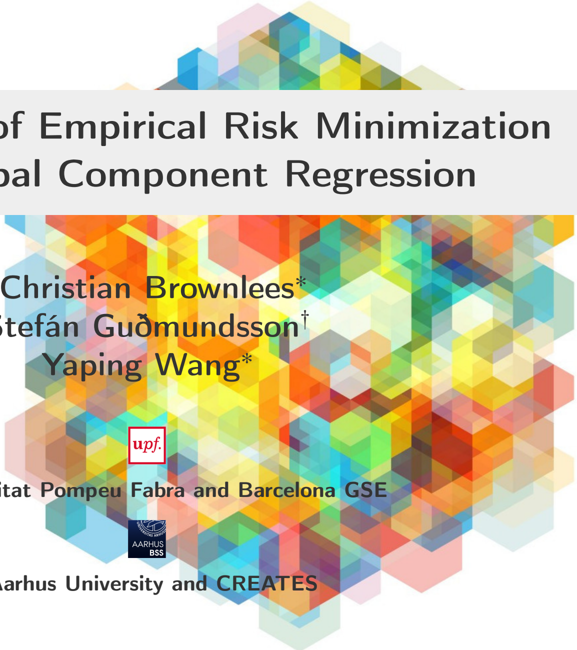
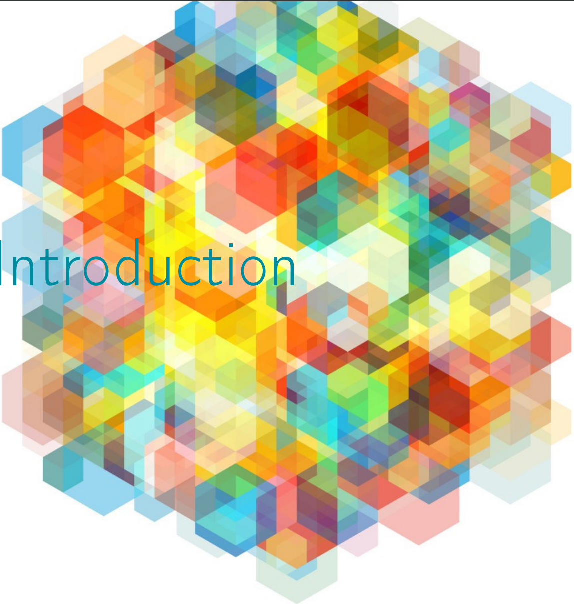
Performance of Empirical Risk Minimization for Principal Component Regression



Christian Brownlees*
Stefán Guðmundsson†
Yaping Wang*

*Universitat Pompeu Fabra and Barcelona GSE

†Aarhus University and CREATES



Introduction

Warm-Up: Statistical Learning Theory

- Learning theory is the theoretical backbone of machine learning.
- A key principle of learning theory is empirical risk minimization.
- Let Y_t be a random variable and $f_{\theta\, t}$ a class of forecasting rules indexed by θ. Empirical risk minimization consists in choosing the forecasting rule $f_{\theta\, t}$ that minimizes the empirical loss for predicting Y_t.
- Key feature of learning theory is its nonparametric nature: The relation between Y_t and $f_{\theta\, t}$ is not specified.
- Objective of learning theory is to establish nonasymptotic bounds on the predictive performance of empirical risk minimization.

Statistical Learning Theory for Time Series

- The bulk of contributions in learning theory focus on setups that are arguably not attractive for economics: i.i.d. and bounded data.
- This paper deals with principal component regression (PCR) and our main contribution consists in establishing nonasymptotic prediction performance guarantees for this regression technique.

Key Features of the Analysis

- **Focus on Predictive Performance.**
  We study the gap between PCR and the best linear predictor given by

$$\mathbb{E}(Y_t - \hat{f}_t^{PCR})^2 - \min_{\theta} \mathbb{E}(Y_t - \boldsymbol{X}_t'\boldsymbol{\theta})^2 \ .$$

- **Nonparametric.**
  The relation between the target Y_t and predictors $\boldsymbol{X}_t$ is not specified. We treat PCR as a regularized regression technique.

- **Strong and Weak Signals.**
  The largest K eigenvalues of the covariance matrix of predictors $\boldsymbol{X}_t$ diverge at the rate p^{α} where
  - $\alpha = 1$ (strong signal case) or
  - $\alpha \in (1/2, 1)$ (weak signal case).

