_n(n)]'\) is the vector of covariances at lags 1, \ldots, \(n\).
**Application II: Linear prediction using the complete history**

**PROBLEM:** Predict $X_{n+1}$ by a linear combination of $X_1, \ldots, X_n$

Assume $\mu = 0$. Then the MSE-optimal predictor is

$$
\tilde{X}_{n+1} = \phi_{n1}X_n + \phi_{n2}X_{n-1} + \ldots + \phi_{nn}X_1,
$$

where the coefficients $\phi_{ni}$ are given by Yule-Walker equations

$$
\phi_n \equiv \begin{bmatrix}
\phi_{11} \\
\vdots \\
\phi_{nn}
\end{bmatrix} = \Gamma_n^{-1} \underline{\gamma}_n.
$$

$\Gamma_n = [\gamma(i-j)]_{i,j=1}^n$ is the covariance matrix of $X_1, \ldots, X_n$, and $\underline{\gamma}_n = [\gamma(1), \ldots, \gamma(n)]'$ is the vector of covariances at lags $1, \ldots, n$. 
The MSE-optimal predictor $\tilde{X}_{n+1}$ is unattainable because $\phi_{n1}, \ldots, \phi_{nn}$ are unknown and must be estimated.
AR–based prediction: finite memory

- The MSE-optimal predictor $\tilde{X}_{n+1}$ is unattainable because $\phi_{n1}, \ldots, \phi_{nn}$ are unknown and must be estimated.
- Typical way out: fit an AR($p$) model with $p$ selected by AIC.
AR–based prediction: finite memory

- The MSE-optimal predictor $\tilde{X}_{n+1}$ is unattainable because $\phi_{n1}, \ldots, \phi_{nn}$ are unknown and must be estimated.
- Typical way out: fit an AR($p$) model with $p$ selected by AIC.
- The resulting predictor is

$$\hat{X}_{n+1}^{AR} = \hat{\phi}_1 X_n + \hat{\phi}_2 X_{n-1} + \ldots + \hat{\phi}_p X_{n-p+1}.$$  

- Finite memory predictor—only uses the last $p$ data points.
Linear prediction using the complete data history

- Under an AR–model, $\Gamma_n$ is determined by $\gamma(1), \ldots, \gamma(p)$
Linear prediction using the complete data history

- Under an AR–model, $\Gamma_n$ is determined by $\gamma(1), \ldots, \gamma(p)$
- Adopting an AR–model is a way out in terms of estimating the $n \times n$ matrix $\Gamma_n$ via estimating a small number of parameters
Under an AR–model, $\Gamma_n$ is determined by $\gamma(1), \ldots, \gamma(p)$.

Adopting an AR–model is a way out in terms of estimating the $n \times n$ matrix $\Gamma_n$ via estimating a small number of parameters.

But we now have consistent estimates of $\Gamma_n$ and its inverse.
Linear prediction using the complete data history

- Under an AR–model, $\Gamma_n$ is determined by $\gamma(1), \ldots, \gamma(p)$
- Adopting an AR–model is a way out in terms of estimating the $n \times n$ matrix $\Gamma_n$ via estimating a small number of parameters
- But we now have consistent estimates of $\Gamma_n$ and its inverse
- USE THEM FOR PREDICTION!
Linear prediction using the complete data history

MSE-optimal predictor: \( \tilde{X}_{n+1} = \phi_1 X_n + \phi_2 X_{n-1} + \ldots + \phi_n X_1 \)
Linear prediction using the complete data history

MSE-optimal predictor: \( \hat{X}_{n+1} = \phi_{n1} X_n + \phi_{n2} X_{n-1} + \ldots + \phi_{nn} X_1 \)

- Estimate the prediction coefficients by \( \hat{\phi}_n = \hat{\Gamma}_n^{-1} \hat{\gamma}_n \)
- \( \hat{\Gamma}_n \) is one of the positive definite versions of the banded and tapered estimate
- \( \hat{\gamma}_n \) is a tapered estimate of the autocovariance function with \( i \)’th coordinate \( \kappa(i/l)\hat{\gamma}(i) \)—or just take the first row of \( \hat{\Gamma}_n \) (leaving out the first element)
MSE-optimal predictor: \( \hat{X}_{n+1} = \phi_{n1}X_n + \phi_{n2}X_{n-1} + \ldots + \phi_{nn}X_1 \)

- Estimate the prediction coefficients by \( \hat{\phi}_n = \hat{\Gamma}_n^{-1} \hat{\gamma}_n \)
- \( \hat{\Gamma}_n \) is one of the positive definite versions of the banded and tapered estimate
- \( \hat{\gamma}_n \) is a tapered estimate of the autocovariance function with \( i \)'th coordinate \( \kappa(i/l)\hat{\gamma}(i) \) —or just take the first row of \( \hat{\Gamma}_n \) (leaving out the first element)

