Large Sample Approximations for Variance-Covariance Matrices of High-Dimensional Time Series

Ansgar Steland

RWTH Aachen University

(with R. v. Sachs, Universite catholique Louvain, Belgium)

Workshop on Data Science 2015, Aachen

Outline

Introduction

Big Data: Sources

- Internet, sensors, cameras, simulations, ...

Aims:

- Extract information, 'knowledge'
- Build predictive models
- Simulate scenarios
- Separate 'structure' and 'noise'
- ...

Complex and high–dimensional data ('big data')

- Functional Data (discretely observed processes, measurement curves, ...)
- Image data, video data
- High–dimensional (vector) correlated data
- Time series structure (temporal correlations)

Overview: Stochastics Group

Overview of Activities Related to Data Science

Questions we address:

- How to monitor (possibly high–dimensional) data streams?
- How to monitor image streams?
- How to analyze spatial-temporal **correlated** data?
- How to analyze **high–dimensional highly-correlated** vector time series? $\rightarrow$ focus of this talk.

Image Data (I)

Example: Preprocessing & analysis of electroluminescence images of solar panels.

PVStatLab–Project: **PV-Scan** (with TÜV Rheinland, ISC Konstanz, Wrocław UoT, BMWi funded),

http://www.pvstatlab.rwth-aachen.de

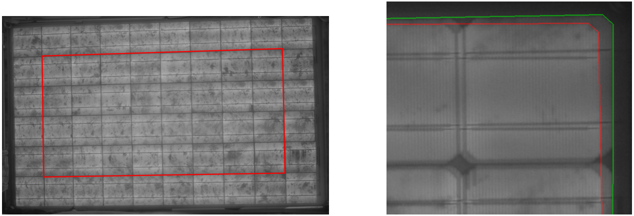

Figure: Example: Preprocessing using robust regression

Image Data (II)

Example: Image Analysis. Image as a random field $\{\xi_{ij}\}$.

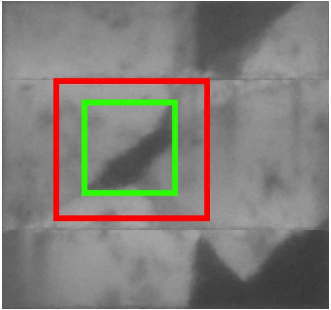

Figure: Regions C and D

$$H_0 : \mathrm{E}(\xi_{ij}) = \mathrm{E}(\xi_{uv}),$$
$$\text{for all } (i,j) \in C,\ (u,v) \in D$$

$$H_1 : \mathrm{E}(\xi_{ij}) \neq \mathrm{E}(\xi_{uv}),$$
$$\text{for all } (i,j) \in C,\ (u,v) \in D$$

Aim: Asymptotic significance test taking into account spatial correlations (ongoing work), detection of defects.

Overview: Stochastics Group

Recent Related Publications:

1. Sovetkin, E. and Steland, A. (2015). On statistical preprocessing of PV field image data using robust regression. In: N. E. Mastorakis, A. Ding & M. V. Shitikova, Advances in Mathematics and Statistical Sciences, Vol. 40.

2. Steland, A. (2015). Vertically weighted averages in Hilbert spaces and applications to imaging: Fixed sample asymptotics and efficient sequential two-stage estimation, *Sequential Analysis*, 34 (3), 295-323.

Overview: Stochastics Group

Monitoring of Multivariate Data and Image Streams

- Aim: Nonparametric detection of changes
- Observe discretely sampled function representing the true signal(s) resp. image(s)
- Approaches: Hilbert-space valued r.e., random fields, Shannon/Whittaker

Recent Related publications:

1. Prause, A. and Steland, A. (2015). Detecting changes in spatial-temporal image data based on quadratic forms. In: Stochastic Models, Statistics and Their Applications, 139-147.

2. Prause, A. and Steland, A. (2015). Sequential detection of three-dimensional signals under dependent noise, *submitted*.

3. Prause, A. (2015). Ph. D. thesis (finished)

Large–Sample Approximations of High–Dimensional Vector Time Series

- project with R. v. Sachs (since 11/2013)
- new DFG project just started

Setting:

Massive data set with observations on a large number of variables (features).

Focus: Analyze **Dependencies**

Introduction

High-dimensional variance-covariance matrices play a crucial role in those areas, since they provide information on the dependence of the coordinates (2*nd* order).