Related Literature

- Principal component regression (PCR) can be traced back to at least [Hotelling, 1957] and [Kendall, 1957].

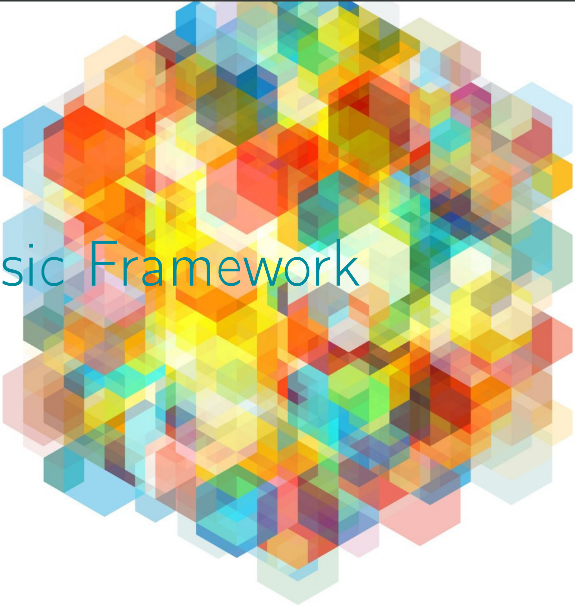
- Humongous list of contributions in econometrics on factor models/principal component regression:

  [Bai and Ng, 2002], [Bai, 2003], [Forni et al., 2000], [Forni et al., 2005], [Fan et al., 2011], [Fan et al., 2013], [Onatski, 2012], [Gagliardini et al., 2020], [Giglio et al., 2023] among others.

- We rely on arguments by Bai, Fan, Liao, Mincheva and Ng.

- [Stock and Watson, 2002] and [Fan et al., 2024] study the prediction properties of PCR. [Fan et al., 2024] is an influential recent key contribution that studies the properties of a large class of high-dimensional models that includes factor models.

Basic Framework



Basic Framework

- Let $\{(Y_t, \boldsymbol{X}_t)'\}_{t=1}^T$ be a stationary sequence of zero-mean random vectors taking values in $\mathcal{Y} \times \mathcal{X} \subset \mathbb{R} \times \mathbb{R}^p$
- We are interested in predicting Y_t using $\boldsymbol{X}_t$ when p is large.
- We rely on principal component regression (PCR) for prediction.
  1. We compute the first K principal components of $\boldsymbol{X}_t$ (denoted by $\widehat{\boldsymbol{P}}_t$)
  2. We run a regression of Y_t on the principal components $\widehat{\boldsymbol{P}}_t$.

Principal Component Regression: Step 1

- The first step consists in computing the $T \times K$ principal components matrix $\widehat{\boldsymbol{P}} = (\widehat{\boldsymbol{P}}_1, \dots, \widehat{\boldsymbol{P}}_T)'$ associated with the $T \times p$ predictor matrix $\boldsymbol{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_T)'$.

- This may be defined as the solution of the problem

$$(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{P}}) = \arg \min_{\substack{\boldsymbol{B} \in \mathbb{R}^{p \times K} \\ \boldsymbol{P} \in \mathbb{R}^{T \times K}}} \|\boldsymbol{X} - \boldsymbol{P}\boldsymbol{B}'\|_F^2 \text{ s.t. } \frac{1}{T}\boldsymbol{P}'\boldsymbol{P} = \boldsymbol{I}_K \, , \frac{1}{p}\boldsymbol{B}'\boldsymbol{B} \text{ is diagonal} \, ,$$

- As it is well known the solution of this problem is given by

$$\widehat{\boldsymbol{P}} = \boldsymbol{X}\widehat{\boldsymbol{V}}_K\widehat{\Lambda}_K^{-1/2}$$

where $\widehat{\Lambda}_K$ and $\widehat{\boldsymbol{V}}_K$ are, respectively, the matrices of the top K eigenvalues and eigenvectors of $\widehat{\Sigma} = \boldsymbol{X}'\boldsymbol{X}/T$.

Principal Component Regression: Step 2

- The regression coefficient for predicting Y_t given $\widehat{\boldsymbol{P}}_t$ is given by

$$\hat{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta}\in\mathbb{R}^K} \| \boldsymbol{Y} - \widehat{\boldsymbol{P}}\boldsymbol{\vartheta}\|_2^2 \ ,$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_T)'$.