The resulting **FSO (finite-sample optimal)** predictor

\[
\hat{X}_{n+1} = \hat{\phi}_{n1}X_n + \hat{\phi}_{n2}X_{n-1} + \ldots + \hat{\phi}_{nn}X_1, \quad (3)
\]

incorporates information from the complete process history, and will be shown to converge to the optimal predictor as \( n \to \infty \).
Linear prediction using the complete data history

Let $|| \cdot ||_2$ denote the matrix 2-norm, and let $|\vec{v}|_2$ denote the vector 2-norm of a vector $\vec{v}$. Then

Lemma (McMurry and Politis (2014))

Under the assumptions of Theorem 1

$$|\hat{\phi}_n - \phi_n|_2 = O_p(r_n)$$

where $r_n = ln^{-1/2} + \sum_{i=1}^{\infty} |\gamma(i)|$. 
Linear prediction using the complete data history

Let $\| \cdot \|_2$ denote the matrix 2-norm, and let $|\vec{v}|_2$ denote the vector 2-norm of a vector $\vec{v}$. Then

Lemma (McMurry and Politis (2014))

Under the assumptions of Theorem 1

$$|\hat{\phi}_n - \phi_n|_2 = O_p(r_n)$$

where $r_n = \ln^{-1/2} + \sum_{i=1}^{\infty} |\gamma(i)|$.

This was previously thought impossible! Bickel and Gel (JRSS, B 2011): “given $n$ observations, it is impossible to estimate $n$ AR parameters sufficiently well for prediction purposes.”
Linear prediction using the complete data history

Convergence of the coefficients, i.e., $|\hat{\phi}_n - \phi_n|_2 = O_p(r_n)$, is not enough to ensure

$$\hat{X}_{n+1} - \tilde{X}_{n+1} = o_p(1)$$

(4)

where $\tilde{X}_{n+1}$ is the ‘oracle’ best possible predictor.
Convergence of the coefficients, i.e., $|\hat{\phi}_n - \phi_n|_2 = O_p(r_n)$, is not enough to ensure

$$\hat{X}_{n+1} - \tilde{X}_{n+1} = o_p(1)$$  \hfill (4)

where $\tilde{X}_{n+1}$ is the ‘oracle’ best possible predictor.

To show (4), we can show that for some appropriate sequence $k_n$

- $\hat{\phi}_1, \ldots, \hat{\phi}_{k_n}$ contain most of the information; and
- $\hat{\phi}_{k_{n+1}}, \ldots, \hat{\phi}_n$ will be small.
Linear prediction using the complete data history

Convergence of the coefficients, i.e., $|\hat{\phi}_n - \phi_n|_2 = O_p(r_n)$, is not enough to ensure

$$\hat{X}_{n+1} - \tilde{X}_{n+1} = o_p(1)$$  \hspace{1cm} (4)

where $\tilde{X}_{n+1}$ is the ‘oracle’ best possible predictor.

To show (4), we can show that for some appropriate sequence $k_n$

- $\hat{\phi}_1, \ldots, \hat{\phi}_{k_n}$ contain most of the information; and
- $\hat{\phi}_{k_n+1}, \ldots, \hat{\phi}_n$ will be small.

We need some additional conditions to ensure the above.
Theorem 2

Theorem (McMurry and Politis (2014))

Assume the conditions of Theorem 1, and let $l$ grow at the optimal rate for estimating $\Gamma_n$. If eventually

- $|\phi_i| \leq C_\phi i^{-k}$ where $k > 3/2$
- $|\gamma_i| \leq C_\gamma i^{-k}$ where $k > 2$

Then, eq. (4) is true, i.e.,

$$|\hat{X}_{n+1} - \tilde{X}_{n+1}| = o_p(1).$$
Theorem 2

Theorem (McMurry and Politis (2014))

Assume the conditions of Theorem 1, and let $l$ grow at the optimal rate for estimating $\Gamma_n$. If eventually

- $|\phi_i| \leq C_\phi i^{-k}$ where $k > 3/2$
- $|\gamma_i| \leq C_\gamma i^{-k}$ where $k > 2$

Then, eq. (4) is true, i.e.,

$$|\hat{X}_{n+1} - \tilde{X}_{n+1}| = o_p(1).$$

FSO $\hat{X}_{n+1}$ is asymptotically equivalent to the ‘oracle’ best predictor!
If $\hat{\Gamma}_n$ is banded with eigenvalues bounded away from 0,

$$|[\hat{\Gamma}_n^{-1}]_{ij}| \leq C_2 \lambda|i-j|/l.$$ 

(Demko, et al. '84)
If $\hat{\Gamma}_n$ is banded with eigenvalues bounded away from 0,

$$\left| (\hat{\Gamma}_n^{-1})_{ij} \right| \leq C_2 \lambda |i-j|/l.$$  

(Demko, et al. '84)

**Consequence:** Entries of $\hat{\phi}(n)$ decay exponentially fast with probability tending to 1.
Partial Sample Optimal Predictors

- Use an AR($p_n$) prediction with arbitrary $p_n \leq n$.
- Regularize $\hat{\Gamma}_{p_n}$ by banding and tapering.
- Similar idea by Bickel and Gel (2011) but with $p_n = o(n)$. 
Partial Sample Optimal Predictors

- Use an AR($p_n$) prediction with arbitrary $p_n \leq n$.
- Regularize $\hat{\Gamma}_{p_n}$ by banding and tapering.
- Similar idea by Bickel and Gel (2011) but with $p_n = o(n)$.