The sample covariance matrix is regarded a poor estimator, since it is not consistent w.r.t. to the operator norm if the dimension is larger than the sample size ($d/n \to c > 0$).

Previous works: Banding/tapering (Bickel & Levina, 2008), Thresholding (Chen at al., 2013), Shrinkage (Böhm and v. Sachs, 2009), ...

Introduction

Basic problem: Observe a large number, $d = d_n$, of variables, n repetitions (over time).

Preliminary data analyses (preprocessing):
Frequently, e.g. by preprocessing methods, one may classify the variables in (a small number of) groups, such that

- the within-group correlation is high but
- the between-group correlation is low/negligible.

We are faced with the problem to model and analyze high-dimensional data for highly correlated variables.

Projection-Based Analysis

Observe $d = d_n$ time series

$$Y_i^{(\nu)}, \dots, Y_i^{(\nu)}, \qquad \nu = 1, \dots, d,\ 1 \leq i \leq n,$$

This means, we are given a vector time series of length n,

$$\mathbf{Y}_{ni} = (Y_i^{(1)}, \dots, Y_i^{(d_n)})', \qquad 1 \leq i \leq n,$$

of dimension d_n, constituting the $(n \times d_n)$–dimensional data matrix

$$\mathcal{Y}_n = \left(Y_i^{(j)} \right)_{1 \leq i \leq n, 1 \leq j \leq d_n}.$$

We focus on second moments and thus assume $E(Y_i^{(j)}) = 0$ for all $i = 1, \dots, n$ and $j = 1, \dots, d_n$.

Projection-Based Analysis

Assume for a moment that $\mathbf{Y}_{n1}, \ldots, \mathbf{Y}_{nn}$ is stationary. Generic copy:

$$\mathbf{Y}_n = (Y^{(1)}, \ldots, Y^{(d_n)})'$$

Unknown $(d_n \times d_n)$-dimensional sample variance–covariance matrix

$$\mathbf{\Sigma}_n = E(\mathbf{Y}_n \mathbf{Y}_n') = \left(E(Y^{(\nu)} Y^{(\mu)}) \right)_{1 \le \nu, \mu \le d_n}$$

Sample variance-covariance matrix

$$\widehat{\mathbf{\Sigma}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_{ni} \mathbf{Y}_{ni}' = \frac{1}{n} \mathcal{Y}_n' \mathcal{Y}_n \qquad (1)$$

Unpleasant properties for $d_n >> n$, when studied as a matrix-valued estimator of $\mathbf{\Sigma}_n$, i.e. in dimension $d_n \times d_n$

Projection-Based Analysis

But typically, one is interested in a (set of) linear combination(s) $\mathbf{w}_n'\mathbf{Y}_n$ of the coordinates. Consider projections

$$T_n = \mathbf{w}_n'\mathbf{Y}_n$$

for weighting vectors

$$\mathbf{w}_n = (w_1, \ldots, w_{d_n})', \qquad n \geq 1,$$

of weights $w_j = w_{d_n j}$, not necessarily non-negative, with

$$\sup_{n\in\mathbb{N}} \|\mathbf{w}_n\|_{\ell_1} = \sup_{n\in\mathbb{N}} \sum_{\nu=1}^{d_n} |w_j| < \infty \tag{2}$$

Amongst others (later), such projections allow to study single covariances between coordinates.

Projection-Based Analysis

The projection $\mathbf{w}_n' \mathbf{Y}_n$ has variance $\mathbf{w}_n' \mathbf{\Sigma}_n \mathbf{w}_n$. Canonical estimator

$$\widehat{\mathrm{Var}}(\mathbf{w}_n' \mathbf{Y}_n) = \mathbf{w}_n' \widehat{\mathbf{\Sigma}}_n \mathbf{w}_n$$

behaves well for weighting vectors which select a finite number of coordinates.

Change-point problem: Test for a change in the variance of such a projection,

$$\sigma_n^2(i) = \mathrm{Var}(\mathbf{w}_n' \mathbf{Y}_{ni}), \qquad 1 \le i \le n.$$

as a consequence of a change of the variance-covariance matrix $\mathbf{\Sigma}_n$ in a high-dimensional setting.

Projection-Based Analysis

To proceed, let us more generally consider the quadratic form

$$Q_n(\mathbf{v}_n, \mathbf{w}_n) = \mathbf{v}_n' \boldsymbol{\Sigma}_n \mathbf{w}_n$$

for such weighting vectors $\mathbf{v}_n$ and $\mathbf{w}_n$.

