Hansen-Jagannathan distance with many assets*

Marine Carrasco† and Cheikh Nokho
Université de Montréal

November 2026

Abstract

This paper studies the evaluation and comparison of asset pricing models in settings with a large number of test assets. Models are specified through a linear stochastic discount factor (SDF). We propose two regularization schemes to extend the Hansen-Jagannathan distance to high-dimensional environments. In addition to stabilizing the inversion of the covariance matrix, the proposed regularizations admit an economic interpretation as relaxing the exact pricing restrictions, thereby accommodating market frictions.

We derive the asymptotic properties of the SDF parameter estimator under a double asymptotic framework in which both the cross-sectional and time dimensions grow. These results allow for inference on individual factor coefficients. We further develop tests for comparing competing asset pricing models under misspecification, providing a formal procedure to identify the least misspecified model. The analysis covers both nested and non-nested specifications.

Monte Carlo simulations show that Tikhonov regularization substantially reduces the severe size distortions of the Moore–Penrose inverse near $N = T$ while retaining good power.

Keywords: asset pricing models, misspecification, ridge, regularization, tests, Tikhonov, stochastic discount factor.

JEL classification: C12, G12

*Carrasco gratefully acknowledges partial financial support from FRQSC and SSHRC. The authors thank the editor Victor Todorov, the associate editor, and two referees for numerous insightful suggestions. We also thank Bertille Antoine, Andrii Babii, Patrick Gagliardini, René Garcia, Eric Ghysels, Christian Gourieroux, Joann Jasiak, Peter Hansen, Ilze Kalnina, Yuan Liao, Denis Pelletier, Benoit Perron, Enrique Sentana, Fabio Trojani, Dacheng Xiu, the participants at NBER-NSF Time-Series conference (2023), CFE-CM Statistics conference (2023), FinEML conference (2024) and seminar participants at UNC Chapel Hill, NC State University, Washington University in St. Louis, and CEMFI for helpful comments. We thank Stella Kontchou Takoukam for excellent research assistance.

†Address: Université de Montréal, Departement de Sciences Economiques, CP 6128, succ Centre Ville, Montreal, QC H3C3J7, Canada. Email: marine.carrasco@umontreal.ca.

1 Introduction

Asset pricing models are typically represented through a parametric stochastic discount factor (SDF), often specified as a linear function of a small number of risk factors. Given the large number of competing models proposed in the literature, it is essential to compare them using an objective and economically meaningful criterion.

A well-known measure of model misspecification is the Hansen-Jagannathan (HJ) distance, which measures the distance between a candidate pricing kernel and the closest valid one (see Hansen and Jagannathan (1997)). This distance is closely related to the Generalized Method of Moments (GMM) criterion, with the key difference that the weighting matrix is given by the inverse second moment matrix of the returns. Empirically, specification tests based on HJ distance are often rejected (Hodrick and Zhang (2001); Ludvigson (2013)), suggesting that most asset pricing models are misspecified. Consequently, rather than testing whether a model is exactly correct, a more relevant objective is to compare competing models and determine which one is less misspecified. In this context, the HJ distance provides a natural tool to rank models and assess whether individual SDF parameters are zero.

A central challenge in this comparison is that many models appear to perform similarly when evaluated on standard sets of test assets, such as the 25 size and book-to-market portfolios of Fama and French (1992). As emphasized by Daniel and Titman (2012) and Lewellen, Nagel, and Shanken (2010), these portfolios span a limited dimension of the return space and exhibit strong covariance structures, making it difficult to discriminate between models. This has led to the view that a larger and more diverse set of test assets is needed. For instance, Kan and Robotti (2009) augment the cross-section with industry portfolios but still find limited power, while Barras (2019) advocate using a very large number of micro portfolios. More generally, increasing the number of test assets improves the ability to detect misspecification, as the unconditional HJ distance approaches its conditional counterpart when the cross-section becomes sufficiently rich.

However, expanding the number of test assets raises some technical issues. The HJ distance relies on the inverse covariance matrix of returns, which becomes ill-conditioned when the number of assets is large relative to the time dimension. In practice, the sample covariance matrix is nearly singular or non-invertible when N is comparable to or exceeds T (Cochrane (2005)). As a result, the implied weighting matrix is unstable and can place excessive weight on directions associated with small eigenvalues, leading to economically meaningless portfolios and unreliable inference. This issue becomes particularly severe when individual stock returns or very large sets of portfolios are used as test assets.

This paper studies the evaluation and comparison of asset pricing models in such high-dimensional settings, explicitly allowing for model misspecification. Our approach is based on regularizing the HJ distance to stabilize the inversion of the covariance matrix. Specifically, we consider ridge and Tikhonov regularizations of the inverse covariance matrix (see Carrasco, Florens, and Renault (2007)). These regularizations yield a modified distance that can be interpreted as allowing for controlled pricing errors, which is consistent with the presence of transaction costs, liquidity fric-

tions, or measurement error. Importantly, our objective is not to test whether the HJ distance is equal to zero, but rather to compare models through differences of (regularized) HJ distances, which remain well-defined even when the cross-section is large.

Our contributions are threefold. First, we introduce regularized versions of the HJ distance that remain well-behaved when the number of test assets is large, and we propose a data-driven procedure to select the regularization parameter. Second, we derive the asymptotic distribution of the SDF parameters obtained by minimizing the regularized distance under a double asymptotic framework where both N and T diverge. We allow for serial dependence and propose a consistent HAC estimator of the asymptotic variance. We show that these estimators are asymptotically normal and can be used to test SDF coefficients. Third, we develop inference procedures for comparing competing models based on differences in regularized HJ distances, treating separately the cases of nested and non-nested models. These results provide a formal basis for ranking models in environments where all models are potentially misspecified.

Monte Carlo simulations compare the power of the tests based on Moore-Penrose inverse with those based on Tikhonov regularization. They show that Tikhonov regularization substantially mitigates the severe size distortions of the Moore–Penrose inverse near $N = T$, while retaining comparable power in many other designs. Finite-sample accuracy nevertheless varies across test statistics: the equal-SDF tests can overreject, whereas the distance-comparison tests tend to be conservative. In an empirical application, we compare several prominent SDF specifications, including the Fama–French three- and five-factor models, the nonlinear model of Dittmar (2002), and the q5 model of Hou, Mo, Xue, and Zhang (2021). Our results indicate that q5 achieves the lowest regularized HJ distance under a common weighting matrix, although its advantage over FF5 is significant only at the 10% level.