- As it is well known the solution of this problem is given by

$$\hat{\boldsymbol{\vartheta}} = (\widehat{\boldsymbol{P}}'\widehat{\boldsymbol{P}})^{-1}\widehat{\boldsymbol{P}}'\boldsymbol{Y} = \frac{1}{T}\widehat{\boldsymbol{P}}'\boldsymbol{Y} \ .$$

PCR as Regularized ERM

- It may not be apparent at first that PCR is an empirical risk minimization procedure, but this is in fact the case.

- Let the regularized empirical risk minimizer $\hat{\boldsymbol{\theta}}_{PCR} \in \mathbb{R}^p$ be given by

$$\hat{\boldsymbol{\theta}}_{PCR} = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \| \boldsymbol{Y} - \boldsymbol{X}' \boldsymbol{\theta} \|_2^2 \ \text{ s.t. } \ \widehat{\boldsymbol{V}}_R' \boldsymbol{\theta} = \boldsymbol{0} \ ,$$

  where $\widehat{\boldsymbol{V}}_R$ is the $p \times p - K$ matrix of eigenvectors corresponding to the $K + 1$ to p eigenvalues of $\widehat{\Sigma}$.

- It is straightforward to establish that

$$\hat{\boldsymbol{\theta}}_{PCR} = \widehat{\boldsymbol{\Lambda}}_K^{-1/2} \widehat{\boldsymbol{V}}_K \hat{\boldsymbol{\vartheta}} \ ,$$

  and that $\hat{\boldsymbol{\theta}}_{PCR}$ and $\hat{\boldsymbol{\vartheta}}$ produce same forecast

$$\hat{f}_t^{PCR} = \hat{\boldsymbol{\theta}}_{PCR}' \boldsymbol{X}_t = \hat{\boldsymbol{\vartheta}}' \widehat{\boldsymbol{P}}_t \ .$$

Performance of Empirical Risk Minimization for PCR

Performance Measure and Benchmark

- We are concerned with establishing an oracle inequality for PCR

- We measure the accuracy of PCR by its conditional risk

$$R(\hat{\boldsymbol{\theta}}_{PCR}) = \mathbb{E}\left[(Y_t - \hat{f}_t^{PCR})^2 \middle| \hat{\boldsymbol{\theta}}_{PCR} = \hat{\boldsymbol{\theta}}_{PCR}(\mathcal{D}) \right] ,$$

  where the observation $(Y_t, \boldsymbol{X}_t)'$ is drawn independently of the data $\mathcal{D}$. This can be interpreted as the risk of the ERM obtained from the "training data" $\mathcal{D}$ over the "validation observation" $(Y_t, \boldsymbol{X}_t)'$.

- The natural benchmark is the risk of the best linear predictor

$$R(\boldsymbol{\theta}^*) = \min_{\boldsymbol{\theta} \in \mathbb{R}^p} R(\boldsymbol{\theta}), \quad \text{where } R(\boldsymbol{\theta}) = \mathbb{E}\left[(Y_t - \boldsymbol{X}_t' \boldsymbol{\theta})^2 \right] .$$

Oracle Inequality

- Our objective is to establish an oracle inequality stating that

$$R(\hat{\boldsymbol{\theta}}_{PCR}) \leqslant R(\boldsymbol{\theta}^*) + B_T(p, K)$$

  holds at least with probability $1 - \delta_T(p, K)$ for all sufficiently large T, where $B_T(p, K)$ and $\delta_T(p, K)$ approach 0 as $T \to \infty$.

- A few "philosophical" remarks to appreciate what this means:
  - We care about prediction accuracy, not about estimation accuracy.
  - We care about achieving optimality relative to the class of forecasting rules (here linear forecasts), not relative to the "true model"
  - We care about finite sample guarantees, not about asymptotic ones.
  - We care about the rate of convergence $B_T(p, K)$.

Optimal Learning Rate

- Key question:
  What is the optimal learning rate $B_T(p, K)$ that can be achieved?

- This in general can be a tough question to answer. In this paper we shed partial light onto this question by comparing the learning rates obtained in this work with the optimal learning rate that could be achieved if the principal components were observed.

- It is well known that in such a case the optimal rate of convergence for linear aggregation is of the order K/T, which is achieved by the least squares estimator [Tsybakov, 2003].