Remark: Observe that even for $\boldsymbol{\Sigma}_n = \sigma \mathbf{1}\mathbf{1}'$ we have

$$|Q_n(\mathbf{v}_n, \mathbf{w}_n)| = \sigma |\mathbf{v}_n' \mathbf{1}\mathbf{1}' \mathbf{w}_n| = \sigma | \sum_i v_{ni} \sum_i |w_{ni}| \le \sigma \|\mathbf{v}_n\|_{\ell_1} \|\mathbf{w}_n\|_{\ell_1}.$$

So, the ℓ_1 condition is a natural one and ensures that even full covariance matrices are not mapped to ∞.

Framework and Assumptions

Model: The coordinates are linear processes

$$Y_k^{(\nu)} = Y_{nk}^{(\nu)} = \sum_{j=0}^{\infty} c_{nj}^{(\nu)} \epsilon_{k-j}, \qquad k = 1, \dots, n,$$

for coefficients $\{c_{nj}^{(\nu)} : j \in \mathbb{N}_0\}$, $\nu = 1, \dots, d_n$, and mean zero independent r.v.s. $\{\epsilon_k\}$ with

$$E|\epsilon_k|^{4+\delta} < \infty$$

for some $\delta > 0$.

Assumption A: The sequences $\{c_{nj}^{(\nu)} : j \in \mathbb{N}_0\}$ satisfy

$$\sup_{n \in \mathbb{N}} \max_{1 \le \nu \le d_n} |c_{nj}^{(\nu)}|^2 << j^{-3/2-\theta/2} \tag{3}$$

for some $0 < \theta < 1/2$.

Strong Approximation

Define

$$\widehat{\boldsymbol{\Sigma}}_{nk} = \left(\sum_{i=1}^{k} Y_i^{(\nu)} Y_i^{(\mu)} \right)_{1 \le \nu, \mu \le d_n}, \qquad (4)$$

$$\boldsymbol{\Sigma}_{nk} = \left(\sum_{i=1}^{k} E Y_i^{(\nu)} Y_i^{(\mu)} \right)_{1 \le \nu, \mu \le d_n}, \qquad (5)$$

for $n, k \geq 1$. To be precise, our results shall deal with

$$D_{nk} = \mathbf{v}_n' (\widehat{\boldsymbol{\Sigma}}_{nk} - \boldsymbol{\Sigma}_{nk}) \mathbf{w}_n, \qquad n, k \geq 1,$$

and the associated càdlàg processes

$$\mathcal{D}_n(t) = \mathbf{v}_n' n^{-1/2} (\widehat{\boldsymbol{\Sigma}}_{n, \lfloor nt \rfloor} - \boldsymbol{\Sigma}_{n, \lfloor nt \rfloor}) \mathbf{w}_n, \qquad t \in [0,1], n \geq 1.$$

If the dependence of the above quantities on $\mathbf{v}_n, \mathbf{w}_n$ matters, we shall indicate this in our notation and then write

$$D_{nk}(\mathbf{v}_n, \mathbf{w}_n), \mathcal{D}_n(t; \mathbf{v}_n, \mathbf{w}_n).$$

Recalling that $\widehat{\mathbf{\Sigma}}_n = n^{-1}\widehat{\mathbf{\Sigma}}_{n,n}$, cf. (1) and (4), we have

$$\mathcal{D}_n(1) = \mathbf{v}_n'\sqrt{n}(\widehat{\mathbf{\Sigma}}_n - \mathbf{\Sigma}_n)\mathbf{w}_n, \qquad n \geq 1,$$

is the *centered and scaled version* of the bilinear form

$$Q(\mathbf{v}_n, \mathbf{w}_n) = \mathbf{v}_n'\widehat{\mathbf{\Sigma}}_n\mathbf{w}_n = \widehat{\mathrm{Cov}}(\mathbf{v}_n'\mathbf{Y}_n, \mathbf{w}_n'\mathbf{Y}_n),$$

where

$$\mathbf{\Sigma}_n = E\widehat{\mathbf{\Sigma}}_n = \frac{1}{n}\sum_{i=1}^{n} E(\mathbf{Y}_{ni}\mathbf{Y}_{ni})',$$

If $\{\mathbf{Y}_{ni} : 1 \leq i \leq n\}$ is stationary, then $\mathbf{\Sigma}_n$ simplifies to $\mathbf{\Sigma}_n = E(\mathbf{Y}_{n1}\mathbf{Y}_{n1}')$ (but our result are more general).