Our work relates to several strands of the literature at the intersection of asset pricing model evaluation and machine learning in finance. Several papers proposed methods to examine asset pricing misspecification (Hansen and Jagannathan (1997); Almeida and Garcia (2012); Almeida and Garcia (2017)). This paper is close to Kan and Robotti (2008) and Kan and Robotti (2009) who derived asymptotic distribution of the SDF parameter and model comparison methods using the HJ-distance under a misspecified setting using a finite number of test assets. Gagliardini and Ronchetti (2020) compare models by conditional Hansen-Jagannathan distances which involves a kernel smoothing step, see also Antoine, Proulx, and Renault (2020). Unlike their approach, we employ the unconditional version of the HJ-distance with many assets. However, when a very large number of test assets is used, the unconditional HJ distance should be close to the conditional one. Several papers also propose methods to either stabilize or improve the estimation of covariance matrices (Brodie, Daubechies, De Mol, and Loris (2009); Carrasco and Rossi (2016); Carrasco, Kone, and Noumon (2019); Ledoit and Wolf (2003); Ledoit and Wolf (2017); D′Hondt, De Winne, Ghysels, and Raymond (2020); Qiu and Otsu (2022); Lönn and Schotman (2024)). Ren and Shimotsu (2009) illustrate via simulations that shrinkage improves the finite sample properties of Hansen-Jagannathan distance. More recently, Chernov, Kelly, Malamud, and Schwab (2025) develop a test of portfolio efficiency in high-dimensional settings using ridge regularization and

derive its asymptotic distribution accounting for regularization bias. Our approach is also related to Korsaye, Quaini, and Trojani (2025), who propose a strictly positive SDF that tolerates pricing errors; in contrast, we relax pricing restrictions without imposing positivity. Barillas and Shanken (2018) study model comparison and show that test assets are irrelevant when only traded factors are used, but remain essential with non-traded factors, a setting we consider. While recent work focuses on high-dimensional factor models (Kozak, Nagel, and Santosh (2020), Didisheim, Ke, Kelly, and Malamud (2024)), we instead compare low-dimensional SDF specifications using a large cross-section of test assets. Finally, our asymptotic results relate to the literature on GMM under misspecification (Hall and Inoue (2003) and Hall and Pelletier (2011)).

The remainder of the paper is organized as follows. Section 2 introduces the framework under which we evaluate models and discusses the issues arising from the weighting matrix. Section 3 presents the regularization methods and derives the asymptotic properties of the SDF parameter estimators. Section 4 develops inference procedure for model comparison. Section 5 reports simulation results and Section 6 compares four empirical asset pricing models using a dataset of 647 test portfolios. Finally, Section 7 concludes. Appendices A to C give some extra details. The proofs are collected in the online appendix.

2 Asset pricing model under misspecification

We believe most asset pricing models are misspecified. We want to compare models and determine which one is the less misspecified. To do this, we compare the differences between Hansen-Jagannathan distances.

2.1 Pricing errors and model specification using excess returns

Let r_t be the excess returns of N assets. Given the availability of K factors f_t, the estimation of asset pricing models can be summarized in finding the expression of the relevant stochastic discount factor y_t. The latter is admissible if it correctly prices the test assets r_t [1]:

$$E[r_t.y_t] = 0.$$

Define $Y_t = [f_t', r_t']'$. Its mean and covariance matrix are given by $\mu = E[Y_t] = [\mu_1', \mu_2']'$ and $V = V(Y_t) = \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}$. We define also the centered variables $\tilde{r}_t = r_t - \mu_2$ and $\tilde{f}_t = f_t - \mu_1$. In this paper, we focus on linear candidate SDF, $y_t(\theta) = 1 - \tilde{f}_t'\theta$. It is common to choose θ by minimizing the aggregate pricing errors $\alpha(\theta) = E[r_t.y_t(\theta)] = \mu_2 - V_{21}\theta$ via

[1] Note that the no-arbitrage constraint, $y_t \geq 0$, is not imposed. Gospodinov, Kan, and Robotti (2016) analyze the benefits and costs of using the HJ-distance with a no-arbitrage constraint and found that preventing the SDF from taking negative values may lower its pricing ability and may be problematic when comparing competing models. Therefore, we focus our attention to unconstrained HJ-distance.

$$\delta_W^2 = \alpha(\theta)' W \alpha(\theta), \tag{1}$$

where W is a positive-definite matrix.

The candidate SDF prices correctly the returns, when one can find θ such that $\delta_W^2(\theta) = 0$. Otherwise, the model is considered globally misspecified. The reason for demeaning the factors is to ensure the invariance to affine transformations of the factors, as explained in Kan and Robotti (2008).

When models are misspecified, the SDF parameter, that minimizes (1), depends on the weighting matrix. Therefore, its choice is paramount.

One possibility is to use the GMM framework. In this case, W is the inverse of the asymptotic variance of $\hat{E}[r_t.y_t(\theta)] = \sum_t r_t y_t(\theta)/T$ and the objective function (1) equates the over-identification test of Hansen (1996). However, using this weighting matrix to compare asset pricing models may be misleading for several reasons. First, it has been shown that this diagnostic is model-dependent and tends to reward models with volatile SDF and pricing errors (Ludvigson (2013)). Second, it tends to produce huge leverage portfolios as W is near singular with many assets (Cochrane (1996)). Alternatively, one could use the identity matrix to circumvent the invertibility issue. Nonetheless, the models estimated will depend on the assets included. This setting is not preferable for researchers looking for results independent of particular dataset.

As established in Kan and Robotti (2008), the use of V_{22}^{-1} as weighting matrix enables the HJ-distance to be model-independent and suitable for asset pricing model comparison. We adopt this approach.