Assumptions

Assumptions

The assumptions of our analysis are a union of

- Assumptions analogous to those used in [Fan et al., 2013] for the analysis of (regularized) covariance estimation when the data is generated by an approximate factor model.
  We work with a weaker version of some of their assumptions.
- The small-ball assumption used in [Lecué and Mendelson, 2016]. This assumption allows to use a proof strategy that leads to sharp rates.
- A mild regularity condition on the predictors' distribution. Analog of a large dimensional version of the bounded density assumption often used in the analysis of nonparametric estimators.

Assumptions

1. **Distribution.**

   Data have subGaussian tails.

2. **Dependence.**

   Data are α-mixing with geometric decaying α-mixing coefficients with rate of decay r_α

3. **Eigenvalues.**

   K largest eigenvalues of the covariance matrix diverge at the rate p^α for $\alpha \in (1/2, 1]$.

4. **Number of predictors and principal components.**

   Number of predictors and principal components grow as a function of the sample size T.

5. **Small-ball condition.**

A.3 Eigenvalues

A.3 Eigenvalues

There is an integer $K \in \{1, \ldots, p\}$, a constant $\alpha \in (1/2, 1]$ and a sequence of non-increasing nonnegative constants $c_1, \ldots, c_p$ with $c_K > 0$ such that, $\lambda_i = c_i p^\alpha$ for $i = 1, \ldots, K$, and $\lambda_i = c_i$ for $i = K + 1, \ldots, p$.

In our analysis we distinguish between

1. strong signal regime when $\alpha = 1$ (analog of strong factor models)
2. weak signal regime when $\alpha \in (1/2, 1)$ (analog of weak factor models).

Note that we allow the non-diverging eigenvalues of Σ to be zero.

A.5 Small-ball Condition

A.5 Small-ball Condition

The sequence $\{\boldsymbol{X}_t\}_{t=1}^{T}$ satisfies, for each $t = 1, \ldots, T$ and for each $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$,

$$\mathbb{P}\left(|f_{\boldsymbol{\theta}_1\,t} - f_{\boldsymbol{\theta}_2\,t}| \geqslant \kappa_1 \|f_{\boldsymbol{\theta}_1\,t} - f_{\boldsymbol{\theta}_2\,t}\|_{L_2}\right) \geqslant \kappa_2 \;,$$

for some $\kappa_1 > 0$ and $\kappa_2 > 0$.

This is an identification condition. It requires the random variable $(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)'\boldsymbol{X}$ not to have excessive mass in a neighbourhood of zero.

Main Result

Empirical Risk Minimization for PCR

Theorem 1 Empirical Risk Minimization for PCR

Suppose A1-A5 are satisfied.
Then for any $\eta > 0$ there exists a constant $C > 0$ such that, for any T sufficiently large,

$$R(\hat{\boldsymbol{\theta}}_{PCR}) \leqslant R(\boldsymbol{\theta}^*) + 2(\boldsymbol{\theta}^*)' \boldsymbol{V}_R \boldsymbol{\Lambda}_R \boldsymbol{V}_R' \boldsymbol{\theta}^* + C \left[\frac{1}{p^{2\alpha-1}} + \left(\frac{p}{Tp^{\alpha}} \right)^2 p^{\frac{2}{r\alpha}} + \frac{K}{T} \right] \log(T),$$

holds with probability at least $1 - T^{-\eta}$.

This is an oracle inequality since $B_T(p, K) \to 0$ as $T \to \infty$.

$B_T(p, K)$ depends on the rate of growth of p and K and signal strength α.

Remarks on Theorem 1

- The gap is made up of two terms. The first can interpreted as the approximation error of PCR and the second as the estimation error.
- The approximation error measures the gap between the performance of the best linear predictor based on the population principal components $\boldsymbol{P}_t$ and the best linear predictor based on the predictors $\boldsymbol{X}_t$.
- The estimation error measures the gap between the performance of PCR relative to the best linear predictor based on the population principal components $\boldsymbol{P}_t$.
- If the contribution of the idiosyncratic component vector $\boldsymbol{u}_t$ is negligible then PCR has a negligible approximation error.

More on the learning rate $B_T(p, K)$

More on the learning rate $B_T(p, K)$ – Strong Signal

For ease of exposition we assume that the approximation error is zero throughout this section.