Strong Approximation

Result:
Within the model framework under certain additional technical conditions, we may approximate the related processes by Brownian motions:

$$|D_{nt} - \alpha_n B_n(t)| = o(t^{1/2}), \qquad \text{for all } t > 0 \text{ a.s.},$$

as $n, t \to \infty$, and

$$\sup_{t \in [0,1]} |\mathcal{D}_n(t) - \alpha_n B_n(\lfloor nt \rfloor / n)| = o(1), \qquad \text{a.s.},$$

as $n \to \infty$, as well as the CLT

$$|\mathcal{D}_n(1) - \alpha_n B_n(1)| = o(1), \qquad \text{a.s.},$$

as $n \to \infty$, i.e. $\mathcal{D}_n(1)$ is asymptotically $\mathcal{N}(0, \alpha_n^2)$.

Strong Approximation

Standardized sequential statistic:
Monitor the sequence of (standardized) deviations from an assumed variance–covariance matrix $\mathbf{\Sigma}_n$ via

$$\mathcal{D}_n^*(t) = \alpha_n^{-1}(\mathbf{v}_n, \mathbf{w}_n)\mathcal{D}_n(t, \mathbf{v}_n, \mathbf{w}_n), \qquad t \in [s_0, 1],$$

which can be approximated by a Brownian motion for large n.

Those results provide a basis for valid statistical inference.

Strong Approximation

Multivariate Extension:
Needed when projecting high-dimensional data in lower-dimensional subspaces!

Theorem

Let $\{\mathbf{v}_{nj}, \mathbf{w}_{nj} : 1 \leq j \leq K\}$ *be weighting vectors of dimension* d_n *satisfying condition (7).*
Then, under the assumptions of the previous theorem, there exists a K*–dimensional Brownian motion* $\{\mathbf{B}^{(n)}(t) : t \in [0,1]\}$ *with coordinates* $B_{ni} = B_n(t; \mathbf{v}_{ni}, \mathbf{w}_{ni})$, $t \in [0,1]$, $i = 1, \dots, K$, *such that*

$$\left\| (\mathcal{D}_n(t; \mathbf{v}_{ni}, \mathbf{w}_{ni}))_{i=1}^K - (B_n(\lfloor nt \rfloor / n; \mathbf{v}_{ni}, \mathbf{w}_{ni}))_{i=1}^K \right\| = o(1), \qquad (6)$$

a.s., as $n \to \infty$, *where* $\| \bullet \|$ *denotes an arbitrary vector norm on* $\mathbb{R}^K$.

Strong Approximation

Corollary

Suppose that $\mathbf{Y}_{n1}, \ldots, \mathbf{Y}_{nn}$ *is a* d_n*–dimensional vector time series satisfying Assumption (A). Then, after redefining the series on a new probability space, there exists a Brownian motion such that*

$$\left| \max_{k \leq n} |\mathcal{D}_n(k/n)| - \max_{k \leq n} |\alpha_n B_n(k/n)| \right| = o(1),$$

as $n \to \infty$.

The proofs rely on generalizations of Kouritzin (1995, SPA), who applied Philipp's (1986) results on strong approximations in Hilbert spaces.

Discussion

...of the ℓ_1-condition:

$$\sup_{n\in\mathbb{N}} \|\mathbf{w}_n\|_{\ell_1} = \sup_{n\in\mathbb{N}} \sum_{\nu=1}^{d_n} |w_j| < \infty \tag{7}$$

Ex. 1: ℓ_0-sparse vectors: $w_i > 0$ only for $i \in \{i_1, \dots, i_L\}$, L fixed.
(classical 'low-dimensional' case)

Ex. 2: $\mathbf{w}_n' = (w_1, \dots, w_{d_n})'$ with $\sum_j |w_j| < \infty$.
(most coordinates receive a negligible weight)

Ex. 3: $w_{ni} = 1/d_n$ for $i = 1, \dots, d_n$.
(all d_n coordinates are taken into account.)