When $W = V_{22}^{-1}$, δ_W^2 is a modified Hansen and Jagannathan (1997) distance, where the mean of the SDF is constrained to 1. Let

$$\delta_{V_{22}}^2 = \delta^2 = (\mu_2 - V_{21}\theta)' V_{22}^{-1} (\mu_2 - V_{21}\theta). \tag{2}$$

We define θ_{HJ} as the solution to the minimization of (2).

$$\theta_{HJ} = \arg\min_{\theta} \delta^2 = (V_{12} V_{22}^{-1} V_{21})^{-1} V_{12} V_{22}^{-1} \mu_2.$$

θ_{HJ} is a pseudo-true value, which coincides with the true θ when the model is correctly specified. It can also be written as $V_{11}^{-1}(\beta' V_{22}^{-1}\beta)^{-1}\beta' V_{22}^{-1}\mu_2 = V_{11}^{-1}\lambda$ where $\beta = V_{21}V_{11}^{-1}$ is the exposure of the returns to the factors f_t and $\lambda = (\beta' V_{22}^{-1}\beta)^{-1}\beta' V_{22}^{-1}\mu_2$ represents the pseudo risk premium[2]. This particular form shows that the SDF parameter can also be estimated via the βs. Such representation is not new as a well-known equivalence between SDF representation, beta-representation and minimum-variance efficiency has already been established (see p.261 of Cochrane (2005), chapter 7 of Ferson (2019) or Goyal (2012)). In this setting, the asset pricing model is misspecified when $\alpha = \mu_2 - V_{21}V_{11}^{-1}\lambda = \mu_2 - \beta\lambda \neq 0$.

[2]The pseudo risk premium coincides with the risk premium when $\alpha = 0$.

We introduce the model in the following assumption.

Assumption 1. *(i)*

$$r_t = \alpha + \beta(\tilde{f}_t + \lambda) + \epsilon_t, \tag{3}$$

where β is a matrix $N \times K$, $\alpha \in \mathbb{R}^N$, $\lambda \in \mathbb{R}^K$, the $N \times 1$ error terms ϵ_t are assumed uncorrelated with the factors. In addition, the errors have mean 0 and variance $V(\epsilon_t \mid f_t) = \Sigma_\epsilon = [\sigma_{i,j}]_{i,j=1,\cdots,N}$ where $\sigma_{i,j} = E[\epsilon_{it}\epsilon_{jt}]$. Moreover, $tr(\Sigma_\epsilon)/N = O(1)$.

(ii) $\frac{1}{N}\sum_{i=1}^{N} \beta_i'\beta_i \to \Sigma_\beta$, *as* $N \to \infty$, *where Σ_β is positive-definite matrix.*

(iii) V_{11} and V_{22} are non singular.

Assumption 1(i) relaxes the approximate factor model of Chamberlain and Rothschild (1983). Indeed, Chamberlain and Rothschild (1983) assume that the eigenvalues of Σ_ϵ are uniformly bounded. The condition $tr(\Sigma_\epsilon)/N = O(1)$ is weaker and allows for the largest eigenvalue to diverge. Diverging largest eigenvalue appears in Gagliardini, Ossola, and Scaillet (2016) and Raponi, Robotti, and Zaffaroni (2020). This is particularly relevant when comparing non-nested models when the DGP includes all factors present in the competing models. In this case, each competing model has missing strong factors. Note also that Assumption 1(i) allows for Σ_ϵ to be singular; this will arise if some assets are spanned by the factors and is consistent with tradable factors.

Assumption 1(ii) is the same as Assumption 2 of Raponi, Robotti, and Zaffaroni (2020). Positive-definite Σ_β excludes spurious factors and cross sectionally constant β_i. Also, this assumption implies that $\beta_k'\beta_k/N < \infty$, $k = 1, \cdots, K$. The factors included in the SDF need to be sufficiently correlated with the returns. However, there could be some weak factors hidden in the error term. This is consistent with what is discussed in Pesaran and Smith (2023): "The idiosyncratic errors, although uncorrelated with $\tilde{f}_t$, could be cross-sectionally correlated due to non-fundamental common factors uncorrelated with the SDF that arise from herding behavior or correlated beliefs for instance at times of crisis." Several authors have warned that using weak factors could yield incorrect econometric inference, see Kleibergen and Zhan (2020) and Manresa, Panaranda, and Sentana (2023) and references therein. Pesaran and Smith (2021) argue that only strong factors should be used to predict returns. So, we stick to the situations where the factors $\tilde{f}_t$ are all strong even though here the aim is not to estimate β but to compare different asset pricing models.

Assumption 1(iii) implies that, in population, none of the factors and none of the assets are redundant. However, in finite sample, $\widehat{V}_{22}$ may be close to singular when N approaches T and will be necessarily singular when $N > T$.

Note that the error term ϵ_t may be serially correlated and the intercept α_i may vary with the asset i.

2.2 Overview of the approach and main results

The covariance matrix of the returns, V_{22}, is often nearly singular when the number of assets N is large, due to the strong cross-sectional dependence induced by common factors. This near-singularity deteriorates the finite-sample properties of both the SDF estimator and the associated misspecification tests. More formally, under a factor structure, the largest eigenvalue of V_{22} is typically of order N, while the smallest eigenvalues remain bounded.

For this reason, it is convenient to work with the normalized covariance matrix $\Sigma = V_{22}/N$, and to express the SDF parameter as

$$\theta_{HJ} = V_{11}^{-1}(\beta'\Sigma^{-1}\beta)^{-1}\beta'\Sigma^{-1}\mu_2$$

While the matrix Σ has bounded largest eigenvalue, its smallest eigenvalues accumulate toward zero as $N \to \infty$. As a result, its condition number $\kappa = \lambda_1/\lambda_N$ diverges, making inversion unstable.

