Double Machine Learning Inference for Conditional Latent Factors*

Patrick Gagliardini† and Hao Ma‡

May 2026

Abstract

Empirical economics often requires inference on low-dimensional objects defined in high-dimensional conditional environments. This paper studies objects in conditional asset pricing, including risk premia, stochastic discount factors, and pricing errors, using conditional latent factor models for large, unbalanced panels of asset returns. We identify them as members of a general class of parameters of interest, represented as functionals of conditional moments of portfolio returns given high-dimensional predictors. We estimate the conditional moments using flexible machine learning methods, and use double machine learning to obtain feasible asymptotically normal inference for serially dependent data. In monthly individual U.S. stock returns, conditional factors are large in number and counter-cyclical; the latent space is only partly spanned by Fama-French factors and dynamically involves leverage, time-series reversal, and liquidity. The implied SDF improves out-of-sample pricing of benchmark test assets, and conditioning information materially affects conditional risk-premium estimates.

Keywords: Large Panel, Latent Factors, Conditioning Information, Double/Debiased Machine Learning, Nonparametric Time Series Regression, Conditional Risk Premia, Stochastic Discount Factor.

*This paper circulated previously under the title "Extracting Statistical Factors When Betas Are Time-Varying". We thank Y. Aït-Sahalia, T. Andersen, F. Carlini, S. Dobrev, R. Engle, A.-P. Fortin, E. Guerre, Y. Li, D. Massacci, M. Pelger, M. Rubin, O. Scaillet, K. Singleton, V. Todorov, F. Trojani, P. Zaffaroni, and participants at the SoFiE 2019 Annual Meeting, EEA & ESEM 2019, CFE 2019, EFA 2020, MFA 2020, ICEEE 2021, FinEML 2025, AMES 2026, the SoFiE seminar series, and seminars at the University of Geneva, North Carolina State University, Tinbergen Institute and EDHEC Business School for useful comments. We gratefully acknowledge financial support from the Swiss National Science Foundation (SNSF) Grants Nr. 162633 and 204106.

†Università della Svizzera italiana (USI, Lugano) and Swiss Finance Institute (SFI). E-mail address: patrick.gagliardini@usi.ch

‡Queen Mary University of London. E-mail address: hao.ma@qmul.ac.uk

1 Introduction

Many empirical questions in economics ask for inference on a low-dimensional object in an environment that is high-dimensional and dynamic. A researcher may care about a policy effect, an elasticity, a heterogeneous treatment effect, or a risk premium, but often the object of interest is defined only after accounting for state variables, latent types, high-dimensional controls, or systematic risk. Asset pricing is one such example: the economic objects of interest include e.g. risk premia, stochastic discount factors and pricing errors; the environment can be a large panel of returns for n assets and T time periods, in which cross-sectional dependence reflects systematic risk. Approximate factor structures (Chamberlain and Rothschild (1983)) provide the natural dimension-reduction method in this setting: a small number of factors summarizes pervasive risks, and Arbitrage Pricing Theory (APT) links compensation to exposures to these risks. Further, in realistic asset-pricing environments, risk exposures and prices of risk vary with macro-financial states and asset characteristics, which call for the need of incorporating conditioning information in the factor model.

The conditional asset-pricing literature has traditionally worked with observable factors, such as consumption growth in intertemporal models or returns on characteristic-sorted portfolios in empirical specifications (Connor and Korajczyk (1989), Cochrane (1996), Ferson and Harvey (1991, 1999), Lettau and Ludvigson (2001), Petkova and Zhang (2005)). Conditional two-pass Fama-MacBeth regressions using large cross-sections of individual stocks extend this logic to richer test-asset environments (Gagliardini, Ossola and Scaillet (2016, 2019, 2020), Chaieb, Langlois and Scaillet (2021), Bakalli, Guerrier and Scaillet (2023)).

Observable factors are useful empirical proxies, but their choice is far from unique, as emphasized by the factor-zoo literature (Cochrane (2011), Feng, Giglio and Xiu (2020)). More importantly, if priced sources of systematic risk are omitted, estimates of risk premia and pricing restrictions can be biased. This concern motivates the use of latent factors extracted statistically from large cross-sections of asset returns (Connor, Hagmann and Linton (2012), Kelly, Pruitt and Su (2017, 2019), Chen, Roussanov and Wang (2023), Zaffaroni (2025)). In an unconditional setting, Giglio and Xiu (2021) show how latent factor methods can address omitted-factor bias when estimating the premium of an observed factor.

This paper studies identification and asymptotically normal inference for a low-dimensional object of interest in conditional latent factor models, when the researcher uses nonparametric methods to estimate conditional moments of returns given investors' information. In the conditional setting, the relevant latent factor space, the projection of observed factors onto that space, and the resulting economic parameters are conditional objects. This step is economically natural but econometrically demanding. The conditioning information can be

high-dimensional, and the relationship between state variables, factor exposures, and conditional moments can be nonlinear. Flexible machine learning methods are therefore natural first-stage tools for estimating the conditional moments needed to identify these objects. However, machine learning estimators may be regularized, biased, and converge at low and method-specific rates. Thus, we need to account for these features when obtaining the standard errors and confidence intervals for the low-dimensional parameters of interest. Our methodology is general with respect to (w.r.t.) the nonparametric regression techniques for estimating conditional moments. It accommodates e.g. Sieve estimators with linear or nonlinear bases (in our empirics we deploy Artificial Neural Networks, ANN), sparse regressions with expanding dictionaries as well as other high-dimensional methods like LASSO.

To fix ideas, consider estimating the average conditional risk premium of an observed factor, such as the market, after controlling for latent systematic risk. In the static omitted-factor setting, the observed factor is projected onto the latent factor space to purge omitted priced risks (Giglio and Xiu (2021)). In the conditional setting, the projection coefficients and the latent prices of risk are themselves state-dependent; the target can be written as a time-series mean, such as $\mathbb{E}(\lambda_t^O)$, where λ_t^O is the conditional premium of the observed factor induced by its projection onto the latent factor space. Identification is nontrivial because this object depends on latent factors and on conditional moments that must be learned from data. We show that this leading example belongs to a general class of low-dimensional parameters (see equation (2.3) below) that can be represented as functionals of identifiable conditional moments of portfolio returns. We then develop double/debiased machine learning inference based on Neyman-orthogonal (locally-robust) moment restrictions, delivering asymptotically normal inference despite first-stage machine learning estimation error. Conditional canonical correlations between latent and observed factors, and conditional pricing errors from the implied stochastic discount factor, are further instances of the same parameter class.

The contributions of this paper are twofold. First, on the methodological side, for $n, T \to \infty$ we prove the nonparametric identification of the path of the latent stochastic process γ_t in equation (2.3) by showing its representation in terms of the conditional mean and variance of a vector of asymptotic portfolio returns. Portfolio returns are characteristic-weighted cross-sectional averages of stock returns. The characteristics instrument conditional betas through full-rank cross-sectional covariation and are orthogonal to idiosyncratic errors. For the conditional risk-premium example, we prove that $c_0 = \mathbb{E}(\lambda_t^O)$ can be written as $\mathbb{E}[\psi(\theta_0(Z_{t-1}))]$, a functional of the conditional means and covariances of portfolio returns. Hence, we take advantage of portfolio formation to circumvent factor unobservability via the surrogate of an unknown function $\theta_0(\cdot)$ involving identifiable conditional moments. We get a semiparametric framework with an infinite-dimensional "nuisance" parameter $\theta_0(\cdot)$ and a

finite-dimensional parameter of interest c_0 that is a functional thereof.

Our identification strategy is constructive and the analogy principle leads to consistent estimators for both the time-varying number of conditional latent factors and unconditional moments such as $c_0 = \mathbb{E}[\gamma_t] = \mathbb{E}[\psi(\theta_0(Z_{t-1}))]$, with process γ_t in our class (2.3), and more general ones. We estimate function $\theta_0(\cdot)$ by nonparametric regression of (squared) portfolio returns onto lagged predictors. Our asymptotic results are general w.r.t. the adopted nonparametric estimator $\hat{\theta}(\cdot)$ as long as L^2-norm convergence $\|\hat{\theta} - \theta_0\|_{L^2} = O_p(T^{-\bar{\delta}})$, for a $\bar{\delta} > 0$, holds. Importantly, we reduce the impact of nuisance parameter estimation on target parameter c_0 by Double/Debiased Machine Learning (DML, Chernozhukov et al. (2018), Chernozhukov et al. (2022), Chernozhukov, Newey and Singh (2022 a,b), Ahrens et al. (2025)). DML uses Neyman orthogonalization to make the moment condition defining c_0 locally robust to first-stage deviations $\hat{\theta} - \theta_0$, and cross-fitting to limit overfitting bias from flexible machine-learning estimators. We extend the theory of DML to cover serially dependent data (β-mixing) and panel double asymptotics [1], and provide a Heteroskedasticity and Autocorrelation Consistent (HAC) estimator of the asymptotic variance. Bonhomme, Jochmans and Weidner (2025) uses Neyman-orthogonalization to address the incidental parameter problem in panel models. Lewis and Syrgkanis (2021) likewise use orthogonalized double machine learning for inference in semiparametric models with complex nuisance components in a dynamic treatment setting. Abadie et al. (2024) deploys ideas familiar to the doubly-robust literature when estimating the Average Treatment Effect assuming a static latent factor structure in the confounders.

Second, in our empirical contribution we estimate the conditional factor space from an unbalanced panel of monthly returns for U.S. individual stocks in the CRSP dataset between July 1964 and December 2023. We build the portfolio returns with 78 company-level characteristics from Chen and Zimmermann (2021). We include more than 400 monthly predictors to generate the conditioning information set using lags of the macro-financial indicators of Goyal, Welch and Zafirov (2024) and of portfolio returns. To summarize the main results, (i) we find evidence that the number of latent conditional factors is large and counter-cyclical. Indeed, we need 7 to 13 factors to exhaust pervasive risks up to 80% of the conditional variance of the systematic risk components. The number of conditional factors is larger during bear market phases suggesting the presence of risk factors related to financial distress during market downturns. (ii) Conditional canonical correlations reveal that traditional observable factor sets, such as FF6 (Fama–French five factors plus momentum),

[1] While most theoretical developments in DML theory are in the i.i.d. setting, a notable exception is Semenova et al. (2023) who cover the panel/time series case with weak dependence. Their focus is however different from ours as Semenova et al. (2023) address estimation of conditional average treatment effects under a sparsity condition.

span at most 2–3 latent-space dimensions. Even a selection of factors from the zoo (like 8 themed factors from Jensen et al. (2023) added to FF6) struggle to span efficiently the remaining dimensions. A more refined analysis sheds light on the unspanned latent factors which are found to load dynamically on e.g. leverage, time-series reversal and liquidity. (iii) Our conditional risk premia estimates for FF6 factors feature time-variation across market phases, whence the importance of augmenting the three-pass methodology of Giglio and Xiu (2021) with conditioning information. Finally, (iv) the SDF implied by our conditional latent factors outperforms IPCA- and FF6-based models in out-of-sample pricing of 78 long-short anomaly and 49 industry portfolios. Hence, latent dimensions beyond those captured by competing approaches contain relevant pricing information.

A recent strand of the literature models time variation of betas in unobserved factor models via instruments. Connor, Hagmann and Linton (2012) estimate nonparametrically latent factor models in which the betas are functions of covariates. Fan, Liao and Wang (2016) develop the Projected PCA (PPCA) method, which accommodates an additive individual-specific component in the characteristic-based betas. Liao and Yang (2017) and Cheng, Liao and Yang (2023) use PPCA for estimating a time-varying beta model with high-frequency data. Kelly, Pruitt and Su (2017, 2019) introduce Instrumented PCA (IPCA), where the cross-sectional projection of betas onto instruments features time-independent coefficients. Chen (2022) models conditional loadings as linear functions of characteristics with stock specific coefficients. Pelger and Xiong (2022) consider kernel estimation of betas which are functions of a small vector of observable state variables. In Regression PCA (RPCA, Chen, Roussanov and Wang (2023)), asset betas are unknown functions of time-varying instruments, and latent factors are extracted by PCA applied to the Sieve regression coefficients. Our paper differs from those works in that it does not rely on modeling how betas depend on asset characteristics, state variables, or unobservable individual heterogeneities, and in fact embeds those specifications as special cases. Hence, our work is complementary to others as we focus on semiparametric inference for functionals of the conditional latent factors, while other methodologies allow for estimating the time-varying betas. [2]

Besides instrumental variables, another approach to address the time-variation of betas consists in estimating static latent factor models over short time windows. Zaffaroni (2025) develops inferential theory in a setting with n large and T fixed assuming a spherical error structure (error sphericity is a sufficient and necessary condition for consistency of PCA with fixed T, see Bai (2003) and Fortin, Gagliardini and Scaillet (2023a)). Fortin, Gagliardini and

[2]Section 4.4 provides an in-depth discussion of the comparison of our method with IPCA and RPCA. Further, machine learning methods expand beyond the structural framework of linear conditional latent factor models and deploy a large number of covariates in nonparametric modeling (e.g., Freyberger, Neuhierl and Weber (2020), Gu, Kelly and Xiu (2020, 2021), Kelly, Malamud and Zhou (2024)).

Scaillet (2023b, 2025) advocate the use of Factor Analysis to cover the diagonal error structures in stock returns. Unlike those papers which estimate ex-post (i.e. realized) quantities, our large T setting allows for estimation of (conditionally) expected quantities.

The rest of the paper is structured as follows. Section 2 introduces the conditional latent factor model. Section 3 proves identifiability of the general class of parameters of interest. In Section 4 we give asymptotically normal estimators via Neyman-orthogonal moment restrictions with serially dependent data and HAC variance estimators. Section 5 provides our empirical application on the CRSP dataset of U.S. single stocks monthly returns. Section 6 concludes. The proofs of the theoretical results are relegated to the Appendix, Sections A and B. Additional theoretical results and a Monte Carlo analysis are provided in the Online Appendix (OA). We present a detailed description of the dataset and implementation of the ANN estimators in the Supplementary Materials for Coding (SMC).

2 Conditional Latent Factor Model

We consider the following linear latent factor model with time-varying coefficients:

$$y_{i,t} = b'_{i,t-1}\lambda_t + b'_{i,t-1}f_t + \varepsilon_{i,t}, \tag{2.1}$$

where $y_{i,t}$ denotes the return on asset i in period t in excess of the risk free rate, for $i = 1, \cdots, n$ and $t = 1, \cdots, T$. The $k \times 1$ stochastic vector f_t gathers the systematic risk factors and is measurable with respect to filtration $\mathscr{F}_t$. As identifying all risk factors with observable variables alone is debatable, the econometrician treats vector f_t as unobservable, and its dimension k as unknown. We normalize the latent factor process such that $\mathbb{E}(f_t|\mathscr{F}_{t-1}) = 0$ by absorbing the conditional mean in the intercept term. Further, the $k \times 1$ stochastic vector $b_{i,t}$ is measurable with respect to $\mathscr{F}_{i,t}$ i.e. the filtration of the relevant information for asset i, with $\mathscr{F}_t \subset \mathscr{F}_{i,t}$ for all i. The idiosyncratic error terms $\varepsilon_{i,t}$ are such that $\mathbb{E}(\varepsilon_{i,t}|\mathscr{F}_{i,t-1}) = 0$ and $\mathbb{C}\text{ov}(\varepsilon_{i,t}, f_t|\mathscr{F}_{i,t-1}) = 0$, which implies that $b_{i,t-1}$ is the vector of the conditional betas (i.e. risk exposures) for asset i. Moreover, the idiosyncratic errors are weakly cross-sectionally correlated as in an approximate factor structure (Chamberlain and Rothschild (1983)) and satisfy Assumption 4 in Section 4.2. Model (2.1) incorporates the restrictions from absence of arbitrage opportunities, i.e. the conditional expected excess return $b'_{i,t-1}\lambda_t$ is linear in the risk exposures. The $k \times 1$ vector λ_t is $\mathscr{F}_{t-1}$-measurable and yields the conditional equity risk premia. [3] For any integrable variable $W_{i,t}$, we assume that the cross-sectional probability

[3] Gagliardini, Ossola and Scaillet (2016) study the APT in conditional factor models for large economies. They build on the framework of Al-Najjar (1995) with a continuum of assets and extend it to a setting with conditional information to establish the no-arbitrage restriction leading to model (2.1). We postpone

limit $\mathbb{E}^{cs}[W_{\cdot,t}] := \underset{n\to\infty}{\text{plim}} \frac{1}{n}\sum_{i=1}^{n} W_{i,t}$ is measurable with respect to $\mathscr{F}_t$ (if it exists for any t), whence the interpretation of $\mathscr{F}_t$ as the sigma-field of undiversifiable risks.

In our conditional setting, the number of active factors - i.e., those with non-vanishing loadings for a non-negligible fraction of assets - can be time-varying and smaller than k at some dates. We define the number k_t of active common factors at date t as the rank of the cross-sectional second-moment matrix of the loadings:

$$k_t = \text{Rank } \mathbb{E}^{cs}\left[b_{\cdot,t-1}b'_{\cdot,t-1}\right], \tag{2.2}$$

which is $\mathscr{F}_{t-1}$-measurable. Whenever $k_t < k$, the loadings of $k-k_t$ factors are zero for almost all assets at date t, possibly after a $\mathscr{F}_{t-1}$-measurable one-to-one transformation of the factor vector. For the active factors, we denote by f_t^* and λ_t^* the $k_t \times 1$ unobservable vectors of factor values and risk premia for date t.

The paper focuses on time-dependent parameters of interest related to the conditional latent factor space (and moments thereof), that have the general representation

$$\gamma_t = \Psi(f_t^*, \mathbb{E}[F_t|\mathscr{F}_{t-1}], \mathbb{V}[F_t|\mathscr{F}_{t-1}]), \tag{2.3}$$

where vector $F_t = \left((f_t^* + \lambda_t^*)' \ , \ f_t^{O\prime}\right)$ stacks both latent and observed quantities, the latter collected in vector f_t^O, and Ψ is a given function. The next section provides concrete examples of parameters of interest for asset pricing model evaluation that belong to such class.

3 Identification of the Parameters of Interest

We establish the nonparametric identification of the path of latent processes in class (2.3) using large panels of asset returns and predictors.

3.1 Conditional PCA on asymptotic managed portfolios

Our identification strategy for the conditional latent factor space relies on the existence of instrumental variables, i.e. asset-level observable variables which satisfy the next assumption.

Assumption 1. *There exists a $K \times 1$ vector of instrumental variables $w_{i,t-1}$ measurable w.r.t. $\mathscr{F}_{i,t-1}$ such that (i) $\mathbb{E}^{cs}[w_{\cdot,t-1}\varepsilon_{\cdot,t}] = 0$, and (ii) the $K \times k$ matrix $\Gamma_{t-1} := \mathbb{E}^{cs}[w_{\cdot,t-1}b'_{\cdot,t-1}]$ has column rank k_t, for any t.*

to Section 3 the possibility of square-summable pricing errors compatible with APT in countable economies.

Instruments are predetermined variables that have zero correlation with errors and full-rank covariance with the active factor loadings cross-sectionally. Lagged firm characteristics such as size, book-to-market ratio, earnings per share, etc. can provide valid instrumental variables (see Section 5.1 for the choice in our empirics). The necessary order condition is $k_t \leq K$ for all t, i.e. K provides an upper bound for the number of conditional factors.

Latent factor models are observationally equivalent under one-to-one affine transformations of the unobservable factors vector. In our conditional setting, the transformation can be time-varying and predetermined, i.e. function of the information in $\mathscr{F}_{t-1}$. [4] We use instruments to define a suitable normalization which deals with the conditional rotational indeterminacy of latent factors.

Normalization Restriction 1. *Without loss of generality, the following normalization restrictions hold for the latent factors: (i)* $\Gamma_{t-1} = [J_{t-1} \ : \ 0_{K\times(k-k_t)}]$*, for any* t*, where the* $K \times k_t$ *matrix* J_{t-1} *is such that* $(J_{t-1})'J_{t-1} = I_{k_t}$*, and (ii) The upper-left* $k_t \times k_t$ *block* Λ_{t-1} *of the conditional variance* $\mathbb{V}[f_t|\mathscr{F}_{t-1}]$ *is diagonal, with diagonal elements ranked in descending order.*

The interpretation of the normalization restriction in part (i) when $k_t < k$ is that we are adopting a transformation of the unobservable factors vector such that the loadings of the last $k - k_t$ components are zero for almost all assets at date t. This yields the right block of zeros $0_{K\times(k-k_t)}$ in matrix Γ_{t-1}. The ortho-normalization of the left block J_{t-1} is achieved by an additional linear transformation of the first k_t components of the factor vector. The normalization restriction in part (i) leaves free a conditional rotation of the upper $k_t \times 1$ block of factor f_t, which we set by diagonalization in part (ii). Hence, Normalization Restriction 1 is not a constraint on the Data Generating Process (DGP) since this normalization can always be achieved by a conditional transformation of the latent factors. [5]

Define the vector of large-cross-section weighted averages of excess returns:

$$\xi_t = \mathbb{E}^{cs}[w_{\cdot,t-1}y_{\cdot,t}], \tag{3.1}$$

which is measurable w.r.t. $\mathscr{F}_t$, and is identifiable being a function of the observed data distribution when the cross-sectional size n tends to infinity. Vector ξ_t consists of the excess returns of K asymptotic characteristics-based managed portfolios using instruments as

[4] Specifically, model (2.1) can be equivalently written in terms of the transformed conditional factor $R_{t-1}f_t$ and transformed loading $(R'_{t-1})^{-1}b_{i,t-1}$, with conditional risk premia $R_{t-1}\lambda_t$, where R_{t-1} is a $\mathscr{F}_{t-1}$-measurable nonsingular matrix.

[5] Normalization Restriction 1 is similar to the condition used in Gagliardini and Gouriéroux (2017) in static latent factor models.

portfolio weights. Then, equation (2.1) and Assumption 1 imply:

$$\xi_t = \Gamma_{t-1}\,(f_t + \lambda_t) = J_{t-1}(f_t^* + \lambda_t^*). \tag{3.2}$$

i.e. the asymptotic portfolio returns are a conditional transformation of the vector of risk-premia adjusted factors. The conditional variance of vector ξ_t given information set $\mathscr{F}_{t-1}$, under Normalization Restriction 1, is

$$\mathbb{V}(\xi_t|\mathscr{F}_{t-1}) = \Gamma_{t-1}\,\mathbb{V}(f_t|\mathscr{F}_{t-1})\,\Gamma_{t-1}' = J_{t-1}\Lambda_{t-1}J_{t-1}', \tag{3.3}$$

which has rank k_t. Hence, the time-varying number of conditional factors equals

$$k_t = \text{Rank }\mathbb{V}(\xi_t|\mathscr{F}_{t-1}), \tag{3.4}$$

and can be identified from the rank-deficiency of the conditional variance of the asymptotic portfolio returns $\mathbb{V}(\xi_t|\mathscr{F}_{t-1})$. The latter matrix is identifiable when the variables generating the information set of aggregate shocks are observables (Assumption 2 in Section 3.3). Moreover, equation (3.3) is the eigenvalue-eigenvector decomposition of $\mathbb{V}(\xi_t|\mathscr{F}_{t-1})$ and J_{t-1} is the matrix of standardized eigenvectors associated with the k_t non-zero eigenvalues (which is identifiable up to sign changes for the columns). From equation (3.2), Normalization Restriction 1, and condition $\mathbb{E}[f_t|\mathscr{F}_{t-1}] = 0$, the upper $k_t \times 1$ subvector of factor values f_t^* and the corresponding risk premia vector λ_t^* are identified as

$$f_t^* = (J_{t-1})'(\xi_t - \mathbb{E}[\xi_t|\mathscr{F}_{t-1}]), \qquad \lambda_t^* = (J_{t-1})'\mathbb{E}[\xi_t|\mathscr{F}_{t-1}]. \tag{3.5}$$

We refer to them as conditional PCA factors and risk premia as they correspond to the conditional transformations of the asymptotic portfolio return vector ξ_t with maximal conditional variance. From (3.5), vectors f_t^* and λ_t^* are identifiable from vector ξ_t and its conditional moments $\mathbb{E}(\xi_t|\mathscr{F}_{t-1})$ and $\mathbb{V}(\xi_t|\mathscr{F}_{t-1})$. The remaining $k-k_t$ factor values and their risk premia are not identifiable at date t, since they do not load on excess returns.

3.2 Invariance to conditional factor rotations

Since the latent conditional factor space is identified only up to a conditional transformation (fixed by Normalization Restriction 1), the conditional PCA factor values and risk premia in (3.5) do not admit an economic interpretation. In the next subsections, we consider three examples of identifiable features of the conditional factor space, which are invariant to factor rotation and belong to the class of parameters of interest given in (2.3). They are inspired

by measuring conditional risk premia, spanning and pricing errors. [6]

Example 1 — Conditional three-pass Fama-MacBeth regression

In an unconditional model setting, Giglio and Xiu (2021) estimate the risk premium of an observable factor by augmenting the two-pass Fama-MacBeth (FM) regression with a third step, namely the projection of the observable factor onto the latent space of Principal Components. [7] Such projection deals with the omitted variable bias that may plague the first pass of the FM procedure. We extend their ideas with conditioning information. Let f_t^O be a vector of k^O observable factors. Let the conditional regression of f_t^O onto the identifiable latent factors be $f_t^O = a_{t-1}^O + B_{t-1}^O f_t^* + u_t$, where the regression matrix is $B_{t-1}^O = \mathbb{C}\text{ov}(f_t^O, f_t^*|\mathscr{F}_{t-1})\mathbb{V}(f_t^*|\mathscr{F}_{t-1})^{-1}$ and u_t is the regression error. Now, we use $\mathbb{V}(f_t^*|\mathscr{F}_{t-1}) = \Lambda_{t-1}$ and $\mathbb{C}\text{ov}(f_t^O, f_t^*|\mathscr{F}_{t-1}) = \mathbb{C}\text{ov}(f_t^O, \xi_t|\mathscr{F}_{t-1})J_{t-1}$ from (3.5). Then, the conditional risk premia vector of the projection is

$$\lambda_t^O := B_{t-1}^O \lambda_t^* = \mathbb{C}\text{ov}(f_t^O, \xi_t|\mathscr{F}_{t-1})\mathbb{V}(\xi_t|\mathscr{F}_{t-1})^\dagger \mathbb{E}(\xi_t|\mathscr{F}_{t-1}), \tag{3.6}$$

where $\mathbb{V}(\xi_t|\mathscr{F}_{t-1})^\dagger = J_{t-1}(\Lambda_{t-1})^{-1}(J_{t-1})'$ is the rank-k_t pseudo-inverse of matrix $\mathbb{V}(\xi_t|\mathscr{F}_{t-1})$ based on Singular Value Decomposition (SVD). Thus, process λ_t^O is identifiable from first- and second-order conditional moments of portfolio returns ξ_t and observable factors f_t^O. It is independent of the normalization of the latent factor.

Example 2 — Conditional spanning with canonical correlations

To get economic interpretation for the latent factors, one often relies on spanning with observed factors (e.g., Bai and Ng (2006)). In the conditional setting, we use the Conditional Canonical Correlations (CCCs) between latent and observed factors f_t^* and f_t^O to measure the extent to which the spaces spanned by those vectors overlap conditionally on $\mathscr{F}_{t-1}$. The CCCs $\rho_{r,t}$, for $r = 1, 2, \cdots, k_t^*$ (where k_t^* is the minimum between k_t and k^O), are defined analogously to their unconditional counterparts (e.g., Anderson (2003) Chapter 12) by replacing conditional moments for unconditional ones. Namely, $\rho_{1,t}$ is the largest conditional correlation between conditional linear transformations of latent and observed factors given $\mathscr{F}_{t-1}$, $\rho_{2,t}$ is the maximal conditional correlation imposing conditional orthogonality to the first conditional canonical directions, and so on. By analogy with the unconditional setting, the squared CCCs equal the k_t^* largest eigenvalues of matrix

[6] In the OA Section C.5 we present further examples inspired by conditional portfolio choice.

[7] See also Bryzgalova et al (2026) for a related proposal of tradable factor risk premia.

$\mathbb{V}(f_t^O|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t^O, f_t^*|\mathscr{F}_{t-1})\mathbb{V}(f_t^*|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t^*, f_t^O|\mathscr{F}_{t-1})$. Using (3.5) we get:

$$\rho_{r,t} = \delta_r \left(\mathbb{V}(f_t^O|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t^O, \xi_t|\mathscr{F}_{t-1})\mathbb{V}(\xi_t|\mathscr{F}_{t-1})^{\dagger}\mathbb{C}\text{ov}(\xi_t, f_t^O|\mathscr{F}_{t-1})\right)^{1/2}, \tag{3.7}$$

for $r = 1, ..., k_t^*$, where $\delta_r(\cdot)$ denotes the rth largest eigenvalue of a symmetric matrix. Equation (3.7) shows that the conditional canonical correlations are identifiable and independent of the chosen normalization of the latent factors.

Example 3 — Implied stochastic discount factor and conditional pricing errors

A conditional linear multi-factor model for returns corresponds to a specification of the Stochastic Discount Factor (SDF) which is conditionally linear in the factors (e.g. Singleton (2006) Chapter 11). Equation $\mathbb{E}(y_{i,t}|\mathscr{F}_{i,t-1}) = b'_{i,t-1}\lambda_t = (b^*_{i,t-1})'\lambda^*_t$ with $b^*_{i,t-1} = \mathbb{V}(f_t^*|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t^*, y_{i,t}|\mathscr{F}_{i,t-1})$ yields $\mathbb{E}\left(m_{t-1,t}y_{i,t}|\mathscr{F}_{i,t-1}\right) = 0$, where the SDF between dates $t-1$ and t is given by:

$$\begin{aligned} m_{t-1,t} &= R_{f,t}^{-1}\left(1 - (\lambda_t^*)'\mathbb{V}(f_t^*|\mathscr{F}_{t-1})^{-1}f_t^*\right) \\ &= R_{f,t}^{-1}\left(1 - \mathbb{E}(\xi_t|\mathscr{F}_{t-1})'\mathbb{V}(\xi_t|\mathscr{F}_{t-1})^{\dagger}[\xi_t - \mathbb{E}(\xi_t|\mathscr{F}_{t-1})]\right), \end{aligned} \tag{3.8}$$

and $R_{f,t} = \mathbb{E}\left(m_{t-1,t}|\mathscr{F}_{t-1}\right)^{-1}$ is the gross return of the conditionally risk-free asset. [8] While the vectors of factor values and risk premia are identifiable only up to conditional transformations, the SDF $m_{t,t-1}$ is invariant to such conditional transformations. If the vector f_t^O consists of tradable returns, the conditional pricing errors vector $\pi_t^O := \mathbb{E}(m_{t,t-1}(f_t^O - r_{f,t}1_{k^O})|\mathscr{F}_{t-1})$ is given by:

$$\begin{aligned} \pi_t^O &= R_{f,t}^{-1}\left(\mathbb{E}(f_t^O - r_{f,t}1_{k^O}|\mathscr{F}_{t-1}) - \mathbb{C}\text{ov}(f_t^O, \xi_t|\mathscr{F}_{t-1})\mathbb{V}(\xi_t|\mathscr{F}_{t-1})^{\dagger}\mathbb{E}(\xi_t|\mathscr{F}_{t-1})\right) \\ &= R_{f,t}^{-1}\left(\mathbb{E}(f_t^O - r_{f,t}1_{k^O}|\mathscr{F}_{t-1}) - \lambda_t^O\right), \end{aligned} \tag{3.9}$$

where $r_{f,t} = R_{f,t} - 1$ is the conditionally risk-free return, and 1_{k^O} is a $k^O \times 1$ vector of ones.

3.3 Main identification result

Let us summarize the results obtained so far and state them in terms of the identifiability of the general class of time-varying parameters of interest in (2.3). Let vector $X_t = (f_t^{O\prime}, \xi_t')'$

[8] The SDF in (3.8) does not necessarily meet the positivity condition. In a SDF model, the cross-section of conditional expected excess returns is spanned by the conditional betas with respect to a single factor, namely the gross return r_t^* of an asset with payoff equal to the SDF (see e.g. Singleton (2006), Chapter 11, and references therein). We stress that this single-factor structure in expected excess returns does not imply that the return conditional factor space itself is one-dimensional. In fact, the implied error terms $y_{i,t} - \beta^*_{i,t-1}r_t^*$, where $\beta^*_{i,t-1}$ is the conditional beta to r_t^*, are not necessarily weakly correlated across assets.

with dimension $L = k^O + K$ gather observable factors and asymptotic portfolio returns. The latter are identifiable by large cross-sections of asset returns and instruments ($n \to \infty$).

Assumption 2. *We have $\mathbb{E}(X_t|\mathscr{F}_{t-1}) = \mu_0(Z_{t-1})$ and $\mathbb{V}(X_t|\mathscr{F}_{t-1}) = \Sigma_0(Z_{t-1})$, for the observed process $\{Z_t\}$ in $\mathbb{R}^d$, and the unknown vector and symmetric matrix functions $\mu_0(\cdot)$ and $\Sigma_0(\cdot)$. We define $\theta_0(\cdot) = (\mu_0(\cdot)', vech(\Sigma_0(\cdot))')'$. Process $(X_t', Z_{t-1}')'$ is strictly stationary.*

Assumption 2 implies that the conditional mean and covariance functions of X_t given Z_{t-1}, that are the components of the vector-valued functional parameter $\theta_0(\cdot)$, are identifiable from a long time series ($T \to \infty$) of processes Z_t and X_t. The time-dependent parameters of interest consisting of the number of conditional latent factors (3.4), the conditional PCA factors and risk premia (3.5), the risk premia of observed factors projected onto the latent space (3.6), the conditional canonical correlations (3.7), the implied SDF (3.8) and the conditional pricing errors (3.9) are all of the form

$$\gamma_t = \psi(\theta_0(Z_{t-1}), \xi_t, k_t), \tag{3.10}$$

for a known function ψ. The unknown function $\theta_0(\cdot)$ subsumes dependence on the latent space and its conditional moments. We single-out argument k_t to highlight that this integer is used in the SVD of matrix $\Sigma_0(Z_{t-1})$ to get the pseudo inverse $\mathbb{V}(\xi_t|Z_{t-1})^\dagger$. In fact, representation (3.10) applies more generally to parameters of interest of the form (2.3), and Normalization Restriction 1 becomes immaterial when γ_t is invariant to conditional factor rotations. Since function $\theta_0(\cdot)$ and processes ξ_t and k_t are identifiable, we get the next result.

Proposition 1. *Under Assumptions 1 and 2, any stochastic process γ_t in class (2.3), which is invariant to conditional factor transformations, admits the form (3.10) for a known function ψ and, as $n, T \to \infty$, its path is nonparametrically identifiable.*

We next turn to the estimation of the parameters of interest in the class (3.10) and establishing feasible asymptotic normality for unconditional expectations thereof.

4 Asymptotic Normality via Double Machine Learning

The identification strategy developed in the previous section leads to a nonparametric estimation methodology by the analogy principle consisting in replacing population quantities with their sample analogues. Data is splitted in two non-overlapping subpanels in the time series dimension, with $\mathcal{T}$ and $\mathcal{I}$ indicating their sets of time indices, and T and I the corresponding time-series lengths. The estimation of the path of process γ_t works in four steps.

1. The cross-sectional expectation $\xi_t = \mathbb{E}^{cs}[w_{\cdot,t-1}y_{\cdot,t}]$ is estimated by the cross-sectional average:

$$\hat{\xi}_t = \frac{1}{n}\sum_{i=1}^{n} w_{i,t-1}y_{i,t}, \qquad t \in \mathcal{T} \cup \mathcal{I}, \tag{4.1}$$

   i.e. the portfolio excess return using instruments as weights.

2. After replacing ξ_t by $\hat{\xi}_t$, we estimate the time-series conditional moments that define parameter $\theta_0(\cdot)$ in Assumption 2 by nonparametric regression of $\hat{X}_t := (f_t^{O\prime}, \hat{\xi}_t')'$ onto Z_{t-1} in the estimation sample $\mathcal{T}$. We denote by $\hat{\theta}(\cdot) = (\hat{\mu}(\cdot)', vech(\hat{\Sigma}(\cdot))')'$ this nonparametric estimator, and $\hat{\Sigma}_{22}(\cdot)$ the $K \times K$ lower-right block of $\hat{\Sigma}(\cdot)$ which yields the portfolio returns' conditional covariance.

3. We estimate the conditional factor dimension at each date t in the test sample $\mathcal{I}$ as:

$$\hat{k}_t^{\alpha} = \min\left\{ r \geq 0 : \frac{\sum_{j=1}^{r}\delta_j(\hat{\Sigma}_{22}(Z_{t-1}))}{Tr(\hat{\Sigma}_{22}(Z_{t-1}))} \geq 1-\alpha \right\}, \qquad t \in \mathcal{I}, \tag{4.2}$$

   where $\alpha \downarrow 0$ as $n, T \to \infty$, i.e. the first integer r such that the r largest estimated eigenvalues of the conditional variance saturate the trace (i.e. the sum of all eigenvalues) up to a residual percentage $\alpha > 0$ shrinking to zero in the asymptotics.

4. We estimate the time-varying parameter (3.10) by plug-in for the test sample:

$$\hat{\gamma}_t = \psi(\hat{\theta}(Z_{t-1}), \hat{\xi}_t, \hat{k}_t^{\alpha}), \qquad t \in \mathcal{I}. \tag{4.3}$$

Several remarks are in order. (i) The estimator in (4.1) is root-n consistent as $n \to \infty$ under our assumptions by a cross-sectional LLN. To ease exposition, we present the methodology for a balanced panel without pricing errors. With an unbalanced panel, the estimator $\hat{\xi}_t$ is obtained by averaging the returns scaled by instruments over the available cross-section of n_t stocks at each date t. Under a missing-at-random condition, if $n := \min\{n_t,\ t \geq 1\}$ diverges and meets our assumptions almost surely, our results remain valid. Moreover, if the DGP has square summable pricing errors like in APT for countable assets markets with information (e.g. in Zaffaroni (2025)), i.e. $y_{i,t} = a_{i,t-1} + b_{i,t-1}'(\lambda_t + f_t) + \varepsilon_{i,t}$ with $\sum_{i=1}^{n} a_{i,t-1}^2 = O_p(1)$ uniformly in t, and we use bounded instruments like in our empirics, the supplementary term in the portfolio returns $\hat{\xi}_t$ is $O_p(n^{-1/2})$, i.e. the same probability order as cross-sectional averaging of idiosyncratic errors, and is negligible in our asymptotics.

(ii) Our theoretical results are independent of the chosen nonparametric regression method, provided a high-level assumption on consistency rate holds. Let $\|\theta\|_{L^2} = (\int \|\theta(Z)\|^2 dP_0(Z))^{1/2}$

denote the L^2 norm w.r.t. the stationary distribution P_0 of process Z_t, where $\|\cdot\|$ is the Frobenious matrix norm.

Assumption 3. *As $n, T \to \infty$ with $n \geq T^\kappa$ and $\kappa > 0$, the estimator $\hat{\theta}$ is such that $\|\hat{\theta} - \theta_0\|_{L^2} = O_p\left(\left(\frac{\log T}{T}\right)^{\bar{\delta}}\right)$, with $\bar{\delta} > 0$.*

Nonparametric estimation is challenging in a framework with high-dimensional conditioning vector Z_{t-1} due to the curse of dimensionality. In our empirics we use Sieve estimators based on Artificial Neural Networks (ANN), namely

$$\hat{\mu}(\cdot) = \underset{\mu \in \Theta_T^\mu}{\arg\min} \frac{1}{T} \sum_{t \in \mathcal{T}} \|\hat{X}_t - \mu(Z_{t-1})\|^2, \tag{4.4}$$

and

$$\hat{\Sigma}(\cdot) = \underset{\Sigma \in \Theta_T^\Sigma}{\arg\min} \frac{1}{T} \sum_{t \in \mathcal{T}} \left\| (\hat{X}_t - \hat{\mu}(Z_{t-1}))(\hat{X}_t - \hat{\mu}(Z_{t-1}))' - \Sigma(Z_{t-1}) \right\|^2, \tag{4.5}$$

where minimization is over the Sieve spaces $\Theta_T^\mu = \left\{\mu(\cdot) \,:\, \mu_j \in \mathscr{N}_{M_T}^d,\ j = 1, ..., L\right\}$ and $\Theta_T^\Sigma = \left\{\Sigma(\cdot) = C(\cdot)C(\cdot)' \,:\, C(\cdot) \text{ lower triangular matrix},\ C_{j,\ell} \in \mathscr{N}_{M_T}^d,\ j, \ell = 1, ..., L,\ j \geq \ell\right\}$. Here, $\mathscr{N}_{M_T}^d = \left\{\varphi(Z) = \sum_{m=1}^{M_T} v_m \phi(w_m' Z + b_m),\ v_m, b_m \in \mathbb{R},\ w_m \in \mathbb{R}^d,\ m = 1, ..., M_T\right\}$ is the single-layer feedforward network with activation function ϕ and the number of neurons M_T grows with sample size T. Estimator $\hat{\Sigma}(\cdot)$ minimizes a matrix distance criterion and uses the Cholesky decomposition in Θ_T^Σ to parameterize the positive-semidefinite conditional covariance matrix. In the OA Proposition 4 we show that under smoothness conditions on function $\theta_0(\cdot)$, for the optimal choice of M_T and with $\kappa > \frac{1}{2}\frac{d+3}{d+2}$, the ANN Sieve estimator (4.4)-(4.5) meets Assumption 3 with $\bar{\delta} = \frac{1}{4}\frac{d+1}{d+2}$ (see Chen and Shen (1998), Chen and White (1999), Farrel et al. (2021)). The lower bound on κ guarantees that convergence rate $\bar{\delta}$ is unaffected by estimation of ξ_t in Step 1 under our double asymptotics. [9]

(iii) From equation (3.4), the factor dimension k_t equals the number of non-zero eigenvalues of matrix $\Sigma_{22,0}(Z_{t-1}) = \mathbb{V}(\xi_t | \mathscr{F}_{t-1})$. This explains the idea behind estimator (4.2) to exhaust the trace of the conditional variance up to a small tolerance level α. The regularization parameter α deals with the fact that the estimates of the nil eigenvalues are non-zero due to statistical error. In Theorem 2 below we give the conditions on the rate of $\alpha \downarrow 0$ to

[9] Proposition 5 in the OA presents similar results for a Lasso estimator under approximate sparsity of the function $\theta_0(\cdot)$ with dimensionality d possibly diverging in the asymptotics (Belloni et al. (2012)). Convergence rates of kernel smoothing, Sieve (series) and random forests estimators also meet Assumption 3, see e.g. Chen (2007) and Syrgkanis and Zampetakis (2020). Moreover, we can impose positive-definiteness of estimator $\hat{\Sigma}$ by lower capping its eigenvalues at $\epsilon > 0$. Lemma 7 in OA shows that, if $\epsilon = O((\frac{T}{\log T})^{-\bar{\delta}})$, such regularization does not affect the convergence rate. See Filipovic and Schneider (2024) for a recent proposal to impose positive definiteness based on kernel learning.

guarantee consistency of estimator $\hat{k}_t^{\alpha}$ as $n, T \to \infty$ uniformly over the test sample. [10]

(iv) We use sample splitting, namely the nonparametric estimator $\hat{\theta}(\cdot)$ is computed on a separate sample $\mathcal{T}$ that is disjoint from the testing sample $\mathcal{I}$ over which we estimate the path of process γ_t. Sample splitting and cross-fitting are standard in Double Machine Learning to reduce the bias from nuisance parameter estimation, see discussion in Section 4.1.