Discussion

...of the ℓ_1-condition:

$$\sup_{n \in \mathbb{N}} \|\mathbf{w}_n\|_{\ell_1} = \sup_{n \in \mathbb{N}} \sum_{\nu=1}^{d_n} |w_j| < \infty \qquad (7)$$

Ex. 1: ℓ_0-sparse vectors: $w_i > 0$ only for $i \in \{i_1, \dots, i_L\}$, L fixed.
(classical 'low-dimensional' case)

Ex. 2: $\mathbf{w}_n' = (w_1, \dots, w_{d_n})'$ with $\sum_j |w_j| < \infty$.
(most coordinates receive a negligible weight)

Ex. 3: $w_{ni} = 1/d_n$ for $i = 1, \dots, d_n$.
(all d_n coordinates are taken into account.)

Discussion

...of the ℓ_1-condition:

$$\sup_{n\in\mathbb{N}} \|\mathbf{w}_n\|_{\ell_1} = \sup_{n\in\mathbb{N}} \sum_{\nu=1}^{d_n} |w_j| < \infty \tag{7}$$

Ex. 1: ℓ_0-sparse vectors: $w_i > 0$ only for $i \in \{i_1, \dots, i_L\}$, L fixed.
(classical 'low-dimensional' case)

Ex. 2: $\mathbf{w}_n' = (w_1, \dots, w_{d_n})'$ with $\sum_j |w_j| < \infty$.
(most coordinates receive a negligible weight)

Ex. 3: $w_{ni} = 1/d_n$ for $i = 1, \dots, d_n$.
(all d_n coordinates are taken into account.)

Discussion

... of Assumption A:

$$\sup_{n\in\mathbb{N}} \max_{1\le\nu\le d_n} |c_{nj}^{(\nu)}|^2 << j^{-3/2-\theta/2} \qquad (8)$$

for some $0 < \theta < 1/2$.
Assumption A...

- $c_{nj}^{(\nu)} = c_j^{(\nu)}$: At time n we observe d_n *sequences* (not depending on n). But we allow for *arrays*.
- covers various **short memory** processes, e.g. ARMA processes.
- covers many **long-range dependent** series such as fractionally integrated noise of order $d \in (-1/2, 1/4 - \theta/2)$,

$$(1-L)^d X_t = \epsilon_t$$

Discussion

Define the *scaled Frobenius norm* by

$$\|\mathbf{A}\|_F^* = \frac{1}{d_n^{1/2}} \left(\sum_{i,j=1}^{d_n} a_{ij}^2 \right)^{1/2} \qquad (s.th. \|\mathbf{I}_{d_n}\|_F^* = 1).$$

Lemma

Suppose Assumption (A) holds true and that, for fixed t, the variances σ_{t-j}^2, $j \geq 0$, of the innovations satisfy

$$\sum_{j=0}^{\infty} j^{-3/2-\theta/2} \sigma_{t-j}^2 < \infty.$$

Then, as $r \to \infty$; we have

$$\sup_{n \in \mathbb{N}} \left\| \boldsymbol{\Sigma}_n[t] - \sum_{j=1}^{r} \sigma_{t-j}^2 \mathbf{c}_{nj} \mathbf{c}_{nj}' \right\|_F^* = o(1).$$

Applications: Portfolio Selection

Consider Assets returns $\mathbf{R}_n = (R_n^{(1)}, \ldots, R_n^{(d_n)})'$ corresponding to the time period $[n-1, n]$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{ij}$.

Since σ_{ij} is the covariance between the return of asset i and asset j, $1 \leq i, j, \leq d_n$, it is not restrictive to assume that the entries of $\boldsymbol{\Sigma}$ neither depend on n nor d_n.

An investor holds at time $n-1$ the position w_{nj} in asset j.
$w_{nj} > 0$ long position, $w_{nj} < 0$ short position.

W.l.o.g. the initial value (capital) at time $n-1$ equals $V = \sum_{j=1}^{d_n} w_{nj} = 1$.
Then the value at time instant n is $\mathbf{w}_n' \mathbf{R}_n$.

Applications: Portfolio Selection

Classical formulation of the portfolio selection problem: Risk = Variance:

$$\min_{\mathbf{w}_n} \operatorname{Var}(\mathbf{w}_n' \mathbf{R}_n) = \mathbf{w}_n' \boldsymbol{\Sigma} \mathbf{w}_n, \qquad \text{subject to } \mathbf{w}_n' \mathbf{1} = 1,$$

whose solution is known to be

$$\mathbf{w}_n^{*\prime} = (\mathbf{1}' \Sigma^{-1} \mathbf{1})^{-1} \mathbf{1}' \Sigma^{-1}.$$

If that solution satisfies the no-short-sales condition, then $\|\mathbf{w}_n^*\|_{\ell_1} = \mathbf{1}' \mathbf{w}_n^* = 1$.