To illustrate this instability, consider the equation

$$\Sigma\varphi = \beta \tag{4}$$

where $\beta \in \mathbb{R}^N$ is given. Let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_N > 0$ denote the eigenvalues of Σ with corresponding orthonormal eigenvectors ϕ_j. The solution of (4) is

$$\varphi = \Sigma^{-1}\beta = \sum_{j=1}^{N} \frac{1}{\lambda_j}\left(\phi_j'\beta\right)\phi_j.$$

Now perturb β by $\tilde{\beta} = \beta + c\phi_N$, where c is scalar. The solution becomes $\tilde{\varphi} = \Sigma^{-1}\tilde{\beta} = \varphi + \dfrac{c}{\lambda_N}\phi_N$. Hence,

$$\|\tilde{\varphi} - \varphi\| \,/\, \left\|\tilde{\beta} - \beta\right\| = 1/\lambda_N = \kappa/\lambda_1$$

which diverges when N goes to infinity. Small perturbations in β are therefore greatly amplified along directions associated with small eigenvalues. This instability arises even when Σ is known, and becomes more severe when it is estimated. It explains why standard HJ-based inference becomes unreliable when the number of assets is large.

To restore stability, we introduce Tikhonov regularization. The regularized solution to (4) is

$$\varphi^{\gamma} = \left(\Sigma^2 + \gamma I\right)^{-1}\Sigma\beta = \sum_{j=1}^{N} \frac{\lambda_j}{\lambda_j^2 + \gamma}\left(\phi_j'\beta\right)\phi_j$$

where $\gamma > 0$ is a regularization parameter. Unlike the Moore-Penrose inverse (obtained as γ goes

to zero), this mapping is continuous. Indeed for a perturbation $\Delta\beta$,

$$\|\tilde{\varphi}^{\gamma} - \varphi\| = \left\| \sum_{j=1}^{N} \frac{\lambda_j}{\lambda_j^2 + \gamma} \left(\phi_j' \Delta\beta\right) \phi_j \right\| \leq \sup_j \frac{\lambda_j}{\lambda_j^2 + \gamma} \|\Delta\beta\| \leq \frac{1}{2\sqrt{\gamma}} \|\Delta\beta\|.$$

Continuity is necessary to establish the convergence of our estimators. Another interesting property of Tikhonov is that the eigenvectors associated with the 0 eigenvalue vanish from the sum.

Building on results from the inverse problem literature (Kress (2014), Carrasco, Florens, and Renault (2007)), we show that $\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\hat{\beta}/N$ and $\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\hat{\mu}_2/N$ are consistent, where $\hat{\Sigma}_{\gamma}^{-1}$ is the regularized inverse of Σ and γ goes to zero at an appropriate rate. It follows that the regularized estimator $\hat{\theta}_{HJ}^{\gamma}$ is consistent for the pseudo-true value θ_{HJ} and $\sqrt{T}\left(\hat{\theta}_{HJ}^{\gamma} - \theta_{HJ}\right)$ is asymptotically normal provided $\gamma T \to \infty$ and $\gamma^2 T \to 0$ (see Proposition 2). The absence of N from these rate conditions results from our normalization of the covariance matrix. We regularize $\Sigma = V_{22}/N$ rather than V_{22} itself, thereby absorbing the dependence on N into the tuning parameter. Indeed, if $\tilde{\gamma}$ denotes the Tikhonov regularization parameter applied directly to V_{22}, while γ denotes that applied to Σ, then $\tilde{\gamma} = N^2\gamma$. Hence, the N-free conditions $\gamma T \to \infty$ and $\gamma^2 T \to 0$ are equivalent to $\tilde{\gamma}T/N^2 \to \infty$ and $\tilde{\gamma}^2 T/N^4 \to 0$ which explicitly display the dependence on N. The asymptotic normality of the estimator is used to test the nullity of individual SDF coefficients. In practice, implementation requires the estimation of variance terms, which we achieve using regularized HAC estimators. We establish their consistency in this high-dimensional framework.

We next discuss the behavior of the regularized Hansen–Jagannathan distance $\hat{\delta}_{\gamma}^2$. It has been shown (see Chernov, Kelly, Malamud, and Schwab (2025)) that $T\hat{\delta}_{\gamma}^2$ tends to diverge as N increases, requiring centering and rescaling for valid inference. However, our objective is not to test whether a model is correctly specified ($\delta^2 = 0$), but rather to compare competing misspecified models.

A key insight of this paper is that differences of regularized HJ distances remain well behaved even when each distance may diverge individually. For a model m, the minimized criterion satisfies

$$\hat{\delta}_m^2 = \frac{\hat{\mu}_2' \hat{\Sigma}_{\gamma}^{-1} \hat{\mu}_2}{N} - \frac{1}{N} \widehat{\theta}_m' \hat{V}_{m,21}' \hat{\Sigma}_{\gamma}^{-1} \hat{V}_{m,21} \widehat{\theta}_m.$$

Provided that the models use the same test assets, observations, and regularization parameter, the first term is identical across models and therefore cancels exactly when their estimated squared HJ distances are subtracted. The comparison thus depends only on the difference between the model-specific fitted components. This cancellation helps explain why inference on differences can be more tractable than inference on the level of an individual HJ distance; the precise limiting distribution depends on the relationship between the competing models.

For nested models, the connection with inference on SDF coefficients is particularly direct. At

the population level, if Model a is nested in Model b, then

$$\delta_a^2 - \delta_b^2 = \theta_2' C_{22}^{-1} \theta_2,$$

where θ_2 contains the coefficients of the additional factors in Model b. So testing the equality of HJ distances is equivalent to testing $H_0 : \theta_2 = 0$. The test takes the form

$$T\left(\hat{\delta}_{a,\gamma}^2 - \hat{\delta}_{b,\gamma}^2\right) = T\widehat{\theta}_{2,\gamma}' \hat{C}_{22,\gamma}^{-1} \widehat{\theta}_{2,\gamma}.$$

Under the null hypothesis, $\sqrt{T}\widehat{\theta}_{2,\gamma} = O_p(1)$, so this statistic remains $O_p(1)$ and converges to a weighted sum of chi-square variables as shown in Proposition 4. Thus, the asymptotic is driven by the estimation error of the finite-dimensional SDF coefficients, rather than by the dimension N of the test-asset space.

For nonnested misspecified models with distinct SDF, $y_a \neq y_b$, the same common-term cancellation applies but the convergence rate is different. Given the population HJ distance is nonzero, each estimated regularized HJ distance itself satisfies a $\sqrt{T}$-central limit theorem, so their difference has the normal limit established in Proposition 7.