In the OA Proposition 3 we show that estimator $\hat{\gamma}_t$ is consistent, for any $t \in \mathcal{I}$, and inherits the convergence rate of the nonparametric estimator $\hat{\theta}(\cdot)$, namely $O_p\left(\left(\frac{\log T}{T}\right)^{\bar{\delta}}\right)$, if $\kappa > 2\bar{\delta}$. In fact, the latter condition ensures that the estimation error at order $O_p(n^{-1/2})$ from cross-sectional averaging used to get $\hat{\xi}_t$ is negligible compared to the estimation error from time-series smoothing used by the nonparametric estimator $\hat{\theta}(\cdot)$.

The slow nonparametric convergence rate of $\hat{\theta}$ in high-dimensions, and the scant pointwise distributional theory for machine learning estimators like ANN, make the use of date-by-date estimates $\hat{\gamma}_t$ unreliable in applications with moderate sample sizes (a few hundreds of monthly observations in our case). Luckily, time-series (cross) moments of process γ_t can be estimated at the parametric rate. In the rest of the section we focus on estimators for such time-series moments and establish their feasible asymptotic normality.

4.1 Double/Debiased Machine Learning estimators

Suppose $\mathbb{E}[\gamma_t|Z_{t-1}] = \psi(\theta_0(Z_{t-1}), k_t)$ (modulo renaming function ψ) and consider the finite-dimensional parameter

$$c_0 = \mathbb{E}(W_{t-1}\gamma_t) = \mathbb{E}[W_{t-1}\psi(\theta_0(Z_{t-1}), k_t)], \tag{4.6}$$

where $W_{t-1} = W(Z_{t-1})$ is a given conformable matrix function of Z_{t-1}. This class of functionals of $\theta_0(\cdot)$, and linear or nonlinear transformations thereof, is broad enough to cover the three examples presented in Section 3.2, namely time-series moments of $\gamma_t \in \{\lambda_t^O, \rho_{r,t}, \pi_t^O\}$; we develop details in Section 4.3. It covers also the regression coefficients of γ_t onto functions of Z_{t-1}, namely $\mathbb{E}[W_{t-1}W_{t-1}']^{-1}\mathbb{E}[W_{t-1}\gamma_t']$. We use such regressions to infer the time-variation of conditional risk premia $\gamma_t = \lambda_t^O$ in our empirics, Section 5.3.2, in particular their (counter)

[10] We can also build on the fact that k_t is the rank of $\Sigma_{22,0}(Z_{t-1})$. Statistical tests for the rank of an unknown matrix, such as Cragg and Donald (1996), Robin and Smith (2000), Kleibergen and Paap (2006), Al-Sadoon (2017), rely on the asymptotic normality of the matrix estimator. Such pointwise distributional results are available for kernel regression and series estimators that are infeasible with high-dimensional conditioning vectors, but are more scarce for machine learning estimators. Moreover, when the test matrix is symmetric, the assumptions underlying rank tests may fail (Donald, Fortuna and Pipiras (2007)). We avoid these difficulties by adopting a model selection approach for the inference on k_t. In the OA Proposition 6 we present an alternative consistent selection method that builds on the insight of Ahn and Horenstein (2013) and we adapt their eigenvalue-ratio selection principle to our setting with conditional factors.

cyclicality with W_{t-1} a lagged business cycle indicator. [11]

The plug-in principle of semi-parametric econometrics provides an estimator of c_0 that is asymptotically normal under regularity assumptions (e.g., Newey (1994), Ichimura and Newey (2022)). It relies on establishing the influence function of the target parameter, which accounts for the effect of estimating the nuisance parameter. Building on that, the literature on Double Machine Learning (DML, see e.g. Chernozhukov et al. (2018), Chernozhukov, Newey and Singh (2018, 2022a, b), Chernozhukov et al. (2022)) provide regularity conditions for modified estimators and i.i.d. data, which have wider applicability in high-dimensional settings. DML has two essential ingredients: Neyman orthogonality and cross-fitting, which address two distinct sources of bias. First, with the terminology in Ahrens (2025), "regularization bias" results from estimation error on the nuisance parameter $\theta_0(\cdot)$ which propagates to c_0. Its impact is alleviated by Neyman orthogonality, i.e. using a "locally robust" moment restriction that is first-order insensitive to deviations $\hat{\theta} - \theta_0$. Such orthogonal moment restriction is obtained by adding a first-step influence function to the original one. Second, "overfitting bias" originates when the same data is used for estimating the nuisance parameter and constructing the orthogonality restriction. Cross-fitting eliminates overfitting bias by using different samples for the two tasks. Here we extend DML theory to a serially dependent framework and panel double asymptotics.

To ease notation, we focus on scalar parameter c_0 (we then apply the Cramer-Wold device to get asymptotic normality in the multivariate case). Let us write the gradient of function ψ w.r.t. argument vector θ such that:

$$\frac{\partial \psi(\theta_0, k)}{\partial \theta'}(\theta - \theta_0) = a(\theta_0, k)'(\mu - \mu_0) + Tr[\Omega(\theta_0, k)(\Sigma - \Sigma_0)], \tag{4.7}$$

for given vector and symmetric matrix functions a and Ω. Then, we consider the locally-robust moment restriction

$$\begin{aligned} \mathbb{E}[h_t(\tau)] \;:=\; & \mathbb{E}\left[W_{t-1}\left\{\psi(\theta(Z_{t-1}), k_t) + a(\theta(Z_{t-1}), k_t)'(X_t - \mu(Z_{t-1}))\right.\right. \\ & \left.\left. + Tr\left[\Omega(\theta(Z_{t-1}), k_t)\left((X_t - \mu(Z_{t-1}))(X_t - \mu(Z_{t-1}))' - \Sigma(Z_{t-1})\right)\right]\right\} - c\right] = 0, \end{aligned} \tag{4.8}$$

for parameter $\tau(\cdot) := (c, \theta(\cdot))$. This moment restriction holds true for $c = c_0$ and $\theta(\cdot) = \theta_0(\cdot)$, and has zero vanishing directional derivative w.r.t. $\theta(\cdot)$ at $\theta_0(\cdot)$, namely from (4.7) we have $\frac{d}{du}\mathbb{E}[h_t(c, \theta_0 + u(\theta - \theta_0))]\big|_{u=0} = 0$, for any θ, c, i.e. Neyman orthogonality holds. Hence, the moment restriction (4.8) is insensitive to deviations of θ from θ_0 at first-order. The DML

[11] Equation (4.6) covers quadratic dependence of $\gamma_t = \psi_0(\theta_0(Z_{t-1}), k_t) + \psi_1(\theta_0(Z_{t-1}), k_t)'\xi_t + \xi_t'\Psi_2(\theta_0(Z_{t-1}), k_t)\xi_t$ on ξ_t, such as for the SDF and its square, because $\mathbb{E}[\gamma_t|Z_{t-1}] = \psi(\theta(Z_{t-1}), k_t)$ by suitable definition of ψ.

estimator of c based on the moment restriction (4.8) is:

$$\begin{aligned}\hat{c} &= \frac{1}{I}\sum_{t\in\mathcal{I}} W_{t-1}\Big\{\psi(\hat{\theta}(Z_{t-1}),\hat{k}_t^{\alpha}) + a(\hat{\theta}(Z_{t-1}),\hat{k}_t^{\alpha})'(\hat{X}_t-\hat{\mu}(Z_{t-1})) \\ &\quad + Tr\left[\Omega(\hat{\theta}(Z_{t-1}),\hat{k}_t^{\alpha})\left((\hat{X}_t-\hat{\mu}(Z_{t-1}))(\hat{X}_t-\hat{\mu}(Z_{t-1}))' - \hat{\Sigma}(Z_{t-1})\right)\right]\Big\} =: \frac{1}{I}\sum_{t\in\mathcal{I}}\hat{\varphi}_t, \quad (4.9)\end{aligned}$$

where $\hat{\theta}(\cdot)$ is estimated on subsample $\mathcal{T}$ disjoint from subsample $\mathcal{I}$ used for averaging, and the number of factors $\hat{k}_t^{\alpha}$ is estimated over interval $\mathcal{I}$ according to (4.2). Here we use the simplest form of cross-fitting. More elaborated versions of cross-fitting can be used at the expense of further technical sophistication in the proofs (see Semenova et al. (2023) for cross-fitting with weakly dependent data in a different framework than ours).

4.2 Feasible asymptotic normality

To establish the asymptotic normality of $\hat{c}$, we use the long-run variance-covariance matrix

$$\begin{aligned}\sigma^2 &= \sum_{j=-\infty}^{\infty}\mathbb{C}\text{ov}(h_t,h_{t-j}) = \sum_{j=-\infty}^{\infty}\mathbb{C}\text{ov}(W_{t-1}\gamma_t, W_{t-1-j}\gamma_{t-j}) \\ &\quad + \mathbb{V}(W_{t-1}\eta_t) + \sum_{j=0}^{\infty}\left\{\mathbb{C}\text{ov}(W_{t-1}\gamma_t, W_{t-j-1}\eta_{t-j}) + \mathbb{C}\text{ov}(W_{t-1}\gamma_t, W_{t-j-1}\eta_{t-j})'\right\}, \quad (4.10)\end{aligned}$$

of the orthogonality vector $h_t \equiv h_t(\tau_0) = W_{t-1}(\gamma_t+\eta_t) - c_0$, where process $\eta_t := a(\theta_0(Z_{t-1}),k_t)'U_t + Tr\left[\Omega(\theta_0(Z_{t-1}),k_t)\left(U_tU_t' - \Sigma_0(Z_{t-1})\right)\right]$ with $U_t = X_t - \mu_0(Z_{t-1})$ is a martingale difference sequence. The first term in the RHS of (4.10) is the long-run variance for estimating the expectation of γ_t with known $\theta_0(\cdot)$, the second term is a contribution of the estimation error for the conditional moments, and the third term is the interaction between the first two. We estimate σ^2 with the HAC estimator

$$\hat{\sigma}^2 = \sum_{j=-I+1}^{I-1} K(j/b_I)\hat{\Gamma}_j, \qquad \hat{\Gamma}_j = \frac{1}{I}\sum_{t=j+1}^{I}\hat{h}_t\hat{h}_{t-j}' \quad (4.11)$$

for kernel function $K(\cdot)$ and bandwidth (or lag truncation parameter) $b_I > 0$, where $\hat{h}_t := \hat{\varphi}_t - \hat{c}$, see equation (4.9). Moreover, we use the following regularity conditions.

Assumption 4. *The errors and the conditional betas satisfy* $\sup_{t\geq 1}\mathbb{E}\left[\|\frac{1}{\sqrt{n}}\sum_{i=1}^n w_{i,t-1}\varepsilon_{i,t}\|^2\right] = O(1)$ *and* $\sup_{t\geq 1}\mathbb{E}\left(\|\frac{1}{\sqrt{n}}\sum_{i=1}^n[w_{i,t-1}b_{i,t-1}' - \mathbb{E}(w_{i,t-1}b_{i,t-1}'|\mathscr{F}_t)]\|^{2r}\right) = O(1)$, *for* $r > 1$.

Assumption 5. *The process* $\{Y_t = (X_t', Z_{t-1}')'\}$ *with* $X_t = (f_t^{O\prime}, \xi_t')'$ *is strictly stationary and* β*-mixing, with* β*-mixing coefficients* $\beta(j) = \mathbb{E}\left[\sup\left\{|\mathbb{P}(B|\mathscr{Y}_{-\infty}^{t}) - \mathbb{P}(B)| : B \in \mathscr{Y}_{t+j}^{\infty}\right\}\right]$, $j \in \mathbb{N}$, *such that* $\beta(j) \leq Ce^{-\bar{m}j}$, *for* $\bar{m} > 0$ *and* $C > 0$, *where* $\mathscr{Y}_t^s = \sigma(Y_u : t \leq u \leq s)$.

Assumption 6. *We have* $\mathbb{E}(\|X_t\|^{2\bar{r}}) < \infty$, *for* $\bar{r} > \max\{\frac{r}{r-1}, 2\}$, *and* $\mathbb{E}[\|X_t\|^4|Z_{t-1}] \leq C$, $\mathbb{P}$*-a.s., for a constant* C.

Assumption 7. *(i) There exists a constant* $\underline{c} > 0$ *such that* $\delta_1(\Sigma_{22,0}(Z_{t-1})) \geq \underline{c}$, $\mathbb{P}$*-a.s. (ii) We have* $\mathbb{P}\left(T^{\bar{\beta}}\delta_{k_t}(\Sigma_{22,0}(Z_{t-1})) \leq 1/a\right) \leq e^{-\bar{c}a}$, *for some constants* $\bar{\beta} \geq 0$ *and* $\bar{c} > 0$, *such that* $\bar{\beta} < \bar{\delta}$, *and for any* T, a *large.*

Assumption 8. *Function* ψ *is such that: (i)* $\int \|W(z)[\psi(\theta(z), k(z)) - \psi(\theta_0(z), k(z))]\|^2 dP_0(z) \leq C\|\theta - \theta_0\|_{L^2}^{\bar{a}}$, *for* $\bar{a} > 0$, *where* $k(Z_{t-1}) = k_t$, *and (ii) we have*

$$\begin{aligned}\| \int W(z) \left[\psi(\theta(z), k(z)) - \psi(\theta_0(z), k(z)) - a(\theta_0(z), k(z))'(\mu(z) - \mu_0(z))\right.\\ \left. -Tr\{\Omega(\theta_0(z), k(z))(\Sigma(z) - \Sigma_0(z))\}\right] dP_0(z)\| \leq C\|\theta - \theta_0\|_{L^2}^2.\end{aligned}$$

Moreover, the vector function $\alpha(\theta, k) = (a(\theta, k)', vec(\Omega(\theta, k))')'$ *is (iii) bounded, and such that (iv)* $\int \|\alpha(\theta(z), k(z)) - \alpha(\theta_0(z), k(z))\|^2 dP_0(z) \leq C\|\theta - \theta_0\|_{L^2}^2$, *for* θ *in a neighborhood of* θ_0, *and a constant* $C > 0$. *Finally, (v)* $E[\|W_{t-1}\psi(\theta_0(Z_{t-1}), k_t)\|^{\bar{r}}] < \infty$.

Assumption 9. *The estimation sample* $\mathcal{T}$ *and the testing sample* $\mathcal{I}$ *are disjoint and such that* $\min_{\tau \in \mathcal{T}, t \in \mathcal{I}} |\tau - t| \to \infty$.

Assumption 10. *The Kernel function* $K(\cdot) : \mathbb{R} \to [-1, 1]$ *is such that* $K(0) = 1$, $K(x) = K(-x)$ *for all* $x \in \mathbb{R}$, *is continuous at* 0 *and at almost all* $x \in \mathbb{R}$, $\int_{-\infty}^{\infty} |K(x)|dx < \infty$ *and* $\phi(\xi) \geq 0$ *for all* $\xi \in \mathbb{R}$, *where* $\phi(\xi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} K(x)e^{-i\xi x}dx$.

Assumption 11. *The bandwidth* $b_I > 0$ *is such that* $b_I^{-1} + b_I I^{-1} \to 0$ *as* $I \to \infty$.

Assumption 4 constrains the cross-sectional dependence of error terms, and of the unsystematic part in assets betas. It can be established with more primitive conditions, such as block-dependence or mixing dependence in the cross-section. We use Assumption 4 to show the root-n convergence of portfolio returns to their asymptotic counterparts as $n \to \infty$. Assumption 5 constrains the time-series dependence of the asymptotic portfolio returns ξ_t, observable factors f_t^O and predictors Z_{t-1}, by requiring that the joint process is β-mixing (absolutely regular). We can relax the condition of geometrically decaying mixing coefficients at the cost of further technicalities in the proofs. Assumption 6 requires existence of higher-order (conditional) moments for process X_t. Conditions such as those in Assumptions 5 and 6 are standard to establish CLT in time-series. Here we need moments beyond

the fourth order because unknown function $\theta_0(\cdot)$ involves the conditional variance of X_t and data are serially dependent. Assumption 7 deals with factor strength conditional on Z_{t-1}. In Part (i), the largest eigenvalue of the portfolio conditional variance $\Sigma_{22,0}(\cdot)$ is bounded away from 0 almost surely, i.e. there is at least one strong factor uniformly across conditioning environments. In Assumption 7 (ii), the stationary distribution of the smallest non-zero eigenvalue $\delta_{k_t}(\Sigma_{22,0}(Z_{t-1}))$ of the conditional covariance is allowed to admit values near 0 to mimic factors that are "weak" in some conditioning environments. If $\bar{\beta} = 0$, Assumption 7 (ii) imposes exponential left tail for such distribution. With $\bar{\beta} > 0$ we allow the smallest eigenvalue to shrink to zero with the sample size, i.e. the factor is "weak" in all conditional environments. We need the bound $\bar{\beta} < \bar{\delta}$, i.e. the decay rate of the smallest eigenvalue is smaller than the non-parametric rate of estimation error, to have the signal dominating the noise in the smallest eigenvalue with probability approaching (w.p.a.) 1. We use Assumption 7 to show consistency of the estimator of the number of conditional factors. Assumption 8 provides smoothness conditions on function ψ defining the parameter of interest. In particular, Assumptions 8 (i)-(ii) yield Lipshitz-type conditions for ψ w.r.t. functional argument $\theta(\cdot)$. They are used to control the effect on γ_t of the estimation error for $\theta_0(\cdot)$. In Assumption 9, the distance between estimation and testing samples grows in the asymptotics to guarantee vanishing effects of the former when computing probabilities for statistics in the latter sample. The growth rate can be as slow as desired without affecting validity of the next theorem (see Lemma 1 in Appendix A for more details). Assumption 10 requires that function $K(\cdot)$ is in the class $\mathcal{K}_2$ of positive definite kernels defined in Andrews (1991), which ensures that $\hat{\sigma}^2$ is positive definite with probability 1. Finally, under Assumption 11 the bandwidth b_I diverges to infinity less fast than the testing sample size I. Assumptions 10 and 11 imply Assumptions 1 and 3 in De Jong and Davidson (2000).

Theorem 2. *Under Assumptions 1-9, and if* $n, T, I \to \infty$ *such that* $n \geq T^{\kappa}$, $I \leq T^{\chi}$ *with* $0 < \chi < \min\{2(\bar{\delta}-\bar{\beta}), \kappa\}$, *and if* $\frac{T^{-\bar{\beta}}}{(\log T)^{1/\bar{c}}} \gg \alpha \gg \sqrt{I}(\frac{T}{\log T})^{-\bar{\delta}}$, *we have (a)* $\mathbb{P}\left[\hat{k}_t^{\alpha} = k_t,\ \forall t \in \mathcal{I}\right] \to 1$ *and (b)* $\sqrt{I}(\hat{c} - c) \Rightarrow N(0, \sigma^2)$. *If additionally* $\chi < \kappa/2$ *and Assumptions 10-11 hold, (c)* $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$.

Under Theorem 2 (a), estimator $\hat{k}_t^{\alpha}$ in (4.2) selects the correct number of conditional factors across interval $\mathcal{I}$ w.p.a. 1. The condition $\alpha \gg \sqrt{I}(\frac{T}{\log T})^{-\bar{\delta}}$ implies that the regularization parameter α dominates over the nonparametric estimation error for nil eigenvalues accumulated over interval $\mathcal{I}$. The condition $\frac{T^{-\bar{\beta}}}{(\log T)^{1/\bar{c}}} \gg \alpha$ guarantees that regularization does not mask the smallest non-zero eigenvalue of the true conditional variance $\Sigma_{22,0}(Z_{t-1})$. In Theorem 2 (b), the asymptotic normality of estimator $\hat{c}$ in (4.9) holds under the general consistency rate condition for the nonparametric estimator $\hat{\theta}(\cdot)$ in Assumption 3. That

generality w.r.t. the nonparametric estimator to plug-in is an advantage of the DML methodology compared to alternative approaches for bias reduction. Theorem 2 covers the case of an expanding testing interval, i.e. $I \to \infty$. We need $I \leq T^{\chi}$ with inequalities $\chi < 2(\bar{\delta} - \bar{\beta})$ and $\chi < \kappa$. The former inequality is to meet the conditions on sequence α. The latter one implies $I/n = o(1)$ and yields no asymptotic effects from estimating ξ_t. When $\bar{\delta} > 1/4$ (as for the ANN Sieve estimator in any dimension $d \geq 1$), and with $n \geq T$ as in our empirical setting, a sufficient condition for the relative growth rates of the sample sizes I and T is $\chi \leq 1/2$ (if $\bar{\beta} = 0$). This condition guarantees that second-order terms in the deviation of $\hat{\theta}(\cdot)$ from $\theta_0(\cdot)$ do not impact the asymptotic distribution. In fact, with known number of latent factors, the weaker condition $\chi < 4\bar{\delta}$ would be enough to get $\sqrt{I}\|\hat{\theta} - \theta_0\|_{L^2}^2 = o_p(1)$. Finally, Theorem 2 (c) shows that the HAC estimator $\hat{\sigma}^2$ is consistent. The proof follows the ideas of De Jong and Davidson (2000) but the arguments need to be substantially expanded because our HAC estimator involves the plug-in of the nonparametric estimator $\hat{\theta}(\cdot)$ with slow convergence rate, as well as of the estimates $\hat{\xi}_t$ and $\hat{k}_t^{\alpha}$ instead of true values ξ_t and k_t.

4.3 Three examples

Let us provide the DML estimators for the moments of processes defined in equations (3.6), (3.7) and (3.9), i.e. Examples 1-3 in Section 3.2. The functions ψ yielding representation (4.6) are given in Table 1, second column. Prior to DML adjustment, the conditional risk premia estimator for the projected observable factors is $\hat{\lambda}_t^O = \hat{\Sigma}_{12}(Z_{t-1})\hat{\Sigma}_{22}(Z_{t-1})^{\dagger}\hat{\mu}_2(Z_{t-1})$, the rth CCC estimator is $\hat{\rho}_{r,t} = \delta_r \left(\hat{\Sigma}_{11}(Z_{t-1})^{-1}\hat{\Sigma}_{12}(Z_{t-1})\hat{\Sigma}_{22}(Z_{t-1})^{\dagger}\hat{\Sigma}_{21}(Z_{t-1})\right)^{1/2}$, and the implied SDF conditional pricing error estimator is $\hat{\pi}_t^O = R_{f,t}^{-1}\left(\hat{\mu}_1(Z_{t-1}) - r_{f,t}1_{k^O} - \hat{\lambda}_t^O\right)$. Blocks $\hat{\mu}_j(\cdot)$ and $\hat{\Sigma}_{j,\ell}(\cdot)$ for $j,\ell = 1,2$ of the nonparametric conditional mean $\hat{\mu}(\cdot)$ and variance $\hat{\Sigma}(\cdot)$ estimates correspond to the partitioning of $X_t = (f_t^{O\prime}, \xi_t')'$, and the pseudo-inverse $\hat{\Sigma}_{22}(Z_{t-1})^{\dagger}$ is computed with the SVD of matrix $\hat{\Sigma}_{22}(Z_{t-1})$ for rank equal to $\hat{k}_t^{\alpha}$.

In the third column of Table 1 we provide the process η_t that is needed for the locally-robust orthogonality restrictions in the three examples (see Appendix B for derivations). To illustrate, let us focus on estimating the expected rth conditional canonical correlation $c_0 = \mathbb{E}[\rho_{r,t}]$. Then, equation (4.7) holds with $a(\theta, k) = 0$ and

$$\Omega(\theta, k) = -\frac{v_r' v_r}{2\rho_r}\begin{pmatrix} \rho_r^2 Q_r & -Q_r\Sigma_{12}\Sigma_{22}^{\dagger} \\ -\Sigma_{22}^{\dagger}\Sigma_{21}Q_r & \Sigma_{22}^{\dagger}\Sigma_{21}Q_r\Sigma_{12}\Sigma_{22}^{\dagger} \end{pmatrix} \tag{4.12}$$

where $Q_r = v_r(v_r'v_r)^{-1}v_r'$ is the rth eigenprojector of $\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{\dagger}\Sigma_{21}$ associated with eigenvalue ρ_r^2, and eigenvector v_r is normalized to have $v_r'\Sigma_{11}v_r = 1$. We have $Tr[\Omega(\theta_0, k)\Sigma_0] = 0$,

because function ψ is invariant to multiplication of its argument Σ by a positive scalar (it applies to the other two examples as well). Moreover, we have:

$$\eta_t = U_t'\Omega(\theta_0(Z_{t-1}), k_t)U_t = \frac{1}{2\rho_{r,t}}\left[(1-\rho_{r,t}^2)(v_{r,t}'U_{1,t})^2 - (v_{r,t}'\tilde{U}_{1,t})^2\right], \tag{4.13}$$

where $\tilde{U}_{1,t} = U_{1,t} - \Sigma_{0,12}(Z_{t-1})(\Sigma_{22,0}(Z_{t-1}))^\dagger U_{2,t}$ is the residual of the conditional regression of $U_{1,t}$ onto $U_{2,t}$, and vector $U_t = (U_{1,t}', U_{2,t}')'$ is partitioned into the k^O- and K-dimensional components. Then, the DML estimator of the expected rth canonical correlation is $\hat{c} = \frac{1}{I}\sum_{t\in\mathcal{I}}\left\{\hat{\rho}_{r,t} + \frac{1}{2\hat{\rho}_{r,t}}\left[(1-\hat{\rho}_{r,t}^2)(\hat{v}_{r,t}'\hat{U}_{1,t})^2 - (\hat{v}_{r,t}'\hat{\tilde{U}}_{1,t})^2\right]\right\}$, where $\hat{U}_t = \hat{X}_t - \hat{\mu}(Z_{t-1})$, and $\hat{v}_{r,t}$ and $\hat{\tilde{U}}_{1,t}$ are defined in terms of $\hat{\Sigma}(Z_{t-1})$. From Theorem 2, the variance of its asymptotic normal distribution is $\sigma^2 = \sum_{j=-\infty}^{\infty}\mathbb{C}\text{ov}(\rho_{r,t}, \rho_{r,t-j}) + \mathbb{V}(\eta_t) + 2\sum_{j=0}^{\infty}\mathbb{C}\text{ov}(\rho_{r,t}, \eta_{t-j})$. [12]

Table 1: **DML Orthogonality Restrictions for Three Parameters of Interest**

c_0	$\psi(\theta, k)$	η_t
$\mathbb{E}(W_{t-1}\lambda_t^O)$	$\Sigma_{12}\Sigma_{22}^\dagger\mu_2$	$U_{1,t} - \tilde{U}_{1,t}R_{f,t}m_{t-1,t}$
$\mathbb{E}(W_{t-1}\rho_{r,t})$	$\delta_r(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^\dagger\Sigma_{21})^{1/2}$	$\frac{1}{2\rho_{r,t}}\left[(1-\rho_{r,t}^2)(v_{r,t}'U_{1,t})^2 - (v_{r,t}'\tilde{U}_{1,t})^2\right]$
$\mathbb{E}(W_{t-1}\pi_t^O)$	$\mu_1 - \Sigma_{12}\Sigma_{22}^\dagger\mu_2$	$\tilde{U}_{1,t}m_{t-1,t}$

The first column reports the parameter of interest c_0 for three cases: the weighted expectation of conditional risk premium, rth conditional canonical correlation, and conditional pricing error. The second column gives the function ψ such that $c_0 = \mathbb{E}[W_{t-1}\psi(\theta_0(Z_{t-1}), k_t)]$, except for the last row, where $c_0 = \mathbb{E}[W_{t-1}R_{f,t}^{-1}(\psi(\theta_0(Z_{t-1}), k_t) - r_{f,t})]$. The third column gives the adjustment term η_t that yields the doubly-robust orthogonality restriction $\mathbb{E}[W_{t-1}(\psi(\theta_0(Z_{t-1}), k_t) + \eta_t) - c_0] = 0$. Processes $m_{t,t-1}$, $U_{1,t}$, $\tilde{U}_{1,t}$, and $v_{r,t}$ are defined in the text.

4.4 Comparison of DML with IPCA and RPCA

The core of Instrumented PCA (IPCA, Kelly, Pruitt and Su (2007), (2009)) is the assumption that the cross-sectional regression of betas onto instruments features time-constant coefficients, namely $b_{i,t} = \mathcal{B}'w_{i,t} + u_{i,t}$, where matrix $\mathcal{B}$ does not depend on date t and the regression residual $u_{i,t}$ is orthogonal to $w_{i,t}$. Then, we have $\Gamma_{t-1} = \mathbb{E}^{cs}[w_{i,t-1}w_{i,t-1}']\mathcal{B}$, and equations (3.1) and (3.2) imply

$$\mathbb{E}^{cs}[w_{\cdot,t-1}w_{\cdot,t-1}']^{-1}\mathbb{E}^{cs}[w_{\cdot,t-1}y_{\cdot,t}] = \mathcal{B}(f_t + \lambda_t). \tag{4.14}$$

[12] When X_t is conditionally Gaussian, and the conditional canonical correlations are time constant, we have $\sigma^2 = \mathbb{V}(\eta_t) = (1-\rho_r^2)^2$, and we recover the asymptotic variance for estimating unconditional canonical correlations with Gaussian or elliptical data (see Bai and Ng (2006) for application in factor analysis).

Thus, in IPCA the cross-sectional regression coefficient vector $\mathbb{E}^{cs}[w_{\cdot,t-1}w'_{\cdot,t-1}]^{-1}\mathbb{E}^{cs}[w_{\cdot,t-1}y_{\cdot,t}]$ of asset returns onto instruments is a linear time-independent transformation of shifted latent factor $f_t + \lambda_t$. In Regression PCA (RPCA), Chen, Roussanov and Wang (2023)) use a Sieve approximation $b_{i,t} = \mathcal{B}'\varphi(w_{i,t-1}) + u_{i,t}$ where $\varphi(\cdot)$ is the vector stacking J additive basis functions, and the Sieve approximation error $u_{i,t}$ shrinks to zero as J grows. Then, equation (4.14) applies replacing $w_{i,t-1}$ with $\varphi(w_{i,t-1})$. Equation (4.14) suggests a simple way to consistently estimate the IPCA model with k factors: first, regress in cross-section asset returns onto instruments by OLS; second, apply standard PCA onto the time series of OLS coefficients and extract the first k Principal Components. In fact, Chen, Roussanov and Wang (2023) advocates this procedure for RPCA on Sieve regression coefficients. Compared to the procedure outlined at the beginning of this section involving the nonparametric estimator $\hat{\theta}(\cdot)$, in RPCA an unconditional variance-covariance matrix is used instead of a conditional one. This simplification is allowed by the time-constancy assumption on matrix $\mathcal{B}$. However, this is a restrictive assumption on the DGP. For instance, the model $b_{i,t-1} = B_i Z_{t-1} + C_i w_{i,t-1}$ with asset-specific coefficients B_i, C_i, that is often employed in empirical specifications (see Gagliardini, Ossola and Scaillet (2016) and references therein) may be incompatible with the assumptions of IPCA or RPCA, while it is coherent with our Assumption 1.

5 Empirical Analysis

5.1 Data

We conduct our empirical analysis on a large unbalanced panel of U.S. equity monthly returns. The sample consists of all common stocks (share codes 10 and 11) listed on the NYSE, AMEX, or NASDAQ from July 1964 through December 2023, a span of 714 months. Delisting returns are incorporated to mitigate survivorship bias (Shumway and Warther (1999)). We get excess returns by subtracting the one-month Treasury bill yield. We obtain firm characteristics from the Open Asset Pricing database (Chen and Zimmermann (2022)), which provides 212 U.S. firm-level predictors, and merge these with monthly stock returns from CRSP to build portfolio returns. To ensure adequate time-series and cross-sectional coverage and satisfy the theoretical requirements for estimation, we retain only $K = 78$ characteristics with sufficiently long non-missing histories across firms (we list them in SMC S.1.1). Following Kelly et al. (2019), we rank-normalize firm characteristics each month to $[-0.5, 0.5]$. At each month t, the characteristics–based portfolio return vector $\hat{\xi}_t$ is computed as the cross–sectional mean of individual stock excess returns times lagged characteristics.

Figure 1 illustrates the cross–sectional dimension of the sample. Light/dark blue bars

display the minimum/maximum number of characteristic–return observations $(w_{i,t-1}, y_{i,t})$ across portfolios (average by year). The max is essentially determined by the number of firms with non–missing returns in a given month, which remains large throughout, ranging between a minimum of about $2,000$ stocks at the beginning of the sample and a maximum of more than $7,000$ before the tech bubble. The 78 characteristics are chosen such that, in every year, each portfolio contains on average at least 500 observations (see the min bars in Figure 1), ensuring that the large–n requirement in Propositions 1 and 2 is satisfied.

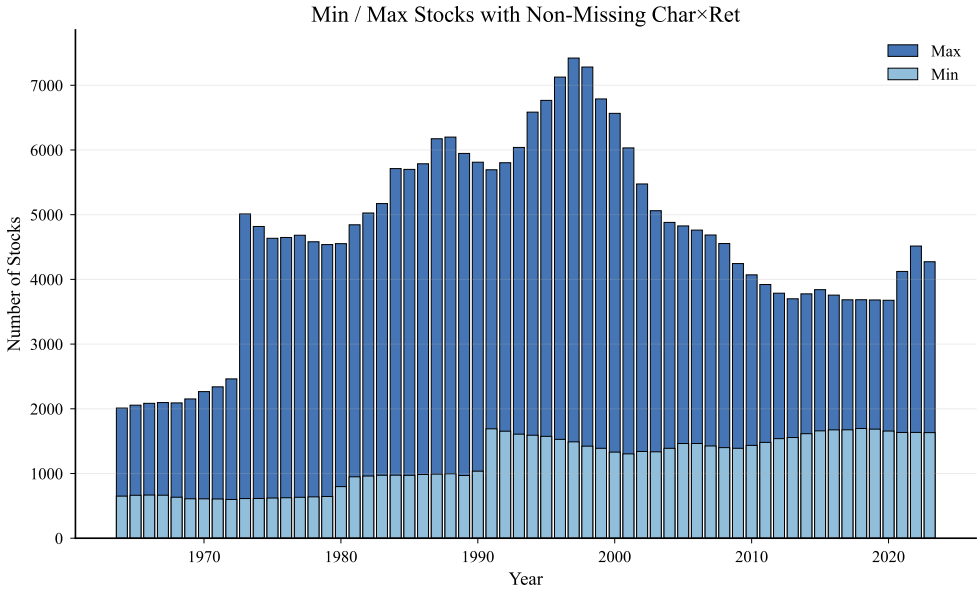

Figure 1: **Cross-sectional sample size over time.** This figure shows the minimum and maximum number of stocks across characteristic-sorted portfolios (averaged by year), where characteristics are lagged one month relative to returns. In particular, from the dark blue bars we read the number of stocks with non-missing return in each cross-section (averaged by year).

We proxy the conditional information vector Z_t with a multivariate process that captures macroeconomic and financial shocks. Our analysis utilizes 25 monthly financial indicators based on the latest collection of predictors compiled by Goyal, Welch and Zafirov (2024), listed in the SMC S.1.2. To further enrich the conditioning information set and capture possible non-Markov persistent dynamics in predictors, we include four lags of both the Goyal, Welch and Zafirov (2024) financial indicators and portfolio return vector $\hat{\xi}_t$. This choice leads to a total of $d = 412$ conditioning variables in vector Z_t.

5.2 Conditional moments estimation via ANN

Once we obtain the portfolio returns $\hat{\xi}_t$, we estimate the elements of vector function $\theta_0(\cdot)$, i.e. the conditional mean $\mu_0(\cdot)$ and conditional variance $\Sigma_0(\cdot)$, by nonparametric regression of vector $\hat{X}_t = (f_t^{O\prime}, \hat{\xi}_t')'$ onto lagged predictor Z_{t-1}. In Section 5.3 we use various combinations

of observable factors, the most extensive one includes the five Fama-French factors, momentum and a selection of eight factors from Jensen, Kelly and Pedersen (2023) that we collect in vector f_t^O of size $k^O = 14$. We use ANN–sieve estimators (see Section 4), and implement feedforward architectures with rectified linear activation (ReLU) and a validated number of hidden layers between 1 and 5. We train them by minimizing mean squared error using the Adam optimizer. To guard against overfitting, early stopping is applied with a patience of 10 epochs. This setup provides a flexible yet computationally tractable specification of $\hat{\mu}(\cdot)$ and $\hat{\Sigma}(\cdot)$ in the expanding–window design described below. Details regarding the configuration and the implementation of the estimation are provided in the SMC S.2.

i) Sample splitting and expanding window strategy. Estimation relies on an expanding–window scheme tailored to the time–series setting. The sample is initially partitioned into three consecutive blocks: a 60–month training set, a subsequent 120–month validation set, and the remaining observations reserved for testing. At each step, the training window expands by one month, the validation window length is held fixed at 10 years, and the subsequent observation is used as the test case. Hyperparameters are selected on the basis of validation performance, after which the chosen network is re–estimated on the combined training and validation sample and employed to produce the one–month–ahead forecast. The minimal sample length for estimation is $T = 180$ months (training + validation) in the initial split, after which the effective sample size grows up to a maximum length of $T = 593$ months for the last window. This expanding design ensures that the implementation aligns with the large–T requirement of the asymptotic theory.

ii) Conditional mean and variance-covariance. Each coordinate $\hat{X}_{t,\ell}$ of $\hat{X}_t$ is modeled separately by an ANN for $\ell = 1, \ldots, L$, where $L = K + k^O = 92$. Stacking the out-of-sample forecasts yields $\widehat{\mu}(Z_{t-1})$, and residuals are defined as $\tilde{X}_{t,\ell} = \hat{X}_{t,\ell} - \widehat{\mu}_\ell(Z_{t-1})$. For $\ell \geq m$, we estimate each conditional covariance entry $\Sigma_{\ell m}(\cdot)$ via an ANN applied to $\tilde{X}_{t,\ell}\tilde{X}_{t,m}$ with predictor Z_{t-1}, and assemble $\widehat{\Sigma}(Z_{t-1}) = \big[\widehat{\Sigma}_{\ell m}(Z_{t-1})\big]_{\ell,m=1}^{L}$. Because residuals rely on out-of-sample mean forecasts, variance estimation excludes the initial training (5 years) and validation (10 years) periods; thus evaluation spans July 1994–December 2023.[13]

5.3 Conditional asset pricing estimates

Our empirical analysis focuses on the number of conditional latent factors and three target parameters inspired by Examples 1-3 (see Sections 3.2 and 4.3).

[13]We also implemented a version of the ANN estimator imposing the positive-semidefinite constraint via the Sieve space like in (4.5). In our Monte Carlo experiments, we observed better performance of the unrestricted and regularized estimator described in this section, which was then used in the empirics.

5.3.1 Number of conditional latent factors and its counter-cyclical patterns

We implement the estimator $\hat{k}_t^{\alpha}$ introduced in Section 4, which selects the smallest number of eigenvalues of the estimated conditional covariance matrix $\hat{\Sigma}_{22}(Z_{t-1})$ required to explain at least a fraction $1-\alpha$ of the total conditional variance. Factor dimensionality is determined entirely out-of-sample, as $\hat{\Sigma}_{22}(Z_{t-1})$ itself is obtained from out-of-sample forecasts. In Figure 2 we report results for tolerance levels $\alpha = 0.10, 0.20, 0.30$, corresponding to target variance coverage of 90%, 80%, and 70%. We display median values of the estimated number of conditional factors by year. Shaded areas in Figure 2 mark bear-market years as identified by the bull–bear dating algorithm of Lunde and Timmermann (2004), which detects alternating trough-to-peak and peak-to-trough phases in cumulative market returns.

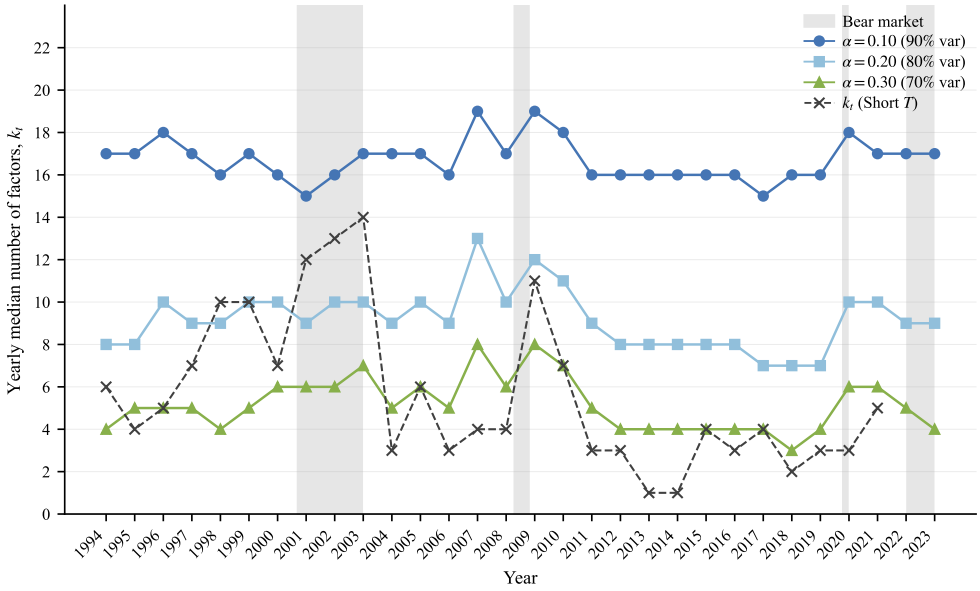

Figure 2: **Yearly median number of conditional factors**. This figure plots the yearly median of the estimated number of conditional latent factors $\hat{k}_t^{\alpha}$ (solid lines), at thresholds $\alpha = 0.10, 0.20, 0.30$ (corresponding to $90\%, 80\%$ and 70% target variance levels). The estimates are computed out-of-sample on the test period only. The dashed line displays the yearly estimate of the number of latent factors obtained in Fortin, Gagliardini and Scaillet (2023b) by fixed-T Factor Analysis. Shaded areas correspond to bear-market years according to Lunde and Timmermann (2004).

Figure 2 highlights two main findings. First, a rather large number of conditional factors is needed to exhaust pervasive risks, e.g., between 7 and 13 factors to capture 80% of the estimated variance of the systematic component. This dimensionality is larger than the number of observable factors typically used in empirical asset pricing, such as the five factors of Fama and French (2015) or the four factors in the models of Carhart (1997) and Hou, Xue and Zhang (2015). It is not due to scaled factors because our methodology captures conditional factors. Second, the estimated dimension of the conditional factor space exhibits counter-cyclical patterns, reflecting shifts in the complexity of the return covariance

structure. The number of conditional factors rises during market downturns, while we see a marked decrease in the decade 2010-2020 between the end of the great financial crisis and the Covid-19 period, after which k_t peaks again. This is compatible with the possibility that some risk factors related to financial distress manifest in bear periods. For comparison purpose, the black dashed line in Figure 2 displays the number of latent factors estimated by the multiple-testing procedure of Fortin, Gagliardini and Scaillet (2023b) that is based on short-T Factor Analysis. While the estimated paths in the two methods do not overlap, our main findings resonate with those in Fortin, Gagliardini and Scaillet (2023b) in pointing to a large and counter-cyclical number of conditional factors. In the rest of the section we select the number of conditional factors with $\alpha = 0.10$. Values of α like 0.10 or larger yield quite good and stable performances across designs in our Monte Carlo experiments in OA Section E. The main empirical findings are robust to the choice of α in range $0.10 - 0.30$.