- For two sequences $\{A_T\}$ and $\{B_T\}$ we use $A_T \lesssim B_T$ to denote that there is a constant $C > 0$ such that $\mathbb{P}(A_T < CB_T) \to 1$ as $T \to \infty$.
- In the strong signal case $(\alpha = 1)$ and independence $(r_\alpha = \infty)$

$$R(\hat{\boldsymbol{\vartheta}}) - R(\boldsymbol{\theta}^*) \lesssim \left[\frac{1}{p} + \frac{1}{T^2} + \frac{K}{T}\right] \log(T) \ .$$

If $p > T$ then we recover the optimal learning rate K/T.

More on the learning rate $B_T(p, K)$ – Weak Signal

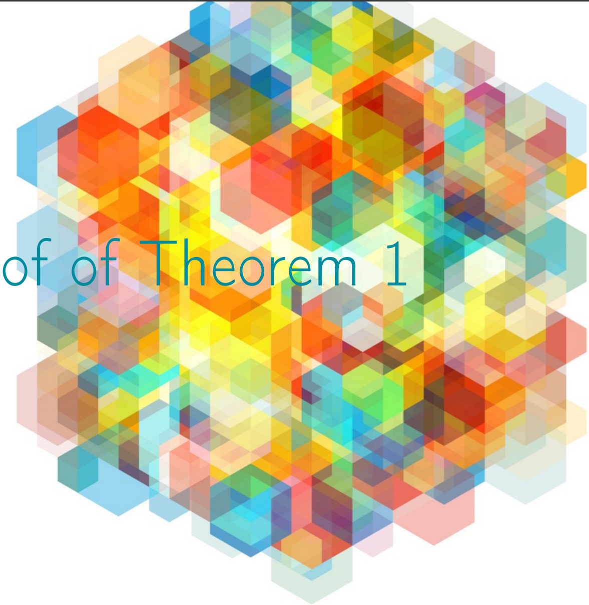
- In the weak signal case ($\alpha \in (1/2, 1)$) and independence ($r_\alpha = \infty$)

$$R(\hat{\boldsymbol{\vartheta}}) - R(\boldsymbol{\theta}^*) \lesssim \left[\frac{1}{p^{2\alpha-1}} + \left(\frac{p}{Tp^\alpha} \right)^2 + \frac{K}{T} \right] \log(T) \ .$$

- Learning is slow when factors are weak.
  Example: when $\alpha = 0.55$ and $p = 10'000$ we have that $\frac{1}{p^{2\alpha-1}} \approx 0.4$.

- When $r_p \in [1/(2\alpha - 1), 1/(2 - 2\alpha)]$, $R(\hat{\boldsymbol{\vartheta}}) - R(\boldsymbol{\theta}^*) \lesssim \frac{K}{T} \log(T)$

- Recovery of the optimal learning rate is only possible when $\alpha > 2/3$.



Proof of Theorem 1

Sketch of Proof

- We conclude with a sketch of the proof of Theorem 1.
- This is a combination of arguments used to establish consistency in the factor model literature and an elegant argument based on [Lecué and Mendelson, 2016].

Sketch of Proof

Define

- the approximate rotation matrix $\boldsymbol{H} = \widehat{\boldsymbol{\Lambda}}_K^{-1/2} \widehat{\boldsymbol{V}}_K' \boldsymbol{V}_K \boldsymbol{\Lambda}_K^{1/2}$
- the vector of coefficients of the optimal linear predictor based on the population principal components $\boldsymbol{P}_t$

$$\boldsymbol{\vartheta}^* = \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^K} \| \boldsymbol{Y}_t - \boldsymbol{P}_t' \boldsymbol{\vartheta} \|_{L_2}^2 ,$$

- the empirical risk minimizer based on the population principal components $\boldsymbol{P}_t$

$$\tilde{\boldsymbol{\vartheta}} = \arg\min_{\boldsymbol{\vartheta} \in \mathbb{R}^K} \| \boldsymbol{Y} - \boldsymbol{P} \boldsymbol{\vartheta} \|_2^2 .$$

Basic Decomposition

Basic Decomposition

Let $\{\boldsymbol{Z}_t\}$ *with* $\boldsymbol{Z}_t = (Y_t, \boldsymbol{X}_t')'$ *be a zero-mean* $(p+1)$*-dimensional stationary process with* $\mathbb{E}(Z_{i\,t}^2) < \infty$ *for all* i*.*
Then it holds that