Overall, regularization stabilizes the estimation of the SDF parameters, while differencing eliminates the common high-dimensional component of the HJ distances and yields well-behaved comparison statistics.

3 Regularized SDF parameter estimator

As stated earlier, inference using the modified HJ-distance may not be robust to a large number of correlated securities that makes the weighting matrix near singular or singular. Relying on the literature on inverse problems in an infinite dimensional space (see Kress (2014) and Carrasco, Florens, and Renault (2007)), we introduce two regularization methods to stabilize the weighting matrix and improve the estimation of asset pricing models.

3.1 Economic and statistical interpretations of regularization

The following lemma sheds light on the properties of the matrix $\Sigma = V_{22}/N$.

Lemma 1. *Under Assumptions 1, assuming (ϵ_t, f_t) is stationary and $\alpha'\alpha/N = O(1)$, we have the following results as $N \to \infty$,*

1. $tr(\Sigma) = O(1)$.
2. $E[r_t' r_t/N] = O(1)$.

Lemma 1 shows that the normalized covariance matrix Σ has bounded trace when N increases. As a consequence, Σ behaves like a compact operator in the limit, in the sense that its eigenvalues are bounded but accumulate at zero. This implies that the inverse of Σ, when it exists, is not a

continuous operator and the associated inverse problem is ill-posed. In particular, small perturbations in the data can lead to large variations in Σ^{-1}, making estimation unstable. Regularization therefore becomes necessary to stabilize the inversion and obtain well-defined estimators. For this reason, it is more natural to regularize Σ rather than V_{22}, as this allows us to directly exploit results from the inverse problem literature.

Let $\hat{\Sigma}$ be the sample counterpart of Σ. Let $\gamma > 0$ be a regularization parameter. We consider two techniques that consist of replacing the singular or nearly singular matrix $\hat{\Sigma}$ by a well-conditioned matrix before inverting the matrix. These two regularization schemes give the following inverses:

Ridge regularization

$$\hat{\Sigma}_\gamma^{-1} = (\hat{\Sigma} + \gamma I_N)^{-1}.$$

Tikhonov regularization

$$\hat{\Sigma}_\gamma^{-1} = (\hat{\Sigma}^2 + \gamma I_N)^{-1}\hat{\Sigma}.$$

Appendix A gives some further details on these regularizations. The regularization parameter γ plays a crucial role in stabilizing the inverse of the matrix Σ. For small γ, the regularized inverse will be close to the actual inverse while being much more stable. In practice, the tuning parameter γ is chosen to go to zero with the sample size. Its choice is discussed later.

We focus on ridge and Tikhonov regularizations for their favorable properties and clear interpretation. While alternative approaches exist—such as nonlinear shrinkage (Ledoit and Wolf (2017)), principal components, Landweber–Fridman regularization (Carrasco, Kone, and Noumon (2019)), thresholding (Fan, Liao, and Mincheva (2013)), graphical lasso (Callot, Caner, Onder, and Ulasan (2021)), or L_2-boosting (Lönn and Schotman (2024))—many rely on sparsity assumptions on Σ or Σ^{-1}, which are unlikely to hold in asset returns. In contrast, ridge and Tikhonov regularizations are well suited to settings with strong cross-sectional dependence, as in our application.

Below, we provide an interpretation of the regularized HJ-distance in terms of a penalization on the norm of the pricing error.

Let L^2 be the set of random variables X such that $E\left(X^2\right) < \infty$. Suppose the candidate SDF y_t belongs to L^2. First, recall that, when N is fixed, Kan and Robotti (2008) showed that δ^2 (not regularized) represents the distance between the proposed SDF y_t and the set of admissible SDFs of mean 1 in L^2.

$$\delta^2 = \min_{m_t \in L^2, E[m_t]=1} E\left(m_t - y_t\right)^2 \quad \text{subject to } E[m_t r_t] = 0. \tag{5}$$

How does this result change when one introduces regularization? Let

$$\begin{aligned} \delta_R^2 &= \alpha'(\Sigma + \gamma I_N)^{-1}\alpha/N, \\ \delta_K^2 &= \alpha'(\Sigma^2 + \gamma I_N)^{-1}\Sigma\alpha/N \end{aligned}$$

be respectively the HJ-distance obtained by Ridge and Tikhonov regularizations. Below, we are going to show that these regularized HJ-distances measure how far y is to the closest valid SDF of mean 1 which prices returns with an error controlled by γ. We use the notation $\|v\|_N^2 = v'v/N$.

Proposition 1. *The following results hold:*

1. For ridge,

$$\delta_R^2 = \inf_{m \in L^2, E[m]=1} E[(m-y)^2] + \frac{1}{\gamma} \parallel E[mr] \parallel_N^2, \tag{6}$$

2. For Tikhonov,

$$\delta_K^2 = \inf_{m \in L^2, E[m]=1} E[(m-y)^2] + \frac{1}{\gamma} \parallel E[mr] \parallel_{N,\Sigma}^2, \tag{7}$$

where $\parallel x \parallel_{N,\Sigma}^2 = \dfrac{x'\Sigma x}{N}$ *for any* $x \in \mathbb{R}^N$.

The previous proposition shows that regularization is equivalent to relaxing the constraint of problem (5). Low values of γ put the emphasis on the asset pricing equation, while high values allow for possible errors in the pricing of assets. Note that the optimal m from (6) also solves

$$\inf_{m \in L^2, E[m]=1} E[(m-y)^2] \text{ s.t. } \parallel E[mr] \parallel_N^2 \leq \tau$$

for an appropriately chosen τ related to γ. Similarly, the optimal m from (7) solves

$$\inf_{m \in L^2, E[m]=1} E[(m-y)^2] \text{ s.t. } \parallel E[mr] \parallel_{N,\Sigma}^2 \leq \tau$$

for an appropriately chosen τ related to γ.

In Appendix B, we also provide an interpretation of regularization as a penalization on the Lagrange multipliers. Note that when $N \to \infty$, duality breaks down if no regularization is applied, see Borwein (1993).

Now we examine how Tikhonov regularization affects the constraint. Since directions associated with small eigenvalues receive disproportionately large weights in the HJ distance, regularization effectively rebalances these contributions.