5.3.2 Conditional risk premia of projected observable factors

In Table 2 we report the estimates for the mean conditional risk premia of the five Fama-French factors and Momentum after projection on our latent space obtained by the DML procedure presented in Sections 4.2-4.3. We provide results for the full test sample (July 1995 – December 2023), for some selected post-crisis periods, as well as for subperiods obtained from the bear-vs-bull and NBER recession-vs-expansion classifications.

Our full–sample estimates are in line with the unconditional risk premia estimates obtained with the Giglio and Xiu (2021) three-pass methodology (see first and last rows in Table 2) albeit the market risk premium estimate with our methodology, equal to 0.54 (monthly %), is substantially larger. However, the sub-period estimates show time-variation. For instance, the mean estimate of market risk premium is double during recessions compared to expansion periods. The difference $0.97 - 0.46 = 0.53$, which estimates the slope coefficient in the regression of conditional risk premium onto a constant and the recession dummy, is positive and significant in a one-sided test at the 10% level. Inference exploits DML and Theorem 2 to get asymptotic normality of that regression coefficient (see SMC S.3). The counter-cyclical behavior is shown by the SMB and CMA factors as well, while the HML and UMD factors premia are pro-cyclical (the latter at 10% significance level). In Figure 3 we display the estimates of the linear projection of conditional risk premium onto a low-dimensional vector of lagged predictors (see caption). Term structure spread, resp. distance to Dow all-time high, load significantly (at the 10%, resp. 1%, level) on the conditional risk premia of MKT factor. An asymptotic DML chi-square test rejects at 10% the null hypothesis of nil regression coefficients, i.e. constant conditional risk premia, for MKT, RMW and UMD factors. The time-variation in the conditional risk premia reveals the importance of

Table 2: **Risk Premia on Projected Observable Factors (Monthly %)**

Subsample		Mkt-RF	SMB	HML	RMW	CMA	UMD
Full Test Sample:	1995-07 to 2023-12	0.54 (0.15)	0.17 (0.08)	0.23 (0.09)	0.19 (0.07)	0.29 (0.07)	0.51 (0.15)
post-Dotcom	2003-06 to 2009-03	0.32	−0.16	0.22	−0.04	0.23	0.41
post-GFC	2009-04 to 2020-02	0.41	0.23	0.06	0.32	0.04	0.54
post-Covid	2020-03 to 2023-12	0.61	0.53	1.12	0.57	0.85	0.69
Bear Periods		0.16	0.55	0.56	0.07	0.70	0.14
Bull Periods		0.60	0.11	0.18	0.21	0.23	0.56
Recession Per.		0.97	2.07	−0.35	−0.99	0.38	−1.35
Non-Recess. Per.		0.46	−0.17	0.34	0.40	0.28	0.84
Cyclicality Test		Counter*	Counter	Pro	Pro	Counter	Pro*
Giglio-Xiu (2021)	1995-07 to 2023-12	0.24	0.16	0.37	0.17	0.19	0.43

This table reports the DML estimates of expected conditional risk premia for Fama–French five factors, i.e. market excess return (Mkt-RF), size (SMB), value (HML), profitability (RMW) and investment (CMA), plus momentum (UMD) projected onto our latent space (upper row, and middle panel). Standard errors via HAC estimator (4.11) are in parentheses (Parzen kernel, lag truncation parameter equal to 9 following Andrews (1991)'s optimal selection rule). Bear and bull periods follow Lunde and Timmermann (2004), recessions follow NBER definitions. The "Cyclicality Test" row reports the sign and significance (10% level denoted by *) of a DML regression-based test of pro- versus counter-cyclical risk premia. The last row reports unconditional estimates obtained with the three-pass FM procedure of Giglio and Xiu (2021).

conditioning information for measuring the reward for dynamic exposure to risk factors.

5.3.3 Conditional spanning with observable factors

In this subsection we investigate the conditional spanning of the latent space by traditional sets of observable risk factors. In Table 3 we report the DML estimates of the mean of the first three conditional canonical correlations (CCC) between latent and observed factors. For the former, we use the results obtained with $\alpha = 0.10$ leading to a number of conditional latent factors as displayed in Figure 2. For the latter, we consider various combinations listed in the second column of Table 3 corresponding to empirical models deployed in the asset pricing literature. We also report in parentheses the standard errors obtained from Proposition 2. The mean first CCC is above 0.90 across models. The mean second CCC is moderately large and above 0.80 for models including at least 6 observable factors like FF6, which includes the five Fama-French factors and momentum. The mean third CCC is substantially smaller for most observable factor models. Adding to FF6 a selection of 8 factors from Jensen, Kelly and Pedersen (2023) as described in the caption of Table 3, which yields a list of 14 observed factors, helps in bringing the third mean CCC at 0.75. However, mean CCC's beyond the third one (not reported) are definitely smaller. Hence, traditional

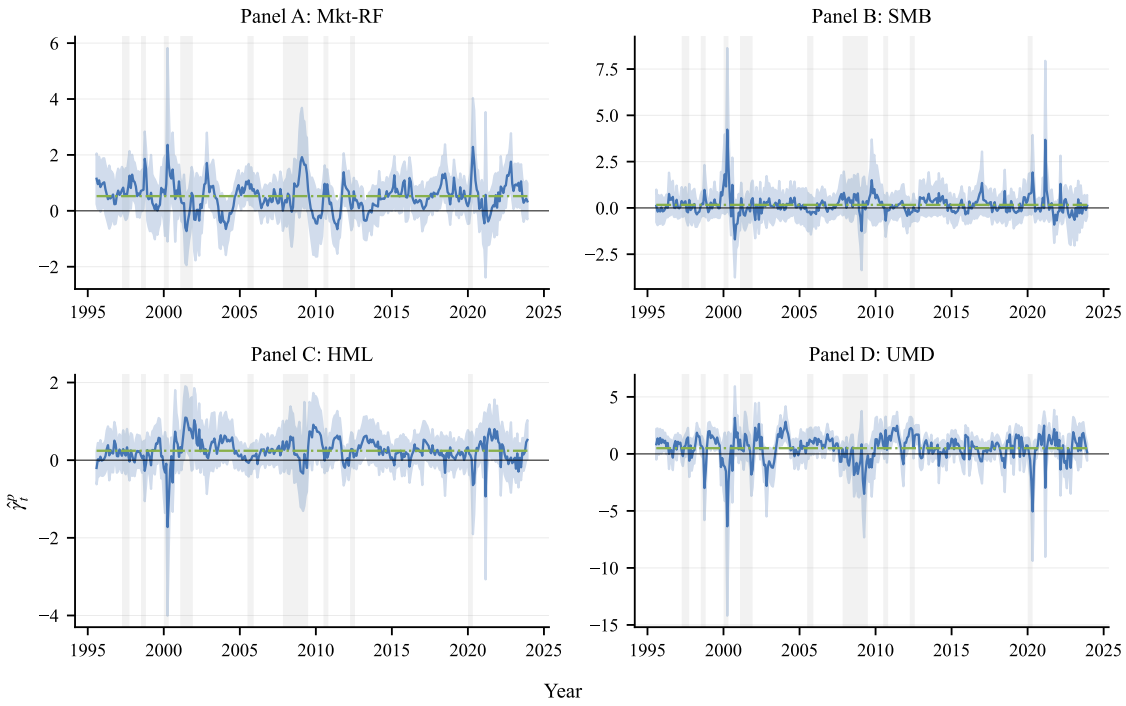

Figure 3: **DML estimates of linearly projected risk premia for observable factors**. In each panel, the solid line displays the path of fitted values in the regression of the conditional risk premium $\gamma_t = \lambda_t^O$ onto the constant and the lagged values of term spread, credit spread, dividend yield, return dispersion, and distance to Dow all-time high in vector W_{t-1}. Covariances in regression coefficients $\mathbb{E}[W_{t-1}W_{t-1}']^{-1}\mathbb{E}[W_{t-1}\lambda_t^{O\prime}]$ are estimated by DML. We get pointwise 10% confidence bands by using the feasible asymptotic normal distribution implied by Proposition 2 (see SMC S.3). We use a time-varying number of factors $\hat{k}_t^\alpha$ estimated at $\alpha = 0.10$. Dashed-dotted horizontal lines are time averages of fitted values. Shaded vertical bars correspond to NBER recessions.

observed risk factors, like Fama-French and momentum factors, span well at most 2–3 latent space dimensions, and even a selection of factors from the zoo is ineffective in improving the spanning for the remaining latent space dimensions, at least not constantly well across time.

What spans latent space dimensions not well explained by the FF6 factors? To shed light on this question, in each month t we consider directions in the latent space conditionally orthogonal to the first three conditional canonical directions with FF6, and focus on the direction with the largest conditional variance. Pre-multiplying this vector by J_{t-1} yields loadings onto the 78 characteristics (Anomalies). Because of asymptotic conditional redundancy in portfolio returns implied by the factor structure, these loadings are not uniquely defined. To obtain interpretable estimates, we impose sparsity using group Lasso and regularize time variation in weights by penalization (implementation details are in SMC S.4). Figure 4 reports ten characteristics-based portfolios that load most heavily on latent factor directions unspanned by the FF6 factors. In period 1995–1999, short-term reversal dominates, indicating near-term price reversals explain most unspanned variation early in the

Table 3: **Average Conditional Canonical Correlations**

Model	**Observable Factors**	$\hat{\mathbb{E}}[\rho_{1,t}]$	$\hat{\mathbb{E}}[\rho_{2,t}]$	$\hat{\mathbb{E}}[\rho_{3,t}]$
FF3	MKT, SMB, HML	0.91	0.74	0.55
		(0.08)	(0.02)	(0.02)
Carhart	MKT, SMB, HML, UMD	0.93	0.82	0.60
		(0.05)	(0.01)	(0.02)
FF5	MKT, SMB, HML, RMW, CMA	0.92	0.72	0.61
		(0.04)	(0.03)	(0.02)
FF6	MKT, SMB, HML, RMW, CMA, UMD	0.93	0.82	0.62
		(0.05)	(0.02)	(0.02)
FF6+JKP8	FF6, and JKP 8 themed factors	0.92	0.80	0.75
		(0.03)	(0.01)	(0.02)

DML estimates of average Conditional Canonical Correlations between latent and observable factors from asset pricing models. Standard errors are reported in parentheses (see details in Table 2). FF3 and FF5 denote the Fama–French three- and five-factor models (Fama and French (1993, 2015)), while FF6 includes also the momentum factor (Fama and French (2018)). Carhart refers to the four-factor Carhart (1994) model. We use eight value-weighted themed factors from Jensen, Kelly, and Pedersen (2023): Accruals, Debt Issuance, Low Leverage, Low Risk, Profit Growth, Quality, Seasonality, and Short-Term Reversal, that are complementary to the characteristics underlying FF6.

sample. During 1999–2003—around the dot-com bubble—industry momentum loads heavily, consistent with sector risk concentrated in technology and internet industries. In the 2003–2007 pre-crisis period, operating leverage, bid–ask spread, and MomVol (momentum in high-volume stocks) emerge as key drivers, with several signals appearing before publication. Throughout the 2008–2010 financial crisis, long-run reversal, bid–ask spread, operating leverage, and industry momentum dominate, reflecting valuation corrections, liquidity stress, and sector repricing. Between 2011–2016, 12-month momentum becomes central, signaling a return to classic momentum patterns during the post-crisis recovery. Finally, in 2017–2023, leverage becomes the primary loading, pointing to an increased role for capital-structure risk. Pronounced time variation in anomaly loadings onto unspanned latent directions implies that a fixed set of observable factors cannot generate stable spanning patterns for pervasive risks. Notably, loadings in the latent dimensions left unspanned by standard observed factors are disproportionately concentrated among characteristics identified before the start of the test sample, while characteristics introduced during the sample period contribute comparatively little to these dimensions.

5.3.4 Implied stochastic discount factor and conditional pricing errors

In this last subsection we establish the relevance of the latent conditional factors extracted with our methodology by evaluating the pricing performance of the implied SDF. In Table 4, middle panel, we report the mean absolute pricing errors (MAE) for 78 Long-Short Anomalies

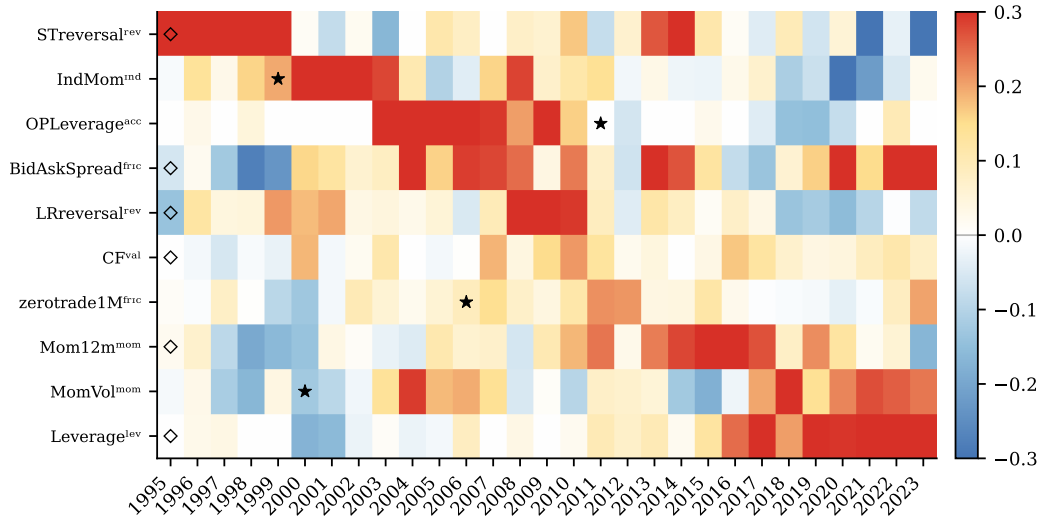

Figure 4: **Portfolio Loadings on Unspanned Factor Direction**. Heatmap of the top 10 characteristics by median weight, displaying annual average weights associated with the latent factor directions that are unspanned by the FF6 factors. Color intensity reflects the magnitude of the estimated weights across characteristics and years. Superscripts on the names denote economic categories: rev (reversal), ind (industry), acc (accruals), fric (trading frictions), val (value), mom (momentum), lev (leverage). Stars ($\star$) mark each anomaly's publication year; diamonds ($\diamond$) at 1995 indicate pre-sample publication.

portfolios corresponding to the characteristics used to build vector ξ_t. These are estimates for $\frac{1}{K}\sum_{k=1}^{K}|\mathbb{E}(\pi^O_{k,t})|$, where $\pi^O_{k,t}$ is the conditional pricing error for the kth component of vector ξ_t (see Example 3). For comparison purpose, we report the MAE for the SDF based on 6 observed factors, namely FF6, and for the SDF based on IPCA factors.

Table 4: **Mean Absolute Pricing Errors (Monthly %)**

		Long-Short Anomalies			Industry Portfolios		
Subsample		**FF6†**	**IPCA†**	**DML‡**	**FF6†**	**IPCA†**	**DML‡**
Full Sample	1995-07 to 2023-12	6.10	1.23	0.24	4.14	0.83	0.43
post-Dotcom	2003-06 to 2009-03	6.56	0.69	0.41	8.37	0.84	0.61
post-GFC	2009-04 to 2020-02	6.92	0.49	0.32	5.51	1.05	0.52
post-Covid	2020-03 to 2023-12	7.01	1.07	0.74	7.50	2.02	0.83
Bear Market		10.78	1.81	0.72	24.48	1.26	1.27
Bull Market		6.49	1.20	0.24	7.91	1.05	0.58

Entries report the estimates of monthly mean absolute pricing errors (MAE) in percent. Long-short anomalies comprise 78 portfolios constructed from the same characteristics. Industry portfolios comprise 49 portfolios from Kenneth French's Data Library. All values are computed *out-of-sample* in the test period. † Conditional pricing errors are computed from the implied stochastic discount factor, and smoothed via shrinkage of the variance–covariance matrix to mitigate matrix inversion instability. ‡ DML uses a time-varying number of factors $\hat{k}_t^{\alpha}$ estimated at $\alpha = 0.10$.

In both full–sample and subperiods estimates, the SDF implied by our method shows

smaller pricing errors. For instance, for the full–sample, the MAE of our SDF is 0.24 (monthly %), against 1.23 for IPCA and 6.10 for the FF6 model. The MAE are larger in bear periods than in bull periods, for all methodologies, with our SDF again showing superior pricing performance. In Table 5 we disaggregate the analysis and estimate the average conditional pricing errors for some canonical Anomalies. Overall, these results show that the latent conditional factors extracted with our methodology contain relevant pricing information over and beyond traditional sets of observable factors, or latent IPCA factors. We corroborate our findings with a similar exercise using the 49 Industry portfolios. This choice offers a fairer benchmark than the 78 anomalies, since none of the competing methodologies use portfolio formation based on industrial sectors. In Table 4, right panel, our method generally has smaller pricing errors for the 49 Industry portfolios compared to FF6 and IPCA.

Table 5: **Average Conditional Pricing Errors of Anomalies (Monthly %)**

Anomaly and Reference	**FF6**	**IPCA**	**DML**	**SE**
Earnings-to-Price, Basu (1977)	0.57	0.01	0.04	(0.09)
Gross Profits / Assets, Novy-Marx (2013)	7.33	0.66	0.19	(0.13)
Illiquidity, Amihud (2002)	−1.50	−0.25	−0.13	(0.23)
Idiosyncratic Volatility, Ang et al. (2006)	16.75	3.13	−0.30	(0.08)
6-Month Momentum, Jegadeesh and Titman (1993)	−6.01	3.02	0.11	(0.24)
Short-Term Reversal, Jegadeesh (1990)	36.54	−1.98	1.53	(0.25)

Entries report monthly average conditional pricing errors for six canonical anomalies, expressed in percent and computed out-of-sample. Standard errors (SE) for DML estimates are in parentheses.

6 Concluding Remarks

We develop a new and general tool for latent factor analysis in the conditional framework. The method is nonparametric with respect to the dynamics of the conditional betas, which may be driven linearly or nonlinearly by either macro variables or company-level covariates and may involve stock-specific unknown coefficients. The identification of a general class of parameters of interest passes through eigen-decompositions of the time-series conditional covariance of instrument-weighted cross-sectional averages of stock returns. The estimators are implemented by first constructing characteristic-sorted portfolio returns and then performing conditional PCA via a nonparametric regression estimator to get conditional moments. The paper establishes feasible asymptotic normality for a Double Machine Learning (DML) estimator of expected conditional features as $n, T \to \infty$ under a β-mixing assumption for the data. The framework is general enough to accommodate an unbalanced panel of returns and time-varying number and strength of the latent conditional factors. In our empirical application we find evidence for a large and counter-cyclical number of conditional factors

in a panel of monthly U.S. individual stock returns. Only 2–3 conditional latent space dimensions are well spanned by e.g. Fama-French factors or a few selected members from the factor zoo. Unspanned dimensions mostly load dynamically on leverage, time-series reversal and liquidity, with time-varying patterns, and play a crucial role in guaranteeing that our implied SDF delivers smaller pricing errors than IPCA or observed-factor SDFs for benchmark universes of test assets. We also find time-variation in conditional risk premia of observed factors projected onto the latent space, whence the importance of supplementing three-pass Fama-MacBeth procedure by conditioning information.

Our methodology can be applied to investigate the dynamic patterns of the conditional factor structure in several settings beyond the one of this paper, e.g. with high-frequency equity returns series (Aït-Sahalia and Xiu (2017), Liao and Yang (2017), Pelger (2019), Cheng, Liao and Yang (2023)) or other asset classes, and deploy other nonparametric estimators for the conditional moments of portfolio returns. One could also investigate the time-variation of the factor strength, in particular of conditional weak factors, more in depth. Further, one could extend the specification to include observed covariates and lagged dependent variables among the regressors to get panel models with interacting effects in a conditional setting. We leave such extensions for future research.

References

Abadie, A., Agarwal, A., Dwivedi, R., and Shah, A. (2024). Doubly Robust Inference in Causal Latent Factor Models. Working Paper.

Ahn, S. C., and Horenstein, A. R. (2013). Eigenvalue Ratio Test for the Number of Factors. *Econometrica*, 81(3), 1203-1227.

Ahrens, A., Chernozhukov, V., Hansen, C., Kozbur, D., Schaffer, M., Wiemann, T. (2025). An Introduction to Double/Debiased Machine Learning. Working Paper.

Aït-Sahalia, Y., and Xiu, D. (2017). Using Principal Component Analysis to Estimate a High Dimensional Factor Model with High-Frequency Data. *Journal of Econometrics*, 201, 384-399.

Al-Najjar, N. (1995). Decomposition and Characterization of Risk With a Continuum of Random Variables. *Econometrica*, 63 (5), 1195-1224.

Al-Sadoon, M. M. (2017). A Unifying Theory of Tests of Rank, *Journal of Econometrics*, 199, 49-62.

Anderson, T. W. (2003). An Introduction to Multivariate Statistical Analysis. Wiley-Interscience.

Andrews, D. (1991). Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation. *Econometrica*, 59(3), 817-858.

Bakalli, G., Guerrier, S., and Scaillet, O. (2023). A Penalized Two-Pass Regression to Predict Stock Returns with Time-Varying Risk Premia. *Journal of Econometrics*, 237(2), 105375.

Bai, J. (2003). Inferential Theory for Factor Models of Large Dimensions. *Econometrica*, 71, 135-171.

Bai, J., and Ng, S. (2006). Confidence Intervals for Diffusion Index Forecasts and Inference for Factor-Augmented Regressions. *Econometrica*, 74(4), 1133-1150.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica*, 80(6), 2369-2429.

Bonhomme, S., Jochmans, K., and Weidner, M. (2025). A Neyman-Orthogonalization Approach to the Incidental Parameter Problem. Working Paper.

Bryzgalova, S., Quaini, A., Trojani, F., and Yuan, M. (2026). Tradable Factor Risk Premia and Oracle Tests of Asset Pricing Models. Forthcoming in *Journal of Financial Economics*.

Carhart, M. M. (1997). On Persistence in Mutual Fund Performance. *The Journal of Finance*, 52 (1), 57-82.

Chaieb, I., Langlois, H., and Scaillet, O. (2021). Factors and Risk Premia in Individual International Stock Returns. *Journal of Financial Economics*, 141 (2), 669-692.

Chamberlain, G., and Rothschild, M. (1983). Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets. *Econometrica*, 51 (5), 1281-1304.

Chen, A., and Zimmermann, T. (2022). Open Source Cross-Sectional Asset Pricing, *Critical Finance Review*, 27(2), 207–264.

Chen, Q. (2022). A Unified Framework for Estimation of High-dimensional Conditional Factor Models. Working Paper.

Chen, Q., Roussanov, N., and Wang, X. (2023). Semiparametric Conditional Factor Models: Estimation and Inference. Working Paper.

Chen, X. (2007). Large Sample Sieve Estimation of Semi-Nonparametric Models. In *Handbook of Econometrics*, Volume 6, Chapter 76, 5549-5632.

Chen, X., and Shen, X. (1998). Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica*, 66, 289-314.

Chen, X., and White, A. (1999). Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators. *IEEE Tran. Information Theory*, 45, 682-691.

Cheng, M., Liao, Y., and Yang, X. (2023). Uniform Predictive Inference for Factor Models with Instrumental and Idiosyncratic Betas. *Journal of Econometrics*, 237, 105373.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C.,

Newey, W., and Robins, J. (2018). Double/Debiased Machine Learning for Treatment and Structural Parameters. *Econometrics Journal*, 21, C1-C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W., and Robins, J. (2022). Locally Robust Semiparametric Estimation. *Econometrica*, 90(4), 1501-1535.

Chernozhukov, V., Newey, W., and Singh, R. (2018). Learning L^2-Continuous Regression Functionals via Regularized Riesz Representers. Working Paper.

Chernozhukov, V., Newey, W., and Singh, R. (2022a). Debiased Machine Learning of Global and Local Parameters Using Regularized Riesz Representers. *Econometrics Journal*, 25(3), 576-601.

Chernozhukov, V., Newey, W., and Singh, R. (2022b). Automated Debiased Machine Learning of Causal and Structural Effects. *Econometrica*, 90(3), 967-1027.

Cochrane, J. H. (1996). A Cross-sectional Test of an Investment-based Asset Pricing Model. *Journal of Political Economy*, 104(3), 572-621.

Cochrane, J. H. (2011). Presidential Address: Discount Rates. *Journal of Finance*, 66(4), 1047-1108.

Connor, G., Hagmann, M. and Linton, O. (2012). Efficient Semiparametric Estimation of the Fama-French Model and Extensions, *Econometrica*, 80, 713-754.

Connor, G., and Korajczyk, R. A. (1989). An Intertemporal Equilibrium Beta Pricing Model. *The Review of Financial Studies*, 2(3), 373-392.

Cragg, J. G. and Donald, S.G. (1996). On the Asymptotic Properties of LDU-Based Tests of the Rank of a Matrix. *Journal of the American Statistical Association*, 91, 1301-1309.

Davidson, J. (1994). Stochastic Limit Theory: An Introduction for Econometricians. Oxford University Press.

De Jong, R., and Davidson, J. (2000). Consistency of Kernel Estimators of Heteroscedastic and Autocorrelated Covariance Matrices. *Econometrica*, 68(2), 407-423.

Donald, S. G., Fortuna, N. and Pipiras, V. (2007). On Rank Estimation in Symmetric Matrices: The Case of Indefinite Matrix Estimators. *Econometric Theory*, 23, 1103-1123.

Eberlein, E. (1984). Weak Convergence of Partial Sums of Absolutely Regular Sequences. *Statistics & Probability Letters*, 2(5), 291-293.

Fama, E. F., and French, K. R. (1993). Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, 33(1), 3-56.

Fama, E. F., and French, K. R. (2015). A Five-Factor Asset Pricing Model. *Journal of Financial Economics,* 116(1), 1-22.

Fama, E. F., and French, K. R. (2018). Choosing Factors. *Journal of Financial Economics,* 128(2), 234-252.

Fan, J., Liao, Y., and Wang, W. (2016). Projected Principal Component Analysis in Factor Models, *Annals of Statistics*, 44, 219-254.

Farrell, M., Liang, T., and Misra, S. (2021). Deep Neural Networks for Estimation and Inference. *Econometrica*, 89(1), 181-213.

Feng, G., Giglio, S., and Xiu, D. (2020). Taming the Factor Zoo: A Test of New Factors, *Journal of Finance*, 75, 1327-1370.

Ferson, W. E., and Harvey, C. R. (1991). The Variation of Economic Risk Premiums. *Journal of Political Economy*, 99(2), 385-415.

Ferson, W. E., and Harvey, C. R. (1999). Conditioning Variables and the Cross Section of Stock Returns. *The Journal of Finance*, 54(4), 1325-1360.

Filipovic, D., and Schneider, P. (2024). Joint Estimation of Conditional Mean and Covariance for Unbalanced Panels. Working Paper.

Fortin, A., Gagliardini, P., and Scaillet, P. (2023a). Eigenvalue Tests for the Number of Latent Factors in Short Panels. *Journal of Financial Econometrics*, 23(1), nbad024.

Fortin, A., Gagliardini, P., and Scaillet, O. (2023b). Latent Factor Analysis in Short Panels. Forthcoming in *Journal of Econometrics*.

Fortin, A., Gagliardini, P., and Scaillet, O. (2025). Optimal Maximin GMM Tests for Sphericity in Latent Factor Analysis of Short Panels. Working Paper.

Freyberger, J., Neuhierl, A., and Weber, M. (2020). Dissecting Characteristics Nonparametrically. *The Review of Financial Studies*, 33, 2326-2377.

Gagliardini, P., and Gourieroux, C. (2017). Double Instrumental Variable Estimation of Interaction Models with Big Data. *Journal of Econometrics*, 201(2), 176-197.

Gagliardini, P., Ossola, E., and Scaillet, O. (2016). Time-Varying Risk Premium in Large Cross-Sectional Equity Data Sets. *Econometrica*, 84(3), 985-1046.

Gagliardini, P., Ossola, E., and Scaillet, O. (2019). A Diagnostic Criterion for Approximate Factor Structure. *Journal of Econometrics*, 212, 503-521.

Gagliardini, P., Ossola, E., and Scaillet, O. (2020). Estimation of Large Dimensional Conditional Factor Models in Finance, in *Handbook of Econometrics*, S. Durlauf, L. Hansen and R. Matzkin Eds., Chapter 3, Volume 7A, 219-282.

Giglio, S., and Xiu, D. (2021). Asset Pricing with Omitted Factors. *Journal of Political Economy*, 129, 1947-1990.

Goyal, A., Welch, I., and Zafirov, A. (2024). A Comprehensive 2022 Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 37(11), 3490-3557.

Gu, S., Kelly, B. T., and Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. *Review of Financial Studies*, 33, 2223-2273.

Gu, S., Kelly, B. T., and Xiu, D. (2021). Autoencoder Asset Pricing Models. *Journal of Econometrics*, 222, 429-450.

Herrndorf, N. (1984). A Functional Central Limit Theorem for Weakly Dependent Sequences of Random Variables, *Annals of Probability*, 12(1), 141-153.

Hou, K., Xue, C., and Zhang, L. (2015). Digesting Anomalies: An Investment Approach. *Review of Financial Studies*, 28(3), 650-705.

Jensen, T., Kelly, B., and Pedersen, L. (2023). Is There a Replication Crisis in Finance? *Journal of Finance*, 78(5), 2465-2518.

Kelly, B., Malamud, S., and Zhou, K. (2024). The Virtue of Complexity in Return Prediction. *The Journal of Finance*, 79(1), 459-503.

Kelly, B., Pruitt, S., and Su, Y. (2017). Instrumented Principal Component Analysis. Working Paper.

Kelly, B., Pruitt, S., and Su, Y. (2019). Characteristics are Covariances: A Unified Model of Risk and Return, *Journal of Financial Economics*, 134, 501-524.

Kleibergen, F., and Paap, R. (2006). Generalized Reduced Rank Tests Using the Singular Value Decomposition, *Journal of Econometrics*, 133, 97-126.

Ichimura, H., and Newey, W. (2022). The Influence Function of Semiparametric Estimators, *Quantitative Economics*, 13, 29-61.

Lettau, M., and Ludvigson, S. (2001). Consumption, Aggregate Wealth, and Expected Stock Returns. *The Journal of Finance*, 56(3), 815-849.

Lewis, G. and Syrgkanis, V. (2021). Double/Debiased Machine Learning for Dynamic Treatment Effects via g-Estimation. *Advances in Neural Information Processing Systems 34*, 20325–20342.

Liao, Y., and Yang, X. (2017). Uniform Inference for Characteristics Effect of Large Continuous-Time Linear Models. Working Paper.

Lunde, A., and Timmermann, A. (2004). Duration Dependence in Stock Prices: An Analysis of Bull and Bear Markets. *Journal of Business & Economic Statistics*, 22(3), 253-273.

Magnus, J., and Neudecker, H. (1989). Matrix Differential Calculus with Applications to Statistics and Econometrics. Wiley.

Newey, W. (1994). The Asymptotic Variance of Semiparametric Estimators. *Econometrica*, 62, 1349-1382.

Pelger, M. (2019). Large-Dimensional Factor Modeling Based on High-Frequency Observations. *Journal of Econometrics*, 4, 23-42.

Pelger, M., and Xiong, R. (2022). State-Varying Factor Models of Large Dimensions. *Journal of Business & Economic Statistics*, 40(3), 1315-1331.

Petkova, R., and Zhang, L. (2005). Is Value Riskier than Growth? *Journal of Financial Economics*, 78(1), 187-202.
Robin, J. M., and Smith, R. (2000). Tests of Rank, *Econometric Theory*, 16, 151-175.
Semenova, M., Goldman, M., Chernozhukov, V., and Taddy, M. (2023). Inference in Heterogeneous Treatment Effects in High-dimensional Dynamic Panels Under Weak Dependence, *Quantitative Economics*, 14, 471-510.
Shumway, T., and Warther, V. (1999). The Delisting Bias in CRSP's Nasdaq Data and Its Implications for the Size Effect, *Journal of Finance*, 54(6), 2361-2379.
Singleton, K. (2006). Empirical Dynamic Asset Pricing. Model Specification and Econometric Assessment. Princeton University Press.
Syrgkanis, V., and Zampetakis, M. (2020). Estimation and Inference with Trees and Forests in High Dimensions, *Proceedings of Machine Learning Research*, 125, 1–2.
Zaffaroni, P. (2025). Factor Models for Conditional Asset Pricing. *Journal of Political Economy*, 133(8), 2615-2642.

APPENDIX

A Proofs of Proposition 1 and Theorem 2

A.1 Proof of Proposition 1

Let us first establish representation (3.10). From the equations in (3.5) (which use Assumption 1 and hold for a specific conditional rotation of the latent factor), f_t^* is a function of vector ξ_t and its conditional mean and variance given $\mathscr{F}_{t-1}$. Moreover, $f_t^* + \lambda_t^* = J_{t-1}'\xi_t$, which yields $F_t = \begin{bmatrix} 0 & J_{t-1}' \\ I_{k^O} & 0 \end{bmatrix} X_t$. Hence, under Assumption 2, the conditional moments $\mathbb{E}(F_t|\mathscr{F}_{t-1})$ and $\mathbb{V}(F_t|\mathscr{F}_{t-1})$ are functions of the elements of vector $\theta_0(Z_{t-1})$. Representation (3.10) follows. The Normalization Restriction 1 is irrelevant as long as γ_t is invariant to conditional rotations of the latent factor. Let us next prove identifiability. With $n \to \infty$, cross-sectional expectations $\mathbb{E}^{cs}[\cdot]$ are identifiable. Thus, the values of process ξ_t are identifiable from (3.1). With $T \to \infty$ and Assumption 2, time series conditional moments of process X_t are identifiable. Thus, function $\theta_0(\cdot)$ is identifiable. Further, from $k_t = \text{Rank}(\Sigma_{22,0}(Z_{t-1}))$ (which is an implication of Assumptions 1 and 2, see equation (3.4)), the time-varying number of conditional factors is identifiable. The conclusion follows.

A.2 Proof of Theorem 2

Throughout the proof, $\mathbb{P}^*$ and $\mathbb{E}^*$ denote the marginal probability measure, and the corresponding expectation, of predictors' realizations Z_t, $t \in \mathcal{I}$, holding the estimate $\hat{\theta}(\cdot)$ fixed, i.e. $\mathbb{E}^*(\hat{d}) = \int d(Z_{\mathcal{I}}, \hat{\theta}) dP_0(Z_{\mathcal{I}})$, where $\hat{d} = d(Z_{\mathcal{I}}, \hat{\theta})$ is a function of the testing sample path $Z_{\mathcal{I}} := (Z_t,\ t \in \mathcal{I})$ with joint distribution P_0, and of estimate $\hat{\theta}(\cdot)$ computed on sample $\mathcal{T}$. Probability orders O_p and o_p are w.r.t. $\mathbb{P}$. We will repeatedly use the next result.

Lemma 1. *Under Assumptions 5 and 9, we have: a) if* $\mathbb{E}^*(|\hat{d}|) \leq \bar{c}\|\hat{\theta} - \theta_0\|_{L^2}$ *for a constant* $\bar{c} > 0$*, then* $\hat{d} = O_p(\|\hat{\theta} - \theta_0\|_{L^2})$*, and b) if* $\mathbb{P}^*(|\hat{d}| > \epsilon) = o_p(1)$*,* $\forall \epsilon > 0$*, then* $\hat{d} = o_p(1)$*.*

The proof of Lemma 1 (in OA) uses the β-mixing condition (Assumption 5) to bound the difference between probabilities under $\mathbb{P}(\cdot|\mathscr{Y}_T)$ and $\mathbb{P}^*$ by a term at order $\beta(a_T)$, where $\mathscr{Y}_T$ is the filtration of data in estimation sample $\mathcal{T}$ and $a_T := \min_{\tau \in \mathcal{T}, t \in \mathcal{I}} |\tau - t|$, and $\beta(a_T) = o(1)$ under Assumptions 5 and 9.

A.2.1 Proof of Part (a) The proof is structured in three steps.

Step 1. We first show that $P_1 := \mathbb{P}\left[\delta_{k_t}(\Sigma_{22,0}(Z_{t-1})) \geq \underline{c}_T,\ \forall t \in \mathcal{I}\right] \to 1$ as $n, T \to \infty$, with $\underline{c}_T := \frac{T^{-\bar{\beta}}}{(\log T)^{1/\bar{c}}}$. By using Assumption 7 (ii) and condition $\chi < 1$, we get

$$\begin{aligned} P_1 &\geq 1 - \sum_{t \in \mathcal{I}} \mathbb{P}\left[\delta_{k_t}(\Sigma_{22,0}(Z_{t-1})) \leq \underline{c}_T\right] = 1 - \sum_{t \in \mathcal{I}} \mathbb{P}\left[T^{\bar{\beta}} \delta_{k_t}(\Sigma_{22,0}(Z_{t-1})) \leq (\log T)^{-1/\bar{c}}\right] \\ &\geq 1 - I e^{-\bar{c}(\log T)^{1/\bar{c}}} = 1 - I/T = 1 - o(1). \end{aligned}$$

Thus, in the following it suffices to focus on the event $\delta_{k_t}(\Sigma_{22,0}(Z_{t-1})) \geq \underline{c}_T$, $\forall t \in \mathcal{I}$.

Step 2. We next show that there exists a constant $C > 0$ independent of t, n, T, such that $\hat{k}_t^\alpha = k_t$ if $\|\hat{\Sigma}(Z_{t-1}) - \Sigma(Z_{t-1})\| \leq C\alpha$. Indeed, $\hat{k}_t^\alpha = k_t$ holds iff $\phi_r(\hat{\Sigma}_{22}(Z_{t-1})) \geq 1 - \alpha$ and $\phi_{r-1}(\hat{\Sigma}_{22}(Z_{t-1})) < 1 - \alpha$ for $r = k_t$, where $\phi_r(\Sigma) := \frac{\sum_{j=1}^r \delta_j(\Sigma)}{Tr(\Sigma)}$. Moreover:

$$\begin{aligned} &\phi_r(\hat{\Sigma}_{22}(Z_{t-1})) - \phi_r(\Sigma_{22,0}(Z_{t-1})) \\ = \ &\frac{\sum_{j=1}^r [\delta_j(\hat{\Sigma}_{22}(Z_{t-1})) - \delta_j(\Sigma_{22}(Z_{t-1}))] - \phi_r(\Sigma_{22,0}(Z_{t-1})) Tr[\hat{\Sigma}_{22}(Z_{t-1}) - \Sigma_{22,0}(Z_{t-1})]}{Tr[\hat{\Sigma}_{22}(Z_{t-1})]}, \end{aligned}$$

which yields:

$$\left|\phi_r(\hat{\Sigma}_{22}(Z_{t-1})) - \phi_r(\Sigma_{22,0}(Z_{t-1}))\right| \leq \frac{2\sqrt{K}\|\hat{\Sigma}(Z_{t-1}) - \Sigma_0(Z_{t-1})\|}{Tr[\Sigma_{22,0}(Z_{t-1})] - \sqrt{K}\|\hat{\Sigma}(Z_{t-1}) - \Sigma_0(Z_{t-1})\|}, \quad \text{(A.1)}$$

because $\sum_{j=1}^K |\delta_j(A) - \delta_j(B)| \leq \sqrt{K}\left(\sum_{j=1}^K |\delta_j(A) - \delta_j(B)|^2\right)^{1/2}$ by the Cauchy-Schwartz

inequality, $\sum_{j=1}^{K} |\delta_j(A) - \delta_j(B)|^2 \leq \|A - B\|^2$ for symmetric $K \times K$ matrices A, B by the Weilandt-Hoffmann inequality, and $\|\hat{\Sigma}_{22}(Z_{t-1}) - \Sigma_{22,0}(Z_{t-1})\| \leq \|\hat{\Sigma}(Z_{t-1}) - \Sigma_0(Z_{t-1})\|$. Now, use $Tr[\Sigma_{22,0}(Z_{t-1})] \geq \underline{c}$ from Assumption 7 (i), and suppose $\|\hat{\Sigma}(Z_{t-1}) - \Sigma_0(Z_{t-1})\| \leq C\alpha$ for $C := \frac{\underline{c}}{4\sqrt{K}}$. From inequality (A.1) we get $\left|\phi_r(\hat{\Sigma}_{22}(Z_{t-1})) - \phi_r(\Sigma_{22,0}(Z_{t-1}))\right| \leq \frac{2\sqrt{K}C\alpha}{\underline{c} - \sqrt{K}C\alpha} \leq \frac{4\sqrt{K}C\alpha}{\underline{c}} = \alpha$, for large n, T, because $\alpha = o(1)$. Then, we get

$$\phi_r(\hat{\Sigma}_{22}(Z_{t-1})) \geq \phi_r(\Sigma_{22,0}(Z_{t-1})) - \left|\phi_r(\hat{\Sigma}_{22}(Z_{t-1})) - \phi_r(\Sigma_{22,0}(Z_{t-1}))\right| \geq 1 - \alpha, \tag{A.2}$$

$$\phi_{r-1}(\hat{\Sigma}_{22}(Z_{t-1})) \leq \phi_{r-1}(\Sigma_{22,0}(Z_{t-1})) + \left|\phi_r(\hat{\Sigma}_{22}(Z_{t-1})) - \phi_r(\Sigma_{22,0}(Z_{t-1}))\right| < 1 - \alpha \tag{A.3}$$

for $r = k_t$, because $\phi_r(\Sigma_{22,0}(Z_{t-1})) = 1$ and $\phi_{r-1}(\Sigma_{22,0}(Z_{t-1})) < 1 - 2\alpha$ by the conditions $\delta_{k_t}(\Sigma_{22,0}(Z_{t-1})) \geq \underline{c}_T$ and $\alpha \ll \frac{T^{-\bar{\beta}}}{(\log T)^{1/\bar{c}}} = \underline{c}_T$. Inequalities (A.2) and (A.3) imply $\hat{k}_t^\alpha = k_t$.