Provided the vector time series of returns satisfies our assumptions, our results provide the asymptotics for the optimal risk

$$\operatorname{Var}((\mathbf{w}_n^*)' \mathbf{Y}_n)$$

associated to the optimal portfolio, when estimating $\boldsymbol{\Sigma}$ (needed for $\mathbf{w}_n^*$) from an independent learning sample.

Applications: Shrinkage

Shrinkage estimation is a well established approach for regularization (Ledoit & Wolf (2004); Sancetta, 2008).

To improve properties such as $E\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_n\|_F^2$ or the condition number, one estimates $\boldsymbol{\Sigma}_n$ by a linear (convex) combination of $\widehat{\boldsymbol{\Sigma}}_n$ and a well-conditioned target such as the identity.

Projecting $\boldsymbol{\Sigma}_n$ onto $\operatorname{span}\{\mathrm{id}_n\}$ leads to the target $\boldsymbol{\Sigma}_n^{(0)} = \mu_n \, \mathrm{id}_n$, where $\mu_n = \operatorname{tr}(\boldsymbol{\Sigma}_n)$ (shrinkage intensity).

We are led to the shrinkage estimator

$$\boldsymbol{\Sigma}_n^s(W_n) = (1 - W_n)\widehat{\boldsymbol{\Sigma}}_n + W_n \mu_n \, \mathrm{id}_n \, .$$

Optimizing W_n w.r.t. the MSE

$$W_n^* = \operatorname{argmin}_{W_n \in [0,1]} d_n^{-1} E\|\boldsymbol{\Sigma}_n^s(W_n) - \boldsymbol{\Sigma}_n\|_F^2$$

leads to explitic formulas for W_n^* and ensures a true improvement

$$E\|\boldsymbol{\Sigma}_n^s - \boldsymbol{\Sigma}_n\|_F^2 < E\|\widehat{\boldsymbol{\Sigma}}_n - \boldsymbol{\Sigma}_n\|_F^2$$

Applications: Sparse Principal Component Analysis LASSO

Let $\mathcal{X}_n$ be a $(n \times d_n)$–dimensional data matrix, independent from $\mathcal{Y}_n$.

SCotLASS (Simplified component technique-lasso), Jolliffee (2003): 1^{st} principal component (pc) solves

$$\max_{\mathbf{v}} \mathbf{v}' \mathcal{X}_n' \mathcal{X}_n \mathbf{v}, \qquad \text{subject to } \|\mathbf{v}\|_{\ell_2}^2 \leq 1,\ \|\mathbf{v}\|_{\ell_1} \leq c.$$

Continue in this way under the additional constraints that further components are orthogonal.

Applications: Sparse Principal Component Analysis

LASSO (Tibshirani, 1996 & 2011): Determine ℓ_1–sparse coefficient vector in a high-dimensional linear regression in dim. p_n

$$Y_t = \mathbf{X}_t' \beta_0 + \epsilon_t, \quad E(\epsilon_t | \mathbf{X}_t) = 0,$$

Given an estimator $\widehat{\beta}_n$, $\pi_n = \mathbf{X}' \widehat{\beta}_n$ is used for prediction.
LASSO minimizes the ℓ_1–constrained least squares criterion

$$\beta \mapsto \sum_t (Y_t - \mathbf{X}_t' \beta)^2, \|\beta\|_{\ell_1} \leq c,$$

for some bound $c > 0$.
Apply results with $\mathbf{w}_n = \widehat{\beta}_n$ estimated from indep. learning sample, $d_n = p_n$, $\mathbf{Y} = \mathbf{X}$, $\mathbf{Y}_t = \mathbf{X}_t$ to infer $\mathrm{Var}(\mathbf{w}_n' \mathbf{X})$ given the learning sample, provided $\{\mathbf{X}_t\}$ satisfies our assumptions.

End

Related Publication:
Steland, A. and R. v. Sachs (2015). Large sample approximations for variance-covariance matrices of high-dimensional time series, *under revision*.

Thanks for your attention.