Let $P'\Lambda P$ be the spectral decomposition of Σ. Then the penalization term can be written as

$$\frac{1}{\gamma} \parallel E[mr] \parallel_{N,\Sigma}^2 = \frac{1}{\gamma N} E[mr]' P' \Lambda P E[mr] = \sum_{j=1}^{N} \omega_j E[m(Pr)_j]^2,$$

where $\omega_j = \dfrac{\lambda_j}{\gamma N}$ and $(Pr)_j$ denotes the jth principal component of returns. This representation shows that regularization entails the repackaging of the assets into N portfolios $(Pr)_j$ where directions associated with small eigenvalues receive less weight.

Korsaye, Quaini, and Trojani (2025) provide a related interpretation in terms of transaction costs. In their framework, pricing errors are penalized for "dubious" assets, leading to a modified SDF that tolerates deviations from exact pricing. In our setting, ridge and Tikhonov regularizations induce a quadratic penalization, which can be interpreted as transaction costs. Unlike their approach, however, we do not impose positivity of the SDF but allow for large N and large T asymptotics.

A Bayesian interpretation of the regularized inverse

We show that the Ridge-regularized inverse arises as the posterior mean of Σ^{-1} under a conjugate Bayesian specification, and that the regularization parameter γ has an explicit interpretation in terms of the prior hyperparameters and the sample size.

Let $x_t = (r_t - \mu_2)/\sqrt{N}$ denote the demeaned and rescaled excess returns, treating μ_2 as known, and suppose that, conditionally on Σ,

$$x_t \mid \Sigma \overset{iid}{\sim} N(0, \Sigma), \qquad t = 1, \ldots, T.$$

Assign Σ an inverse-Wishart prior,

$$\Sigma \sim \mathrm{IW}(\Psi_0, \nu_0),$$

with scale matrix $\Psi_0 = \tau I_N$, $\tau > 0$, and degrees of freedom $\nu_0 > N - 1$. The scaled-identity choice for Ψ_0 reflects a prior belief that, in absence of data, there is no information distinguishing one asset from another.

By Gaussian–inverse-Wishart conjugacy (e.g., Zellner (1971)), the posterior is

$$\Sigma \mid \text{data} \sim \mathrm{IW}(\Psi_0 + S, \nu_0 + T),$$

with $S = \sum_{t=1}^{T} x_t x_t'$. Since $\Sigma^{-1} \mid \text{data} \sim \text{Wishart}\left((\Psi_0 + S)^{-1}, \nu_0 + T\right)$, we have

$$E[\Sigma^{-1} \mid \text{data}] = (\nu_0 + T)(\Psi_0 + S)^{-1} = (\nu_0 + T)\left(\tau I_N + S\right)^{-1}.$$

The matrix $\Sigma = V_{22}/N$ can be estimated by its sample analog $\widetilde{\Sigma} = S/T$. Therefore,

$$\begin{aligned} E[\Sigma^{-1} \mid \text{data}] &= (\nu_0 + T)\left(\tau I_N + T\widetilde{\Sigma}\right)^{-1} = \frac{\nu_0 + T}{T}\left(\gamma I_N + \widetilde{\Sigma}\right)^{-1} \\ &= \frac{\nu_0 + T}{T}\widetilde{\Sigma}_\gamma^{-1} \end{aligned}$$

with $\gamma = \tau/T$ and $\widetilde{\Sigma}_\gamma^{-1} = (\widetilde{\Sigma} + \gamma I_N)^{-1}$ which is the ridge-regularized inverse.

Because the scalar coefficient $(\nu_0 + T)/T > 0$ does not affect the minimizer of the quadratic criterion $(\hat{\mu}_2 - \hat{V}_{21}\theta)' W (\hat{\mu}_2 - \hat{V}_{21}\theta)$ in θ, using $W = \mathbb{E}[\Sigma^{-1} \mid \text{data}]$ as the weighting matrix in (1)

yields our Ridge[3] estimator $\hat{\theta}_{HJ}^{\gamma}$ with regularization parameter $\gamma = \frac{\tau}{T}$. Thus, the regularization parameter is determined by the ratio of the prior scale parameter τ relative to the sample size.

Thus, the inverse-Wishart prior provides an exact Bayesian interpretation of the Ridge regularization. While this correspondence is not exact for the Tikhonov inverse, both regularization schemes attenuate the influence of directions associated with small eigenvalues and are closely related, as discussed in the online Appendix.

3.2 Asymptotic distribution of the regularized SDF parameter of misspecified models

Let $R = [r_1, \cdots, r_T]^{'}$ and $F = [f_1, \cdots, f_T]^{'}$ be respectively the $T \times N$ and $T \times K$ matrices of returns and factors. The OLS estimates of β is given by

$$\hat{\beta} = (\bar{R}^{'}\bar{F})(\bar{F}^{'}\bar{F})^{-1} = \hat{V}_{21}\hat{V}_{11}^{-1}$$

where $\bar{R} = R - 1_T\hat{\mu}_2^{'}$ and $\bar{F} = F - 1_T\hat{\mu}_1^{'}$. $\bar{R} = \begin{bmatrix} \bar{r}_1^{'} \\ \vdots \\ \bar{r}_T^{'} \end{bmatrix}$ and $\bar{F} = \begin{bmatrix} \bar{f}_1^{'} \\ \vdots \\ \bar{f}_T^{'} \end{bmatrix}$ with $\bar{r}_t = r_t - \hat{\mu}_2$ and $\bar{f}_t = f_t - \hat{\mu}_1$. $\hat{\mu}_1 = \frac{1}{T}\sum_{t=1}^{T} f_t$ and $\hat{\mu}_2 = \frac{1}{T}\sum_{t=1}^{T} r_t$ are respectively the estimators of μ_1 and μ_2.