Step 3. We establish a lower bound for the probability $\mathbb{P}\left[\hat{k}_t^\alpha = k_t,\ \forall t \in \mathcal{I}\right]$. First, from Step 2, we have $\hat{k}_t^\alpha = k_t$, $\forall t \in \mathcal{I}$, if $d(\hat{\theta}, \theta_0; \mathcal{I}) \leq \frac{C\alpha}{\sqrt{I}}$, where $d(\hat{\theta}, \theta_0; \mathcal{I}) := (\frac{1}{I}\sum_{t\in\mathcal{I}} \|\hat{\Sigma}(Z_{t-1}) - \Sigma_0(Z_{t-1})\|^2)^{1/2}$. Thus:

$$\mathbb{P}\left[\hat{k}_t^\alpha = k_t,\ \forall t \in \mathcal{I}\right] \geq 1 - \mathbb{P}\left[d(\hat{\theta}, \theta_0; \mathcal{I}) \geq CI^{-1/2}\alpha\right]. \tag{A.4}$$

We have $E^*[d(\hat{\theta}, \theta_0; \mathcal{I})] \leq E^*[d(\hat{\theta}, \theta_0; \mathcal{I})^2]^{1/2} = \|\hat{\Sigma} - \Sigma_0\|_{L^2} \leq \|\hat{\theta} - \theta_0\|_{L^2}$. Then, by Lemma 1 (a) and Assumption 3, $d(\hat{\theta}, \theta_0; \mathcal{I}) = O_p(\|\hat{\theta} - \theta_0\|_{L^2}) = O_p((\frac{T}{\log T})^{-\bar{\delta}})$. From the condition $\frac{\alpha}{\sqrt{I}} \gg (\frac{T}{\log T})^{-\bar{\delta}}$, the probability on the RHS of (A.4) is $o(1)$. The conclusion follows.

A.2.2 Proof of Part (b) We rewrite estimator (4.9) as $\hat{c} = \frac{1}{I}\sum_{t\in\mathcal{I}} W_{t-1}\left\{\psi(\hat{\theta}(Z_{t-1}), \hat{k}_t^\alpha) + \bar{a}(\hat{\theta}(Z_{t-1}), \hat{k}_t^\alpha)'(\hat{X}_t - \hat{\mu}(Z_{t-1})) + Tr\left[\Omega(\hat{\theta}(Z_{t-1}), \hat{k}_t^\alpha)\left(\hat{X}_t\hat{X}_t' - \hat{\Lambda}(Z_{t-1})\right)\right]\right\}$, for $\bar{a}(\theta, k) = a(\theta, k) - 2\Omega(\theta, k)\mu$, and $\Lambda(\cdot) = \Sigma(\cdot) + \mu(\cdot)\mu(\cdot)'$ is the conditional second-order moment matrix parameter, and $\hat{\Lambda}(\cdot)$ is its estimator obtained from $\hat{\theta}(\cdot)$. We work conditionally on the event $\{\hat{k}_t^\alpha = k_t,\ \forall t \in \mathcal{I}\}$, which has probability approaching 1 as $n, T \to \infty$ by Part (a). Then, by defining $\alpha(\theta, k) = [\bar{a}(\theta, k)', vec(\Omega(\theta, k))']'$ and $\vartheta(\cdot) = [\mu(\cdot)', vec(\Lambda(\cdot))']'$, and using $Tr(A'B) = vec(A)'vec(B)$ for matrices A, B, we can write the DML estimator as:

$$\hat{c} \;=\; \frac{1}{I}\sum_{t\in\mathcal{I}} W_{t-1}\left\{\psi(\hat{\theta}(Z_{t-1}), k_t) + \alpha(\hat{\theta}(Z_{t-1}), k_t)'[\hat{\zeta}_t - \hat{\vartheta}(Z_{t-1})]\right\}, \tag{A.5}$$

where $\hat{\zeta}_t = (\hat{X}_t', \hat{X}_t' \otimes \hat{X}_t')'$ is an estimator of $\zeta_t = (X_t', X_t' \otimes X_t')'$, and $\hat{\vartheta}(\cdot)$ is based on $\hat{\theta}(\cdot)$. In the rest of the proof we set $W_{t-1} = 1$, and omit the argument k_t in function α, to ease notation. We build on the proof of Theorem 10 in Chernozhukov, Newey and Singh (2018), expanding the arguments to cover β-mixing dependent data, the fact that vector $\hat{\zeta}_t$

is estimated, and that $\alpha(\theta)$ is a known vector function of θ (see also the published version Chernozhukov, Newey and Singh (2022b), and Lemma 8 in Chernozhukov et al. (2022)).

STEP 1. We first break $\sqrt{I}(\hat{c} - c_0)$ into an asymptotically Gaussian term and four reminders. From (A.5) we have:

$$\begin{aligned}\sqrt{I}(\hat{c} - c_0) &= \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} h_t + \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \left[m(Z_{t-1}, \hat{\theta}) - m(Z_{t-1}, \theta_0) - \bar{m}(\hat{\theta}) \right] \\ &+ \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \left[\alpha(\hat{\theta}(Z_{t-1}))'(\zeta_t - \hat{\vartheta}(Z_{t-1})) - \alpha(\theta_0(Z_{t-1}))'(\zeta_t - \vartheta_0(Z_{t-1})) \right] \\ &+ \sqrt{I}\bar{m}(\hat{\theta}) + \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \alpha(\hat{\theta}(Z_{t-1}))'(\hat{\zeta}_t - \zeta_t),\end{aligned}$$

where $h_t(\theta) := m(Z_{t-1}, \theta) + \alpha(\theta(Z_{t-1}))'(\zeta_t - \vartheta(Z_{t-1}))$ with $m(Z_{t-1}, \theta) := \psi(\theta(Z_{t-1}), k_t) - c_0$ and $\bar{m}(\theta) := \mathbb{E}[m(Z_{t-1}, \theta)]$, and $h_t \equiv h_t(\theta_0)$. Thus, we get:

$$\sqrt{I}(\hat{c} - c_0) = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} h_t + R_1 + R_2 + R_3 + R_4, \tag{A.6}$$

where $R_1 = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \left[m(Z_{t-1}, \hat{\theta}) - m(Z_{t-1}, \theta_0) - \bar{m}(\hat{\theta}) \right] + \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \left[\phi(Z_{t-1}, \hat{\theta}) - \bar{\phi}(\hat{\theta}) \right]$ $+ \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} [\alpha(\hat{\theta}(Z_{t-1})) - \alpha(\theta_0(Z_{t-1}))]' u_t$ with $\phi(Z_{t-1}, \theta) := -\alpha(\theta_0(Z_{t-1}))'[\vartheta(Z_{t-1}) - \vartheta_0(Z_{t-1})]$, $\bar{\phi}(\theta) = \mathbb{E}[\phi(Z_{t-1}, \theta)]$ and $u_t = \zeta_t - \vartheta_0(Z_{t-1})$, $R_2 = -\frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} [\alpha(\hat{\theta}(Z_{t-1})) - \alpha(\theta_0(Z_{t-1}))]'[\hat{\vartheta}(Z_{t-1}) - \vartheta_0(Z_{t-1})]$, $R_3 = \sqrt{I} \left[\bar{m}(\hat{\theta}) + \bar{\phi}(\hat{\theta}) \right]$ and $R_4 = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \alpha(\hat{\theta}(Z_{t-1}))'(\hat{\zeta}_t - \zeta_t)$. Let us next show that the remainder terms R_1, R_2, R_3 and R_4 are $o_p(1)$ under our assumptions.

STEP 2. We prove $R_1 = o_p(1)$. Let us start with term $R_{11} = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t$ where $\Delta_t = m(Z_{t-1}, \hat{\theta}) - m(Z_{t-1}, \theta_0) - \bar{m}(\hat{\theta})$. For any $\epsilon > 0$, let us prove

$$\mathbb{P}^* \left[\left| \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t \right| \geq \epsilon \right] = o_p(1) \tag{A.7}$$

Then, Lemma 1 b) yields $R_{11} = o_p(1)$. To show (A.7) we use an independent block construction similar as in Chen and Shen (1998), p. 306. We partition interval $\mathcal{I}$ into $2a_{2,T}$ blocks $H_{1,j}$, $H_{2,j}$, $j = 1, ..., a_{2,T}$, each of length $a_{1,T}$, with $a_{2,T} = \lfloor I/(2a_{1,T}) \rfloor$ and $a_{1,T}$ such that:

$$a_{1,T} \ll T^{\bar{a}\bar{\delta}}, \quad a_{1,T} \ll I^{1/2}, \quad I \ll a_{1,T} e^{\bar{m} a_{1,T}}, \tag{A.8}$$

where $\bar{m}$ and $\bar{a}$ are defined in Assumptions 5 and 8 (i), plus a residual block: $\mathcal{I} = \bigcup_{j=1}^{a_{2,T}} (H_{1,j} \cup$

$H_{2,j}) \cup H_3$. By Eberlein (1984), we have:

$$\left| \mathbb{P}^* \left[\left| \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t \right| \geq \epsilon \right] - \mathbb{P}^* \left[\left| \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t^* \right| \geq \epsilon \right] \right| \leq (a_{2,T} - 1)\beta(a_{1,T}), \tag{A.9}$$

where Z_t^*, $t \in \mathcal{I}$ be a random sequence which is independent of Z_t, $t \in \mathcal{I}$, has independent blocks and has the same joint distribution as the corresponding block of the Z_t sequence, and $\Delta_t^* := m(Z_{t-1}^*, \hat{\theta}) - m(Z_{t-1}^*, \theta_0) - \bar{m}(\hat{\theta})$. Let us write $\sum_{t \in \mathcal{I}} \Delta_t^* = \sum_{j=1}^{a_{2,T}} V_{1,j} + \sum_{j=1}^{a_{2,T}} V_{2,j} + \sum_{t \in H_3} \Delta_t^*$, where $V_{1,j} = \sum_{t \in H_{1,j}} \Delta_t^*$ and $V_{2,j} = \sum_{t \in H_{2,j}} \Delta_t^*$. Then, by stationarity:

$$\mathbb{P}^* \left[\left| \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t^* \right| \geq \epsilon \right] \leq 2p + q \tag{A.10}$$

where $p := \mathbb{P}^* \left[\left| \frac{1}{\sqrt{I}} \sum_{j=1}^{a_{2,T}} V_{1,j} \right| \geq \frac{\epsilon}{3} \right]$ and $q := \mathbb{P}^* \left[\left| \frac{1}{\sqrt{I}} \sum_{t \in H_3} \Delta_t^* \right| \geq \frac{\epsilon}{3} \right]$. By the Chebyshev inequality, $p \leq \frac{9}{\epsilon^2} \mathbb{E}^* \left[\left(\frac{1}{\sqrt{I}} \sum_{j=1}^{a_{2,T}} V_{1,j} \right)^2 \right]$. Thus, since the $V_{1,j}$ are independent across j, we have $\mathbb{E}^*[(\frac{1}{\sqrt{I}} \sum_{j=1}^{a_{2,T}} V_{1,j})^2] = \frac{a_{2,T}}{I} \mathbb{E}^*[(\sum_{t \in H_{1,j}} \Delta_t)^2] \leq \frac{a_{1,T} a_{2,T}}{I} \sum_{t \in H_{1,j}} \mathbb{E}^* [\Delta_t^2] \leq \frac{a_{1,T}}{2} \mathbb{E}^* [\Delta_t^2]$, where we use the Cauchy-Schwartz inequality, and that $Z_{t-1}^* \stackrel{d}{=} Z_{t-1}$ within blocks. Moreover from Assumption 8 (i):

$$\begin{aligned} \mathbb{E}^* \left[\Delta_t^2 \right] &\leq \int [m(z, \hat{\theta}) - m(z, \theta_0)]^2 dP_0(z) = \int [\psi(\hat{\theta}(z), k(z)) - \psi(\theta_0(z), k(z))]^2 dP_0(z) \\ &\leq C \|\hat{\theta} - \theta_0\|_{L^2}^{\bar{a}}. \end{aligned}$$

Thus, from $\|\hat{\theta} - \theta_0\|_{L^2} = O_p((\frac{T}{\log T})^{-\bar{\delta}})$, we get $p = O_p(a_{1,T} (\frac{T}{\log T})^{-\bar{a}\bar{\delta}})$. Further, $q = O_p(a_{1,T}^2 / I)$. Then, by the bounds in (A.9) and (A.10), and using the conditions in (A.8), we get $\mathbb{P}^* \left[\left| \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t \right| \geq \epsilon \right] = O_p(a_{1,T} T^{-\bar{a}\bar{\delta}} + a_{2,T} \beta(a_{1,T}) + a_{1,T}^2 / I) = O_p(a_{1,T} T^{-\bar{a}\bar{\delta}} + I e^{-\bar{m} a_{1,T}} / a_{1,T} + a_{1,T}^2 / I) = o_p(1)$, which yields (A.7).

We use similar independent block arguments to control the other two terms $R_{12} = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t^{(2)}$ and $R_{13} = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t^{(3)}$ in R_1, with $\Delta_t^{(2)} = \phi(Z_{t-1}, \hat{\theta}) - \bar{\phi}(\hat{\theta})$ and $\Delta_t^{(3)} = [\alpha(\hat{\theta}(Z_{t-1})) - \alpha(\theta_0(Z_{t-1}))]' u_t$. We have $E^*[(\Delta_t^{(2)})^2] \leq \int \left\{ \alpha(\theta_0(z))' [\hat{\vartheta}(z) - \vartheta_0(z)] \right\}^2 dP_0(z) \leq \bar{C} \|\hat{\theta} - \theta_0\|_{L^2}^2$ and $E^*[(\Delta_t^{(3)})^2] \leq \int [\alpha(\hat{\theta}(z)) - \alpha(\theta_0(z))]' \mathbb{V}(u_t | Z_{t-1} = z) [\alpha(\hat{\theta}(z)) - \alpha(\theta_0(z))] dP_0(z) \leq \bar{C} \|\hat{\theta} - \theta_0\|_{L^2}^2$, from Assumptions 6 and 8 (iii) and (iv). With $\|\hat{\theta} - \theta_0\|_{L^2} = O_p((\frac{T}{\log T})^{-\bar{\delta}})$, and using the Markov inequality to verify the condition of Lemma 1 b), we get $\frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} \Delta_t^{(k)} = o_p(1)$, for $k = 2, 3$. Thus, $R_1 = o_p(1)$.

STEP 3. We prove $R_2 = o_p(1)$. We have $E^*[|R_2|] = O_p(\sqrt{I} \|\hat{\theta} - \theta_0\|_{L^2}^2) = o_p(1)$ under Assumption 8 (iv) and the conditions $\|\hat{\theta} - \theta_0\|_{L^2} = O_p((\frac{T}{\log T})^{-\bar{\delta}})$ and $I \leq T^{\chi}$ with $\chi < 4\bar{\delta}$.

Using the Markov inequality and Lemma 1 b), we get $R_2 = o_p(1)$.

STEP 4. We have $R_3 \leq \sqrt{I}\left|\int[\psi(\hat{\theta}(z),k) - c_0 - \alpha_0(z)'(\hat{\vartheta}(z) - \vartheta_0(z))]dP_0(z)\right|$ $= \sqrt{I}\left|\int[\psi(\hat{\theta}(z),k) - \psi(\theta_0(z),k) - a(\theta_0)'(\hat{\mu}(z) - \mu_0(z)) - Tr\{\Omega(\theta_0)(\hat{\Sigma}(z) - \Sigma_0(z))\}]dP_0(z)\right|$ $\leq \sqrt{I}C\|\hat{\theta} - \theta_0\|_{L^2}^2 = o_p(1)$, by Assumptions 3 and 8 (ii) and condition $I \leq T^\chi$ with $\chi < 4\bar{\delta}$.

STEP 5. We prove $R_4 = o_p(1)$. We have

$$\left|\frac{1}{\sqrt{I}}\sum_{t\in\mathcal{I}}\alpha(\hat{\theta}(Z_{t-1}))'(\hat{\zeta}_t - \zeta_t)\right| \leq \frac{C}{\sqrt{I}}\sum_{t\in\mathcal{I}}\|\hat{\zeta}_t - \zeta_t\| \leq \frac{\tilde{C}}{\sqrt{I}}\sum_{t\in\mathcal{I}}(\|\hat{\xi}_t - \xi_t\| + \|\hat{\xi}_t - \xi_t\|^2), \quad \text{(A.11)}$$

by Assumption 8 (iii). To ease control of $\|\hat{\xi}_t - \xi_t\|$, we use $\xi_t = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[w_{i,t-1}y_{i,t}|\mathscr{F}_t]$. Then, $\hat{\xi}_t - \xi_t = \frac{1}{n}\sum_{i=1}^n w_{i,t-1}\varepsilon_{i,t} + (\frac{1}{n}\sum_{i=1}^n \eta_{i,t-1})'(f_t + \lambda_t)$, with $\eta_{i,t} := w_{i,t-1}b_{i,t-1}' - \mathbb{E}(w_{i,t-1}b_{i,t-1}'|\mathscr{F}_t)$. We get $\sup_t \mathbb{E}[\|\hat{\xi}_t - \xi_t\|^2] = O(1/n)$ from Assumptions 4 and 6 using the triangular inequality, the Hölder inequality, and the bound $\|f_t + \lambda_t\| \leq \|\xi_t\| \leq \|X_t\|$. Then, from (A.11), we have $\frac{1}{\sqrt{I}}\sum_{t\in\mathcal{I}}\alpha(\hat{\theta}(Z_{t-1}))'(\hat{\zeta}_t - \zeta_t) = O_p(\sqrt{I/n})$. Using $I \leq T^\chi$ and $n \geq T^\kappa$, with $\kappa > \chi$, we get $I/n = o(1)$, which yields $R_4 = o_p(1)$.

From (A.6) and Steps 2-5, $\sqrt{I}(\hat{c} - c_0) = \frac{1}{\sqrt{I}}\sum_{t\in\mathcal{I}} h_t + o_p(1)$. From Assumptions 6 and 8 (iii), (v), we have $\mathbb{E}[\|h_t\|^{\bar{r}}] < \infty$ with $\bar{r} > 2$. Then, from Assumption 5 and condition $0 < \sigma^2 < \infty$, and the CLT for times-series data in Hernndorf (1984), the conclusion follows.

A.2.3 Proof of Part (c) To show consistency of $\hat{\sigma}^2$, we use the infeasible estimator $\tilde{\sigma}^2 = \sum_{j=-I+1}^{I-1} K(j/b_I)\Gamma_j(\tau_0)$ with $\Gamma_j(\tau) := \frac{1}{I}\sum_{t=j+1}^T h_t(\tau)h_{t-j}(\tau)'$, where

$$h_t(\tau) = W_{t-1}[\psi(\theta(Z_{t-1})) + \alpha(\theta(Z_{t-1}))'(\zeta_t - \vartheta(Z_{t-1}))] - c =: \varphi_t(\theta) - c, \quad \text{(A.12)}$$

and $\tau_0(\cdot) = (c_0, \theta_0(\cdot))$. Then, $\tilde{\sigma}^2 - \sigma^2 = o_p(1)$ from our Assumptions 5, 8 (iii) and (iv), 10 and 11, and Theorem 2.1 in De Jong and Davidson (2000). In order to prove $\hat{\sigma}^2 - \tilde{\sigma}^2 = o_p(1)$, we have to account for the fact that $\hat{h}_t$ in (4.11) involves the estimated number of factors $\hat{k}_t^\alpha$ and the estimated $\hat{\xi}_t$ in $\hat{\zeta}_t$. To tackle these issues, first, from Part (a), without loss of generality we work conditionally on the event $\hat{k}_t^\alpha = k_t$ for all $t \in \mathcal{I}$. Second, we use $\hat{\sigma}^2 - \breve{\sigma}^2 = O_p(b_I/\sqrt{n}) = o_p(I/\sqrt{n}) = o_p(1)$, for $\breve{\sigma}^2 := \sum_{j=-I+1}^{I-1} K(j/b_I)\Gamma_j(\hat{\tau})$, if $\chi \leq \kappa/2$. Then, it remains to show $\breve{\sigma}^2 - \tilde{\sigma}^2 = o_p(1)$. For this purpose we use the decomposition

$$\breve{\sigma}^2 - \tilde{\sigma}^2 = \frac{1}{I}\sum_{t,s=1}^I K((t-s)/b_I)[h_t(\hat{\tau})h_s(\hat{\tau})' - h_t(\tau_0)h_s(\tau_0)'] = \quad J_1 + J_1' + J_2,$$

where $J_1 := \frac{1}{I}\sum_{t,s=1}^I K((t-s)/b_I)(h_t(\hat{\tau}) - h_t(\tau_0))h_s(\tau_0)'$ and $J_2 := \frac{1}{I}\sum_{t,s=1}^I K((t-s)/b_I)(h_t(\hat{\tau}) - h_t(\tau_0))(h_s(\hat{\tau}) - h_s(\tau_0))'$. We now bound both terms.

STEP 1. We prove that $J_2 = o_p(1)$. By applying twice the Cauchy-Schwartz inequality, we have $\|J_2\| \leq [\sum_{t,s=1}^I K(\frac{t-s}{b_I})^2]^{1/2} \frac{1}{I} \sum_{t=1}^I \|h_t(\hat{\tau}) - h_t(\tau_0)\|^2$. In order to control term $\frac{1}{I} \sum_{t=1}^I \|h_t(\hat{\tau}) - h_t(\tau_0)\|^2$, we use $h_t(\hat{\tau}) - h_t(\tau_0) = \varphi_t(\hat{\theta}) - \varphi_t(\theta_0) - (\hat{c} - c_0)$ from (A.12), where

$$\varphi_t(\hat{\theta}) - \varphi_t(\theta_0) = W_{t-1} \Big\{ [\psi(\hat{\theta}(Z_{t-1})) - \psi(\theta_0(Z_{t-1}))] - \alpha(\theta_0(Z_{t-1}))' \Delta\hat{\vartheta}(Z_{t-1})$$
$$- [\alpha(\hat{\theta}(Z_{t-1})) - \alpha(\theta_0(Z_{t-1}))]' \Delta\hat{\vartheta}(Z_{t-1}) + [\alpha(\hat{\theta}(Z_{t-1}) - \alpha(\theta_0(Z_{t-1})]'[\zeta_t - \vartheta_0(Z_{t-1}))] \Big\},$$

where $\Delta\hat{\vartheta}(Z_{t-1}) := \hat{\vartheta}(Z_{t-1}) - \vartheta_0(Z_{t-1})$. We have $\hat{c} - c_0 = O_p(I^{-1/2})$ from Part (b), and by similar arguments as in Parts (a)-(b) and Assumptions 8 (i), (iii) and (iv), we have $\mathbb{E}^*[\frac{1}{I} \sum_{t=1}^I \|\varphi_t(\hat{\theta}) - \varphi_t(\theta_0)\|^2] \leq C\|\theta - \theta_0\|_{L^2}^2 = o_p(T^{-2\bar{\delta}+\epsilon})$, for any $\epsilon > 0$, which leads to $\frac{1}{I} \sum_{t=1}^I \|h_t(\hat{\tau}) - h_t(\tau_0)\|^2 = o_p(T^{-2\bar{\delta}+\epsilon}) + O_p(I^{-1})$ using Lemma 1 b). Moreover, $\sum_{t,s=1}^I K(\frac{t-s}{b_I})^2 = O(Ib_I)$ under Assumption 10. Thus, from Assumption 11, we have $J_2 = o_p(\sqrt{Ib_I} T^{-2\bar{\delta}+\epsilon}) + O_p(\sqrt{b_I/I}) = o_p(1)$ if ϵ is small enough and $\chi < 2\bar{\delta}$.

STEP 2. We prove that $J_1 = o_p(1)$. Let $\mathbb{E}[h_t(\tau)] = \mathbb{E}[\varphi_t(\theta)] - c =: \bar{M}(\theta) - c$. We split $J_1 = J_{11} + J_{12}$, where $J_{11} := [\bar{M}(\hat{\theta}) - \bar{M}(\theta_0) - (\hat{c} - c_0)][\frac{1}{I} \sum_{t,s=1}^I K((t-s)/b_I) h_s(\tau_0)]'$ and $J_{12} := \frac{1}{I} \sum_{t,s=1}^I K((t-s)/b_I)(\varphi_t(\hat{\theta}) - \varphi_t(\theta_0) - (\bar{M}(\hat{\theta}) - \bar{M}(\theta_0)))h_s(\tau_0)'$. We can prove $\frac{1}{I} \sum_{t,s=1}^I K((t-s)/b_I) h_s(\tau_0) = O_p(\frac{b_I}{\sqrt{I}})$ and $\bar{M}(\hat{\theta}) - \bar{M}(\theta_0) - (\hat{c} - c_0) = O_p(T^{-\bar{\delta}} + I^{-1/2})$, so that we get $J_{11} = o_p(1)$ if $\chi < 2\bar{\delta}$. To control J_{12}, we use Fourier inversion $K(x) = \int_{-\infty}^{\infty} e^{-i\xi x} \phi(\xi) d\xi$ to get that J_{12} equals

$$\int_{-\infty}^{\infty} \phi(\xi) \left[\frac{1}{\sqrt{I}} \sum_{t=1}^I e^{-i\xi t/b_I} (\varphi_t(\hat{\theta}) - \varphi_t(\theta_0) - (\bar{M}(\hat{\theta}) - \bar{M}(\theta_0))) \right] \left[\frac{1}{\sqrt{I}} \sum_{s=1}^I e^{i\xi s/b_I} h_s(\tau_0)) \right]' d\xi.$$

Then, by Cauchy-Schwartz inequality we get $\|J_{12}\| \leq \sqrt{J_{121} J_{122}}$ where

$$J_{121} := \int_{-\infty}^{\infty} \phi(\xi) \left\| \frac{1}{\sqrt{I}} \sum_{t=1}^I e^{-i\xi t/b_I} (\varphi_t(\hat{\theta}) - \varphi_t(\theta_0) - (\bar{M}(\hat{\theta}) - \bar{M}(\theta_0))) \right\|^2 d\xi, \quad \text{and,}$$

$$J_{122} := \int_{-\infty}^{\infty} \phi(\xi) \left\| \frac{1}{\sqrt{I}} \sum_{s=1}^I e^{i\xi s/b_I} h_s(\tau_0)) \right\|^2 d\xi = \frac{1}{I} \sum_{t,s=1}^I K(\frac{t-s}{b_I}) h_t(\tau_0)' h_s(\tau_0).$$

We have $\mathbb{E}(J_{122}) = \sum_{j=-(I-1)}^{I-1} \frac{I-|j|}{I} K(\frac{j}{b_I}) Tr(\mathbb{E}[\Gamma_j(\tau_0)]) = O(1)$ so that $J_{122} = O_p(1)$.

What remains to show is $J_{121} = o_p(1)$. Write $J_{121} = \frac{1}{I} \sum_{t,s=1}^I K(\frac{t-s}{b_I})(\varphi_t(\hat{\theta}) - \varphi_t(\theta_0) - (\bar{M}(\hat{\theta}) - \bar{M}(\theta_0)))'(\varphi_s(\hat{\theta}) - \varphi_s(\theta_0) - (\bar{M}(\hat{\theta}) - \bar{M}(\theta_0)))$. By Assumption 5, and Corollary 14.3

in Davidson (1994) under $\mathbb{P}^*$, we have:

$$\mathbb{E}^*(J_{121}) = \frac{1}{I}\sum_{t,s=1}^{I} K(\frac{t-s}{b_I}) Tr\left(\mathbb{C}\text{ov}^*(\varphi_t(\hat{\theta}) - \varphi_t(\theta_0), \varphi_s(\hat{\theta}) - \varphi_s(\theta_0))\right)$$

$$\leq C\frac{1}{I}\sum_{t,s=1}^{I} \beta(t-s)^{1-2/\breve{r}} \mathbb{E}^*[\|\varphi_t(\hat{\theta}) - \varphi_t(\theta_0)\|^{\breve{r}}]^{2/\breve{r}} = O_p(\mathbb{E}^*[\|\varphi_t(\hat{\theta}) - \varphi_t(\theta_0)\|^{\breve{r}}]^{2/\breve{r}}), \quad \text{(A.13)}$$

for $\breve{r}$ such that $\bar{r} > \breve{r} > 2$. From Assumptions 8 (i) and (v), we can show

$$\mathbb{E}^*[\|\varphi_t(\hat{\theta}) - \varphi_t(\theta_0)\|^{\breve{r}}] = \int \|\varphi(z,\hat{\theta}) - \varphi(z,\theta_0)\|^{\breve{r}} dP_0(z) \leq C(\|\hat{\theta} - \theta_0\|_{L^2}^{\bar{a}}) = o_p(1). \quad \text{(A.14)}$$

Thus, (A.13) and (A.14) imply $\mathbb{E}^*(J_{121}) = o_p(1)$ and, from Lemma 1 b), we get $J_{121} = o_p(1)$.

B Orthogonality restrictions for DML estimators

In this appendix we detail the derivation of the locally-robust orthogonality restrictions for three parameters of interest defined in Section 4.3. Again we set $W_{t-1} = 1$ to ease notation.

B.1 Average conditional risk premia. For the j-th component of the expected conditional risk premia vector, we have $\mathbb{E}(\lambda_{j,t}^O) = \mathbb{E}[\psi(\theta_0(Z_{t-1}), k_t)]$ with $\psi(\theta, k) = e_j'\Sigma_{12}\Sigma_{22}^\dagger\mu_2$, where the pseudo-inverse $\Sigma_{22}^\dagger$ has rank k and e_j denotes the j-th unit vector in $\mathbb{R}^{k^O}$ (see equation (3.6)). To compute the gradient of function ψ w.r.t. parameter θ, we use the matrix differential notation (Magnus and Neudecker (1989)). We have $d\psi = e_j'(d\Sigma_{12})\Sigma_{22}^\dagger\mu_2 + e_j'\Sigma_{12}(d\Sigma_{22}^\dagger)\mu_2 + e_j'\Sigma_{12}\Sigma_{22}^\dagger(d\mu_2)$. The next lemma (proved in the OA) yields the differential of the pseudo-inverse.

Lemma 2. *We have* $d\Sigma_{22}^\dagger = \mathscr{P}_0(d\Sigma_{22})(\Sigma_{22}^\dagger)^2 + (\Sigma_{22}^\dagger)^2(d\Sigma_{22})\mathscr{P}_0 - (\Sigma_{22}^\dagger)(d\Sigma_{22})(\Sigma_{22}^\dagger)$, *where* $\mathscr{P}_0$ *is the projection matrix onto the null space of matrix* Σ_{22}.

We note that $\Sigma_{12}\mathscr{P}_0 = 0$ and $\mathscr{P}_0\mu_2 = 0$ once these quantities are evaluated at the true parameter values, from the reduced-rank structure implied by equation (3.2). Then, Lemma 2 yields $d\psi = e_j'(d\Sigma_{12})\Sigma_{22}^\dagger\mu_2 - e_j'\Sigma_{12}\Sigma_{22}^\dagger(d\Sigma_{22})\Sigma_{22}^\dagger\mu_2 + e_j'\Sigma_{12}\Sigma_{22}^\dagger(d\mu_2) = Tr(\Sigma_{22}^\dagger\mu_2 e_j' d\Sigma_{12}) - Tr(\Sigma_{22}^\dagger\mu_2 e_j'\Sigma_{12}\Sigma_{22}^\dagger d\Sigma_{22}) + e_j'\Sigma_{12}\Sigma_{22}^\dagger d\mu_2 = a'd\mu + Tr(\Omega d\Sigma)$, where

$$a = a(\theta,k) = \begin{pmatrix} 0 \\ \Sigma_{22}^\dagger\Sigma_{21}e_j \end{pmatrix}, \quad \Omega = \Omega(\theta,k) = \frac{1}{2}\begin{pmatrix} 0 & e_j\mu_2'\Sigma_{22}^\dagger \\ \Sigma_{22}^\dagger\mu_2 e_j' & -\Sigma_{22}^\dagger(\mu_2 e_j'\Sigma_{12} + \Sigma_{21}e_j\mu_2')\Sigma_{22}^\dagger \end{pmatrix}.$$

We have $Tr(\Omega\Sigma) = 0$ because function $\psi(\theta, k)$ is homogeneous of degree 0 with respect to

multiplication of argument Σ by a scalar. Then, $Tr\left[\Omega(\theta_0(Z_{t-1}),k_t)\Sigma_0(Z_{t-1})\right]=0$ implies that the locally-robust orthogonality restriction is $\mathbb{E}(\psi(\theta_0(Z_{t-1}),k_t)+\eta_t-c_0)=0$, where:

$$\begin{aligned}\eta_t &= a(\theta_0(Z_{t-1}),k_t)'U_t+U_t'\Omega(\theta_0(Z_{t-1}),k_t)U_t\\ &= e_j'\Sigma_{0,12}(Z_{t-1})\Sigma_{22,0}(Z_{t-1})^\dagger U_{2,t}+e_j'\tilde{U}_{1,t}\left(\mu_{0,2}(Z_{t-1})'\Sigma_{22,0}(Z_{t-1})^\dagger U_{2,t}\right)\\ &= e_j'U_{1,t}-e_j'\tilde{U}_{1,t}\left(1-\mu_{0,2}(Z_{t-1})'\Sigma_{22,0}(Z_{t-1})^\dagger U_{2,t}\right)=e_j'U_{1,t}-e_j'\tilde{U}_{1,t}R_{f,t}m_{t-1,t},\end{aligned}$$

(see equation (3.8)) and $\tilde{U}_{1,t}=U_{1,t}-\Sigma_{0,12}(Z_{t-1})\Sigma_{22,0}(Z_{t-1})^\dagger U_{2,t}$.

B.2 Average r-th conditional canonical correlation. We have $\mathbb{E}(\rho_{r,t})=\mathbb{E}[\psi(\theta_0(Z_{t-1}),k_t)]$ with $\psi(\theta,k)=\rho_r=[\delta_r(A)]^{1/2}$ and $A:=\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^\dagger\Sigma_{21}$. We have $d\rho_r=\frac{1}{2\rho_r}d\delta_r(A)=\frac{1}{2\rho_r}v_r'\Sigma_{11}(dA)v_r$, where dA is the differential of matrix function A with respect to matrix Σ, and v_r is the eigenvector of A associated to eigenvalue $\delta_r(A)$ normalized such that $v_r'\Sigma_{11}v_r=1$. [14] Now, we use $dA=(d\Sigma_{11}^{-1})\Sigma_{12}\Sigma_{22}^\dagger\Sigma_{21}+\Sigma_{11}^{-1}(d\Sigma_{12})\Sigma_{22}^\dagger\Sigma_{21}+\Sigma_{11}^{-1}\Sigma_{12}(d\Sigma_{22}^\dagger)\Sigma_{21}+\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^\dagger(d\Sigma_{21})$, where $d\Sigma_{11}^{-1}=-\Sigma_{11}^{-1}(d\Sigma_{11})\Sigma_{11}^{-1}$. Then, from Lemma 2, and $\Sigma_{12}\mathscr{P}_0=0$ and $\mathscr{P}_0\Sigma_{21}=0$, we get $dA=-\Sigma_{11}^{-1}(d\Sigma_{11})A+\Sigma_{11}^{-1}(d\Sigma_{12})\Sigma_{22}^\dagger\Sigma_{21}-\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^\dagger(d\Sigma_{22})\Sigma_{22}^\dagger\Sigma_{21}+\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^\dagger(d\Sigma_{21})$. Thus:

$$\begin{aligned}&d\rho_r=\frac{1}{2\rho_r}v_r'\Sigma_{11}(dA)v_r=-\frac{\rho_r}{2}v_r'(d\Sigma_{11})v_r+\frac{1}{2\rho_r}v_r'(d\Sigma_{12})\Sigma_{22}^\dagger\Sigma_{21}v_r+\frac{1}{2\rho_r}v_r'\Sigma_{12}\Sigma_{22}^\dagger(d\Sigma_{21})v_r\\ &-\frac{1}{2\rho_r}v_r'\Sigma_{12}\Sigma_{22}^\dagger(d\Sigma_{22})\Sigma_{22}^\dagger\Sigma_{21}v_r=-\frac{v_r'v_r}{2\rho_r}\Big(\rho_r^2Tr(Q_rd\Sigma_{11})\\ &-Tr(\Sigma_{22}^\dagger\Sigma_{21}Q_rd\Sigma_{12})-Tr(Q_r\Sigma_{12}\Sigma_{22}^\dagger d\Sigma_{21})+Tr(\Sigma_{22}^\dagger\Sigma_{21}Q_r\Sigma_{12}\Sigma_{22}^\dagger d\Sigma_{22})\Big)=Tr(\Omega d\Sigma),\end{aligned}$$

where $Q_r=v_r(v_r'v_r)^{-1}v_r'$ and $\Omega=\Omega(\theta,k)$ as in equation (4.12). Process η_t is given in (4.13).

B.3 Average pricing errors of observed factors. For the j-th component, $\mathbb{E}[(f_{j,t}^O-r_{f,t})m_{t-1,t}]=\mathbb{E}[R_{f,t}^{-1}(\psi(\theta_0(Z_{t-1}),k_t)-r_{f,t})]$ with $\psi(\theta,k)=e_j'\mu_1-e_j'\Sigma_{12}\Sigma_{22}^\dagger\mu_2$ (see equation (3.9)). Note that $\mathbb{E}[(f_{j,t}^O-r_{f,t})m_{t-1,t}]=\mathbb{E}[R_{f,t}^{-1}(e_j'\mu_1(Z_{t-1})-\lambda_{j,t}^O-r_{f,t})]$. Then, from the calculations in Subsection i), the locally-robust orthogonality restriction is $\mathbb{E}[R_{f,t}^{-1}(\psi(\theta_0(Z_{t-1}),k_t)-r_{f,t})+\eta_t-c_0]=0$, where $\eta_t=R_{f,t}^{-1}e_j'\tilde{U}_{1,t}\left(1-\mu_{0,2}(Z_{t-1})'\Sigma_{22,0}(Z_{t-1})^\dagger U_{2,t}\right)=e_j'\tilde{U}_{1,t}m_{t-1,t}$.

[14] Here, the equation $d\delta_r(A)=v_r'\Sigma_{11}(dA)v_r$ follows by computing the differential of the identity $Av_r=\delta_r(A)v_r$, left-multiplying the resulting equation by $v_r'\Sigma_{11}$, and using $v_r'\Sigma_{11}dv_r=0$ and $\Sigma_{11}A=A'\Sigma_{11}$.

ONLINE APPENDIX

Double Machine Learning Inference for Conditional Latent Factors

Patrick Gagliardini and Hao Ma

In Section C we provide additional theoretical results. They concern the consistency rate of estimators of process γ_t (Section C.1), the proof that ANN-based Sieve estimators and Lasso estimators relying on approximate sparsity both meet Assumption 3 on norm consistency rate of $\hat{\theta}(\cdot)$ (Sections C.2 and C.3), a consistent estimator for the number of conditional factors relying on the max eigenvalue ratio (C.4), and further examples of parameters of interest in class (2.3) related to conditional portfolio choice (C.5). We give the proofs of additional Propositions and the supporting technical Lemmas in Section D. Finally, Section E reports our Monte Carlo experiments.

C Additional theory

C.1 Consistency rate of the time-varying parameters of interest

We establish the consistency rate of estimator $\hat{\gamma}_t$ defined in (4.3) in terms of mean error along the path $t \in \mathcal{I}$. We need additional regularity conditions on function ψ.

Assumption 12. *We have (i)* $\|\psi(\theta, \xi^*, k) - \psi(\theta, \xi, k)\| \leq \bar{\psi}(\theta, k)\|\xi^* - \xi\|$ *for a scalar function* $\bar{\psi} > 0$ *such that* $\sup_\theta \bar{\psi}(\theta(Z_{t-1}), k_t) \leq v_t$ *and* $E(v_t^2) \leq C$ *for all* t *and a constant* $C > 0$*, where the* $\sup_\theta$ *is with respect to* $\theta \in \Theta_T$ *with* $\|\theta - \theta_0\|_{L^2} \leq T^{-\delta^*}$*, for* $\delta^* < \bar{\delta}$*, and* $\Theta_T \subset \Theta$ *such that* $\hat{\theta} \in \Theta_T$*,* $\mathbb{P}$*-a.s. Moreover, (ii)* $\int \|\psi(\theta(z), \xi, k(z)) - \psi(\theta_0(z), \xi, k(z))\| dP_0(z, \xi) \leq c\|\theta - \theta_0\|_{L^2}$*, for a constant* $c > 0$*, where* $k(Z_{t-1}) = k_t$ *and* P_0 *is the stationary distribution of* $(Z'_{t-1}, \xi'_t)'$*.*

Proposition 3. *Under Assumptions 1-7, 9 and 12, if* $n, T \to \infty$ *such that* $n \geq T^\kappa$ *and* $I \leq T^\chi$ *with* $\kappa > 2\bar{\delta} > \chi + 2\bar{\beta}$*, and if* $\frac{T^{-\bar{\beta}}}{(\log T)^{1/\bar{c}}} \gg \alpha \gg \sqrt{I}(\frac{T}{\log T})^{-\bar{\delta}}$ *, we have* $\frac{1}{I}\sum_{t\in\mathcal{I}} \|\hat{\gamma}_t - \gamma_t\| = O_p\left((\frac{T}{\log T})^{-\bar{\delta}}\right)$*, where* $\bar{\delta} > 0$ *is defined in Assumption 3.*

Hence, $\hat{\gamma}_t$ is a consistent estimator for γ_t, for any $t \in \mathcal{I}$ and in average error along the path, and inherits the convergence rate of the nonparametric estimator $\hat{\theta}(\cdot)$. The condition $\kappa > 2\bar{\delta}$ implies that the estimation error from cross-sectional averaging used to get $\hat{\xi}_t$ is asymptotically negligible.

C.2 Estimating conditional expectations and variances via ANN

We prove that under primitive conditions on function $\theta_0(\cdot)$ the ANN-based Sieve estimators (4.4) and (4.5) meet Assumption 3 and establish the convergence rate $\bar{\delta}$.

Proposition 4. *Assume that $\mu_{0,j} \in \mathcal{H}^d$, and $\Sigma_0 = C_0 C_0'$ with $C_{0,j\ell} \in \mathcal{H}^d$, where $\mathcal{H}^d := \{f : \mathbb{R}^d \to \mathbb{R},\ f(x) = \int e^{i\zeta' x} d\sigma(\zeta),\ \int \max\{\|\zeta\|, 1\} d|\sigma|_{tv}(\zeta) < \infty\}$ and $|\sigma|_{tv}$ is the total variation of complex measure σ. Under Assumptions 1-2 and 4-6, if the activation function ϕ defining the single-layer network $\mathscr{N}_{M_T}^d$ is Lipshitz continuous, as $n, T \to \infty$ and $M_T \to \infty$ such that $\frac{T}{M_T \log M_T} = o(n)$, we have $\|\hat{\theta} - \theta_0\|_{L^2} = O_p(\max\{M_T^{-b}, \sqrt{\frac{M_T \log M_T}{T}}\})$, where $b = 1/2 + 1/(d+1)$.*

The fastest convergence rate for the ANN-Sieve estimator $\hat{\theta}$ in Proposition 4 is obtained when the number of neurons grows with sample size as $M_T \asymp (T/\log T)^{\frac{1}{2b+1}} = (T/\log T)^{\frac{1}{2}\frac{d+1}{d+2}}$, which yields Assumption 3 with consistency rate $\bar{\delta} = \frac{1}{4}\frac{d+3}{d+2}$. The latter coincides with the rate in Chen and Shen (1999) for nonparametric regression with ANN. The rate $\bar{\delta}$ is comprised between 1/4 and 1/3, for any $d \geq 1$. In particular, the lower bound 1/4 holding for any dimension d shows some resilience of the ANN estimator to the curse of dimensionality. The convergence rate is established under the condition $(\frac{T}{\log T})^{\frac{1}{2}\frac{d+3}{d+2}} = o(n)$ on the relative growth rate of n and T, namely for $n \geq T^{\kappa}$ with $\kappa > \frac{1}{2}\frac{d+3}{d+2}$.