Partial Sample Optimal (PSO) Predictor has coefficients given by

$$\hat{\phi}(p_n) = \hat{\Gamma}_{p_n}^{-1} \hat{\gamma}(p_n).$$
Partial Sample Optimal Predictors

- Use an AR($p_n$) prediction with arbitrary $p_n \leq n$.
- Regularize $\hat{\Gamma}_{p_n}$ by banding and tapering.
- Similar idea by Bickel and Gel (2011) but with $p_n = o(n)$.

Partial Sample Optimal (PSO) Predictor has coefficients given by

$$\hat{\phi}(p_n) = \hat{\Gamma}_{p_n}^{-1} \hat{\gamma}(p_n).$$

Corollary (McMurry and Politis (2014))

Under the conditions of last Theorem, for any $p_n \leq n$

$$|\hat{\phi}(p_n) - \phi(p_n)|_2 = O_p(r_n) \text{ and } |\hat{X}^{p_n}_{n+1} - \tilde{X}^{p_n}_{n+1}| = o_p(1).$$
Simulations AR1

<table>
<thead>
<tr>
<th>$\phi$</th>
<th>FSO-Th</th>
<th>FSO-Sh</th>
<th>PSO-Th</th>
<th>AR</th>
<th>BG</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.9</td>
<td>1.0998</td>
<td>1.1061</td>
<td>1.0217</td>
<td>1.0006</td>
<td>1.0553</td>
<td>0.9913</td>
</tr>
<tr>
<td>−0.5</td>
<td>1.0430</td>
<td>1.0304</td>
<td>1.0431</td>
<td>1.0207</td>
<td>1.0558</td>
<td>1.0043</td>
</tr>
<tr>
<td>−0.1</td>
<td>0.9996</td>
<td>0.9995</td>
<td>0.9996</td>
<td>0.9981</td>
<td>1.0355</td>
<td>0.9953</td>
</tr>
<tr>
<td>0.1</td>
<td>1.0043</td>
<td>1.0044</td>
<td>1.0043</td>
<td>1.0023</td>
<td>1.0236</td>
<td>1.0009</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0522</td>
<td>1.0372</td>
<td>1.0513</td>
<td>1.0141</td>
<td>1.0314</td>
<td>1.0005</td>
</tr>
<tr>
<td>0.9</td>
<td>1.1304</td>
<td>1.1027</td>
<td>1.0519</td>
<td>1.0179</td>
<td>1.1143</td>
<td>1.0101</td>
</tr>
</tbody>
</table>

**Table**: Root mean square prediction errors for AR(1) processes.

AR uses $p_{AIC}$; PSO uses $p = \sqrt{np_{AIC}}$. BG gets $p$ via cross-validation.
Simulations MA1

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>FSO-Th</th>
<th>FSO-Sh</th>
<th>PSO-Th</th>
<th>AR</th>
<th>BG</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td>−0.9</td>
<td>1.0584</td>
<td>1.0568</td>
<td>1.0586</td>
<td>1.0172</td>
<td>1.0202</td>
<td>0.9815</td>
</tr>
<tr>
<td>−0.5</td>
<td>1.0511</td>
<td>1.0464</td>
<td>1.0511</td>
<td>1.0528</td>
<td>1.0876</td>
<td>1.0334</td>
</tr>
<tr>
<td>−0.1</td>
<td>1.0224</td>
<td>1.0223</td>
<td>1.0224</td>
<td>1.0251</td>
<td>1.0522</td>
<td>1.0188</td>
</tr>
<tr>
<td>0.1</td>
<td>1.0269</td>
<td>1.0268</td>
<td>1.0269</td>
<td>1.0257</td>
<td>1.0536</td>
<td>1.0248</td>
</tr>
<tr>
<td>0.5</td>
<td>1.0298</td>
<td>1.0266</td>
<td>1.0303</td>
<td>1.0333</td>
<td>1.0714</td>
<td>1.0216</td>
</tr>
<tr>
<td>0.9</td>
<td><strong>1.0757</strong></td>
<td>1.0831</td>
<td><strong>1.0757</strong></td>
<td>1.0769</td>
<td>1.0783</td>
<td>1.0158</td>
</tr>
</tbody>
</table>

**Table**: Root mean square prediction errors for MA(1) processes.