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}_{PCR}) - R(\boldsymbol{\theta}^*) \leqslant{}& 2 \max_{1\leqslant s\leqslant T}\{Y_s^2\}\mathbb{E}(\|\widehat{\boldsymbol{P}}_t - \boldsymbol{HP}_t\|_2^2|\mathcal{D}) \\ &+ 4\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2 + 4\|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2^2 + 2\|\boldsymbol{u}_t'\gamma^*\|_{L_2}^2 \ , \end{aligned}$$

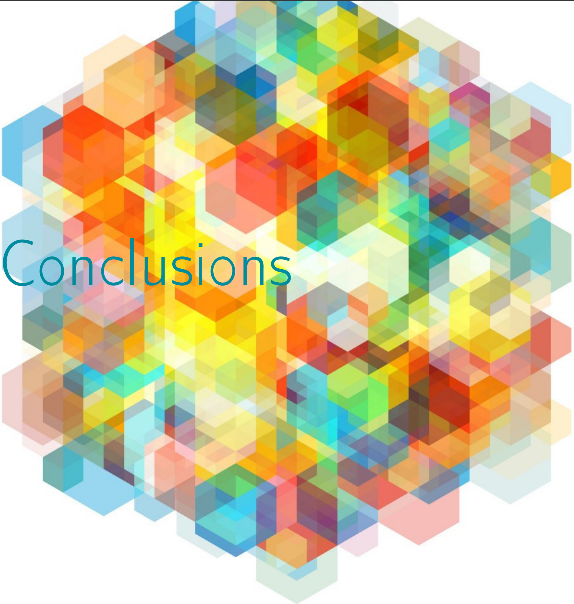
Basic Decomposition

- $\mathbb{E}(\|\widehat{\boldsymbol{P}}_t - \boldsymbol{H}\boldsymbol{P}_t\|_2^2|\mathcal{D})$ and $\|\tilde{\boldsymbol{\vartheta}} - \boldsymbol{H}'\hat{\boldsymbol{\vartheta}}\|_2^2$ are controlled using arguments analog to the ones used in the factor model literature, in particular [Fan et al., 2013].

- $\|\boldsymbol{\vartheta}^* - \tilde{\boldsymbol{\vartheta}}\|_2^2$ is controlled using the small-ball method of [Lecué and Mendelson, 2016]

- $\|\boldsymbol{u}_t'\gamma^*\|_{L_2}^2$ is the approximation error of PCR. However under A.6 this is at most of the same order of magnitude of the estimation error.

Conclusions



Conclusion

- We establish prediction performance guarantees for empirical risk minimization for principal component regression.
- We establish oracle inequalities under strong and weak signal regimes.
- Analysis is carried out in a nonparametric framework. In particular the target variable Y_t is not assumed to be generated by a factor model.

Thanks!

References I

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70:191–221.

Fan, J., Liao, Y., and Mincheva, M. (2011). High Dimensional Covariance Matrix Estimation in Approximate Factor Models. *The Annals of Statistics*, 39:3320–3356.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680.

Fan, J., Masini, R. P., and Medeiros, M. C. (2024). Bridging Factor and Sparse Models. *The Annals of Statistics*, forthcoming.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2000). The generalized dynamic factor model: Identification and estimation. *The Review of Economics and Statistics*, 82:540–554.

Forni, M., Hallin, M., Lippi, M., and Reichlin, L. (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *Journal of the American Statistical Association*, 100(471):830–840.

Gagliardini, P., Ossola, E., and Scaillet, O. (2020). Estimation of large dimensional conditional factor models in finance. In *Handbook of Econometrics*, volume 7A. North-Holland.

Giglio, S., Xiu, D., and Zhang, D. (2023). Prediction when factors are weak. Technical report.

Hotelling, H. (1957). the relations of the newer multivariate statistical methods to factor analysis. *British Journal of Statistical Psychology*, 10(2):69–79.

Kendall, M. (1957). *A Course In Multivariate Statistics*. Griffin, London.

References II

Lecué, G. and Mendelson, S. (2016). Performance of empirical risk minimization in linear aggregation. *Bernoulli*, 22:1520–1534.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics*, 168(2):244–258.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179.

Tsybakov, A. B. (2003). Optimal Rate of Linear Aggregation. In *Learning Theory and Kernel Machines (COLT-2003)*, volume 2777 of *Lecture Notes in Artificial Intelligence*, pages 303–313. Springer, Heidelberg.