C.3 Estimating conditional expectations and variances via Lasso

This subsection presents an alternative estimation method for vector function $\theta_0(\cdot)$, that applies when its components are sparse functions. We build on Belloni et al. (2012) and estimate nonparametrically each component function of $\mu_0(\cdot)$ and $\Sigma_0(\cdot)$ by Lasso using an expanding set of basis functions (dictionary) as regressors. Let $h(Z) = (h_1(Z), ..., h_p(Z))'$ be a $p \times 1$ dictionary of functions of argument vector $Z \in \mathbb{R}^d$. Let us define the estimators $\hat{\mu}(\cdot) = \hat{P}_\mu h(\cdot)$ and $\hat{\Sigma}(\cdot) = [\tilde{\Sigma}(\cdot)]^\epsilon$ where $vech[\tilde{\Sigma}(\cdot)] = \hat{P}_\Sigma h(\cdot)$, the rows of matrices $\hat{P}_\mu = [\hat{\rho}_{\mu,1} : \cdots : \hat{\rho}_{\mu,L}]'$ and $\hat{P}_\Sigma = [\hat{\rho}_{\Sigma,1} : \cdots : \hat{\rho}_{\Sigma,L(L+1)/2}]'$ are defined by

$$\hat{\rho}_{\mu,l} = \underset{\rho \in \mathbb{R}^p}{\arg\min}\ \frac{1}{T}\sum_{t \in \mathcal{T}} [\hat{X}_{t,l} - \rho' h(Z_{t-1})]^2 + 2\alpha\|\rho\|_1,$$

for $l = 1, ..., L$, and

$$\hat{\rho}_{\Sigma,l} = \underset{\rho \in \mathbb{R}^p}{\arg\min}\ \frac{1}{T}\sum_{t \in \mathcal{T}} [\hat{\zeta}_{t,l} - \rho' h(Z_{t-1})]^2 + 2\alpha\|\rho\|_1,$$

for $l = 1, ..., L(L+1)/2$, with $\hat{\zeta}_{t,l}$ being the lth component of $\hat{\zeta}_t = vech[(\hat{X}_t - \hat{\mu}(Z_{t-1}))(\hat{X}_t - \hat{\mu}(Z_{t-1}))']$. Here, $\|\cdot\|_1$ denotes the L^1 norm in $\mathbb{R}^p$, and the vanishing positive scalar α is the (possibly random) penalization coefficient. Moreover, $[\tilde{\Sigma}]^\epsilon := UD^\epsilon U'$ with $D^\epsilon = diag(\max\{d_{jj}, \epsilon\})$ is the regularization of symmetric matrix $\tilde{\Sigma} = UDU'$ based on its SVD, where $D = diag(d_{jj})$ is the diagonal matrix of eigenvalues, and $\epsilon > 0$ is the regularization threshold. We use regularization to get a positive-definite estimated conditional covariance matrix. We set $\hat{\theta}(\cdot) = [\hat{\mu}(\cdot)', vech(\hat{\Sigma}(\cdot))']'$. We work with the following assumptions.

Assumption 13. *For any component $\theta_{0,l}(\cdot)$ of function $\theta_0(\cdot)$ and integer $\bar{s} \leq p$, there exists $\bar{\rho}_l \in \mathbb{R}^p$ with $\|\bar{\rho}_l\|_0 \leq \bar{s}$ and $\|\theta_{0,l} - \bar{\rho}_l' h\|_{L^2} \leq C\bar{s}^{-a-1}$, with $a > 3/2$ and $C > 0$, where the L^0 norm $\|\cdot\|_0$ is the number of nonzero elements of a vector.*

Assumption 14. *(i) For a constant $\bar{B}_h > 0$, we have with probability one, $\max_{1\leq j\leq p} |h_j(Z_{t-1})| \leq \bar{B}_h$. (ii) Matrix $G = E[h(Z_{t-1})h(Z_{t-1})']$ is non-singular and its largest eigenvalue is uniformly bounded in p. (iii) There is $\bar{\kappa} > 3$ such that*

$$\inf_{\delta\neq 0: \sum_{j\in\mathcal{J}_{\bar{\rho}_l}^c} |\delta_j| \leq \bar{\kappa} \sum_{j\in\mathcal{J}_{\bar{\rho}_l}} |\delta_j|} \frac{\delta' G \delta}{\sum_{j\in\mathcal{J}_{\bar{\rho}_l}} \delta_j^2} > 0,$$

for any l, where $\bar{\rho}_l$ is defined in Assumption 13, and $\mathcal{J}_\rho$ (resp. $\mathcal{J}_\rho^c$) is the set of the $j \in \{1, ..., p\}$ with $\rho_j \neq 0$ (resp. $\rho_j = 0$) for any vector $\rho \in \mathbb{R}^p$. (iii) The same condition holds replacing $\bar{\rho}_l$ with the population Lasso coefficients vector $\rho_{L,l} = \arg\min_{\rho\in\mathbb{R}^p} \left(\|\theta_{0,l} - h'\rho\|_{L^2}^2 + 2\alpha_L\|\rho\|_1\right)$, where $\alpha_L = \frac{\log T}{\sqrt{T}}$, for any l.

Assumption 13 defines the sparsity condition required for any component of $\theta_0(\cdot)$, namely these functions are well approximated by a finite number $\bar{s}$ of functions in the dictionary (whose identity is unknown to the econometrician). Parameter a controls how fast the approximation error vanishes when allowing $\bar{s}$ to grow. Assumption 14 is used in Chernozhukov, Newey and Singh (2018, 2022b). In particular, condition (iii) is a population version of the "restricted eigenvalue condition" of Bickel, Ritov and Tsybakov (2009).

Proposition 5. *Let $n, T, p \to \infty$ such that $p = O(T^{\bar{b}})$, $T = O(p^{1/\underline{b}})$, for $0 < \underline{b} \leq \bar{b}$, and $\frac{T}{\log^2 T} = O(n)$; moreover, $\alpha = o_p(1)$ such that $\alpha \gg \left(\frac{\log T}{T}\right)^{\frac{1}{2}\frac{2a+1}{2a+3}}$, and $\epsilon = O\left(\left(\frac{\log T}{T}\right)^{\frac{a}{2a+3}}\right)$. Then, under Assumptions 1-2, 4-6, 13 and 14, the Lasso estimator $\hat{\theta}$ is consistent and $\|\hat{\theta} - \theta_0\|_{L^2} = O_p\left(\left(\frac{\log^2 T}{T}\right)^{\frac{a}{2a+3}}\right)$.*

The convergence rate of Lasso estimator $\hat{\theta}$ depends only on parameter a and not explicitly on dimension d, which can grow with n and T in the asymptotics. Proposition 5 shows that

Assumption 3 is met for the Lasso estimator with $\kappa = 1$ and $\bar{\delta} = \frac{a}{2a+3}$. The condition on ϵ guarantees that eigenvalue regularization does not impact on the convergence rate.

C.4 Max eigenvalue ratio to select the number of conditional factors

In this section we provide an approach for estimating the number of conditional factors that is alternative to estimator (4.2). We build on the insight of Ahn and Horenstein (2013) and adapt their eigenvalue-ratio selection principle to our setting with conditional factors. The idea is that, if the eigenvalues of the empirical counterpart of $\Sigma_{22,0}(Z_{t-1})$ feature a scree-plot decay behavior, the number of conditional factors can be estimated by the integer r such that the ratio between the rth and $(r+1)$th largest eigenvalues is maximal.

Suppose $k_t \leq k_{\max}$ for some known upper bound $k_{\max}$ (which exists by Assumption 1 and $k_{\max} \leq K$). Let $\hat{\varrho}_{r,t} = \frac{[\delta_r(\hat{\Sigma}_{22}(Z_{t-1}))]_\epsilon}{[\delta_{r+1}(\hat{\Sigma}_{22}(Z_{t-1}))]_\epsilon}$ be the ratio of two consecutive regularized eigenvalues, with $[x]_\epsilon = \max\{x, \epsilon\}$ for $\epsilon > 0$ tending to 0 as a function of n, T, for any integer $r \leq k_{\max}$. The regularization is needed to control small eigenvalue estimates in the denominator of the ratio. Then, the estimator of the number of conditional factors at date t is:

$$\hat{k}_t^\epsilon = \underset{1 \leq r \leq k_{\max}}{\arg\max}\ \hat{\varrho}_{r,t}, \tag{C.1}$$

i.e. the order corresponding to the largest eigenvalue ratio.

Proposition 6. *Under Assumptions 1-3, 5, 7 and 9, if $n, T \to \infty$ such that $n \geq T^\kappa$ for $\kappa > 0$, $I \leq T^\chi$ for $\chi > 0$, and $\epsilon \asymp T^{-\bar{\alpha}}$ for $2\bar{\beta} < \bar{\alpha} < 2(\bar{\delta} - \bar{\beta}) - \chi$, where $\bar{\delta} > 0$ is the convergence rate of estimator $\hat{\theta}$ in Assumption 3, we have $\mathbb{P}\left(\hat{k}_t^\epsilon = k_t,\ \forall t \in \mathcal{I}\right) \to 1$.*

We stress that the consistency of estimator $\hat{k}_t^\epsilon$ does not follow from the theory developed in Ahn, Horenstein (2013). Indeed, matrix $\hat{\Sigma}_{22}(Z_{t-1})$ is the nonparametrically estimated time-series *conditional* variance-covariance matrix of a vector of cross-sectional averages, and not a sample unconditional variance-covariance matrix of a large panel. Instead of truncation by a lower bound, regularization of the eigenvalues could be implemented by adding a vanishing positive quantity in the spirit of the perturbed eigenvalue criterion of Pelger (2019). Moreover, other estimators could be developed building on the approaches of e.g. Bai and Ng (2002) and Onatski (2009).

C.5 Parameters of interest from conditional portfolio choice

In this subsection we provide additional parameters of interest in class (2.3) (with representation as in (3.10)) that are inspired from conditional portfolio choice. The DML approach

proposed in this paper can be applied to estimate their unconditional moments.

i) Conditional tangency portfolios and mean–variance optimality. The implied SDF in (3.8) indicates that, up to scale, the centered SDF $m_{t-1,t} - \mathbb{E}(m_{t-1,t} \mid \mathscr{F}_{t-1})$ is proportional to the centered payoff of any portfolio with weights $\omega_t \propto \mathbb{V}(\xi_t \mid \mathscr{F}_{t-1})^\dagger \mathbb{E}(\xi_t \mid \mathscr{F}_{t-1}) =: v_t$, i.e. the conditional tangency direction in the ξ-span. [15] Using (3.3)–(3.5), process v_t is invariant to conditional transformations of the latent factor space. Two common normalizations of weights ω_t are

$$\text{unit-variance:} \qquad w_t^{(\sigma=1)} := \frac{v_t}{\sqrt{\mathbb{E}(\xi_t \mid \mathscr{F}_{t-1})' \, \mathbb{V}(\xi_t \mid \mathscr{F}_{t-1})^\dagger \, \mathbb{E}(\xi_t \mid \mathscr{F}_{t-1})}}, \tag{C.2}$$

$$\text{unit-budget:} \qquad w_t^{(1)} := \frac{v_t}{1_K' v_t} \quad \text{(when a budget constraint is meaningful).} \tag{C.3}$$

The unconstrained conditional mean–variance problem (for portfolio excess returns) with risk aversion $A > 0$ is $\max_{w \in \mathbb{R}^K} \; \mathbb{E}(\xi_t \mid \mathscr{F}_{t-1})' w - \frac{A}{2} w' \mathbb{V}(\xi_t \mid \mathscr{F}_{t-1}) w$. Using the conditional factor structure and equations (3.2) and (3.3), the solution with minimum norm is:

$$w_t^\star(A) = \frac{1}{A} \, \mathbb{V}(\xi_t \mid \mathscr{F}_{t-1})^\dagger \mathbb{E}(\xi_t \mid \mathscr{F}_{t-1}), \tag{C.4}$$

which yields another normalization of the conditional tangency portfolio.

ii) Conditional maximum Sharpe ratio. The conditional maximal Sharpe ratio in the ξ-span is

$$\mathrm{SR}_{\max,t}^2 = (\lambda_t^*)' \Lambda_{t-1}^{-1} \lambda_t^* = \mathbb{E}(\xi_t \mid \mathscr{F}_{t-1})' \, \mathbb{V}(\xi_t \mid \mathscr{F}_{t-1})^\dagger \, \mathbb{E}(\xi_t \mid \mathscr{F}_{t-1}). \tag{C.5}$$

The tight conditional Hansen–Jagannathan bound holds: $\frac{\mathbb{V}\left(m_{t-1,t} \mid \mathscr{F}_{t-1}\right)^{1/2}}{\mathbb{E}\left(m_{t-1,t} \mid \mathscr{F}_{t-1}\right)} = \frac{1}{\mathrm{SR}_{\max,t}}$.

Processes (C.2)-(C.5) are in class (2.3) and are invariant to latent factor normalization.

D Proofs of Propositions 3-6 and Lemmas 1-11

D.1 Proof of Proposition 3

From Assumption 3, 5, 7 and 9, and Proposition 2 a), the event with $\|\hat{\theta} - \theta_0\|_{L^2} \leq T^{-\delta^*}$, for $\delta^* < \bar{\delta}$, and $\hat{k}_t^\alpha = k_t$, $\forall t \in \mathcal{I}$, has probability approaching 1. Thus, without loss of generality we can work on that event in the rest of the proof. Then, by the triangular inequality and

[15] The SDF in (3.8) solves the dual (minimum-variance SDF) problem, which is equivalent to the primal (maximum-Sharpe / mean–variance) portfolio problem; see Hansen and Jagannathan (1991).

Assumption 12 (i), we get:

$$\frac{1}{I}\sum_{t\in\mathcal{I}}\|\hat{\gamma}_t-\gamma_t\|\leq\frac{1}{I}\sum_{t\in\mathcal{I}}v_t\|\hat{\xi}_t-\xi_t\|+\frac{1}{I}\sum_{t\in\mathcal{I}}\|\psi(\hat{\theta}(Z_{t-1}),\xi_t,k_t)-\psi(\theta_0(Z_{t-1}),\xi_t,k_t)\|. \quad \text{(D.1)}$$

By the Cauchy-Schwarz inequality, Assumption 12 (i), and bound $\sup_t \mathbb{E}[\|\hat{\xi}_t-\xi_t\|^2]=O(1/n)$ (see Step 5 in the proof of Proposition 2 b)), the first term on the RHS is $O_p(n^{-1/2})$. Let us now consider the second term on the RHS of (D.1) and show that

$$D(\hat{\theta},\mathcal{I}):=\frac{1}{I}\sum_{t\in\mathcal{I}}\|\psi(\hat{\theta}(Z_{t-1}),\xi_t,k_t)-\psi(\theta_0(Z_{t-1}),\xi_t,k_t)\|=O_p(\|\hat{\theta}-\theta_0\|_{L^2}). \quad \text{(D.2)}$$

We use $E^*\left[D(\hat{\theta},\mathcal{I})\right]=\int\|\psi(\hat{\theta}(z),\xi,k(z))-\psi_1(\theta_0(z),\xi,k(z))\|dP_0(z,\xi)\leq c\|\hat{\theta}-\theta_0\|_{L^2}$, for a constant c, from Assumption 12 (ii). Then, Lemma 1 a) yields (D.2). From (D.1) we get $\frac{1}{I}\sum_{t\in\mathcal{I}}\|\hat{\gamma}_t-\gamma_t\|=O_p\left(\frac{1}{\sqrt{n}}+\|\hat{\theta}-\theta_0\|_{L^2}\right)$. Then, from Assumption 3 we have $\|\hat{\theta}-\theta_0\|_{L^2}=O_p((\frac{\log T}{T})^{\bar{\delta}})$, and $n\geq T^\kappa$ with $\kappa>2\bar{\delta}$ implies that $\frac{1}{\sqrt{n}}=o((\frac{\log T}{T})^{\bar{\delta}})$. The conclusion follows.

D.2 Proof of Proposition 4

We focus on estimator $\hat{\Sigma}$ (the arguments for estimator $\hat{\mu}$ are similar). We build on the arguments in Chen and Shen (1998) and Chen and White (1999) which establish the convergence rate of Sieve estimators in time series. We use an argument as in the proof of Theorem 3.1 in Ai and Chen (2003) to control the estimation error induced by using $\hat{\xi}_t$ and $\hat{\mu}(\cdot)$ instead of ξ_t and $\mu_0(\cdot)$. Specifically, we first establish a slow convergence rate (Step 1) and then show that this rate can be improved by a refinement using the local curvature of the criterion function (Step 2). By iterating this argument, the desired rate is established (Step 3).

Define criteria $Q_T(\Sigma)=\frac{1}{T}\sum_{t\in\mathcal{T}}\|(X_t-\mu_{t-1})(X_t-\mu_{t-1})'-\Sigma(Z_{t-1})\|^2=:\frac{1}{T}\sum_{t\in\mathcal{T}}l(Y_t,\Sigma)$ and $\hat{Q}_T(\Sigma)=\frac{1}{T}\sum_{t\in\mathcal{T}}\left\|(\hat{X}_t-\hat{\mu}_{t-1})(\hat{X}_t-\hat{\mu}_{t-1})'-\Sigma(Z_{t-1})\right\|^2$ for $\Sigma\in\Theta_T^\Sigma$, where $\mu_{t-1}=\mu_0(Z_{t-1})$ and $\hat{\mu}_{t-1}=\hat{\mu}(Z_{t-1})$. The criterion $\hat{Q}_T$ is minimized by $\hat{\Sigma}$, while criterion Q_T is an infeasible version based on the true values of $X_t-\mu_{t-1}$. Moreover, we have:

$$K(\Sigma,\Sigma_0):=\mathbb{E}[l(Y_t,\Sigma)-l(Y_t,\Sigma_0)]=\mathbb{E}[\|\Sigma(Z_{t-1})-\Sigma_0(\Sigma_{t-1})\|^2]=\|\Sigma-\Sigma_0\|_{L^2}^2. \quad \text{(D.3)}$$

The next Lemma is proved with results in Chen and Shen (1998), Chen and White (1999).

Lemma 3. *Under the assumptions of Proposition 4, the Sieve space Θ_T^Σ is such that: (i) $\|\Sigma_0-\pi_T\Sigma_0\|_{L^2}=O(M_T^{-b})$ where $b=1/2+1/(d+1)$, for a $\pi_T\Sigma_0\in\Theta_T^\Sigma$, and integer $M_T\to\infty$, and (ii) the L^2 ϵ-bracketing metric entropy of set $\mathcal{F}_T=\{l(\cdot,\Sigma)-l(\cdot,\Sigma_0):\Sigma\in\Theta_T^\Sigma\}$ is*

$H(\epsilon, \mathcal{F}_T) \leq AM_T \log(AM_T/\epsilon)$, *where* A *is a constant.*

We give the proof of Proposition 4 for $M_T \to \infty$ such that the two components of the convergence rate $\delta_T = \max\{M_T^{-b}, (\frac{M_T \log M_T}{T})^{1/2}\}$ are balanced, i.e.

$$M_T^{-b} \asymp (\frac{M_T \log M_T}{T})^{1/2}, \tag{D.4}$$

which holds true for the optimal sequence M_T achieving the faster convergence rate.

STEP 1 Let $\delta_{0T} := \max\{M_T^{-b}, (\frac{M_T \log M_T}{T})^{1/4}\}$ (note that $\delta_{0,T} = (\frac{M_T \log M_T}{T})^{1/4}$ under condition (D.4)). We show that:

$$\|\hat{\Sigma} - \Sigma_0\|_{L^2} = O_p(\delta_{0T}). \tag{D.5}$$

For any $x > 0$ we have:

$$\begin{aligned}
&\mathbb{P}[\|\hat{\Sigma} - \Sigma_0\|_{L^2} \geq x\delta_{0,T}] \leq \mathbb{P}\left[\inf_{\Sigma \in \Theta_T^\Sigma : \|\Sigma - \Sigma_0\|_{L^2} \geq x\delta_{0,T}} \hat{Q}_T(\Sigma) - \hat{Q}_T(\Sigma_0) \leq \hat{Q}_T(\pi_T\Sigma_0) - \hat{Q}_T(\Sigma_0)\right] \\
&\leq \mathbb{P}\left[\inf_{\Sigma \in \Theta_T^\Sigma : \|\Sigma - \Sigma_0\|_{L^2} \geq x\delta_{0,T}} Q_T(\Sigma) - Q_T(\Sigma_0) \leq Q_T(\pi_T\Sigma_0) - Q_T(\Sigma_0) + 2\delta_{0T}^2\right] \\
&\quad + \mathbb{P}\left[\sup_{\Sigma \in \Theta_T^\Sigma} |\hat{Q}_T(\Sigma) - \hat{Q}_T(\Sigma_0) - (Q_T(\Sigma) - Q_T(\Sigma_0))| \geq \delta_{0T}^2\right] =: P_1 + P_2.
\end{aligned} \tag{D.6}$$

Consider first P_1, and use $Q_T(\Sigma) - Q_T(\Sigma_0) = \mu_T(l(\cdot, \Sigma) - l(\cdot, \Sigma_0)) + K(\Sigma, \Sigma_0)$, where $\mu_T(f(\cdot)) := \frac{1}{T}\sum_{t \in \mathcal{T}}(f(Y_t) - \mathbb{E}[f(Y_t)])$. Then:

$$P_1 \leq \mathbb{P}\left[\inf_{\Sigma \in \Theta_T^\Sigma : \|\Sigma - \Sigma_0\|_{L^2} \geq x\delta_{0,T}} K(\Sigma, \Sigma_0) \leq K(\pi_T\Sigma_0, \Sigma_0) + 2\sup_{\Sigma \in \Theta_T^\Sigma} \mu_T[l(\cdot, \Sigma) - l(\cdot, \Sigma_0)] + 2\delta_{0T}^2\right].$$

Now, using $\inf_{\Sigma \in \Theta_T^\Sigma : \|\Sigma - \Sigma_0\| \geq x\delta_{0,T}} K(\Sigma, \Sigma_0) = (x\delta_{0T})^2$ and $K(\pi_T\Sigma_0, \Sigma_0) = \|\Sigma_0 - \pi_T\Sigma_0\|_{L^2}^2 \leq c\delta_{0T}^2$ from (D.3) and Lemma B.3 (i), we get $P_1 \leq \mathbb{P}\left[\sup_{\Sigma \in \Theta_T^\Sigma} \mu_T[l(\cdot, \Sigma) - l(\cdot, \Sigma_0)] \geq c(x)\delta_{0T}^2\right]$, where $c(x) = \frac{1}{2}(x^2 - 2 - c) > 0$ for $x > \sqrt{c+2}$. By using Lemma 3 (ii) and Assumption 5, condition $\delta_{0T} \geq (\frac{M_T \log M_T}{T})^{1/4}$ and Lemma 1 in Chen and Shen (1998), we can show

$$\mathbb{P}\left[\sup_{\Sigma \in \Theta_T^\Sigma} \mu_T[l(\cdot, \Sigma) - l(\cdot, \Sigma_0)] \geq c(x)\delta_{0T}^2\right] \leq 6\exp\left(-\frac{Tc(x)^2\delta_{0T}^4}{18\sigma^2(1 + a_{T1}C)}\right) + 2(a_{T2} - 1)\beta(a_{T1}),$$

for constants $C, \sigma^2 > 0$, where $a_{T2} = [\frac{T}{2a_{T1}}] \to \infty$ and $a_{T1} = T^{\bar{c}}$, $\bar{c} > 0$, such that $a_{T1} \leq T\delta_{0T}^4$. On the RHS, the second term is $o(1)$ by Assumption 5, and the first term can be made small by choosing x large. Thus, for any $\epsilon > 0$, we have $P_1 \leq \epsilon$ for x and T large enough.

Let us now consider P_2, and control $\hat{Q}_T(\Sigma) - Q_T(\Sigma)$ uniformly w.r.t. $\Sigma \in \Theta_T^\Sigma$. Let us write $\hat{X}_t - \hat{\mu}_{t-1} = X_t - \mu_{t-1} + \Delta_t$, where $\Delta_t = (0', \hat{\xi}_t' - \xi_t')' - [\hat{\mu}(Z_{t-1}) - \mu_0(Z_{t-1})]$. Then:

$$\begin{aligned}
& \left\| (\hat{X}_t - \hat{\mu}_{t-1})(\hat{X}_t - \hat{\mu}_{t-1})' - \Sigma(Z_{t-1}) \right\|^2 \\
= \; & \| (X_t - \mu_{t-1})(X_t - \mu_{t-1})' - \Sigma(Z_{t-1}) + (X_t - \mu_{t-1})\Delta_t' + \Delta_t (X_t - \mu_{t-1})' + \Delta_t \Delta_t' \|^2 \\
= \; & \| (X_t - \mu_{t-1})(X_t - \mu_{t-1})' - \Sigma(Z_{t-1}) \|^2 + \| (X_t - \mu_{t-1})\Delta_t' + \Delta_t (X_t - \mu_{t-1})' + \Delta_t \Delta_t' \|^2 \\
& + 4(X_t - \mu_{t-1})'[(X_t - \mu_{t-1})(X_t - \mu_{t-1})' - \Sigma(Z_{t-1})]\Delta_t \\
& + 2\Delta_t'[(X_t - \mu_{t-1})(X_t - \mu_{t-1})' - \Sigma(Z_{t-1})]\Delta_t.
\end{aligned}$$

Thus we get:

$$\hat{Q}_T(\Sigma) - \hat{Q}_T(\Sigma_0) - [Q_T(\Sigma) - Q_T(\Sigma_0)] = -4\frac{1}{T}\sum_{t\in\mathcal{T}} (X_t - \mu_{t-1})'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})]\Delta_t$$

$$-2\frac{1}{T}\sum_{t\in\mathcal{T}} \Delta_t'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})]\Delta_t = \sum_{j=1}^{5} \hat{I}_j(\Sigma), \tag{D.7}$$

where $\hat{I}_1(\Sigma) := -4\frac{1}{T}\sum_{t\in\mathcal{T}}\{(X_t - \mu_{t-1})'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})]\}_2(\hat{\xi}_t - \xi_t)$, $\hat{I}_2(\Sigma) := 4\frac{1}{T}\sum_{t\in\mathcal{T}}(X_t - \mu_{t-1})'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})][\hat{\mu}(Z_{t-1}) - \mu_0(Z_{t-1})]$, $\hat{I}_3(\Sigma) := -2\frac{1}{T}\sum_{t\in\mathcal{T}}(\hat{\xi}_t - \xi_t)'\{\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})\}_{22}(\hat{\xi}_t - \xi_t)$, $\hat{I}_4(\Sigma) := +4\frac{1}{T}\sum_{t\in\mathcal{T}}(\hat{\xi}_t - \xi_t)'\{[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})][\hat{\mu}(Z_{t-1}) - \mu_0(Z_{t-1})]\}_2$, $\hat{I}_5(\Sigma) := -2\frac{1}{T}\sum_{t\in\mathcal{T}}[\hat{\mu}(Z_{t-1}) - \mu_0(Z_{t-1})]'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})][\hat{\mu}(Z_{t-1}) - \mu_0(Z_{t-1})]$, and $\{\cdot\}_2$ and $\{\cdot\}_{22}$ denote the lower $K \times 1$ vector block, and the lower-right $K \times K$ matrix block.

Lemma 4. *Under Assumptions 4-6, we have:* $\hat{I}_1(\Sigma) = O_p(\frac{1}{\sqrt{n}}\epsilon_T)$, $\hat{I}_2(\Sigma) = o_p(\epsilon_T\delta_T)$, $\hat{I}_3(\Sigma) = O_p(\frac{1}{n}\epsilon_T)$, $\hat{I}_4(\Sigma) = O_p(\frac{1}{\sqrt{n}}\epsilon_T\delta_T)$, $\hat{I}_5(\Sigma) = O_p(\delta_T^2)$, *uniformly in* $\Sigma \in \Theta_T^\Sigma$ *such that* $\|\Sigma - \Sigma_0\|_{L^2} \leq \epsilon_T$, *where* $\delta_T = \max\{M_T^{-b}, (\frac{M_T \log M_T}{T})^{1/2}\}$ *and* $\epsilon_T > 0$ *with* $\delta_T = o(\epsilon_T)$.

Note that condition (D.4) implies that $\delta_T = O((\frac{M_T \log M_T}{T})^{1/2}) = O(\delta_{0,T}^2)$, and condition $\frac{T}{M_t \log M_T} = o(n)$ implies $\frac{1}{\sqrt{n}} = o((\frac{M_T \log M_T}{T})^{1/2}) = o(\delta_{0,T}^2)$. Then, from Lemma 4 (with ϵ_T a positive constant large enough), we get $\hat{Q}_T(\Sigma) - \hat{Q}_T(\Sigma_0) - [Q_T(\Sigma) - Q_T(\Sigma_0)] = o_p(\delta_{0,T}^2)$, uniformly w.r.t. $\Sigma \in \Theta_T^\Sigma$. Then, $P_2 = o(1)$. From (D.6) we get (D.5).

Step 2 Let us define $\delta_{1,T} = \max\{[\sqrt{\frac{M_T \log M_T}{T}}\delta_{0,T}]^{1/2}, M_T^{-b}\} \ll \delta_{0T}$. Then for $x > 0$:

$$\begin{aligned}
& \mathbb{P}[\|\hat{\Sigma} - \Sigma_0\|_{L^2} \geq x\delta_{1,T}] = \mathbb{P}[\|\hat{\Sigma} - \Sigma_0\|_{L^2} \geq x\delta_{0,T}] + \mathbb{P}[x\delta_{0,T} \geq \|\hat{\Sigma} - \Sigma_0\|_{L^2} \geq x\delta_{1,T}] \\
\leq \; & \bar{\epsilon} + o(1) + \mathbb{P}\left[\inf_{\Sigma\in\Theta_T^\Sigma : x\delta_{0,T} \geq \|\Sigma - \Sigma_0\|_{L^2} \geq x\delta_{1,T}} \hat{Q}_T(\Sigma) - \hat{Q}_T(\Sigma_0) \leq \hat{Q}_T(\pi_T\Sigma_0) - \hat{Q}_T(\Sigma_0) \right] \\
\leq \; & \bar{\epsilon} + o(1) + \mathbb{P}\left[\inf_{\Sigma\in\Theta_T^\Sigma : x\delta_{0,T} \geq \|\Sigma - \Sigma_0\|_{L^2} \geq x\delta_{1,T}} Q_T(\Sigma) - Q_T(\Sigma_0) \leq Q_T(\pi_T\Sigma_0) - Q_T(\Sigma_0) + 2\delta_{1,T}^2 \right]
\end{aligned}$$

$$+\mathbb{P}\left[\sup_{\Sigma\in\Theta_T^\Sigma:\|\Sigma-\Sigma_0\|_{L^2}\leq x\delta_{0,T}}|\hat{Q}_T(\Sigma)-\hat{Q}_T(\Sigma_0)-(Q_T(\Sigma)-Q_T(\Sigma_0))|\geq\delta_{1,T}^2\right]=:\bar{\epsilon}+o(1)+P_3+P_4,$$

for any $\bar{\epsilon}>0$, and x and T large. By similar arguments as in Step 1, and using Lemma 1 in Chen and Shen (1998) and the condition $\delta_{1T}^2\geq\sqrt{\frac{M_T\log M_T}{T}}\delta_{0,T}$, we have:

$$\begin{aligned}P_3 &\leq \mathbb{P}\left[\sup_{\Sigma\in\Theta_T^\Sigma:\|\Sigma-\Sigma_{L^2}\|\leq x\delta_{0T}}\mu_T[l(\cdot,\Sigma)-l(\cdot,\Sigma_0)]\geq c(x)\delta_{1,T}^2\right]\\ &\leq 6\exp\left(-\frac{Tc(x)^2\delta_{1,T}^4}{18C_1x^2\delta_{0T}^2(1+a_{T1}C)}\right)+2(a_{T2}-1)\beta(a_{T1})\leq\bar{\epsilon}+o(1),\end{aligned}$$

for any $\bar{\epsilon}>0$, and large x and T, if $a_{T2}=[T/(2a_{T1})]\to\infty$ and $a_{1T}\to\infty$ such that $a_{1,T}\leq T\delta_{1T}^4/\delta_{0T}$. Moreover, by (D.7), Lemma 4 (with $\epsilon_T=x\delta_{0T}$), and the conditions $\frac{T}{M_t\log M_T}=o(n)$ and (D.4) we have: $\sup_{\Sigma\in\Theta_T^\Sigma:\|\Sigma-\Sigma\|_{L^2}\leq x\delta_{0T}}|\hat{Q}_T(\Sigma)-\hat{Q}_T(\Sigma_0)-(Q_T(\Sigma)-Q_T(\Sigma_0))|=O_p\left(\frac{\delta_{0T}}{\sqrt{n}}\right)+o_p\left(\delta_{0T}\delta_T\right)=o_p(\delta_{1,T}^2)$, which yields $P_4=o(1)$. Thus, $\|\hat{\Sigma}-\Sigma_0\|_{L^2}=O_p(\delta_{1,T})$.

Step 3 By iterating the argument in Step 2, we establish the sequence of probability bounds $\|\hat{\Sigma}-\Sigma_0\|_{L^2}=O_p(\delta_{j,T})$, where the rates $\delta_{j,T}$, $j=1,...,$ are defined recursively as $\delta_{j,T}=\max\{[\sqrt{\frac{M_T\log M_T}{T}}\delta_{j-1,T}]^{1/2},M_T^{-b}\}$, for $j=1,2,\cdots$. If $\delta_{j,T}\leq M_T^{-b}$ for some j, the conclusion follows. If not, the iteration is continued. Since $[\sqrt{\frac{M_T\log M_T}{T}}\delta_{j-1,T}]^{1/2}$ is the geometric average of $\sqrt{\frac{M_T\log M_T}{T}}$ and $\delta_{j-1,T}$, the rate $\delta_{j,T}$ approaches $\delta_T=\max\{\sqrt{\frac{M_T\log M_T}{T}},M_T^{-b}\}$ as j grows. The conclusion follows.

D.3 Proof of Proposition 5

We present the proof for the matrix Lasso estimator $\hat{\Sigma}(\cdot)$ (the proof for $\hat{\mu}(\cdot)$ is similar).

Step 1 We show $\|\tilde{\Sigma}-\Sigma_0\|_{L^2}=O_p\left((\frac{\log^2(T)}{T})^{\frac{a}{2a+3}}\right)$. We focus on a given component l, and to ease notation we set $\hat{\rho}_l\equiv\hat{\rho}_{\Sigma,l}=\arg\min_{\rho\in\mathbb{R}^p}\frac{1}{T}\sum_{t\in\mathcal{T}}\left(\hat{\zeta}_{t,l}-h(Z_{t-1})'\rho\right)^2+2\alpha\|\rho\|_1$ with $\hat{\theta}_l(\cdot)=h(\cdot)'\hat{\rho}_l$ and $\theta_{0,l}(\cdot)=[vech(\Sigma_0(\cdot))]_l$. The proof is based on the arguments in Chernozhukov, Newey and Singh (2018) with some changes in the set of assumptions. Write $\hat{\rho}_l=\arg\min_{\rho\in\mathbb{R}^p}\left(-2\hat{M}_l'\rho+\rho'\hat{G}\rho+2\alpha\|\rho\|_1\right)$, where $\hat{M}_l:=\frac{1}{T}\sum_{t\in\mathcal{T}}\hat{\zeta}_{t,l}h(Z_{t-1})$ and $\hat{G}:=\frac{1}{T}\sum_{t\in\mathcal{T}}h(Z_{t-1})h(Z_{t-1})'$. We show the next result using β-mixing and controlling for the effects of estimating $\zeta_{t,l}=[vech((X_t-\mu_{t-1})(X_t-\mu_{t-1})')]_l$ by $\hat{\zeta}_{t,l}$.

Lemma 5. *Under Assumptions 4-6 and 14, and conditions $\frac{T}{\log^2 T}=O(n)$ and $p=O(T^{\bar{b}})$ for $\bar{b}>0$, we have a) $\|\hat{G}-G\|_\infty=O_p(\eta_T)$ and b) $\|\hat{M}_l-M_l\|_\infty=O_p(\eta_T)$, for $M_l=\mathbb{E}[\zeta_{t,l}h(Z_{t-1})]$,*

where $\eta_T = \frac{\log T}{\sqrt{T}}$ *and* $\|\cdot\|_\infty$ *denotes the uniform matrix norm.*

Then, by the arguments in the proof of Theorem 3 in Chernozhukov, Newey and Singh (2018), we can show the next lemma.

Lemma 6. *Under the Assumptions of Lemma 5, Assumption 13, and the conditions* $\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 \leq \bar{s}\eta_T^2$ *and* $\eta_T(1+\bar{B}) = o_p(\alpha)$ *with* $\bar{B} := \|\bar{\rho}_l\|_1$*, we have* $\|\hat{\theta}_l - \theta_{0,l}\|_{L^2}^2 = O_p\left(\alpha^2\bar{s}\right)$.

Now, we use the sparsity condition $\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2} \leq C\bar{s}^{-a-1}$ for $a, C > 0$ (Assumption 13), and $\bar{B} = O(\bar{s})$. Then, the condition $\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 \leq \bar{s}\eta_T^2$ in Lemma 6 holds with $\bar{s} \geq C\eta_T^{-\frac{2}{2a+3}}$, and condition $\eta_T(1+\bar{B}) = o_p(\alpha)$ in Lemma 6 is met if α tends to zero slightly slower than $\eta_T^{\frac{2a+1}{2a+3}}$. We conclude from Lemma 6 that $\|\hat{\theta}_l - \theta_{0,l}\|_{L^2} = O_p\left(\eta_T^{\frac{2a}{2a+3}} l_T^{-1}\right) = O_p\left((\frac{\log^2 T}{T})^{\frac{a}{2a+3}} l_T^{-1}\right)$, for any $l_T = o(1)$. This bound holds for any l, and the statement in Step 1 follows.

Step 2 We account for the effect of regularization. We use $\hat{\Sigma}(\cdot) = [\tilde{\Sigma}(\cdot)]^\epsilon$ and the next result, where $\|\cdot\|_2$ is the 2-norm of a matrix, that is equal to its largest singular value.

Lemma 7. *We have* $\|[A]^\epsilon - A_0\|_2 \leq 2(\|A - A_0\| + \epsilon)$ *for symmetric matrices* A, A_0.

From Lemma 7 and the equivalence of matrix norms, $\|\hat{\Sigma} - \Sigma_0\|_{L^2} = O_p\left(\|\tilde{\Sigma} - \Sigma_0\|_{L^2} + \epsilon\right)$. Then, from Step 1 and the condition $\epsilon = O\left((\frac{\log^2 T}{T})^{\frac{a}{2a+3}}\right)$, the conclusion follows.

D.4 Proof of Proposition 6

For a symmetric matrix Σ with SVD $\Sigma = UDU'$, let us define $[\Sigma]^\epsilon = UD^\epsilon U'$ and $(D^\epsilon)_{jj} = \max\{D_{jj}, \epsilon\} = [D_{jj}]_\epsilon$, where $\epsilon > 0$ tends to zero with n, T. To ease exposition, we use the following short notation: $\hat{\Sigma}_t^\epsilon \equiv [\hat{\Sigma}_{22}(Z_{t-1})]^\epsilon$, $\hat{\Sigma}_t \equiv \hat{\Sigma}_{22}(Z_{t-1})$, $\Sigma_t \equiv \Sigma_{22,0}(Z_{t-1})$, and $\hat{\Phi}_t^\epsilon = \hat{\Sigma}_t^\epsilon - \Sigma_t$ and $\hat{\Phi}_t = \hat{\Sigma}_t - \Sigma_t$. Then, the eigenvalue ratio is $\hat{\varrho}_{j,t} = \delta_j(\hat{\Sigma}_t^\epsilon)/\delta_{j+1}(\hat{\Sigma}_t^\epsilon)$.

Step 1: We first show that there exists a constant $C > 0$ such that we have $\hat{k}_t^\epsilon = k_t$ for $t \in \mathcal{I}$ whenever $\|\hat{\Phi}_t^\epsilon\|_2 \leq C\underline{c}_T\sqrt{\epsilon}$, where $\|\cdot\|_2$ is the 2-norm and $\underline{c}_T = \frac{T^{-\bar{\beta}}}{(\log T)^{1/\bar{c}}}$, with $\bar{\beta}, \bar{c}$ defined in Assumption 7. For this purpose, we first note that from the Weyl's inequalities we have $|\delta_j(\hat{\Sigma}_t^\epsilon) - \delta_j(\Sigma_t)| \leq \|\hat{\Phi}_t^\epsilon\|_2$, for any j. Moreover, we have $\delta_j(\Sigma_t) = 0$ for any $j \geq k_t + 1$, and $\bar{C} \geq \delta_j(\Sigma_t) \geq \underline{c}_T$, for all $j \leq k_t$, for any $t \in \mathcal{I}$, with probability approaching 1 by Assumption 7 (see Step 1 in the proof Proposition 2 a)). We then establish some inequalities for the eigenvalue ratios depending on the lag j.

(i) Let $j \leq k_t - 1$ and $t \in \mathcal{I}$. We have $\delta_j(\hat{\Sigma}_t^\epsilon) \leq \bar{C} + \|\hat{\Phi}_t^\epsilon\|_2$. Moreover, we have $\delta_{j+1}(\hat{\Sigma}_t^\epsilon) \geq \underline{c}_T/2$ if $\|\hat{\Phi}_t^\epsilon\|_2 \leq \underline{c}_T/2$, and $\delta_{j+1}(\hat{\Sigma}_t^\epsilon) \geq \epsilon$ otherwise. Thus, we have $\hat{\varrho}_{j,t} \leq \frac{2}{\underline{c}_T}(\bar{C} + \|\hat{\Phi}_t^\epsilon\|_2) + 1\{\|\hat{\Phi}_t^\epsilon\|_2 \geq \underline{c}_T/2\}\frac{1}{\epsilon}(\bar{C} + \|\hat{\Phi}_t^\epsilon\|_2)$. Now we use that $1\{x \geq \tau\} \leq \frac{1}{\tau}x$, for

any $x \geq 0$ and $\tau > 0$. We get:

$$\hat{\varrho}_{j,t} \leq \frac{2}{\underline{c}_T}(\bar{C} + \|\hat{\Phi}_t^\epsilon\|_2) + \frac{2}{\epsilon \underline{c}_T}\|\hat{\Phi}_t^\epsilon\|_2(\bar{C} + \|\hat{\Phi}_t^\epsilon\|_2) = \frac{1}{\underline{c}_T}\left(c_1 + c_2\frac{1}{\epsilon}\|\hat{\Phi}_t^\epsilon\|_2 + c_3\frac{1}{\epsilon}\|\hat{\Phi}_t^\epsilon\|_2^2\right), \quad \text{(D.8)}$$

for constants $c_1, c_2, c_3 > 0$ (assuming $\epsilon < 1$).

(ii) Let $j = k_t$. Then, $\delta_j(\hat{\Sigma}_t^\epsilon) \geq \underline{c}_T - \|\hat{\Phi}_t^\epsilon\|_2$ and $\epsilon \leq \delta_{j+1}(\hat{\Sigma}_t^\epsilon) \leq \|\hat{\Phi}_t^\epsilon\|_2$. Thus $\hat{\varrho}_{j,t} \geq \underline{c}_T\|\hat{\Phi}_t^\epsilon\|_2^{-1} - 1$.