For any regularization scheme, the estimator of the pseudo-true value θ_{HJ} is given by

$$\hat{\theta}_{HJ}^{\gamma} = \arg\min_{\theta}(\hat{\mu}_2 - \hat{V}_{21}\theta)^{'}\hat{\Sigma}_{\gamma}^{-1}(\hat{\mu}_2 - \hat{V}_{21}\theta). \tag{8}$$

$$\hat{\theta}_{HJ}^{\gamma} = \hat{V}_{11}^{-1}(\hat{\beta}^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{\beta})^{-1}\hat{\beta}^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{\mu}_2$$

and the regularized HJ-distance is

$$\begin{aligned} \hat{\delta}_{\gamma}^2 &= \frac{\hat{\mu}_2^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{\mu}_2}{N} - \frac{\hat{\mu}_2^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{V}_{21}}{N}\left(\frac{\hat{V}_{12}\hat{\Sigma}_{\gamma}^{-1}\hat{V}_{21}}{N}\right)^{-1}\frac{\hat{V}_{12}\hat{\Sigma}_{\gamma}^{-1}\hat{\mu}_2}{N} \\ &= \frac{\hat{\mu}_2^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{\mu}_2}{N} - \frac{\hat{\mu}_2^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{\beta}}{N}\left(\frac{\hat{\beta}^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{\beta}}{N}\right)^{-1}\frac{\hat{\beta}^{'}\hat{\Sigma}_{\gamma}^{-1}\hat{\mu}_2}{N} \end{aligned}$$

where $\hat{\Sigma}_{\gamma}^{-1}$ is the regularized inverse of $\hat{\Sigma}$ obtained either by Ridge or Tikhonov regularization.

Using the definition of the asset pricing model, the average of the excess returns can be rewritten as

$$\hat{\mu}_2 = \hat{\beta}(\hat{\mu}_1 - \mu_1 + \lambda) + (\beta - \hat{\beta})(\hat{\mu}_1 - \mu_1 + \lambda) + \alpha + \bar{\epsilon},$$

[3] In practice, μ_2 is unknown and replaced by $\widehat{\mu}_2$.

where $\bar{\epsilon} = \frac{1}{T}\sum_{t=1}^{T} \epsilon_t$. Remark that $\mu_2 = \alpha + \beta\lambda$ and $\beta' V_{22}^{-1}\alpha = 0$ by the first order condition of (2). Then, $\hat{\theta}_{HJ}^{\gamma} - \theta_{HJ}$ can be decomposed as

$$\begin{aligned}\hat{\theta}_{HJ}^{\gamma} - \theta_{HJ} = (\hat{V}_{11}^{-1} - V_{11}^{-1})\lambda + \hat{V}_{11}^{-1}(\hat{\mu}_1 - \mu_1) \\ + \hat{V}_{11}^{-1}(\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\hat{\beta})^{-1}\left[\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}(\beta - \hat{\beta})(\lambda + \hat{\mu}_1 - \mu_1)\right. \\ \left. + \hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\alpha + \hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\bar{\epsilon}\right].\end{aligned} \tag{9}$$

Appendix A shows that Ridge and Tikhonov regularizations are closely related. For this reason, we focus on Tikhonov regularization. From now on, $\hat{\theta}_{HJ}^{\gamma}$ and $\hat{\delta}_{\gamma}^{2}$ correspond to the estimators obtained by Tikhonov regularization.

Assumption 2. *(i) The process $x_{Nt} = (\epsilon_{1t}, ..., \epsilon_{Nt}, f_t')_{t=1,2,\cdots,T}$ is stationary and strong mixing with mixing coefficients $\alpha_{x,N}(l)$ uniformly bounded, with $\alpha_{x,N}(l) \leq a(l)$ and*

$$\sum_{l=1}^{\infty} l^s a(l)^{\frac{\rho}{2+\rho}} < \infty,$$

for some $s > 0$ and $\rho > 0$. $\alpha_x(l) = \sup_{i,k\geq 1}\sup_{A,B}\left[|P(A\cap B) - P(A)P(B)| : A \in \mathcal{F}_{-\infty}^{0}, B \in \mathcal{F}_{l}^{\infty}\right]$, where $l \geq 1$ and $\mathcal{F}_u^v = \sigma(x_t : u \leq t \leq v)$ is the σ-field generated by the data from time u to v.

(ii) $E[\epsilon_{it}^{4+2\rho}] < c$, for $i = 1, 2, \cdots, N$, where c is a constant and ρ is given in (i).

(iii) $E[\|f_t\|^{4+2\rho}] < \infty$.

Assumption 2(i) specifies the rate of decay for the mixing coefficient in terms of a parameter ρ. When the data are independent, $\alpha_x = 0$ and this condition is automatically satisfied for all $\rho > 0$. If x_t is exponentially strong mixing, then A2(i) is also valid for any $\rho > 0$. Assumption 2(ii) implies that $E[\epsilon_t'\epsilon_t/N] = O(1)$ as $N \to \infty$.

Definition 1. *The matrix Σ can be viewed as an operator $\Sigma : \mathbb{R}^N \to \mathbb{R}^N$ where $\mathbb{R}^N$ is endowed with norm $\|v\|_N$ and inner product $\langle v_1, v_2\rangle_N = \frac{v_1' v_2}{N}$. We define the range of Σ^{ω} (for some scalar $\omega > 0$), denoted $\mathcal{R}(\Sigma^{\omega})$, as the set $\{\psi_N \in \mathbb{R}^N : \psi_N = \Sigma^{\omega}\phi_N$ for some $\phi_N \in \mathbb{R}^N$ such that $\|\phi_N\|_N < \infty\}$.*

As the eigenvalues λ_j (associated with the eigenvectors ϕ_j) of Σ decline to zero as N goes to infinity, the elements of $\mathcal{R}(\Sigma^{\omega})$ are such that the inner products $\langle \psi, \phi_j\rangle_N$ decline to zero faster than λ_j^{ω}. The sequence $(\mathcal{R}(\Sigma^{\omega}))_{\omega>0}$ is a decreasing family of subspaces of $\mathbb{R}^N$ as ω increases. The parameter ω permits to quantify the regularization bias and determine conditions on γ.

We introduce the following assumption on α and the elements, β_k, of β.