(iii) Let $j \geq k_t + 1$ and $j \leq k_{\max}$. Then, $\delta_j(\hat{\Sigma}_t^\epsilon) \leq \|\hat{\Phi}_t^\epsilon\|_2$ and $\delta_{j+1}(\hat{\Sigma}_t^\epsilon) \geq \epsilon$. Thus $\hat{\varrho}_{j,t} \leq \frac{1}{\epsilon}\|\hat{\Phi}_t^\epsilon\|_2$.

By the criterion of the maximum eigenvalue ratio, and the inequality (D.8) in case (i), and those in cases (ii)-(iii), we deduce that $\hat{k}_t^\epsilon = k_t$, if the sufficient condition

$$\begin{cases} \underline{c}_T\|\hat{\Phi}_t^\epsilon\|_2^{-1} - 1 \geq \frac{1}{\underline{c}_T}\left(c_1 + c_2\frac{1}{\epsilon}\|\hat{\Phi}_t^\epsilon\|_2 + c_3\frac{1}{\epsilon}\|\hat{\Phi}_t^\epsilon\|_2^2\right) \\ \underline{c}_T\|\hat{\Phi}_t^\epsilon\|_2^{-1} - 1 \geq \frac{1}{\epsilon}\|\hat{\Phi}_t^\epsilon\|_2, \end{cases} \quad \text{(D.9)}$$

holds. The system of inequalities (D.9) for $\|\hat{\Phi}_t^\epsilon\|_2$ holds if $\|\hat{\Phi}_t^\epsilon\|_2 \leq C\underline{c}_T\sqrt{\epsilon}$, for a constant $C > 0$ small enough (we use the condition $\epsilon \ll \underline{c}_T^2$).

Step 2: We link the 2-norm of the matrix estimation error with regularization, i.e. $\|\hat{\Phi}_t^\epsilon\|_2$, to the Frobenius norm without regularization, i.e. $\|\hat{\Phi}_t\|$. Using Lemma 7 we have $\|\hat{\Phi}_t^\epsilon\|_2 \leq 2(\|\hat{\Phi}_t\| + \epsilon)$. Using that $\epsilon \ll \underline{c}_T\sqrt{\epsilon}$ (which follows from $\epsilon \ll \underline{c}_T^2$), and Step 1, we get that there exists a constant $\tilde{C} > 0$ such that we have $\hat{k}_t^\epsilon = k_t$ whenever $\|\hat{\Phi}_t\| \leq \tilde{C}\underline{c}_T\sqrt{\epsilon}$.

Step 3: We use $\|\hat{\Phi}_t\| \leq \|\hat{\Sigma}(Z_{t-1}) - \Sigma_0(Z_{t-1})\|$, and the result in Step 2, to get

$$\mathbb{P}[\hat{k}_t^\varepsilon = k_t, \ \forall t \in \mathcal{I}] \geq 1 - \mathbb{P}\left(d(\hat{\theta}, \theta_0; \mathcal{I}) \geq \tilde{C}\frac{\underline{c}_T\sqrt{\epsilon}}{\sqrt{I}}\right), \quad \text{(D.10)}$$

where $d(\hat{\theta}, \theta_0; \mathcal{I}) = (\frac{1}{I}\sum_{t\in\mathcal{I}}\|\hat{\Sigma}(Z_{t-1}) - \Sigma_0(Z_{t-1})\|^2)^{1/2}$. Step 3 in the proof of Proposition 2 a) shows that the probability on the RHS of (D.10) vanishes asymptotically under Assumption 3, i.e. $\mathbb{P}\left(d(\hat{\theta}, \theta_0; \mathcal{I}) \geq \tilde{C}\frac{\underline{c}_T\sqrt{\epsilon}}{\sqrt{I}}\right) = o(1)$, if $\frac{\underline{c}_T\sqrt{\epsilon}}{\sqrt{I}} \gg (\frac{T}{\log T})^{-\bar{\delta}}$. Both conditions $\epsilon \ll \underline{c}_T^2$ and $\frac{\underline{c}_T\sqrt{\epsilon}}{\sqrt{I}} \gg (\frac{T}{\log T})^{-\bar{\delta}}$ are satisfied if $\epsilon \asymp T^{-\bar{\alpha}}$ with $2\bar{\beta} < \bar{\alpha} < 2(\bar{\delta} - \bar{\beta}) - \chi$. The conclusion follows from (D.10).

D.5 Proof of Lemma 1

Without loss of generality, $\hat{d} \geq 0$. a) By the β-mixing condition in Assumption 5, we have $\mathbb{P}(\hat{d} \geq C\|\hat{\theta} - \theta_0\|_{L^2}|\mathscr{Y}_T) = \mathbb{P}^*(\hat{d} \geq C\|\hat{\theta} - \theta_0\|_{L^2}) + r_T$, for any constant $C > 0$, where

$\mathbb{E}(r_T) \leq \beta(a_T)$ and $\mathscr{Y}_T$ denotes the sigma-field of Y_t, $t \in \mathcal{T}$. Then:

$$\mathbb{P}(\hat{d} \geq C\|\hat{\theta} - \theta_0\|_{L^2}) = \mathbb{E}\left(\mathbb{P}(\hat{d} \geq C\|\hat{\theta} - \theta_0\|_{L^2}|\mathscr{Y}_T)\right) \leq \mathbb{E}\left(\mathbb{P}^*(\hat{d} \geq C\|\hat{\theta} - \theta_0\|_{L^2})\right) + \beta(a_T).$$

Moreover, $\mathbb{P}^*\left(\hat{d} \geq C\|\hat{\theta} - \theta_0\|_{L^2})\right) \leq \frac{1}{C\|\hat{\theta}-\theta_0\|_{L^2}}\mathbb{E}^*(\hat{d}) \leq \frac{\bar{c}}{C}$ by the Markov inequality for $\mathbb{P}^*$ and the hypothesis of part a). Then, $\mathbb{P}(\hat{d} \geq C\|\hat{\theta} - \theta_0\|_{L^2}) \leq \frac{\bar{c}}{C} + \beta(a_T)$, where the first term can be made small by choosing C large, and the second term on RHS is $o(1)$ by Assumption 9. Part (a) follows. To prove part b), a similar argument based on Assumptions 5 and 9 yields $\mathbb{P}(\hat{d} \geq \epsilon) \leq \mathbb{E}\left(\mathbb{P}^*(\hat{d} \geq \epsilon)\right) + \beta(a_T) = \mathbb{E}\left(\mathbb{P}^*(\hat{d} \geq \epsilon)\right) + o(1)$, for any constant $\epsilon > 0$. Because the random variable $\mathbb{P}^*(\hat{d} \geq \epsilon)$ is bounded and $o_p(1)$ by hypothesis, we get $\mathbb{E}\left(\mathbb{P}^*(\hat{d} \geq \epsilon)\right) = o(1)$, which yields $\mathbb{P}(\hat{d} \geq \epsilon) = o(1)$, for any $\epsilon > 0$. The conclusion follows.

D.6 Proof of Lemma 2

We use $\Sigma_{22}^{\dagger} = \sum_{j=1}^{k} \frac{1}{\lambda_j} w_j w_j'$, where the λ_j are the eigenvalues and the w_j are the eigenvectors of Σ_{22}. The differential is $d\Sigma_{22}^{\dagger} = \sum_{j=1}^{k}(-\frac{1}{\lambda_j^2})(d\lambda_j)w_j w_j' + \sum_{j=1}^{k} \frac{1}{\lambda_j}(dw_j)w_j' + \sum_{j=1}^{k} \frac{1}{\lambda_j} w_j (dw_j)'$, where $d\lambda_j = w_j'(d\Sigma_{22})w_j$ and $dw_j = \sum_{\ell=0,\ell\neq j}^{k} \frac{1}{\lambda_j - \lambda_\ell}\mathscr{P}_\ell(d\Sigma_{22})w_j$, with $\mathscr{P}_\ell = w_\ell w_\ell'$ the eigenprojector for eigenvalue λ_ℓ, and the sum includes the null eigenvalue $\lambda_0 = 0$ with eigenprojector $\mathscr{P}_0$. Then, we get:

$$\begin{aligned} d\Sigma_{22}^{\dagger} &= -\sum_{j=1}^{k} \frac{1}{\lambda_j^2}\mathscr{P}_j(d\Sigma_{22})\mathscr{P}_j + \sum_{j=1}^{k}\sum_{\ell=0,\ell\neq j}^{k} \frac{1}{\lambda_j(\lambda_j - \lambda_\ell)}\left[\mathscr{P}_\ell(d\Sigma_{22})\mathscr{P}_j + \mathscr{P}_j(d\Sigma_{22})\mathscr{P}_\ell\right] \\ &= -\sum_{j=1}^{k} \frac{1}{\lambda_j^2}\left(\mathscr{P}_j(d\Sigma_{22})\mathscr{P}_j - \mathscr{P}_0(d\Sigma_{22})\mathscr{P}_j - \mathscr{P}_j(d\Sigma_{22})\mathscr{P}_0\right) \\ &\quad + \sum_{j=1}^{k}\sum_{\ell=1,\ell\neq j}^{k} \frac{1}{\lambda_j(\lambda_j - \lambda_\ell)}\mathscr{P}_\ell(d\Sigma_{22})\mathscr{P}_j + \sum_{j=1}^{k}\sum_{\ell=1,\ell\neq j}^{k} \frac{1}{\lambda_j(\lambda_j - \lambda_\ell)}\mathscr{P}_j(d\Sigma_{22})\mathscr{P}_\ell. \end{aligned}$$

The sum of the last two terms on the RHS equals $-\sum_{j=1}^{k}\sum_{\ell=1,\ell\neq j}^{k} \frac{1}{\lambda_j\lambda_\ell}\mathscr{P}_\ell(d\Sigma_{22})\mathscr{P}_j$. Then:

$$\begin{aligned} d\Sigma_{22}^{\dagger} &= -\sum_{j=1}^{k} \frac{1}{\lambda_j^2}\left(\mathscr{P}_j(d\Sigma_{22})\mathscr{P}_j - \mathscr{P}_0(d\Sigma_{22})\mathscr{P}_j - \mathscr{P}_j(d\Sigma_{22})\mathscr{P}_0\right) - \sum_{j=1}^{k}\sum_{\ell=1,\ell\neq j}^{k} \frac{1}{\lambda_j\lambda_\ell}\mathscr{P}_\ell(d\Sigma_{22})\mathscr{P}_j \\ &= \mathscr{P}_0(d\Sigma_{22})(\sum_{j=1}^{k} \frac{1}{\lambda_j^2}\mathscr{P}_j) + (\sum_{j=1}^{k} \frac{1}{\lambda_j^2}\mathscr{P}_j)(d\Sigma_{22})\mathscr{P}_0 - (\sum_{j=1}^{k} \frac{1}{\lambda_j}\mathscr{P}_j)(d\Sigma_{22})(\sum_{\ell=1}^{k} \frac{1}{\lambda_\ell}\mathscr{P}_\ell) \\ &= \mathscr{P}_0(d\Sigma_{22})(\Sigma_{22}^{\dagger})^2 + (\Sigma_{22}^{\dagger})^2(d\Sigma_{22})\mathscr{P}_0 - (\Sigma_{22}^{\dagger})(d\Sigma_{22})(\Sigma_{22}^{\dagger}), \quad \text{and the conclusion follows.} \end{aligned}$$

D.7 Proof of Lemma 3

(i) Under the conditions of Proposition 4, we have $\Sigma_0 = C_0 C_0'$ with C_0 lower triangular and $C_{0,jk} \in \mathcal{H}^d$. Then, $\|C_{jk} - C_{0,jk}\|_{L^2} = O(M_T^{-b})$ with $b = 1/2 + 1/(d+1)$ for a $C_{j,k} \in \mathscr{N}_{M_T,B}^d$, where we set $\mathscr{N}_{M,B}^d = \Big\{\varphi(Z) = \sum_{m=1}^M v_m \phi(w_m' Z + b_m),\ v_m, b_m \in \mathbb{R},\ w_m \in \mathbb{R}^d,\ m = 1, ..., M,\ \sum_{m=1}^M |v_m| \leq B\Big\}$, $M \in \mathbb{N}$, for a constant B large enough (Chen and White (1999), Theorem 2.1). We use $\|\Sigma(z) - \Sigma_0(z)\| \leq \sqrt{\dim(\Sigma)}(\|C(z)\| + \|C_0(z)\|)\|C(z) - C_0(z)\|$. Under Assumption 6 we have that $\|C_0(z)\|$ is bounded uniformly in z. Further, $C \in \mathscr{N}_{M_T,B}^d$ is uniformly bounded too, since the activation function ϕ is bounded. Hence, for a $\Sigma \in \Theta_T^\Sigma$ we have $\|\Sigma - \Sigma_0\|_{L^2} \leq const \times \|C - C_0\|_{L^2} = O(M_T^{-b})$. (ii) The bound is shown along the lines of the proof of Proposition 1 of Chen and Shen (1998), p. 311.

D.8 Proof of Lemma 4

(i) Bound on $\hat{I}_1(\Sigma) = -4\frac{1}{T}\sum_{t\in\mathcal{T}}\{(X_t - \mu_{t-1})'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})]\}_2(\hat{\xi}_t - \xi_t)$. We have $|\hat{I}_1(\Sigma)| \leq 4[\frac{1}{T}\sum_{t\in\mathcal{T}}\|X_t - \mu_{t-1}\|^2\|\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})\|^2]^{1/2}[\frac{1}{T}\sum_{t\in\mathcal{T}}\|\hat{\xi}_t - \xi_t\|^2]^{1/2}$ and $\frac{1}{T}\sum_{t\in\mathcal{T}}\|\hat{\xi}_t - \xi_t\|^2 = O_p(\frac{1}{n})$ by Assumptions 4 and 6 (see Step 5 in the proof of Proposition 2 b)). Write $\frac{1}{T}\sum_{t\in\mathcal{T}}\|X_t - \mu_{t-1}\|^2\|\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})\|^2 = \frac{1}{T}\sum_{t\in\mathcal{T}}[\psi_t(\Sigma - \Sigma_0) - \mathbb{E}[\psi_t(\Sigma - \Sigma_0)]] + \mathbb{E}[\psi_t(\Sigma - \Sigma_0)]$, where $\psi_t(\Sigma - \Sigma_0) := \|X_t - \mu_{t-1}\|^2\|\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})\|^2$. We use $\mathbb{E}[\|X_t - \mu_{t-1}\|^2 | Z_{t-1}] \leq C$ by Assumption 6 to get $\mathbb{E}[\psi_t(\Sigma - \Sigma_0)] \leq C\|\Sigma - \Sigma_0\|_{L^2}^2 \leq C\epsilon_T^2$, uniformly in $\Sigma \in \Theta_{T,\epsilon}^\Sigma := \{\Sigma \in \Theta_T^\Sigma : \|\Sigma - \Sigma_0\|_{L^2} \leq \epsilon_T\}$. We show $\sup_{\Sigma\in\Theta_{T,\epsilon}^\Sigma} \frac{1}{T}\sum_{t\in\mathcal{T}}(\psi_t(\Sigma - \Sigma_0) - \mathbb{E}[\psi_t(\Sigma - \Sigma_0)]) = o_p(\epsilon_T)$ using Lemma 1 in Chen and Shen (1998). We get $\hat{I}_1(\Sigma) = O_p(\epsilon_T \frac{1}{\sqrt{n}})$ uniformly in $\Theta_{T,\epsilon}^\Sigma$.

(ii) Bound on $\hat{I}_2(\Sigma) := 4\frac{1}{T}\sum_{t\in\mathcal{T}}(X_t - \mu_{t-1})'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})][\hat{\mu}(Z_{t-1}) - \mu_0(Z_{t-1})]$. From analogous arguments applied to the consistency rate of Sieve estimator $\hat{\mu}$, we have $\hat{\mu} \in \Theta_{T,\delta}^\mu := \{\mu \in \Theta_T^\mu : \|\mu - \mu_0\|_{L^2} \leq x\delta_T\}$ with probability at least $1 - \eta$, where we can make η as small as desired by setting x large enough. Then, we get $\sup_{\Sigma\in\Theta_{T,\epsilon}^\Sigma}|\hat{I}_2(\Sigma)| \leq 4 \sup_{\mu\in\Theta_{T,\delta}^\mu, \Sigma\in\Theta_{T,\epsilon}^\Sigma} \left|\frac{1}{T}\sum_{t\in\mathcal{T}}\psi_t(\theta)\right|$, with probability at least $1-\eta$, where $\psi_t(\theta) = (X_t - \mu_{t-1})'[\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})][\mu(Z_{t-1}) - \mu_0(Z_{t-1})]$. Note that $\mathbb{E}[\psi_t(\theta)] = 0$, for any $\theta \in \Theta_T$ by the law of iterated expectation. We show that:

$$\sup_{\mu\in\Theta_{T,\delta}^\mu, \Sigma\in\Theta_{T,\epsilon}^\Sigma} \left|\frac{1}{T}\sum_{t\in\mathcal{T}}\psi_t(\theta)\right| = o_p(\epsilon_T\delta_T) \tag{D.11}$$

by using Lemma 1 in Chen and Shen (1998). We get $\hat{I}_2(\Sigma) = o_p(\epsilon_T\delta_T)$ uniformly in $\Theta_{T,\epsilon}^\Sigma$.

To prove (D.11), we write $f(Y_t) := \psi_t(\theta)$ with $f \in \mathcal{F}_T$, where $\mathcal{F}_T$ denotes the function set

obtained by letting $\mu \in \Theta^{\mu}_{T,\delta}, \Sigma \in \Theta^{\Sigma}_{T,\epsilon}$ vary. We have $\sup_{f \in \mathcal{F}_T} f(Y_t) \leq \bar{B}$ for a constant $\bar{B}$, by supposing the X_t vector bounded to ease the proof (in the unbounded case, we use a truncation argument). Further, we can prove $\sup_{f \in \mathcal{F}_T} \mathbb{V}(\frac{1}{\sqrt{T}} \sum_{t \in \mathcal{T}} f(Y_t)) \leq \bar{\sigma}^2$, where $\bar{\sigma}^2 = C\mathbb{E}[\|\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})\|^2 \|\mu(Z_{t-1}) - \mu_0(Z_{t-1})\|^2] \leq C\mathbb{E}[\|\Sigma(Z_{t-1}) - \Sigma_0(Z_{t-1})\|^{2p}]^{1/p} \mathbb{E}[\|\mu(Z_{t-1}) - \mu_0(Z_{t-1})\|^{2\bar{p}}]^{1/\bar{p}} = O(\epsilon_T^{2/p} \delta_T^{2/\bar{p}})$, for $\frac{1}{p} + \frac{1}{\bar{p}} = 1$. Let $a_{1,T} \to \infty$ and $a_{2,T} \to \infty$ be diverging integer sequences, such that $a_{2,T} = [T/(2a_{1,T})]$. Further, let ξ be such that $0 < \xi < 1$. Then, under the β-mixing condition in Assumption 5, if conditions (a.1), (a.2) and (a.3) of Lemma 1 in Chen and Shen (1998) hold, for $z > 0$ we have

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}_T} \frac{1}{T} \sum_{t \in \mathcal{T}} f(Y_t) \geq z \right) \leq 6 \exp\left(- \frac{(1-\xi) T z^2}{9 \bar{\sigma}^2 (1 + a_{1,T} \bar{B} \xi / 12)} \right) + 2(a_{2,T} - 1) \beta(a_{1,T}). \quad \text{(D.12)}$$

We take $z = \bar{z} \epsilon_T \delta_T$, for a constant $\bar{z} > 0$. We now check conditions (a.1), (a.2) and (a.3).

(a.1) $z \leq \xi \bar{\sigma}^2 / 4$. If we take $p < 2$ (and hence $\bar{p} > 2$), we have $\bar{\sigma}^2 \gg \epsilon_T \delta_T$, because $\epsilon_T \gg \delta_T$. Then, condition (a.1) applies for large T, for any $\bar{z}, \xi$.

(a.2) $\int_{\xi z/32}^{\bar{\sigma} \bar{B}^{1/2}} H^{1/2}(w, \mathcal{F}_T) dw \leq z T^{1/2} \xi^{3/2} / 2^{10}$, where $H(w, \mathcal{F}_T)$ is the bracketing metric entropy. To show this bound, we use $H(w, \mathcal{F}_T) \leq C M_T \log(C M_T / w)$ (see also Chen and Shen (1998), p. 311). Moreover, we have the integral

$$I(w, K) := \int \sqrt{\log(K/w)} dw = w \sqrt{\log(K/w)} - K \sqrt{\pi} \Phi\left(\sqrt{2 \log(K/w)} \right),$$

where $\Phi(\cdot)$ is the standard Gaussian cdf (we use the change of variable $y = \sqrt{2 \log(K/w)}$, and partial integration). Further, using $\Phi(x) \sim 1 - \frac{1}{x} \phi(x)$ as $x \to \infty$, we get:

$$\begin{aligned} I(w, K) &\sim w \sqrt{\log(K/w)} - K \sqrt{\pi} \left(1 - \frac{1}{\sqrt{2 \log(K/w)}} \phi\left(\sqrt{2 \log(K/w)} \right) \right) \\ &= K \sqrt{\pi} + w \sqrt{\log(K/w)} \left(1 - \frac{1}{2 \log(K/w)} \right) \end{aligned}$$

as $w \to 0$. Then, with $a = \xi z / 32$, $b = \bar{\sigma} \bar{B}^{1/2}$ and $K = C M_T$, we have

$$\begin{aligned} \int_{\xi z/32}^{\bar{\sigma} \bar{B}^{1/2}} H^{1/2}(w, \mathcal{F}_T) dw &\lesssim \sqrt{K} \left(b \sqrt{\log(K/b)} - a \sqrt{\log(K/a)} \right) \leq b \sqrt{K \log(K/b)} \\ &\leq \bar{C} \epsilon_T^{1/p} \delta_T^{1/\bar{p}} \sqrt{M_T \log(M_T)} = o\left(\epsilon_T \sqrt{M_T \log(M_T)} \right) \end{aligned}$$

for a constant $\bar{C} > 0$, because $\delta_T = o(\epsilon_T)$ and $1/p + 1/\bar{p} = 1$. With $z T^{1/2} \xi^{3/2} / 2^{10} \sim \epsilon_T \delta_T \sqrt{T}$, and $\delta_T \sim \sqrt{\frac{M_T \log(M_T)}{T}}$ for the optimal rate (see equation (D.4)), condition (a.2) follows.

(a.3) $z/3 > (2\bar{B}/a_{T,2})$. We have $z \gg \delta_T^2$, and we can take $a_{1,T}$ slowly growing, so that $a_{2,T}$ grows close to as fast as T. Then, condition (a.3) holds because $\delta_T^2 \gg 1/T$.

(iii) The proofs of the bounds for $\hat{I}_j(\Sigma)$, with $j = 3, 4, 5$, are established using similar arguments, and are omitted.

D.9 Proof of Lemma 5

Write $\hat{G} - G = \frac{1}{T}\sum_{t\in\mathcal{T}}(h(Z_{t-1})h(Z_{t-1})' - \mathbb{E}[h(Z_{t-1})h(Z_{t-1})']) =: \frac{1}{T}\sum_{t\in\mathcal{T}} x_t$. Then, for any constant $C > 0$, we have:

$$\mathbb{P}(\|\hat{G} - G\|_\infty \geq C\eta_T) \leq p^2 \max_{i,j=1,...,p} \mathbb{P}(|\frac{1}{T}\sum_{t\in\mathcal{T}}(x_t)_{i,j}| \geq C\eta_T). \tag{D.13}$$

We apply the Hoeffding's inequality with β-mixing (Bosq (1998) Theorem 1.3). From Assumption 14 (i), $|(x_t)_{i,j}| \leq \bar{B}_h^2$. Then, for any $i, j = 1, ..., p$ and integer $N_T \in [1, \frac{T}{2}]$:

$$\mathbb{P}(|\frac{1}{T}\sum_{t\in\mathcal{T}}(x_t)_{i,j}| \geq C\eta_T) \leq 4\exp\left(-\frac{C^2\eta_T^2 N_T}{8\bar{B}^4}\right) + 22(1 + \frac{4\bar{B}}{C\eta_T})^{1/2} N_T \beta(\lfloor\frac{T}{2N_T}\rfloor). \tag{D.14}$$

With $\eta_T = \frac{\log T}{\sqrt{T}}$ and $N_T = \frac{\bar{c}T}{2\log T}$, for a constant $\bar{c} > 0$, $p = O(T^{\bar{b}})$ for $\bar{b} > 0$, and Assumption 5, the inequalities (D.13) and (D.14) yield $\mathbb{P}(\|\hat{G} - G\|_\infty \geq C\eta_T) = o(1)$ for any C large enough, and $\bar{c} > 0$ small enough. Part (a) follows. To prove part b), we use

$$\hat{M}_l - M_l = \frac{1}{T}\sum_{t\in\mathcal{T}}(\zeta_{t,l}h(Z_{t-1}) - \mathbb{E}[\zeta_{t,l}h(Z_{t-1})]) + \frac{1}{T}\sum_{t\in\mathcal{T}}(\hat{\zeta}_{t,l} - \zeta_{t,l})h(Z_{t-1}). \tag{D.15}$$

We have $\frac{1}{T}\sum_{t\in\mathcal{T}}(\zeta_{t,l}h(Z_{t-1}) - \mathbb{E}[\zeta_{t,l}h(Z_{t-1})]) = O_p(\eta_T)$ similarly as in part (a). To bound the second term on the RHS of (D.15), we use $\hat{X}_t - \hat{\mu}(Z_{t-1}) = X_t - \mu_0(Z_{t-1}) + \Delta_{1,t} - \Delta_{2,t}$, where $\Delta_{1,t} := (0', (\hat{\xi}_t - \xi_t)')'$ and $\Delta_{2,t} := \hat{\mu}(Z_{t-1}) - \mu_0(Z_{t-1})$, and that $\sup_t \mathbb{E}[\|\Delta_{1,t}\|^2] = O(1/n)$ under Assumptions 4 and 6 (see Step 5 in the proof of Proposition 2 b)). Then, $\|\frac{1}{T}\sum_{t\in\mathcal{T}}(\hat{\zeta}_{t,l} - \zeta_{t,l})h(Z_{t-1})\|_\infty \leq 2J_1 + J_2 + O_p(n^{-1/2})$, where $J_1 := \|\frac{1}{T}\sum_{t\in\mathcal{T}}[(X_t - \mu_0(Z_{t-1}))\otimes\Delta_{2,t}]h(Z_{t-1})'\|_\infty$ and $J_2 := \|\frac{1}{T}\sum_{t\in\mathcal{T}}[\Delta_{2,t}\otimes\Delta_{2,t}]h(Z_{t-1})'\|_\infty$. We have $n^{-1/2} = O(\eta_T)$ if $\frac{T}{\log^2 T} = O(n)$. To bound J_1 we use $\Delta_{2,t} = \Delta_{21,t} + \Delta_{22,t}$, for $\Delta_{21,t} := \hat{\mu}(Z_{t-1}) - \mu_L(Z_{t-1}) = (\hat{P}_\mu - P_{L,\mu})h(Z_{t-1})$ and $\Delta_{22,t} := \mu_L(Z_{t-1}) - \mu_0(Z_{t-1})$, where $\mu_L(Z_{t-1}) = P_{L,\mu}h(Z_{t-1})$ and $P_{L,\mu} = [\rho_{L,\mu,1} : \cdots : \rho_{L,\mu,L}]'$ for the population Lasso solution. Then, $J_1 \leq \max_{l=1,...,L}\|\hat{\rho}_{\mu,l} - \rho_{L,\mu,l}\|_1 \|\frac{1}{T}\sum_{t\in\mathcal{T}}[(X_t - \mu_0(Z_{t-1}))\otimes h(Z_{t-1})]h(Z_{t-1})'\|_\infty + \|\frac{1}{T}\sum_{t\in\mathcal{T}}[(X_t - \mu_0(Z_{t-1}))\otimes(\mu_L(Z_{t-1}) - \mu_0(Z_{t-1}))]h(Z_{t-1})'\|_\infty = O_p(\eta_T)$,

using $\mathbb{E}(X_t - \mu_0(Z_{t-1})|Z_{t-1}) = 0$ and similar arguments as in the proof of part (a). Moreover

$$J_2 \leq \frac{1}{T}\sum_{t\in\mathcal{T}} \|\Delta_{2,t}\|^2 \bar{B}_h = O_p\left(\max_{l=1,\dots,L} |(\hat{\rho}_{\mu,l} - \rho_{L,\mu,l})'\hat{G}(\hat{\rho}_{\mu,l} - \rho_{L,\mu,l})| + \max_{l=1,\dots,L} \|\mu_{L,l} - \mu_{0,l}\|_{L^2}^2\right)$$

and $|(\hat{\rho}_{\mu,l}-\rho_{L,\mu,l})'\hat{G}(\hat{\rho}_{\mu,l}-\rho_{L,\mu,l})| \leq \|\hat{\rho}_{\mu,l}-\rho_{L,\mu,l}\|_1\|G(\hat{\rho}_{\mu,l}-\rho_{L,\mu,l})\|_\infty + \|\hat{G}-G\|_\infty\|\hat{\rho}_{\mu,l}-\rho_{L,\mu,l}\|_1^2$. Repeating the arguments in the proof of Lemma 6 applied to the Lasso estimator $\hat{\mu}$, we can show that $\|\hat{\rho}_{\mu,l}-\rho_{L,\mu,l}\|_1 = O_p(\alpha\bar{s})$, $\|G(\hat{\rho}_{\mu,l}-\rho_{L,\mu,l})\|_\infty = O_p(\alpha)$ and $\|\mu_{L,\mu,l}-\mu_{0,l}\|_{L^2}^2 = O(\eta_T^2\bar{s})$, while we have $\|\hat{G} - G\|_\infty = O_p(\eta_T)$ from part (a). Hence, $J_2 = O_p(\alpha^2\bar{s} + \eta_T\alpha^2\bar{s}^2 + \eta_T^2\bar{s})$. By taking $\bar{s} = C\eta_T^{-\frac{2}{2a+3}}$ and $\alpha = \eta_T^{\frac{2a+1}{2a+3}} l_T^{-1}$, with $l_T = o(1)$, it follows $J_2 = O_p(\eta_T^{\frac{4a}{2a+3}} l_T^{-2} + \eta_T^{\frac{6a+1}{2a+3}} l_T^{-2} + \eta_T^{\frac{4(a+1)}{2a+3}}) = O_p(\eta_T)$, if $a > 3/2$. The conclusion follows.

D.10 Proof of Lemma 6

We use:

$$\|\hat{\theta}_l - \theta_{0,l}\|_{L^2}^2 \leq 2\|\hat{\theta}_l - \theta_{L,l}\|_{L^2}^2 + 2\|\theta_{L,l} - \theta_{0,l}\|_{L^2}^2, \tag{D.16}$$

where $\theta_{L,l}(\cdot) = \rho_{L,l}'h(\cdot)$. Hence, we have to bound the two terms on the RHS.

Step 1: Bound of $\|\theta_{L,l} - \theta_{0,l}\|_{L^2}$.

Lemma 8. *Under the condition $\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 \leq \bar{s}\eta_T^2$, and Assumption 14, we have $\|\theta_{L,l} - \theta_{0,l}\|_{L^2}^2 = O\left(\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 + \alpha_L^2\bar{s}\right)$.*

Hence, the conditions $\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 \leq \bar{s}\eta_T^2$ and $\alpha_L = \eta_T$ (Assumption 14 (iv)) yield $\|\theta_{L,l} - \theta_{0,l}\|_{L^2}^2 = O(\eta_T^2\bar{s})$.

Step 2: Bound of $\|\hat{\theta}_l - \theta_{L,l}\|_{L^2}$. We use:

$$\|\hat{\theta}_l - \theta_{L,l}\|_{L^2}^2 = (\hat{\rho}_l - \rho_{L,l})'G(\hat{\rho}_l - \rho_{L,l}) \leq \|\hat{\rho}_l - \rho_{L,l}\|_1\|G(\hat{\rho}_l - \rho_{L,l})\|_\infty. \tag{D.17}$$

Moreover, let $\hat{\delta} := \hat{\rho}_l - \rho_{L,l}$.

Lemma 9. *Under the condition $\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 \leq \bar{s}\eta_T^2$, and if $\eta_T(1+\bar{B}) = o_p(\alpha)$, then vector $\hat{\delta}$ satisfies $\sum_{j\in\mathcal{J}_{\rho_L}^c} |\hat{\delta}_j| \leq 3\sum_{j\in\mathcal{J}_{\rho_L}} |\hat{\delta}_j|$, with probability approaching (w.p.a.) 1.*

From Lemma 9 and Assumption 14 (iii) we have:

$$\begin{aligned}
\|\hat{\delta}\|_1^2 &= \left(\sum_{j\in\mathcal{J}_{\rho_L}^c} |\hat{\delta}_j| + \sum_{j\in\mathcal{J}_{\rho_L}} |\hat{\delta}_j|\right)^2 \leq \left(4\sum_{j\in\mathcal{J}_{\rho_L}} |\hat{\delta}_j|\right)^2 \\
&\leq 16\|\rho_{L,l}\|_0 \sum_{j\in\mathcal{J}_{\rho_L}} |\hat{\delta}_j|^2 \leq \frac{16}{\phi}\|\rho_{L,l}\|_0\hat{\delta}'G\hat{\delta} \leq \frac{16}{\phi}\|\rho_{L,l}\|_0\|\hat{\delta}\|_1\|G\hat{\delta}\|_\infty,
\end{aligned}$$

where $\phi := \inf_{\delta \neq 0: \sum_{j \in \mathcal{J}_{\rho_L}^c} |\delta_j| \leq \bar{\kappa} \sum_{j \in \mathcal{J}_{\rho_L}} |\delta_j|} \frac{\delta' G \delta}{\sum_{j \in \mathcal{J}_{\rho_L}} \delta_j^2} > 0$. By dividing both sides by $\|\hat{\delta}\|_1$ we get $\|\hat{\delta}\|_1 \leq \frac{16}{\phi} \|\rho_{L,l}\|_0 \|G\hat{\delta}\|_\infty$, i.e. $\|\hat{\rho}_l - \rho_{L,l}\|_1 \leq \frac{16}{\phi} \|\rho_{L,l}\|_0 \|G(\hat{\rho} - \rho_L)\|_\infty$. From (D.17) we get:

$$\|\hat{\theta}_l - \theta_{L,l}\|_{L^2}^2 \leq \frac{16}{\phi} \|\rho_{L,l}\|_0 \|G\hat{\delta}\|_\infty^2. \tag{D.18}$$

Hence, the important quantities to bound $\|\hat{\theta}_l - \theta_{L,l}\|_{L^2}$ are $\|G\hat{\delta}\|_\infty$ and $\|\rho_{L,l}\|_0$. Regarding the former, the equation $G(\hat{\rho}_l - \rho_{L,l}) = (G - \hat{G})\hat{\rho}_l + (\hat{G}\hat{\rho}_l - \hat{M}_l) + (\hat{M}_l - M_l) + (M_l - G\rho_{L,l})$, the KKT conditions $\|\hat{G}\hat{\rho}_l - \hat{M}_l\|_\infty \leq \alpha$ and $\|G\rho_{L,l} - M_l\|_\infty \leq \alpha_L$, and Lemma 5 imply

$$\|G(\hat{\rho}_l - \rho_{L,l})\|_\infty \leq \eta_T \|\hat{\rho}_l\|_1 + \alpha + \alpha_L + O_p(\eta_T). \tag{D.19}$$

Lemma 10. *Under the condition $\eta_T(1 + \bar{B}) + \alpha_L = o_p(\alpha)$, we have $\|\hat{\rho}_l\|_1 = O_p(1 + \bar{B})$.*

From (D.19), Lemma 10 and the condition $\eta_T(1 + \bar{B}) + \alpha_L = o_p(\alpha)$, it follows $\|G(\hat{\rho}_l - \rho_{L,l})\|_\infty = O_p(\alpha)$. From (D.18) we get:

$$\|\hat{\theta}_l - \theta_{L,l}\|_{L^2}^2 = O_p\left(\|\rho_{L,l}\|_0 \alpha^2\right). \tag{D.20}$$

Lemma 11. *We have $\|\rho_{L,l}\|_0 \leq \delta_1(G) \frac{1}{\alpha_L^2} \|\theta_{0,l} - \theta_{L,l}\|_{L^2}^2$, where $\delta_1(G)$ denotes the largest eigenvalue of symmetric matrix G.*

From Lemmas 8 and 11, the conditions $\|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 \leq \bar{s}\eta_T^2$ and $\alpha_L = \eta_T$, we get $\|\rho_{L,l}\|_0 = O_p(\bar{s})$. Then, from the latter bound and (D.20), we get:

$$\|\hat{\theta}_l - \theta_{L,l}\|_{L^2}^2 = O_p\left(\bar{s}\alpha^2\right). \tag{D.21}$$

From (D.16), Step 1 and (D.21), and $\eta_T = o_p(\alpha)$, we get $\|\hat{\theta}_l - \theta_{0,l}\|_{L^2}^2 = O_p\left(\alpha^2 \bar{s}\right)$.

D.11 Proof of Lemma 7

By the triangular inequality we have $\|[A]^\epsilon - A_0\|_2 \leq \|A - A_0\|_2 + \|[A]^\epsilon - A\|_2$. Moreover, by the spectral decomposition $A = U\Lambda U'$ and the orthogonality of U, we have $\|[A]^\epsilon - A\|_2 = \|[\Lambda]^\epsilon - \Lambda\|_2$. To bound the RHS, we need a bound for the maximal difference between the capped and uncapped eigenvalues of A. This difference is $\delta_j([A]^\epsilon) - \delta_j(A) = 0$ if $\delta_j(A) \geq \epsilon$, while it is equal to $\epsilon - \delta_j(A)$ otherwise. Thus, we have $|\delta_j([A]^\epsilon) - \delta_j(A)| \leq 2\epsilon$ if $\delta_j(A) \geq -\epsilon$. Moreover, if $\delta_j(A) \leq -\epsilon$, we have that $\delta_j([A]^\epsilon) - \delta_j(A) \geq 2\epsilon$ and $\delta_j([A]^\epsilon) - \delta_j(A) \leq \epsilon - (\delta_j(A_0) - \|A - A_0\|_2) \leq \epsilon + \|A - A_0\|_2$, which together imply $|\delta_j([A]^\epsilon) - \delta_j(A)| \leq \epsilon + \|A - A_0\|_2$. Therefore we have $|\delta_j([A]^\epsilon) - \delta_j(A)| \leq 2\epsilon + \|A - A_0\|_2$ for all j. This yields $\|[\Lambda]^\epsilon - \Lambda\|_2 \leq 2\epsilon + \|A - A_0\|_2$.

By gathering the bounds above, and using that $\|A - A_0\|_2 \leq \|A - A_0\|$, where $\|\cdot\|$ is the Frobenius norm, we get $\|[A]^\epsilon - A_0\|_2 \leq 2(\|A - A_0\| + \epsilon)$.

D.12 Proof of Lemma 8

This lemma essentially corresponds to Lemma A4 of Chernozhukov, Newey and Singh (2018); we report the proof as it should be modified at some steps. We start by noting that

$$\|\theta_{0,l} - h'\rho_{L,l}\|_{L^2}^2 + 2\alpha_L\|\rho_{L,l}\|_1 \leq \|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 + 2\alpha_L\|\bar{\rho}_l\|_1. \tag{D.22}$$

Let us define $\bar{\delta}_l := \bar{\rho}_l - \rho_{L,l}$. Then we get:

$$\begin{aligned} & \|\theta_{0,l} - h'\rho_{L,l}\|_{L^2}^2 + \alpha_L\|\bar{\delta}_l\|_1 \leq \|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 + 2\alpha_L\|\bar{\rho}_l\|_1 - 2\alpha_L\|\rho_{L,l}\|_1 + \alpha_L\|\bar{\delta}_l\|_1 \\ \leq\ & \|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 + 2\alpha_L\left(\|\bar{\rho}_l\|_1 - \|\rho_{L,l}\|_1 + \|\bar{\rho}_l - \rho_{L,l}\|_1\right) = \|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 \\ & + 2\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}_l}} \left(|\bar{\rho}_{l,j}| - |\rho_{L,l,j}| + |\bar{\rho}_{l,j} - \rho_{L,l,j}|\right) \leq \|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 + 4\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}_l}} |\bar{\rho}_{l,j} - \rho_{L,l,j}|. \end{aligned}$$

Thus, we get:

$$\|\theta_{0,l} - h'\rho_{L,l}\|_{L^2}^2 + \alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}_l}^c} |\bar{\delta}_{l,j}| \leq \|\theta_{0,l} - h'\bar{\rho}_l\|_{L^2}^2 + 3\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}_l}} |\bar{\delta}_{l,j}|. \tag{D.23}$$

We distinguish two cases. Let $\xi := 3/(\bar{\kappa} - 3)$, where $\bar{\kappa} > 3$ is defined in Assumption 14 (iii). (i) Suppose $3\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}}} |\bar{\delta}_j| \leq \xi\|\theta_0 - h'\bar{\rho}\|^2$. Thus, from (D.23) we get:

$$\|\theta_0 - h'\rho_L\|_{L^2}^2 \leq (1 + \xi)\|\theta_0 - h'\bar{\rho}\|_{L^2}^2. \tag{D.24}$$

(ii) Suppose $3\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}}} |\bar{\delta}_j| > \xi\|\theta_0 - b'\bar{\rho}\|^2$. Then from (D.23):

$$\alpha_L \sum_{j\in J_{\bar{\rho}}^c} |\bar{\delta}_j| \ \leq\ \|\theta_0 - h'\bar{\rho}\|_{L^2}^2 + 3\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}}} |\bar{\delta}_j| \leq (1 + 1/\xi)3\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}}} |\bar{\delta}_j| = \bar{\kappa}\alpha_L \sum_{j\in\mathcal{J}_{\bar{\rho}}} |\bar{\delta}_j|.$$

By canceling α_L on both sides, the latter inequality and Assumption 14 (iii) imply $\sum_{j\in\mathcal{J}_{\bar{\rho}}} \bar{\delta}_j^2 \leq \frac{1}{\bar{\phi}}\bar{\delta}'G\bar{\delta} = \frac{1}{\bar{\phi}}\|\theta_L - h'\bar{\rho}\|_{L^2}^2$, where $\bar{\phi} := \inf_{\delta\neq 0:\sum_{j\in\mathcal{J}_{\bar{\rho}^c}} |\delta_j| \leq \bar{\kappa}\sum_{j\in\mathcal{J}_{\bar{\rho}}} |\delta_j|} \frac{\delta'G\delta}{\sum_{j\in\mathcal{J}_{\bar{\rho}}} \delta_j^2} > 0$. Moreover, by the Cauchy-Schwarz inequality, we get:

$$\sum_{j\in\mathcal{J}_{\bar{\rho}}} |\bar{\delta}_j| \leq \sqrt{\mathcal{M}(\bar{\rho})}\sqrt{\sum_{j\in\mathcal{J}_{\bar{\rho}}} \bar{\rho}_j^2} \leq \sqrt{\frac{\mathcal{M}(\bar{\rho})}{\bar{\phi}}}\|\theta_L - h'\bar{\rho}\|_{L^2} \leq \sqrt{\frac{\mathcal{M}(\bar{\rho})}{\bar{\phi}}}\left(\|\theta_0 - h'\bar{\rho}\|_{L^2} + \|\theta_0 - \theta_L\|_{L^2}\right),$$

for $\mathcal{M}(\bar{\rho}) := \|\bar{\rho}\|_0$. Thus, from (D.23) we get:

$$\begin{aligned}\|\theta_0 - h'\rho_L\|_{L^2}^2 &\leq \|\theta_0 - h'\bar{\rho}\|_{L^2}^2 + 3\alpha_L \sum_{j\in J_{\bar{\rho}}} |\bar{\delta}_j| \\ &\leq \|\theta_0 - h'\bar{\rho}\|_{L^2}^2 + 3\alpha_L \sqrt{\frac{\mathcal{M}(\bar{\rho})}{\bar{\phi}}} \left(\|\theta_0 - h'\bar{\rho}\|_{L^2} + \|\theta_0 - \theta_L\|_{L^2}\right). \qquad \text{(D.25)}\end{aligned}$$

Now, by using that $ab \leq \frac{1}{4}a^2 + b^2$, we have:

$$\begin{aligned}3\alpha_L \sqrt{\frac{\mathcal{M}(\bar{\rho})}{\bar{\phi}}} \|\theta_0 - h'\bar{\rho}\|_{L^2} &\leq \frac{9}{4}\alpha_L^2(\mathcal{M}(\bar{\rho})/\bar{\phi}) + \|\theta_0 - h'\bar{\rho}\|_{L^2}^2, \\ 3\alpha_L \sqrt{\frac{\mathcal{M}(\bar{\rho})}{\bar{\phi}}} \|\theta_0 - \theta_L\|_{L^2} &= 6\alpha_L \sqrt{\frac{\mathcal{M}(\bar{\rho})}{\bar{\phi}}} (\frac{1}{2}\|\theta_0 - \theta_L\|_{L^2}) \leq \frac{9\alpha_L^2 \mathcal{M}(\bar{\rho})}{\bar{\phi}} + \frac{\|\theta_0 - \theta_L\|_{L^2}^2}{4}.\end{aligned}$$

By plugging these inequalities into the RHS of (D.25), and rearranging terms, we get:

$$\|\theta_0 - \theta_L\|_{L^2}^2 \leq \frac{8}{3}\|\theta_0 - h'\bar{\rho}\|_{L^2}^2 + 15\alpha_L^2 \frac{\mathcal{M}(\bar{\rho})}{\bar{\phi}}. \qquad \text{(D.26)}$$

From inequalities (D.24) and (D.26) we get $\|\theta_0 - \theta_L\|_{L^2}^2 \leq C\left(\|\theta_0 - h'\bar{\rho}\|_{L^2}^2 + \alpha_L^2 \mathcal{M}(\bar{\rho})\right)$, with $C = \max\{1 + \xi, 8/3, 15/\bar{\phi}\}$. Finally, with $\mathcal{M}(\bar{\rho}) \leq \bar{s}$, the conclusion follows.

D.13 Proof of Lemma 9

This result corresponds to Lemma A6 of Chernozhukov, Newey and Singh (2018).

D.14 Proof of Lemma 10

This result corresponds to Lemma A3 of Chernozhukov, Newey and Singh (2018). From Lemma 5 and using $\|M_l - G\rho_{L,l}\| \leq \alpha_L$ and the condition $\eta_T(1 + \bar{B}) + \alpha_L = o_p(\alpha)$, we have:

$$\|\hat{M}_l - \hat{G}\rho_{L,l}\|_\infty \leq \|\hat{M}_l - M_l\|_\infty + \|M_l - G\rho_{L,l}\|_\infty + \|(G - \hat{G})\rho_{L,l}\|_\infty = O_p(\eta_T + \alpha_L + \eta_T\bar{B}) = o_p(\alpha).$$

Since the Lasso estimator minimizes the Lasso criterion, we have $-2\hat{M}_l'\hat{\rho}_l + \hat{\rho}_l'\hat{G}\hat{\rho}_l + 2\alpha\|\hat{\rho}_l\|_1 \leq -2\hat{M}_l'\rho_{L,l} + \rho_{L,l}'\hat{G}\rho_{L,l} + 2\alpha\|\rho_{L,l}\|_1$. Then, we get:

$$\begin{aligned}&2\alpha\|\hat{\rho}_l\|_1 \leq 2\hat{M}_l'(\hat{\rho}_l - \rho_{L,l}) - [\hat{\rho}_l'\hat{G}\hat{\rho}_l - \rho_{L,l}'\hat{G}\rho_{L,l}] + 2\alpha\|\rho_{L,l}\|_1 \\ =\ &2\hat{M}_l'(\hat{\rho}_l - \rho_{L,l}) - [(\hat{\rho}_l - \rho_{L,l})'\hat{G}(\hat{\rho}_l - \rho_{L,l}) + 2\rho_{L,l}'\hat{G}(\hat{\rho}_l - \rho_{L,l})] + 2\alpha\|\rho_{L,l}\|_1 \\ \leq\ &2(\hat{M}_l - \hat{G}\rho_{L,l})'(\hat{\rho}_l - \rho_{L,l}) + 2\alpha\|\rho_{L,l}\|_1 \leq 2\|\hat{M}_l - \hat{G}\rho_{L,l}\|_\infty \|\hat{\rho}_l - \rho_{L,l}\|_1 + 2\alpha\|\rho_{L,l}\|_1.\end{aligned}$$

Then, by dividing both sides by α, and using $\|\hat{M}_l - \hat{G}\rho_{L,l}\|_\infty = o_p(\alpha)$, we get $2\|\hat{\rho}_l\|_1 = o_p(1)(\|\hat{\rho}_l\|_1 + \|\rho_{L,l}\|_1) + 2\|\rho_{L,l}\|_1$, which implies $\|\hat{\rho}_l\|_1 = O_p(\|\rho_{L,l}\|_1) = O_p(\bar{B})$.

D.15 Proof of Lemma 11

This result essentially corresponds to Lemma A5 of Chernozhukov, Newey and Singh (2018). The KKT conditions for the population Lasso imply $\mathbb{E}[h_j(Z_{t-1})(\theta_0(Z_{t-1}) - \theta_L(Z_{t-1}))] = \alpha_L sign(\rho_{L,j})$, if $\rho_{L,j} \neq 0$. Then, for $e_L := \theta_0 - \theta_L$ we have:

$$\begin{aligned}
\|\theta_0 - \theta_L\|_{L^2}^2 &= \mathbb{E}[e_L(Z_{t-1})^2] \geq \mathbb{E}[e_L(Z_{t-1})h(Z_{t-1})']G^{-1}\mathbb{E}[h(Z_{t-1})e_L(Z_{t-1})] \\
&\geq \delta_1(G)^{-1}\mathbb{E}[e_L(Z_{t-1})h(Z_{t-1})']\mathbb{E}[h(Z_{t-1})e_L(Z_{t-1})] \\
&\geq \delta_1(G)^{-1}\sum_{j\in\mathcal{J}_{\rho_L}}\mathbb{E}[e_L(Z_{t-1})h_j(Z_{t-1})]^2 = \delta_1(G)^{-1}\mathcal{M}(\rho_L)\alpha_L^2.
\end{aligned}$$

By rearranging terms, the conclusion follows.

E Monte Carlo experiments

We first define the Monte Carlo design and the target parameters, before presenting the results of the simulations.

E.1 Data Generating Process and Target Parameters

We consider a panel of excess returns $\{y_{i,t}\}_{i\leq N, t\leq T}$ generated according to

$$y_{i,t} = b_{i,t-1}'g_t + \varepsilon_{i,t}, \qquad \varepsilon_{i,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \tag{E.1}$$

where errors $\varepsilon_{i,t}$ are i.i.d. across i and t, and independent of $(g_t, b_{i,t-1}, Z_{t-1})$. The factor vector of dimension k satisfies

$$g_t = \lambda(Z_{t-1}) + f_t, \qquad f_t \mid Z_{t-1} \sim \mathcal{N}(0, \Sigma_f), \tag{E.2}$$

with conditional risk premium $\lambda(Z_{t-1}) = \Lambda Z_{t-1}$, where $\Lambda \in \mathbb{R}^{k\times d}$ is a fixed coefficients matrix. The components of the $d \times 1$ state vector Z_t beyond the constant ($Z_{1,t} = 1$) follow independent AR(1) processes:

$$Z_{j,t} = \rho Z_{j,t-1} + \nu_{j,t}, \qquad \nu_{j,t} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_Z^2), \quad |\rho| < 1, \tag{E.3}$$

for $j = 2, \cdots, d$. The initial condition is drawn from its stationary distribution $Z_{j,0} \sim \mathcal{N}\left(0, \frac{\sigma_Z^2}{1-\rho^2}\right)$. Factor loadings evolve as

$$b_{i,t-1} = Bw_{i,t-1} + \tilde{b}_{i,t-1}, \tag{E.4}$$

where $B \in \mathbb{R}^{k \times K}$ is a fixed coefficients matrix. The $K \times 1$ vector of instruments $w_{i,t-1}$ and the $k \times 1$ vector $\tilde{b}_{i,t-1}$ are mutually independent, such that

$$w_{i,t-1} \mid Z_{t-1} \sim \mathcal{N}(0, \Sigma_w(Z_{t-1})), \qquad \tilde{b}_{i,t-1} \sim \mathcal{N}(0, \sigma_b^2 I_k), \tag{E.5}$$

with $\Sigma_w(Z_{t-1}) = \sigma_w^2 \exp(\gamma_w Z_{2,t-1}) I_K$ yielding a conditional heteroskedasticity effect. The vectors $\{w_{i,t}\}$ are conditionally independent across i given Z_{t-1}. Finally, the k^O-dimensional observed factors satisfy

$$f_t^O = Ag_t + u_t, \qquad u_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \Sigma_u), \tag{E.6}$$

where the "measurement errors" u_t are independent of Z_t, and $A \in \mathbb{R}^{k^O \times k}$ is a fixed loading matrix. If the observable factors are returns of well-diversified portfolios (as in our empirics), u_t represents the residual idiosyncratic risk for finite n.

Baseline parameter values are $k = 4$, $k^O = 2$, $K = 10$, $d = 15$, $\rho = 0.9$, $\sigma_Z = 0.05$, $\gamma_w = 1$, $\sigma_w = 0.3$, $\sigma_b = 0.1$. The factor covariance matrix is $\Sigma_f = 0.0025\, I_k$, and the measurement error covariance matrix is $\Sigma_u = 0.0001\, I_{k^O}$. The idiosyncratic error variance is calibrated as $\sigma_\varepsilon^2 = \mathrm{Var}(b_{i,t-1}' g_t) \frac{1-R^2}{R^2}$ where the unconditional explanatory R-square of the common component in returns is set to $R^2 = 0.2$. The matrix $\Lambda \in \mathbb{R}^{k \times d}$ is dense, with the first column drawn from $|\mathcal{N}(0.06, 0.01)|$ and the remaining columns drawn from $\mathcal{N}(0, 0.15)$. The matrix $A \in \mathbb{R}^{k^O \times k}$ is dense with positive entries drawn from $|\mathcal{N}(0, 1/\sqrt{k})|/10$. The values of Λ and A are kept constant across Monte Carlo replications. The matrix $B \in \mathbb{R}^{k \times K}$ is given by $[I_k : 0_{k \times (K-k)}]$, with I_k in the first k columns and zeros elsewhere.

Our experiments focus on estimation of two target parameters: (a) the (time-invariant) number of conditional factors k, and (b) the expected projected conditional risk premium $c_0 = \mathbb{E}[\lambda_t^O]$, where $\lambda_t^O = \Sigma_{12}(Z_{t-1})\Sigma_{22}(Z_{t-1})^\dagger \mu_2(Z_{t-1}) = A\lambda(Z_{t-1})$ (see Example 1 in Section 3.2). Then, (a) we estimate k by $\hat{k}_t^\alpha$ in (4.2) for values of regularization parameter α in a grid between 0.1% and 20%, and (b) we estimate c_0 by the DML estimator

$$\hat{c} = \frac{1}{I} \sum_{t \in I} \left(\hat{\lambda}_t^O + \hat{\eta}_t \right), \tag{E.7}$$

where $\hat{\lambda}_t^O = \hat{\Sigma}_{12}(Z_{t-1})' \hat{\Sigma}_{22}(Z_{t-1})^\dagger \hat{\mu}_2(Z_{t-1})$ and $\hat{\eta}_t$ is the sample counterpart of η_t defined in

Table 1 first line. We estimate conditional moments by feedforward neural networks (see Section 5.2 and SMC S.2).

We define the total sample size as $T_{\text{total}} = T_1 + T_2 + I$, where T_1 and T_2 denote the in-sample periods used for two-stage training and I denotes the out-of-sample testing period. Consistent with the empirical analysis, feedforward neural networks are trained to estimate conditional matrices in an element-wise manner: the first 40% of T_{total} (i.e. T_1) is used to estimate conditional means of managed portfolio returns and observable factors, and the second 40% (i..e T_2) is used to de-mean the data and estimate conditional covariance matrices. To mitigate overfitting, we employ 5-fold time-series cross-validation within each training stage. All results are evaluated out of sample over the remaining 20% of the observations (i.e. I). Sample sizes are $N \in \{500, 1000, 2500\}$ and $T_{total} \in \{400, 600, 1000\}$, with 1,000 Monte Carlo replications.

E.2 Selection of the Number of Conditional Factors k

Table E1 summarizes the behavior of the estimated factor dimension $\hat{k} \equiv \hat{k}_t^{\alpha}$ across nine combinations of N and T_{total}, in terms of mode and accuracy, i.e. percentage of correct selection. At each testing date, the estimator selects the smallest $\hat{k}$ such that the leading eigenvalues of $\hat{\Sigma}_{22}(Z_{t-1})$ explain at least a fraction $(1-\alpha)$ of total variance. For small cross-sectional sizes, we achieve high accuracy in selection for $\alpha = 10\%$, while sampling noise in the managed returns $\hat{\xi}_t$ induces upward bias in $\hat{k}$ when α is too small. For example, when $N = 500$ and $T_{total} = 600$, setting $\alpha = 0.1\%$ leads to severe overestimation (mode $\hat{k} = 9$), whereas $\alpha = 10\%$ correctly selects $\hat{k} = 4$ with 99% accuracy. In contrast, for larger cross sections, estimation is markedly more stable: when $N = 2500$ and $T_{total} = 600$, the true factor dimension is recovered with accuracy near 100% for α between 2% and 10%.

Table E1 indicates that factor number estimation is robust for threshold values $\alpha \geq 6\%$. Viewed through a signal-to-noise lens, α acts as a lower bound on the strength required for a component to be classified as a factor, so that higher thresholds increasingly suppress weak common variation. When α is set too high, the procedure selects $\hat{k} = 3$ in the great majority of the runs [16], yielding a one-factor underestimation relative to the truth $k = 4$. By contrast, low values of α lead to substantial and unstable overestimation, with $\hat{k}$ ranging between 5 and 9, consistent with idiosyncratic noise being misclassified as factors. Given this pronounced asymmetry—where modest underestimation is far less consequential than severe overestimation—we adopt a conservative threshold of $\alpha = 10\%$ in the empirical analysis (Section 5), which delivers stable performance across different cross-sectional sample sizes.

[16] Unreported results show that $\hat{k} = 2$ or smaller is selected only in few cases, (typically much) less than 10% for all but one sample sizes combination.

Table E1: **Conditional Factor Number Estimation: Mode of $\hat{k}$ and Accuracy (%)**

		α										
N	T_{total}	0.1%	2%	4%	6%	8%	10%	12%	14%	16%	18%	20%
500	400	8 (0)	6 (0)	5 (11)	4 (94)	4 (97)	4 (84)	4 (57)	3 (28)	3 (8)	3 (0)	3 (0)
	600	9 (0)	6 (0)	5 (1)	4 (87)	4 (100)	4 (99)	4 (94)	4 (74)	3 (38)	3 (7)	3 (0)
	1000	8 (0)	6 (0)	5 (9)	4 (99)	4 (99)	4 (86)	4 (53)	3 (17)	3 (2)	3 (0)	3 (0)
1000	400	8 (0)	5 (5)	4 (99)	4 (90)	4 (70)	3 (46)	3 (22)	3 (5)	3 (0)	3 (0)	3 (0)
	600	9 (0)	5 (0)	4 (100)	4 (100)	4 (97)	4 (87)	4 (64)	3 (29)	3 (5)	3 (0)	3 (0)
	1000	8 (0)	5 (3)	4 (100)	4 (94)	4 (70)	3 (35)	3 (11)	3 (2)	3 (0)	3 (0)	3 (0)
2500	400	8 (0)	4 (98)	4 (84)	4 (66)	3 (43)	3 (20)	3 (7)	3 (0)	3 (0)	3 (0)	3 (0)
	600	8 (0)	4 (100)	4 (99)	4 (96)	4 (88)	4 (68)	3 (34)	3 (7)	3 (0)	3 (0)	3 (0)
	1000	7 (0)	4 (100)	4 (87)	4 (64)	3 (36)	3 (13)	3 (2)	3 (0)	3 (0)	3 (0)	3 (0)

Each cell reports the mode of estimator $\hat{k}$ and its accuracy (%, in parenthesis), for different values of regularization parameter α. Accuracy is defined as percentage of correct selection. Results are evaluated out of sample (the last 20% of total sample period) based on 1000 Monte Carlo replications per configuration for the DGP defined in Section E.1 with true number of conditional factors $k = 4$.

E.3 Average Projected Conditional Risk Premia

Table E2 reports Monte Carlo results for the estimation of the average projected conditional risk premia c_0 for the two observable factors. All results are evaluated out of sample using the last 20% of the observations. Entries report bias, and standard deviation in parentheses, expressed in percentage points, across different combinations of N and T_{total}, for both the original and DML estimators. The latter is defined in (E.7), while the former is obtained by setting $\hat{\eta}_t = 0$ in (E.7).

The uncorrected ("Original") estimator exhibits noticeable finite-sample bias, particularly when the cross-sectional dimension is small. Bias decreases as either N or T_{total} increases, but remains non-negligible for moderate sample sizes. The DML correction reduces bias in all cases, with especially large gains for small cross sections, and yields comparable bias when the original estimator is already nearly unbiased ($N = 2500$).

Given that our empirical application features $T_{\text{total}} \approx 700$ with around 500 in-sample observations, and average cross-sectional size N in characteristic portfolios is between $N = 1000$ and a few thousand for most of the dates, the configurations with either $N = 1000$, or $N = 2500$, and $T_{\text{total}} = 600$ are the most representative. In this empirically relevant setup, Table E2 shows that the DML procedure delivers substantial bias reduction relative to the original estimator while maintaining moderate dispersion. In some Monte Carlo replications,

the DML correction $\hat{\eta}_t$ admits large values at some dates due to small estimated eigenvalues, which may explain part of the standard deviation of the DML estimator. In unreported results we have experienced that moderate levels of clipping in the DML correction can improve the finite sample performance of the estimator.

Figure 5 reports Monte Carlo distributions of the standardized statistic $\sqrt{I}(\hat{c}_1 - c_{1,0})/\hat{\sigma}_{11}$ for Factor 1 across different cross-sectional sizes N and total sample lengths T_{total}. The estimated standard deviation $\hat{\sigma}_{11}$ is obtained via a HAC estimator with Parzen kernel and optimal selection of the truncation parameter (Andrews (1991)). Each panel displays the histogram of the statistic computed from 1000 Monte Carlo replications, with the red curve indicating the standard normal density $N(0,1)$. Across all configurations, the simulated distributions are centered near zero and quite closely track the $N(0,1)$ benchmark, albeit with standard deviation a bit larger than 1. These findings provide evidence in favor of asymptotic normality of the DML estimator and sufficient accuracy of the HAC variance estimator for feasibility. As either N or T_{total} increases, the approximation improves further, with tighter dispersion and closer alignment to the $N(0,1)$ distribution, consistent with the expected large-sample behavior of the estimator.

Table E2: **Average Projected Conditional Risk Premia: Bias and Std (%)**

	$T_{total} = 400$		$T_{total} = 600$		$T_{total} = 1000$	
	Original	**DML**	**Original**	**DML**	**Original**	**DML**
Factor 1 (true value $c_{1,0} = 0.87\%$)						
$N = 500$	-0.19 (0.30)	-0.04 (0.60)	-0.14 (0.23)	-0.03 (0.34)	-0.18 (0.21)	-0.05 (0.39)
$N = 1000$	-0.13 (0.34)	-0.03 (0.68)	-0.09 (0.25)	0.02 (0.41)	-0.11 (0.24)	-0.03 (0.52)
$N = 2500$	-0.08 (0.39)	-0.05 (0.92)	-0.02 (0.28)	-0.01 (0.48)	-0.06 (0.29)	-0.02 (0.79)
Factor 2 (true value $c_{2,0} = 0.70\%$)						
$N = 500$	-0.15 (0.25)	-0.06 (0.51)	-0.10 (0.17)	-0.05 (0.28)	-0.15 (0.17)	-0.02 (0.34)
$N = 1000$	-0.10 (0.27)	-0.04 (0.62)	-0.06 (0.20)	-0.03 (0.34)	-0.09 (0.18)	-0.02 (0.45)
$N = 2500$	-0.04 (0.33)	-0.07 (0.85)	-0.01 (0.24)	-0.04 (0.47)	-0.04 (0.23)	0.01 (0.63)

Each cell reports bias, and standard deviation in parentheses, expressed in percentage points, for the estimates of average projected conditional risk premia. Labels "DML" and "Original" refer to the estimator in (E.7), respectively the estimator in (E.7) with $\hat{\eta}_t = 0$. Results are evaluated out of sample (the last 20% of total sample period) based on 1000 Monte Carlo replications per configuration for the DGP defined in Section E.1 with true number of conditional factors $k = 4$.

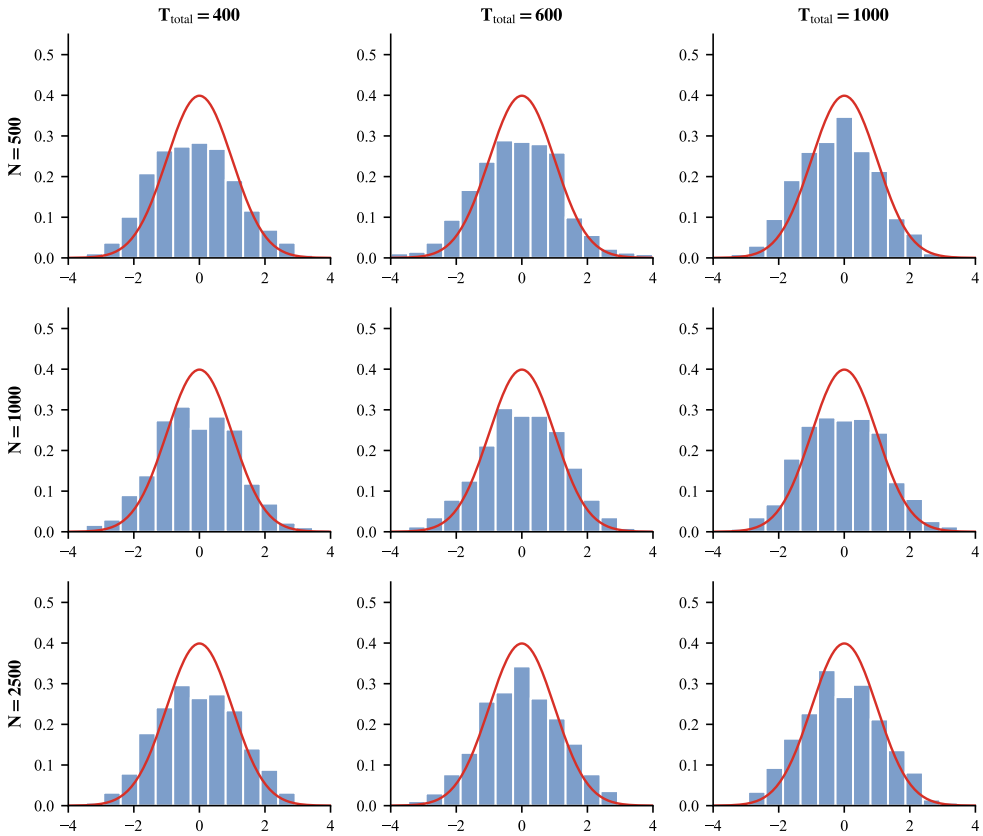

Figure 5: **Standardized Statistic for Factor 1**. Blue bars are histograms of the standardized statistic $\sqrt{I}(\hat{c}_1 - c_{1,0})/\hat{\sigma}_{11}$ across different sample sizes. The red curve shows the standard normal density $N(0,1)$. Results are evaluated out of sample (the last 20% of total sample period) based on 1000 Monte Carlo replications per configuration for the DGP defined in Section E.1 with true number of conditional factors $k = 4$.

Additional References

Ai, C., and Chen, X. (2003). Efficient Estimation of Models with Conditional Moment Restrictions Containing Unknown Functions. *Econometrica*, 71 (6), 1795-1843.

Bai, J., and Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica*, 70(1), 191-221.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *Annals of Statistics*, 37, 1705-1732.

Hansen, L., and Jagannathan, R. (1991). Implications of Security Market Data for Models of Dynamic Economies. *Journal of Political Economy*, 99(2), 225–262.

Onatski, A. (2009). Testing Hypotheses about the Number of Factors in Large Factor Models. *Econometrica*, 77(5), 1447-1479.

SUPPLEMENTARY MATERIALS FOR CODING

Double Machine Learning Inference for Conditional Latent Factors

Patrick Gagliardini and Hao Ma

Section 1 provides the list of firm-level characteristics and predictors in the dataset used for the empirics, and their description. In Section 2 we give details for the implementation of the ANN-based Sieve estimators used in the empirics. Section 3 reformulates linear projections of the conditional risk premium as DML-estimable population moments and develops valid asymptotic inference supporting Table 2 and Figure 3. Section 4 develops a procedure to identify and economically interpret latent factor directions that are orthogonal to observable factor models supporting Figure 4.

S.1 Dataset

S.1.1 Firm-Level Characteristics and Returns

In Table S.1 we provide the list and description of the 78 characteristics used to build our portfolio returns $\hat{\xi}_t$. The complete csv file for all characteristics is available at the Open Asset Pricing database (Chen and Zimmermann (2022)). The 78 characteristics are chosen to ensure that each portfolio comprises at least 500 cross-sectional lagged instrument $\times$ stock return observations (in yearly average) throughout our sample.

We apply some data standardization for numerical stability reasons. Specifically, we clip the monthly stock excess returns to -100% and 100%. We apply the volatility standardization approach introduced in Kelly et al. (2024) to manage scaling issues in predictive tasks. Predictors in Z_t are standardized to zero mean and unit variance using mean and variance estimates from the training window only, which are then held fixed when transforming the validation and test samples, ensuring out-of-sample validity and preventing data leakage.

S.1.2 Predictors

In Table S.2 we list the 25 predictors used to generate the conditioning information vector. Recall that the conditioning vector Z_{t-1} comprises four lags of the predictors and of the portfolio returns.

The collection of our conditional variables includes several variables introduced in recent studies: the illiquidity measure (`lzrt`) from Chen, Eaton, and Paye (2018), the output gap (`ogap`) from Cooper and Priestley (2009), oil price changes (`wtexas`) examined by Driesprong, Jacobsen, and Maat (2008), average stock skewness (`skvw`) analyzed by Jondeau, Zhang,

and Zhu (2019), and cross-sectional tail-risk (`tail`) as described by Kelly and Jiang (2014). Additionally, we incorporate a factor extracted from the cross-section of book-to-market ratios (`fbm`) following Kelly and Pruitt (2013), the scaled current difference to the 52-week high of the Dow Jones Index (`dtoy`) and the current difference to the lifetime high (`dtoat`), both from Li and Yu (2012), stock return dispersion (`rdsp`) from Maio (2016), the average correlation of stock returns (`avgcor`) documented by Pollett and Wilson (2010), and the lagged 1-month market return (`mkt`).

We also include 14 variables originally examined in Goyal and Welch (2008): the net issuing activity (`ntis`), the Treasury-bill rate (`tbl`), the dividend-price ratio (`d/p`), the dividend-yield (`d/y`), the earnings-price ratio (`e/p`), the dividend-payout ratio (`d/e`), stock volatility (`svar`), the book-to-market ratio (`b/m`), the long-term yield (`lty`), the long-term bond rate of return (`ltr`), the term spread (`tms`), the default yield (`dfy`), the default rate of return (`dfr`), and the inflation rate (`infl`). For precise definitions, we refer to Goyal and Welch (2008).

These 25 conditioning variables encompass a broad range of macroeconomic indicators, financial market measures, and cross-sectional risk factors. Integrating these variables into our framework allows for the robust estimation of conditional expectations and variances, grounded in both economic theory and empirical evidence.

Table S.1: Summary of Firm-Level Predictors

Acronym	Name	Definition	Reference
AM	Fama and French (1992)	Total assets to market	Total assets (at) divided by market value of equity.
BM	Stattman (1980)	Book to market, original (Stattman 1980)	Log of tangible book equity (ceqt) over market equity matched at FYE
Beta	Fama and MacBeth (1973)	CAPM beta	Coefficient of a 60-month rolling window regression of monthly stock returns minus the riskfree rate on market return minus the risk free rate (ewretd - rf). Exclude if estimate based on less than 20 months of returns.
BetaFP	Frazzini and Pedersen (2014)	Frazzini-Pedersen Beta	Using daily data, call tempRi the sum of today's return and the past 2 days, call tempRm the sum of today's mkt return and the past 2 days mkt return. Regress return on tempRi and tempRm using past 5 years of data, with a minimum of 3 years. BetaFP is the square root of the r square from this regression times the ratio of the idiosyncratic stock vol and the market vol.
BetaTailRisk	Kelly and Jiang (2014)	Tail risk beta	Each month, compute the 5th percentile over daily returns over all firms. For all daily return observations with return below that 5th percentile, compute the average of (log(ret/5th percentile of cross-sectional return distribution). Call that average tailEX. BetaTailRisk is the coefficient of a 120-month rolling regression of a firm's stock return on tailEX. Exclude if price less than 5 or share code greater than 11.
BidAskSpread	Amihud and Mendelson (1986)	Bid-ask spread	Effective bid ask spread based on Corwin-Schulz scaled by stock price.
BookLeverage	Fama and French (1992)	Book leverage (annual)	Total assets (at) divided by book value of equity plus deferred taxes (txditc) and preferred stock. Equity is shareholder equity (seq) if available, or book equity (ceq) plus preferred stock (pstk, if missing pstkrv, if missing pstkl), or total assets minus total liabilities (lt).
CF	Lakonishok, Shleifer, Vishny (1994)	Cash flow to market	Net income (ib) plus depreciation (dp) divided by market equity. Exclude NASDAQ stocks.
CashProd	Chandrashekar and Rao (2009)	Cash Productivity	Calculate market value of equity (mve_c) as absolute price (prc) times number of shares outstanding (shrout). Cash productivity is equal to the difference between mve_c and total assets (at) divided by cash and short-term investments (che).
CompEquIss	Daniel and Titman (2006)	Composite equity issuance	5 year growth rate of market value of equity minus 5 year stock return.
ConvDebt	Valta (2016)	Convertible debt indicator	Binary variable equal to 1 if deferred charges (dc) greater than 0 or common shares reserved for convertible debt (cshrc) greater than 0.
Coskewness	Harvey and Siddique (2000)	Coskewness	Signal is the sample counterpart of $E[\tilde{r}_{it}\tilde{r}_{mt}^2]/(SD[\tilde{r}_{it}]SD[\tilde{r}_{mt}]^2$ where $\tilde{r}_{it}$ is the de-meaned stock excess return and $\tilde{r}_{mt}$ is the de-meaned market excess return. Signal is computed using the past 60 months of monthly data, and using the NYSE/AMEX CRSP VW index for the market (msic). See code for details.
DivInit	Michaely, Thaler and Womack (1995)	Dividend Initiation	Keep only distcd 2nd digit = 2 or 3. Define dividend initiation as having paid a dividend in month t (divamt >0), and not having paid a dividend in the last 24 months. DivInit is equal to 1 if a dividend was initiated in the past 6 months, and 0 for all other stocks.
DivOmit	Michaely, Thaler and Womack (1995)	Dividend Omission	Keep only distcd 2nd digit = 2 or 3. Define firms as quarterly, semi-annual, or annual payers based on payment history. Define a consistent payer as a firm that paid quarterly or semi-annual dividends for the past 18 months, or annual dividends for the past 2 years. An omission is the first month where a consistent payer failed to pay a dividend over the past quarter/6-months/year. Finally, DivOmit = 1 if there is an omission in the past 2 months and 0 otherwise
DivSeason	Hartzmark and Salomon (2013)	Dividend seasonality	Drop if 3rd digit of distcd = 2 or >= 6. Assign DivSeason = 0 if there was a dividend paid in the last 12 months. Replace DivSeason = 1 if the third digit of disctcd is 3, 0, or 1, and a positive dividend was paid 2, 5, 8, or 11 months ago, if the third digit is 4 and a dividend was paid 5 or 11 months ago, or if the third digit is 5 and a dividiend was paid 11 months ago.

(Continued on next page)

(Table continued from previous page)

Acronym	Reference	Long Description	Detailed Definition
DivYieldST	Litzenberger and Ramaswamy (1979)	Predicted div yield next month	Using CRSP distributions, keep only distcd beginning in 12 and if the third digit of distcd is 3, 4, or 5. Also keep only if stock paid a dividend in the last 12 months. Define Ediv1 = div 2 months ago if distcd's 3rd digit is 0, 1, or 3. Define Ediv1 = div 5 months ago if the distcd 3rd digit is 4. Define Ediv = div 11 months ago if the distcd 3rd digit is 5. Define Edy1 as Ediv1/abs(prc). Finally, discretize Edy1 to smooth around the huge mass at 0 as follows: DivYieldST = 0 if Edy1 = 0, 1 if Edy is between 0 and 0.005, 2 if Edy1 is between 0.005 and 0.010, and 3 if Edy >0.010.
DolVol	Brennan, Chordia, Subra (1998)	Past trading volume	Log of two-month lagged trading volume (vol) times two-month lagged price (prc).
EP	Basu (1977)	Earnings-to-Price Ratio	ib / lag(market value of equity, 6 months). NYSE stocks only. Exclude if EP <0. Lag simulates the Dec 31 market equity used in original paper
EntMult	Loughran and Wellman (2011)	Enterprise Multiple	Market value of equity + long-term debt (dltt) + debt in current liabilities (dlc) + deferred charges (dc) - cash and short-term investments (che) , divided by operating income (oibdp). Exclude if missing book equity or negative operating income.
ExchSwitch	Dharan and Ikenberry (1995)	Exchange Switch	Binary variable equal to 1 if a firm switched from AMEX or NASDAQ to NYSE within the past year, or from NASDAQ to AMEX within the past year.
FirmAge	Barry and Brown (1984)	Firm age based on CRSP	Months since start of CRSP coverage.
FirmAgeMom	Zhang (2006)	Firm Age - Momentum	6 month return, restricted to the bottom quintile of the cross-sectional firm age distribution. Exclude if price less than 5 or firm younger than 12 months.
GP	Novy-Marx (2013)	gross profits / total assets	Revenue (sale) - cost of goods solds (cogs), divided by total assets (at). Drop if financial.
Herf	Hou and Robinson (2006)	Industry concentration (sales)	Three-year rolling average of the three digit industry Herfindahl index based on firm revenue (sale). Exclude regulated industries (4011, 4210, 4213 & year $\leq$ 1980; 4512 & year $\leq$ 1978, 4812, 4813 & year $\leq$ 1982, 4900-4999 in any year)
High52	George and Hwang (2004)	52 week high	Price (prc/cfacshr) divided by the maximum price over the previous 12 months.
IdioVol3F	Ang et al. (2006)	Idiosyncratic risk (3 factor)	Standard deviation of residuals from Fama-French three factor regressions using the past month of daily data. Value weighted
IdioVolAHT	Ali, Hwang, and Trombley (2003)	Idiosyncratic risk (AHT)	Standard deviation of residuals from CAPM regressions using the past year of daily data. Require at least 100 non-missing observations.
Illiquidity	Amihud (2002)	Amihud's illiquidity	Past twelve month average of: daily return (abs(ret)) divided by turnover((abs(prc)*vol)
IndIPO	Ritter (1991)	Initial Public Offerings	1 if IPO in the past 6-36 months. 0 otherwise. IPO dates are taken from Jay Ritter's IPO data available at: http://bear.warrington.ufl.edu/ritter/ipodata.htm. Missing IPO dates imply IndIPO = 0
IndMom	Grinblatt and Moskowitz (1999)	Industry Momentum	Weighted average of firm-level 6 month buy-and-hold return. Average is taken over two digit industries each month and weights are based on market value of equity.
IndRetBig	Hou (2007)	Industry return of big firms	Average monthly return (ret) of the 30% largest companies by market value of equity in the same Fama-French 48 industry. Exclude the largest 30% of companies for IndRetBig (not to compute the anomaly!)
IntMom	Novy-Marx (2012)	Intermediate Momentum	Stock return between months t-12 and t-6
LRreversal	De Bondt and Thaler (1985)	Long-run reversal	Stock return between months t-36 and t-13.
Leverage	Bhandari (1988)	Market leverage	Total liabilities (lt) divided by market value of equity.
MRreversal	De Bondt and Thaler (1985)	Medium-run reversal	Stock return between months t-18 and t-13.
MaxRet	Bali, Cakici, and Whitelaw (2011)	Maximum return over month	Maximum of daily returns (ret) over the previous month
Mom12m	Jegadeesh and Titman (1993)	Momentum (12 month)	Stock return between months t-12 and t-1
Mom12mOffSeason	Heston and Sadka (2008)	Momentum without the seasonal part	Average return in other months over the previous year.
Mom6m	Jegadeesh and Titman (1993)	Momentum (6 month)	Stock return between months t-6 and t-1
MomOffSeason	Heston and Sadka (2008)	Off season long-term reversal	Average return in other months over the preceding 2-5 years.
MomOffSeason06YrPlus	Heston and Sadka (2008)	Off season reversal years 6 to 10	Average return in other months over the preceding 6-10 years.
MomOffSeason11YrPlus	Heston and Sadka (2008)	Off season reversal years 11 to 15	Average return in other months over the preceding 11-15 years.
MomOffSeason16YrPlus	Heston and Sadka (2008)	Off season reversal years 16 to 20	Average return in other months over the preceding 16-20 years.
MomRev	Chan and Ko (2006)	Momentum and LT Reversal	Binary variable equal to 1 if firm is in the highest Mom6m quintile and the lowest Mom36m quintile, and equal to 0 if firm is in the lowest Mom6m quintile and the highest Mom36m quintile. Exclude if price less than 5.
MomSeason	Heston and Sadka (2008)	Return seasonality years 2 to 5	Average return in the same month over the preceding 2-5 years.

(Continued on next page)

(Table continued from previous page)

Acronym	Reference	Long Description	Detailed Definition
MomSeason06YrPlus	Heston and Sadka (2008)	Return seasonality years 6 to 10	Average return in the same month over the preceding 6-10 years.
MomSeason11YrPlus	Heston and Sadka (2008)	Return seasonality years 11 to 15	Average return in the same month over the preceding 11-15 years.
MomSeason16YrPlus	Heston and Sadka (2008)	Return seasonality years 16 to 20	Average return in the same month over the preceding 16-20 years.
MomSeasonShort	Heston and Sadka (2008)	Return seasonality last year	Average return in the same monthin the previous year.
MomVol	Lee and Swaminathan (2000)	Momentum in high volume stocks	Define momentum as Mom6m, and volume as the rolling average of the past 6 months of monthly turnover (minimum 5 months). Independent sort stocks into 10 momentum ports and 3 volume ports. Keep if volume is in the top port, and assign signal = momentum port. Drop if less than 2 years on CRSP.
OPLeverage	Novy-Marx (2011)	Operating leverage	Sum of administrative expenses (xsga) and cost of goods sold (cogs), scaled by total assets (at). Use xsga = 0 if xsga is missing.
Price	Blume and Husic (1973)	Price	Log of absolute value of price (prc).
PriceDelayRsq	Hou and Moskowitz (2005)	Price delay r square	Each July regress stock return on day t on market return in $t, t-1, \ldots, t-4$ using observations from July 1 of the previous year to June 30 of the current year. Then regress again with no market return lags. PriceDelay Rsq = 1 - [Rsq from second regression]/[Rsq from first regression]
PriceDelaySlope	Hou and Moskowitz (2005)	Price delay coeff	Each July regress daily stock return (ret) on market return (mktrf) in $t, t-1, \ldots, t-4$ with observations over the previous year. Define PriceDelaySlope as the ratio of]1*beta on $\text{mktrf}_t - 1$ + 2*beta on $\text{mktrf}_t - 2$ + 3*beta on $\text{mktrf}_t - 3$ + 4*beta on $\text{mktrf}_t - 4$], and [beta on mktrf_t + beta on $\text{mktrf}_t - 1$ + beta on $\text{mktrf}_t - 2$ + beta on $\text{mktrf}_t - 3$ + beta on $\text{mktrf}_t - 4$].
PriceDelayTstat	Hou and Moskowitz (2005)	Price delay SE adjusted	Each July regress daily stock return (ret) on market return (mktrf) in $t, t-1, \ldots, t-4$ using observations from July 1 of the previous year to June 30 of the current year. Trim the highest and lowest 1% of estimated coefficients, standard errors, and t-stats. Define PriceDelayTstat as the ratio of [1*tstat on $\text{mktrf}_t - 1$ + 2*tstat on $\text{mktrf}_t - 2$ + 3*tstat on $\text{mktrf}_t - 3$ + 4*tstat on $\text{mktrf}_t - 4$], and [tstat on mktrf_t + tstat on $\text{mktrf}_t - 1$ + tstat on $\text{mktrf}_t - 2$ + tstat on $\text{mktrf}_t - 3$ + tstat on $\text{mktrf}_t - 4$]. Finally winsor the extreme 10 and 90 percentiles each month.
RDIPO	Gou, Lev and Shi (2006)	IPO and no R&D spending	Binary variable equal to 1 if R&D expense (xrd) = 0 and IndIPO = 1. 0 otherwise.
RIO_Turnover	Nagel (2005)	Inst Own and Turnover	Follow RIO_MB, except define turnover as vol/shrout. Let RIO_Turnover = lagged RIO quintile if the Volatility quintile == 5
RIO_Volatility	Nagel (2005)	Inst Own and Idio Vol	Follow RIO_MB, except define volatility as the rolling standard deviation of the last 12 months of the stock return. Let RIO_Disp = lagged RIO quintile if the Volatility quintile == 5
RealizedVol	Ang et al. (2006)	Realized (Total) Volatility	Standard deviation of residuals from CAPM regressions using the past month of daily data. Value weighted
ResidualMomentum	Blitz, Huij and Martens (2011)	Momentum based on FF3 residuals	Run a rolling regression over 36 months of excess return (retrf) on excess market return (mktrf), size and value factors (smb, hml) and compute idiosyncratic returns as the one-month lagged residual. ResidualMomentum is the rolling mean of the residual divided by the rolling standard deviation of the residual, both computed over the past 11 months.
ReturnSkew	Bali, Engle and Murray (2015)	Return skewness	Skewness of daily returns (ret) over previous month.
ReturnSkew3F	Bali, Engle and Murray (2015)	Idiosyncratic skewness (3F model)	Skewness of idiosyncratic returns computed as residuals from regression of daily excess returns (ret - rf) on Fama-French factors (mktrf, smb, hml) over the previous month. We require at least 15 non-missing observations.
SP	Barbee, Mukherji and Raines (1996)	Sales-to-price	Ratio of annual sales (sale) to market value of equity.
STreversal	Jegadeesh (1990)	Short term reversal	Stock return (ret) over the previous month.
ShareIss1Y	Pontiff and Woodgate (2008)	Share issuance (1 year)	Growth in number of shares between t-18 and t-6. Number of shares is calculated as shrout/cfacshr to adjust for splits.
ShareIss5Y	Daniel and Titman (2006)	Share issuance (5 year)	5-year growth in number of shares. Number of shares is calculated as shrout/cfacshr to adjust for splits.
ShareRepurchase	Ikenberry, Lakonishok, Vermaelen (1995)	Share repurchases	Binary variable equal to 1 if stock repurchase indicated in cash flow statement (prstkc >0), and 0 if prstkc = 0.

(Continued on next page)

(Table continued from previous page)

Acronym	Reference	Long Description	Detailed Definition
ShareVol	Datar, Naik and Radcliffe (1998)	Share Volume	Let tempvol = sum of monthly share trading volume (vol) over the previous three months, scaled by 3 times common shares outstanding (shrout). Let ShareVol = 1 if tempvol >10%, and ShareVol = 0 if tempvol <5%.
Size	Banz (1981)	Size	Log of monthly market value of equity (abs(prc)*shrout)).
Spinoff	Cusatis, Miles and Woolridge (1993)	Spinoffs	Spinoffs are identified as all observations in the CRSP acquisition file with valid acperm entry. Spinoff is a binary variable equal to 1 if a firm is identified in the CRSP Acquisition data and if it has at most two years of history in the CRSP stock return data. Spinoff is equal to 0 otherwise.
TrendFactor	Han, Zhou, Zhu (2016)	Trend Factor	See paper section 2.1 and 2.2
VolMkt	Haugen and Baker (1996)	Volume to market equity	Average monthly dollar trading volume (vol*abs(prc)) over the previous 12 months, scaled by market value of equity. Exclude if price less than 5.
VolSD	Chordia, Subra, Anshuman (2001)	Volume Variance	Rolling standard deviation of monthly trading volume (vol) over the past 36 months (require at least 24 observations). Include only NYSE stocks.
VolumeTrend	Haugen and Baker (1996)	Volume Trend	Rolling coefficient from regressing monthly trading volume on a linear time trend over a window of 60 months (require that at least 30 exist). Scale coefficient by 60-month average of trading volume.
std_turn	Chordia, Subra, Anshuman (2001)	Share turnover volatility	Standard deviation of turnover (vol/shrout) over the past 36 months.
zerotrade12M	Liu (2006)	Days with zero trades	In each month, count the number of days with no trades. Define zerotrade as the number of days without trades plus (the sum of monthly turnover (vol/shrout) divided by $48*10^5$), multiplied by 21/number of trading days per month. Take 12-month average.
zerotrade1M	Liu (2006)	Days with zero trades	In each month, count the number of days with no trades. Define zerotrade as the number of days without trades plus (the sum of monthly turnover (vol/shrout) divided by $48*10^5$), multiplied by 21/number of trading days per month. Take 1-month average.
zerotrade6M	Liu (2006)	Days with zero trades	In each month, count the number of days with no trades. Define zerotrade as the number of days without trades plus (the sum of monthly turnover (vol/shrout) divided by $48*10^5$), multiplied by 21/number of trading days per month. Zerotrade is the 6-month average of that variable.

Table S.2: Summary of Macroeconomic and Market Indicators

Acronym	Name	Definition	Reference
avgcor	Pollett and Wilson (JFE 2010)	Average Correlation of Daily Stock Returns	Measures the average correlation between daily stock returns, which can serve as a market-wide risk indicator.
b/m	Kothari and Shanken (1997), Pontiff and Schall (1998)	Book-to-Market Ratio	The ratio of the book value of a firm's equity to its market value, often used as an indicator of value.
d/e	Campbell (1987)	Dividend-Payout Ratio	The proportion of earnings paid out as dividends to shareholders, typically expressed as a percentage.
d/p	Campbell (1987)	Dividend-Price Ratio	The ratio of dividends paid out by a company relative to its stock price.
d/y	Campbell (1987)	Dividend-Yield	The percentage of a company's stock price paid out in dividends annually.
dfr	Fama and French (1989)	Default Rate of Return	The rate of return adjusted for the likelihood of default by the issuer.
dfy	Fama and French (1989)	Default Yield	The yield spread between corporate bonds and government bonds of similar maturity, reflecting the credit risk of corporate issuers.
dtoy, dtoat	Li and Yu (JFE 2012)	Nearness to Dow 52-Week High	Represents the proximity of the stock price to its 52-week high, indicating investor sentiment and market pressure.
e/p	Campbell (1987)	Earnings-Price Ratio	The ratio of a company's earnings to its stock price, indicating the earnings generated per share.
fbm	Kelly and Pruitt (JF 2013)	Single Factor from B/M Cross-Section	A single factor derived from the cross-section of book-to-market ratios, indicating market expectations.
infl	Fama and Schwert (1977)	Inflation Rate	The rate at which the general level of prices for goods and services is rising, and subsequently, eroding purchasing power.
ltr	Fama and French (1989)	Long-Term Bond Rate of Return	The total return on long-term government or corporate bonds, including both interest payments and changes in value.
lty	Fama and French (1989)	Long-Term Yield	The interest rate on long-term government or corporate bonds, typically reflecting expectations of future economic conditions.
lzrt	Chen, Eaton, Paye (JFE 2018)	Illiquidity Measures	A set of nine aggregate illiquidity measures reflecting market frictions.
mkt		Market returns	CRSP value-weighted market returns
ntis	Boudoukh et al. (2007)	Net Issuing Activity	Measures the net issuance of stocks and bonds by corporations, reflecting their financing decisions.
ogap	Cooper and Priestley (RFS 2009)	Output Gap of Industrial Production	Measures the deviation of industrial production from its long-term trend.
rdsp	Maio (JFM 2016)	Stock-Return Dispersion	Measures the variation in returns across different stocks, which can indicate market volatility and risk.
skvw	Jondeau, Zhang, Zhu (JFE 2019)	Average Stock Skewness	Represents the average skewness of stock returns, capturing asymmetry in the distribution of asset returns.
svar	Guo (2006)	Stock Volatility	A measure of the dispersion of returns for a given security or market index.
tail	Kelly and Jiang (RFS 2014)	Tail Risk from Cross-Section	Measures the risk associated with extreme market movements or 'tail events,' reflecting the distribution's downside.
tbl	Campbell (1987)	Treasury-Bill Rate	The yield on short-term U.S. government securities, reflecting the cost of borrowing for the government.
tms	Fama and French (1989)	Term-Spread	The difference between long-term and short-term interest rates, often used as an indicator of economic expectations.
wtexas	Driesprong, Jacobsen, Maat (JFE 2008)	Oil Price Changes	Represents fluctuations in the price of West Texas Intermediate (WTI) crude oil.

S.2 Implementation and Computational Details

S.2.1 Codebase Structure and Execution Order

The implementation is organized as a sequential pipeline with four steps, executed in the following order:

1. Step 1: conditional-mean prediction (validation, then final prediction with ensembling)
2. Step 2: residual construction and quadratic-target construction
3. Step 3: conditional second-moment prediction on quadratic targets (validation, then final prediction with ensembling)
4. Step 4: matrix reconstruction and eigenvalue regularization

The scripts referenced below follow these naming conventions:

- `py1aValidate_*.py`: Step 1 validation (hyperparameter selection)
- `py1bPredict_*_ensemble.py`: Step 1 final prediction (ensemble)
- `py2Create_*TildeQuad.py`: Step 2 residual and quadratic-target construction
- `py3aValidate_*TildeQuad.py`: Step 3 validation (hyperparameter selection)
- `py3bPredict_*TildeQuad_ensemble.py`: Step 3 final prediction (ensemble)
- `py4CondEandVC_*.py`: Step 4 assembly and eigenvalue adjustment

S.2.2 Learning Model, Software, and Fixed Settings

All prediction steps use `scikit-learn` `MLPRegressor` with the following fixed settings:

- activation: ReLU
- solver: Adam
- learning rate schedule: adaptive
- `max_iter`: 2000
- early stopping: enabled
- early-stopping patience: 10 iterations
- internal validation fraction (for early stopping): 0.10

(see Pedregosa et al. (2011)).

S.2.3 Hyperparameter Grid

For each validation step, a grid search is run over:

- number of hidden layers: {1,2,3,4,5}
- L2 regularization parameter `alpha`: {10, 100, 1000}

This yields 15 configurations per validation.

Hidden-layer widths follow a pyramid rule:

- base width: $\lfloor \text{input\_dim}/2 \rfloor$
- each subsequent layer: half the previous layer (integer-rounded)
- minimum layer width: $\max(\text{output\_dim}, 10)$

S.2.4 Time-Series Cross-Validation Protocol

Model selection is performed using an expanding-window protocol with:

- `T_min`: minimum training length
- `T_val`: validation window length, fixed over time (10 years)

For each evaluation date t, the data split is:

- training: all observations up to $t -$ `T_val`
- validation: the `T_val`-year window ending at t
- test: the single observation at t

The expanding-window estimation strategy is illustrated in Figure S.1.

Hyperparameters are selected using the Campbell–Thompson out-of-sample R^2 computed over the validation window, as implemented in the codebase.

S.2.5 Feature Scaling (Fold-by-Fold)

At each evaluation date t, scaling is performed as follows:

1. Fit `StandardScaler` using the training sample only.
2. Transform training, validation, and test samples using the training-sample mean and standard deviation.
3. Convert predictions back to the original scale using the inverse transformation.

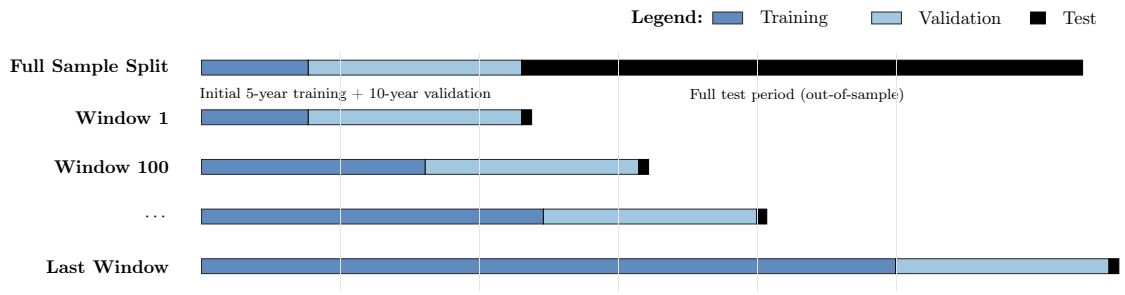

Figure S.1: **Time-Series Sample Split.** This figure depicts the expanding-window cross-validation scheme used for estimation and evaluation. The sample runs from July 1964 to December 2023, with an initial window of 5 years for training, 10 years for validation, and a one-month test period. After the initial split, the training window expands by one month at each step, while the validation window remains fixed.

S.2.6 Ensembling and Fallback Rule

Final predictions are produced using an ensemble of `N_ensemble = 10` independently initialized networks (different random seeds) with the selected hyperparameters. Ensemble predictions are computed by averaging across the 10 fitted models.

A validation-score threshold is applied before producing the final test-date prediction:

- if the selected-model validation R^2 is strictly greater than `R2_threshold` (default 0), the ensemble is refit on training+validation and evaluated on the test observation;
- otherwise, the prediction for that test date is set to the historical mean computed from the expanding training sample (as implemented in the codebase).

S.2.7 Step 1 Outputs (Conditional-Mean Prediction)

Step 1 is run separately for each target index. For each target, the implementation writes a CSV file containing:

- test dates
- realized target values
- predicted values (ensemble or fallback)
- historical-mean values used by the fallback
- validation R^2 used for hyperparameter selection
- indicator flags for whether the ensemble or fallback was used

S.2.8 Step 2 (Residual and Quadratic-Target Construction)

Step 2 constructs:

- residual series for each target index, using Step 1 predictions
- quadratic targets formed as pairwise products of residual series

Only the upper-triangular elements (including the diagonal) are retained for storage and subsequent prediction. The resulting quadratic targets are saved as a single vectorized dataset merged with predictors.

S.2.9 Step 3 Outputs (Conditional Second-Moment Prediction)

Step 3 repeats the Step 1 workflow (Sections 2.2–2.6) on the quadratic targets:

- expanding-window validation over the 15-configuration grid
- 10-model ensemble prediction using the selected configuration
- the same threshold-based fallback rule (default `R2_threshold = 0`)

Outputs are written as CSV files containing predicted quadratic-target values by date and target index.

S.2.10 Step 4 (Matrix Reconstruction and Eigenvalue Regularization)

For each date t, predicted quadratic targets are mapped back to a symmetric matrix using the inverse of the upper-triangular vectorization used in Step 2.

An eigenvalue adjustment is then applied:

1. compute an eigendecomposition of the reconstructed symmetric matrix;
2. keep the `K_max` largest eigenvalues (user-specified);
3. set all remaining eigenvalues to `eigen_threshold` (default 10^{-10});
4. reconstruct the adjusted matrix using the original eigenvectors and the adjusted eigenvalues.

S.2.11 Parallelization and Blocking

Two levels of parallelization are used:

- an outer level across test dates, default 5 parallel jobs;
- an inner level across ensemble members, default 2 parallel jobs.

Test dates are processed in blocks (default block size: 50 dates) with incremental writes to disk.

S.2.12 Software Dependencies

The implementation requires:

- `numpy`
- `pandas`
- `scikit-learn`
- `joblib`

S.2.13 Command-Line Interface

Validation scripts accept command-line arguments in the following format:

```
python py1aValidate_xi.py <target_idx> <num_targets> <num_pred_lags>

<min_train_years> <val_years>
```

Example:

```
python py1aValidate_xi.py 0 78 24 5 10
```

S.2.14 Replication Parameters (Defaults)

Table S.3 lists the default parameter values used in the provided scripts.

Table S.3: Default implementation parameters.

Parameter	Default value
`N_ensemble`	10
`alpha_grid`	{10, 100, 1000}
`hidden_layers`	{1,2,3,4,5}
`max_iter`	2000
`early_stopping`	enabled
`patience`	10
`validation_fraction`	0.10
`R2_threshold`	0
`eigen_threshold`	10^{-10}
`block_size`	50
`parallel_outer_jobs`	5
`parallel_inner_jobs`	2

S.3 Projection Coefficients as DML Estimable Objects

This section provides a formal DML-compatible framework for interpreting linear projections of the conditional risk premium (or other identifiable stochastic processes). While smoothed time-series fitted values are not themselves valid DML targets, the *projection coefficients* can be written via population moments of the form $E[Z_{t-1}\gamma_t]$ and hence are estimable via DML. The section derives the asymptotic distribution of the projection coefficients and projected risk-premium process using joint long-run variance estimation and the delta method, and develops Wald-type tests for economically relevant hypotheses such as time-constancy and counter-cyclicality of conditional risk premia. The methodology is implemented in the code that produces testing results in Section 5.3.2, in particular Table 2 and Figure 3.

S.3.1 Methodology

Suppose γ_t is scalar to ease notation. We project linearly γ_t onto the predictor vector Z_{t-1}, i.e.

$$\gamma_t = B_0' Z_{t-1} + u_t \tag{S.1}$$

where u_t is the linear projection error. In our application, vector Z_{t-1} consists of a subset of the predictors used for nonparametric estimation of conditional moments. For ease of

exposition, we still denote Z_{t-1} such subset. Coefficient vector B_0 is

$$B_0 = \mathbb{E}[Z_{t-1}Z'_{t-1}]^{-1}\mathbb{E}[Z_{t-1}\gamma_t] =: \Sigma_Z^{-1}c_0 \tag{S.2}$$

where $\Sigma_Z = \mathbb{E}[Z_{t-1}Z'_{t-1}]$, and $c_0 = \mathbb{E}[Z_{t-1}\gamma_t]$ is in the form of an object that can be estimated with DML. We define the projected process as $\gamma_t^p = B_0'Z_{t-1}$, that is a function of a DML estimable parameter and an observable process.

Let $\hat{c}$ be the DML estimator of c_0, namely $\hat{c} = \frac{1}{I}\sum_{t\in\mathcal{I}} Z_{t-1}(\hat{\gamma}_t + \hat{\eta}_t)$, where $\hat{\gamma}_t$ and $\hat{\eta}_t$ are estimators of the γ_t and η_t processes based on $\hat{\theta}(\cdot)$ computed on sample $\mathcal{T}$ (when $\gamma_t = \lambda_t^O$, process η_t is given in the first row of Table 1). Its asymptotic distribution is $\sqrt{I}(\hat{c} - c_0) \Rightarrow N(0, \sigma^2)$, where σ^2 is the long run covariance of vector process $h_t = Z_{t-1}(\gamma_t + \eta_t) - c_0$. Let $\hat{\Sigma}_Z = \frac{1}{I}\sum_{t\in\mathcal{I}} Z_{t-1}Z'_{t-1}$, with $\sqrt{I}vec(\hat{\Sigma}_Z - \Sigma_Z) \Rightarrow N(0, V_{\Sigma_Z})$, where V_{Σ_Z} is the long-run variance of process $Z_{t-1} \otimes Z_{t-1}$. Then, the estimator

$$\hat{B} = \hat{\Sigma}_Z^{-1}\hat{c} \tag{S.3}$$

is asymptotically normal $\sqrt{I}(\hat{B} - B_0) \Rightarrow N(0, \Sigma_B)$, and we get Σ_B by the delta method (see below). Moreover, the estimator of the projected process value γ_t^p is

$$\hat{\gamma}_t^p = \hat{B}'Z_{t-1} \tag{S.4}$$

and again by delta method we have $\sqrt{I}(\hat{\gamma}_t^p - \gamma_t^p) \Rightarrow N(0, Z'_{t-1}\Sigma_B Z_{t-1})$, pointwise for any t.

Let us develop the delta method arguments for the RHS of equation (S.3). We have:

$$\sqrt{I}vec(\hat{\Sigma}_Z^{-1} - \Sigma_Z^{-1}) = -(\Sigma_Z^{-1} \otimes \Sigma_Z^{-1})\sqrt{I}vec(\hat{\Sigma}_Z - \Sigma_Z) + o_p(1). \tag{S.5}$$

Moreover:

$$\begin{aligned}
\sqrt{I}(\hat{B} - B_0) &= \sqrt{I}(\hat{\Sigma}_Z^{-1} - \Sigma_Z^{-1})c_0 + \Sigma_Z^{-1}\sqrt{I}(\hat{c} - c_0) + o_p(1) \\
&= (c_0' \otimes I_d)\sqrt{I}vec(\hat{\Sigma}_Z^{-1} - \Sigma_Z^{-1}) + \Sigma_Z^{-1}\sqrt{I}(\hat{c} - c_0) + o_p(1) \\
&= -(c_0'\Sigma_Z^{-1} \otimes \Sigma_Z^{-1})\sqrt{I}vec(\hat{\Sigma}_Z - \Sigma_Z) + \Sigma_Z^{-1}\sqrt{I}(\hat{c} - c_0) + o_p(1) \\
&= \left[\Sigma_Z^{-1} \;:\; -(c_0'\Sigma_Z^{-1} \otimes \Sigma_Z^{-1})\right] \begin{bmatrix} \sqrt{I}(\hat{c} - c_0) \\ \sqrt{I}vec(\hat{\Sigma}_Z - \Sigma_Z) \end{bmatrix} + o_p(1)
\end{aligned}$$

Hence, we need the joint asymptotic distribution of estimators $\hat{c}$ and $\hat{\Sigma}_Z$. From the proof of

Proposition 2 Part (b) in Appendix A.2.2, $\sqrt{I}(\hat{c} - c_0) = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} h_t + o_p(1)$. Then:

$$\left[\begin{array}{c} \sqrt{I}(\hat{c} - c_0) \\ \sqrt{I} vec(\hat{\Sigma}_Z - \Sigma_Z) \end{array} \right] = \frac{1}{\sqrt{I}} \sum_{t \in \mathcal{I}} H_t + o_p(1) \Rightarrow N(0, \Sigma_H)$$

where $H_t = (h_t', Z_{t-1}' \otimes Z_{t-1}' - vec(\Sigma_Z)')'$. The asymptotic variance of estimator $\hat{B}$ is

$$\Sigma_B = \left[\Sigma_Z^{-1} \ : \ -(c_0' \Sigma_Z^{-1} \otimes \Sigma_Z^{-1}) \right] \Sigma_H \left[\Sigma_Z^{-1} \ : \ -(c_0' \Sigma_Z^{-1} \otimes \Sigma_Z^{-1}) \right]'.$$

The long-run matrix Σ_H can be estimated by a HAC estimator, and we get a consistent estimator of Σ_B by plug-in.

S.3.2 Testing for time-constancy of the conditional risk premia

Let B_1 denote the $p \times 1$ lower subvector of B consisting of the coefficients of the p predictors in Z_{t-1} (i.e. except the constant). Then, if the conditional risk premium is time constant, we have $B_1 = 0$. To test this null hypothesis, we use $\sqrt{I}(\hat{B}_1 - B_{1,0}) \Rightarrow N(0, \Sigma_{B,11})$, where $\Sigma_{B,11}$ is the lower-right $p \times p$ block of Σ_B. The Wald test statistic is $I \hat{B}_1' (\hat{\Sigma}_{B,11})^{-1} \hat{B}_1$. It is asymptotically distributed as χ_p^2 under the null. If the Wald statistic is larger than the chi-square quantile, we reject the null of time-constancy.

S.3.3 Testing for counter-cyclicality of the conditional risk premia

Let $W_t = 1$ be the indicator for recession month (and $W_t = 0$ otherwise), and assume it is a (unknown) function of Z_t. Write the regression:

$$\gamma_t = a_0 + b_0 W_{t-1} + u_t \tag{S.6}$$

where the error u_t has mean zero and is orthogonal to W_{t-1}. Then, one can easily check that

$$\begin{aligned} a_0 &= \mathbb{E}(\gamma_t | W_{t-1} = 0) = \frac{\mathbb{E}[\gamma_t (1 - W_{t-1})]}{\mathbb{E}(1 - W_{t-1})} \\ b_0 &= \mathbb{E}(\gamma_t | W_{t-1} = 1) - \mathbb{E}(\gamma_t | W_{t-1} = 0). \end{aligned}$$

Hence, a_0 is the average conditional risk premium in non-recession periods, while b_0 is the difference between average conditional risk premia in recession vs non-recession periods. We can test for counter-cyclicality of the conditional risk premium by testing the null hypothesis $b_0 \leq 0$ and check if we find evidence to reject it.

Regression (S.6) fits in the general framework of equation (S.1), with $B_0 = (a_0, b_0)'$ and $Z_{t-1} = (1, W_{t-1})'$. The estimator with DML adjustment is

$$\hat{B} = \begin{bmatrix} 1 & \hat{q} \\ \hat{q} & \hat{q} \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{I}\sum(\hat{\gamma}_t + \hat{\eta}_t) \\ \frac{1}{I}\sum W_{t-1}(\hat{\gamma}_t + \hat{\eta}_t) \end{bmatrix} = \begin{bmatrix} \frac{\sum(1-W_{t-1})(\hat{\gamma}_t+\hat{\eta}_t)}{\sum(1-W_{t-1})} \\ \frac{\sum W_{t-1}(\hat{\gamma}_t+\hat{\eta}_t)}{\sum W_{t-1}} - \frac{\sum(1-W_{t-1})(\hat{\gamma}_t+\hat{\eta}_t)}{\sum(1-W_{t-1})} \end{bmatrix}$$

where $\hat{q} = \frac{1}{I}\sum W_{t-1}$ is the proportion of recession periods within sample $\mathcal{I}$, and we use the short notation $\sum$ for the sum $\sum_{t\in\mathcal{I}}$. Hence, the estimator of the slope $\hat{b} = \frac{\sum W_{t-1}(\hat{\gamma}_t+\hat{\eta}_t)}{\sum W_{t-1}} - \frac{\sum(1-W_{t-1})(\hat{\gamma}_t+\hat{\eta}_t)}{\sum(1-W_{t-1})}$ is simply the difference between the estimators of average conditional risk premium computed in subperiods of recessions vs non-recessions.

We get the variance in the asymptotic Gaussian distribution of $\sqrt{I}(\hat{b} - b_0)$ from the lower-right element of matrix Σ_B, denoted $(\Sigma_B)_{22}$. Equivalently (and easier), we can get it by directly obtaining the asymptotic expansion of estimator $\hat{b}$. We have:

$$\begin{aligned} \hat{b} &= \frac{1}{\hat{q}}\frac{1}{I}\sum W_{t-1}(\hat{\gamma}_t + \hat{\eta}_t) - \frac{1}{1-\hat{q}}\frac{1}{I}\sum(1 - W_{t-1})(\hat{\gamma}_t + \hat{\eta}_t) \\ &= \frac{1}{q}(1 - \frac{\hat{q}-q}{q} + o_p(I^{-1/2}))\frac{1}{I}\sum W_{t-1}(\hat{\gamma}_t + \hat{\eta}_t) \\ &\quad -\frac{1}{1-q}(1 + \frac{\hat{q}-q}{1-q} + o_p(I^{-1/2}))\frac{1}{I}\sum(1 - W_{t-1})(\hat{\gamma}_t + \hat{\eta}_t), \end{aligned}$$

where $q = \mathbb{E}(W_{t-1})$ is the unconditional probability of recession, which yields:

$$\begin{aligned} \hat{b} - b_0 &= \frac{1}{q}\left[\frac{1}{I}\sum W_{t-1}(\hat{\gamma}_t + \hat{\eta}_t) - \mathbb{E}(W_{t-1}\gamma_t)\right] \\ &\quad -\frac{1}{1-q}\left[\frac{1}{I}\sum(1 - W_{t-1})(\hat{\gamma}_t + \hat{\eta}_t) - \mathbb{E}[(1 - W_{t-1})\gamma_t]\right] \\ &\quad -\left[\frac{1}{q}\mathbb{E}(\gamma_t|W_{t-1} = 1) + \frac{1}{1-q}\mathbb{E}(\gamma_t|W_{t-1} = 0)\right](\hat{q} - q) + o_p(I^{-1/2}). \end{aligned}$$

Using the asymptotic expansion of DML estimators as in Proposition 2, and arranging terms, we get:

$$\sqrt{I}(\hat{b}-b_0) = \left[\frac{1}{q} : -\frac{1}{1-q} : \left(\frac{\mathbb{E}(\gamma_t|W_{t-1}=1)}{q} + \frac{\mathbb{E}(\gamma_t|W_{t-1}=0)}{1-q}\right)\right]\frac{1}{\sqrt{I}}\sum(H_t - \mathbb{E}(H_t)) + o_p(1),$$

where $H_t := [W_{t-1}(\gamma_t + \eta_t), (1 - W_{t-1})(\gamma_t + \eta_t), W_{t-1}]'$, and $\frac{1}{\sqrt{I}}\sum(H_t - \mathbb{E}(H_t)) \Rightarrow N(0, \sigma_H^2)$, where we can estimated σ_H^2 by a HAC estimator. Then, we get $\sqrt{I}(\hat{b}-b_0) \Rightarrow N(0, \sigma_b^2)$, where $\sigma_b^2 = S'\sigma_H^2 S$ and $S := \left[\frac{1}{q} \ : \ -\frac{1}{1-q} \ : \ \left(\frac{1}{q}\mathbb{E}(\gamma_t|W_{t-1}=1) + \frac{1}{1-q}\mathbb{E}(\gamma_t|W_{t-1}=0)\right)\right]'$.

The critical region for a test of the null hypothesis $b_0 \leq 0$ at asymptotic level 5% is $\hat{b} >$

$\frac{1.64}{\sqrt{I}}\hat{\sigma}_b$. If we reject, this is evidence for counter-cyclicality of the conditional risk premium.

S.4 Interpreting Unspanned Latent Factors

This section develops a formal procedure to identify and interpret latent factor directions that are not spanned by observable factor models, by maximizing conditional variance subject to conditional orthogonality with respect to observed factors. It maps these unspanned latent directions into economically interpretable characteristic-portfolio weights using sparsity and smoothness regularization to address non-uniqueness and time variation. The methodology is implemented in the code that produces output displayed in Figure 4.

S.4.1 Methodology

Our empirical findings imply that only a few dimensions of the latent factor space are spanned well by traditional sets of factors like FF5 and MOM. In order to understand better what drives the latent factor space, in Section 5.3.3 we focus on the dimensions in that space that are poorly conditionally correlated with the observed factors. For this purpose, we consider the constrained maximization problem

$$\max_{b\in\mathbb{R}^k} \mathbb{V}(b'f_t|\mathscr{F}_{t-1}), \qquad \text{s.t.} \quad \mathbb{C}\text{ov}(b'f_t, F_t^{(r)}|\mathscr{F}_{t-1}) = 0,\ b'b = 1, \tag{S.7}$$

for each month t, where $F_t^{(r)}$ is the vector of the first r conditional canonical directions in the latent space (i.e., those directions with the r largest conditional correlations with linear transformations of the observed factor f_t^O), with r a given integer, and $k_t \geq k^O > r$ (here we write $f_t \equiv f_t^*$ for the active latent factor to ease notation). In words, we seek the dimension in the latent factor space with the largest conditional variance, among those that are conditionally orthogonal to the space spanned by the canonical directions that have the r largest conditional correlations with the observed factor space. In fact, the condition $\mathbb{C}\text{ov}(b'f_t, F_t^{(r)}|\mathscr{F}_{t-1}) = 0$ means that $b'f_t$ is spanned by the $k^O - r$ conditional canonical directions of order $r+1, \cdots, k^O$ and the $k - k^O$ directions in the latent space that are conditionally orthogonal to the observed factors. The condition $b'b = 1$ is for normalization purpose and is irrelevant for our construction.

Let the solution to the maximization problem in (S.7) be $b = b_{t-1}$, i.e. $\mathscr{F}_{t-1}$ measurable (we characterize it below). Using $f_t = J'_{t-1}(\xi_t - \mathbb{E}[\xi_t|\mathscr{F}_{t-1}])$, then we get

$$b'_{t-1}f_t = (J_{t-1}b_{t-1})'(\xi_t - \mathbb{E}[\xi_t|\mathscr{F}_{t-1}]) \tag{S.8}$$

which yields the vector of weights

$$w_{t-1} = \frac{1}{1_K' J_{t-1} b_{t-1}} J_{t-1} b_{t-1}. \tag{S.9}$$

These weights give how the poorly spanned direction of the latent space with maximal conditional variance loads on the portfolios. In fact, we could also get the weights on the individual stocks, but the interpretation is easier for the characteristic-based portfolios.

The conditional factor structure induces a singularity in the covariance matrix of the portfolio returns ξ_t. This redundancy in the portfolio returns implies that the weight vector generating the latent space direction $b'f_t$ is not unique. Indeed, if Q_{t-1} is a $K \times (K - k_t)$ matrix whose columns span the orthogonal complement of the range of J_{t-1}, then we have $(J_{t-1}b_{t-1})'(\xi_t - \mathbb{E}[\xi_t|\mathscr{F}_{t-1}]) = (J_{t-1}b_{t-1} + Q_{t-1}c_{t-1})'(\xi_t - \mathbb{E}[\xi_t|\mathscr{F}_{t-1}])$ for any $c_{t-1} \in \mathbb{R}^{K-k_t}$. Then, any vector $w_{t-1} = J_{t-1}b_{t-1} + Q_{t-1}c_{t-1}$ yields valid weights. Equivalently, $J_{t-1}'w_{t-1} = b_{t-1}$. Then, we can search for such vector w_{t-1} that is sparse. This suggests the L^1-penalized least square problem

$$\min_{w_{t-1} \in \mathbb{R}^K} (b_{t-1} - J_{t-1}'w_{t-1})'(b_{t-1} - J_{t-1}'w_{t-1}) + \alpha |w_{t-1}|_1 \tag{S.10}$$

for $\alpha > 0$. I.e., the L^1-penalized LS regression of outcome $b_{t-1} \in \mathbb{R}^{k_t}$ onto design matrix $J_{t-1}' \in \mathbb{R}^{k_t \times K}$.

We can impose some regularization for limiting the time variation. First, we can assume that the vector of weights is constant within a year. Then we define the vector of weights w_y for year y from

$$\min_{w_y \in \mathbb{R}^K} (b_y - J_y'w_y)'(b_y - J_y'w_y) + \alpha |w_y|_1 \tag{S.11}$$

where the $\bar{k}_y \times K$ matrix J_y' stacks vertically the matrices J_{t-1}' for the 12 months $t \in y$, with $\bar{k}_y := \sum_{t \in y} k_t$ and similarly b_y stacks vertically the b_{t-1} for $t \in y$. Second we can impose some Ridge penalization for differences $w_y - w_{y-1}$, i.e. we solve

$$\min_{\{w_y \in \mathbb{R}^K\}_{y=1,\ldots,N}} \sum_{y=1}^{N} (b_y - J_y'w_y)'(b_y - J_y'w_y) + \alpha \sum_{y=1}^{N} |w_y|_1 + \beta \sum_{y=2}^{N} (w_y - w_{y-1})'(w_y - w_{y-1}) \tag{S.12}$$

where α and β are the regularization parameters.

Let us finally characterize the solution b_t to the constrained maximization problem in (S.7). We use that the vector of the first r conditional canonical directions is $F_t^{(r)} = C_{t-1}^{(r)\prime} f_t$, where $C_{t-1}^{(r)}$ is the $k \times r$ matrix of the eigenvectors associated with the r largest eigenvalues

of matrix

$$\begin{aligned} & \mathbb{V}(f_t|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t, f_t^O|\mathscr{F}_{t-1})\mathbb{V}(f_t^O|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t^O, f_t|\mathscr{F}_{t-1}) \\ = \; & (\Lambda_{t-1})^{-1}J_{t-1}'\mathbb{C}\text{ov}(\xi_t, f_t^O|\mathscr{F}_{t-1})\mathbb{V}(f_t^O|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t^O, \xi_t|\mathscr{F}_{t-1})J_{t-1}, \end{aligned}$$

with the normalization $(C_{t-1}^{(r)})'\Lambda_{t-1}C_{t-1}^{(r)} = I_r$. [17] Further, using $\mathbb{V}(b'f_t|\mathscr{F}_{t-1}) = b'\Lambda_{t-1}b$ and $\mathbb{C}\text{ov}(b'f_t, F_t^{(r)}|\mathscr{F}_{t-1}) = b'\Lambda_{t-1}C_{t-1}^{(r)} =: b'B_{t-1}$, problem (S.7) is equivalent to:

$$\max_{b\in\mathbb{R}^k} b'\Lambda_{t-1}b, \qquad \text{s.t.} \quad b'B_{t-1} = 0, \; b'b = 1. \tag{S.13}$$

The solution of (S.13) yields that b_{t-1} is the standardized eigenvector associated with the largest eigenvalue of matrix $M_{B_{t-1}}\Lambda_{t-1}M_{B_{t-1}}$.

S.4.2 Implementation

This section summarizes the steps of the procedure highlighted above for the purpose of implementation. We write the procedure in terms of population quantities for convenience of notation, the implementation involves the corresponding estimated quantities.

Step 1: Recover Latent Factor Covariance Structure

Given the conditional variance of the K-dimensional vector of characteristic-based portfolio returns,

$$\mathbb{V}(\xi_t|\mathscr{F}_{t-1}) = J_{t-1}\Lambda_{t-1}J_{t-1}',$$

we obtain J_{t-1} and Λ_{t-1} by eigen-decomposition of the symmetric matrix $\mathbb{V}(\xi_t|\mathscr{F}_{t-1})$ using the first k_t non-zero eigenvalues. Time-varying integer k_t is selected with the procedure introduced in the paper, equation (4.2).

Step 2: Compute Conditional Canonical Directions

Take f_t^O to be the vector of observable factors, i.e. FF6 for Figure 4. Using the estimated $\mathbb{V}(f_t^O|\mathscr{F}_{t-1})$, $\mathbb{C}\text{ov}(\xi_t, f_t^O|\mathscr{F}_{t-1})$ and (J_{t-1}, Λ_{t-1}), form the matrix

$$\mathcal{R}_{t-1} = (\Lambda_{t-1})^{-1/2}J_{t-1}'\,\mathbb{C}\text{ov}(\xi_t, f_t^O|\mathscr{F}_{t-1})\,\mathbb{V}(f_t^O|\mathscr{F}_{t-1})^{-1}\,\mathbb{C}\text{ov}(f_t^O, \xi_t|\mathscr{F}_{t-1})\,J_{t-1}(\Lambda_{t-1})^{-1/2}.$$

[17] i.e., $C_{t-1}^{(r)} = (\Lambda_{t-1})^{-1/2}\tilde{C}_{t-1}^{(r)}$, where $\tilde{C}_{t-1}^{(r)}$ is the $k \times r$ matrix with unit-length columns that are the standardized eigenvectors for the r largest eigenvalues of symmetric matrix $(\Lambda_{t-1})^{-1/2}J_{t-1}'\mathbb{C}\text{ov}(\xi_t, f_t^O|\mathscr{F}_{t-1})\mathbb{V}(f_t^O|\mathscr{F}_{t-1})^{-1}\mathbb{C}\text{ov}(f_t^O, \xi_t|\mathscr{F}_{t-1})J_{t-1}(\Lambda_{t-1})^{-1/2}$.

Let $\tilde{C}_{t-1}^{(r)}$ be the $k_t \times r$ matrix with unit-length columns that are the standardized eigenvectors for the r largest eigenvalues of symmetric matrix $\mathcal{R}_{t-1}$. Obtain matrix $C_{t-1}^{(r)} = (\Lambda_{t-1})^{-1/2}\tilde{C}_{t-1}^{(r)}$, whose columns represent the conditional canonical directions, while the eigenvalues correspond to the *squared* canonical correlations between latent and observable factors. We use $r = 3$.

Step 3: Estimate Poorly Spanned Latent Directions

Define $B_{t-1} = \Lambda_{t-1}C_{t-1}^{(r)}$ and compute the projection matrix onto the orthogonal complement of B_{t-1}:

$$M_{B_{t-1}} = I_{k_t} - B_{t-1}(B_{t-1}'B_{t-1})^{-1}B_{t-1}'.$$

The unspanned latent direction b_{t-1} with maximal conditional variance is the leading eigenvector of $M_{B_{t-1}}\Lambda_{t-1}M_{B_{t-1}}$.

Step 4: Compute Portfolio Weights

Get the weights w_y for $y = 1, ..., N$ by solving the Ridge-Lasso regularized problem (S.12).

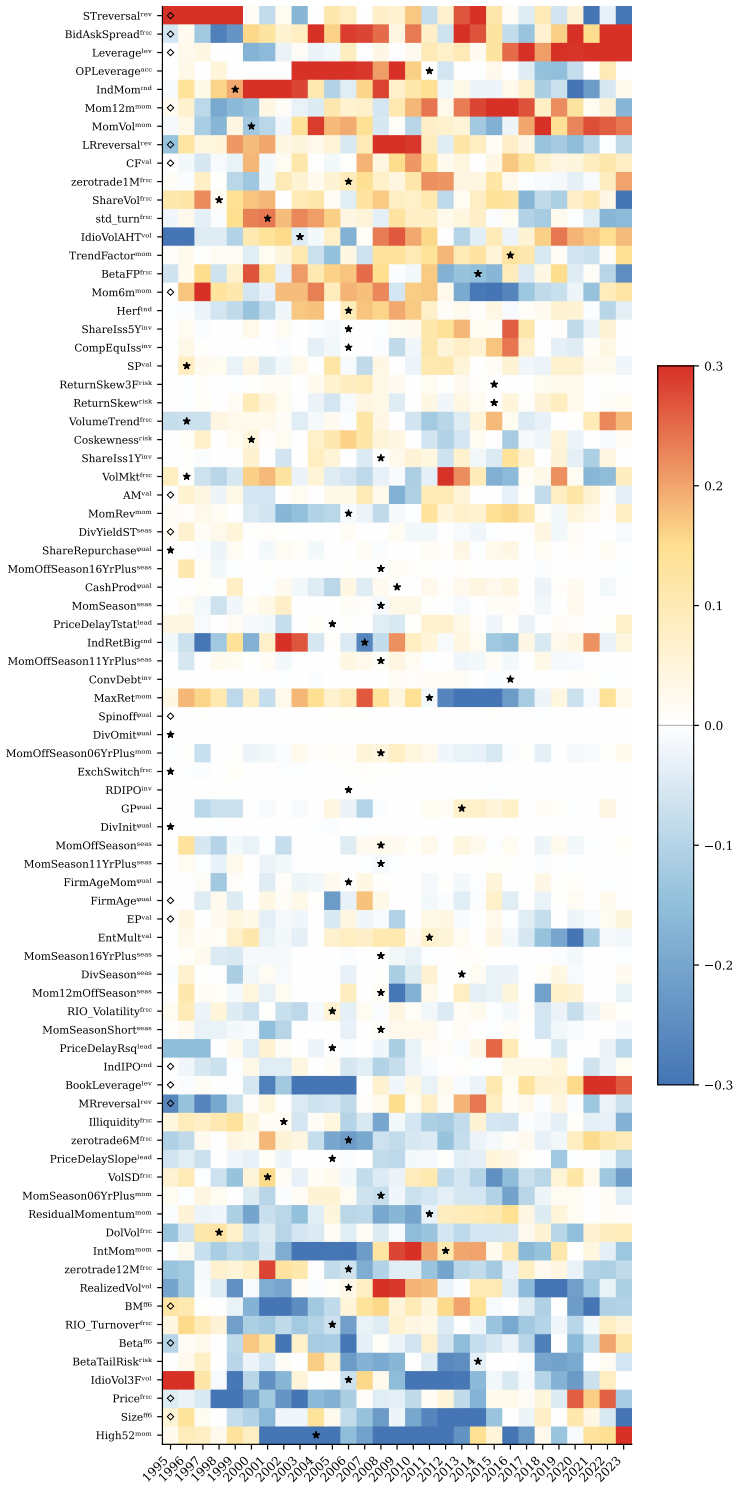

Figure S.2: Annual average median characteristic weights for latent factor directions not spanned by the Fama–French six factors. Colors indicate the sign and magnitude of the estimated weights by characteristic and year. Superscripts denote economic categories: rev (reversal), ind (industry), acc (accruals), fric (trading frictions), val (value), mom (momentum), and lev (leverage). Stars mark publication years; diamonds at 1995 indicate pre-sample publication. Persistent positive exposure is concentrated in friction- and leverage-related characteristics.

Additional References

Campbell, J. Y., and S. B. Thompson (2008). Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? *Review of Financial Studies*, 21(4), 1509–1531.

Goyal, A. and Welch, I. (2008). A Comprehensive Look at the Empirical Performance of Equity Premium Prediction. *The Review of Financial Studies*, 21(4), 1455-1508.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.