Assumption 3. *(i)* $\beta_k, \alpha \in \mathcal{R}(\Sigma^2)$*, uniformly in* N*, for* $k = 1, 2, ..., K$*, i.e. there exist* ϕ_k *and* ϕ_α *such that* $\beta_k = \Sigma^2\phi_k$ *and* $\alpha = \Sigma^2\phi_\alpha$*,* $\sup_N \|\phi_k\|_N < \infty$ *and* $\sup_N \|\phi_\alpha\|_N < \infty$*.*

(ii) As $N \to \infty$*,* $C_\beta = \frac{1}{N}\beta'\Sigma^{-1}\beta = \left\langle \Sigma^{-\frac{1}{2}}\beta, \Sigma^{-\frac{1}{2}}\beta \right\rangle_N \to C$*, where* C *is a positive-definite matrix.*

Assumption 3(i) states that β_k and α belong to the range of Σ^ω with $\omega = 2$ so that objects $\Sigma^{-2}\beta_k$ and $\Sigma^{-2}\alpha$ are well defined even when N goes to infinity. This condition rules out concentration of β_k and α on directions associated with very small eigenvalues and helps control the regularization bias. It ensures that quadratic forms such as $\beta'\Sigma^{-1}\beta/N$, $\alpha'\Sigma^{-1}\alpha/N$ and hence δ^2 remain bounded when $N \to \infty$. In the online Appendix, we show that $\hat{\beta}$ belongs to the range of $\hat{\Sigma}^{1/2}$.

The following assumption is needed to derive the distribution of the regularized SDF parameter when N and T go to ∞.

Assumption 4. *For* $\rho > 0$ *defined in Assumption 2 (i), we assume:*

(i) $\lim_{N\to\infty} E[\|r_t\|_N^{4+2\rho}] < \infty$*.*

(ii) Let h_t *be as defined in Proposition 2. We have* $S_{N,T} = Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_t\right) \to S$ *as* $T, N \to \infty$*, where* S *is finite and positive definite.*

Proposition 2. *Suppose Assumptions 1-4 are satisfied.*

As T*,* N *go to infinity and* γ *goes to zero, if* γ *is chosen such that* $\gamma T \to \infty$ *and* $\gamma^2 T \to 0$ *, we have the following results for Tikhonov regularization*

1. $\hat{\theta}_{HJ}^{\gamma} \xrightarrow{P} \theta_{HJ}$

2. $\sqrt{T}(\hat{\theta}_{HJ}^{\gamma} - \theta_{HJ}) \xrightarrow{d} \mathcal{N}(0_K, V_{11}^{-1} S V_{11}^{-1})$

where $S = \lim_{N,T\to\infty} var\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} h_t\right]$ *and* h_t *is defined as*

$$h_t = \widetilde{f}_t y_t + \lambda + C_\beta^{-1}\frac{\beta'\Sigma^{-1}}{N}(\epsilon_t y_t - \tilde{r}_t\tilde{u}_t) + C_\beta^{-1}V_{11}^{-1}\widetilde{f}_t\frac{\epsilon_t'\Sigma^{-1}\alpha}{N} \text{ with } \tilde{u}_t = \frac{\tilde{r}_t'\Sigma^{-1}\alpha}{N}.$$

Proposition 2 holds for both correctly specified (case where $\alpha = 0$) and misspecified models.

Note that there is no condition connecting N and T so that N can be smaller or larger than T. Even though N does not explicitly appear in the proposition, N is present in the weighting matrix $\Sigma = V_{22}/N$, so that γ depends on N.

Let us discuss the impact of regularization on the properties of the estimator. It is well-known that when the number of moments increase, the variance of the standard GMM estimator decreases but its bias increases (Newey and Smith (2004)) and regularization permits to mitigate the bias at the cost of a higher variance (Carrasco (2012)). Here, we are not in the standard setting because of the misspecification and the fact that the weighting matrix is not the optimal

one. A formal investigation of the bias and variance as a function of the number of test assets would require a higher-order expansion and is beyond the scope of this paper. The primary role of the regularization is to stabilize the inverse of Σ when the number of test assets is large, hence reducing the variance of the estimator. However, the regularization introduces a regularization bias which vanishes asymptotically when $\gamma\sqrt{T} \to 0$ (see term (33) in the proof). Proposition 2 shows that $\hat{\theta}^{\gamma}_{HJ}$ is a consistent estimator of the pseudo-true value provided the regularization parameter γ converges to zero at an appropriate rate as $T \to \infty$. This rate is satisfied if $\gamma \sim \frac{1}{T^{\kappa}}$, with $\kappa \in (1/2, 1)$. In practice, we select γ using a data-driven procedure.

The results of Proposition 2 provide the basis for comparing competing asset pricing models. They can also be used to test whether a given factor has a zero coefficient in the SDF representation, conditional on the other factors, via a $t-$test on individual components of $\hat{\theta}^{\gamma}_{HJ}$. To implement the test, we estimate the standard error by replacing Σ^{-1} in h_t by the regularized inverse $\hat{\Sigma}^{-1}_{\gamma}$. So regularization is also useful to get a reliable standard error estimate. If h_t is autocorrelated, we need to rely on a HAC estimator.

3.3 HAC estimator

The long-run variance S can be estimated using the HAC estimator of Andrews (1991) as explained below.

Let $\widehat{h}_t$ be an estimator of h_t:

$$\widehat{h}_t = \bar{f}_t\hat{y}_t + \hat{V}_{11}\hat{\theta} + \widehat{C}^{-1}_{\beta}\frac{\hat{\beta}^{'}\hat{\Sigma}^{-1}_{\gamma}}{N}\left(\hat{\epsilon}_t\hat{y}_t - \overline{r}_t\widehat{u}_t\right) + \widehat{C}^{-1}_{\beta}\widehat{V}^{-1}_{11}\overline{f}_t\frac{\widehat{\epsilon}^{\prime}_t\hat{\Sigma}^{-1}_{\gamma}\widehat{\alpha}}{N}$$

with $\hat{y}_t = 1 - \bar{f}_t\hat{\theta}$, $\hat{\epsilon}_t = \overline{r}_t - \hat{\beta}\bar{f}_t$, $\widehat{\alpha} = \hat{\mu}_2 - \widehat{V}_{21}\hat{\theta}$, $\widehat{u}_t = \frac{\overline{r}^{\prime}_t\hat{\Sigma}^{-1}_{\gamma}\widehat{\alpha}}{N}$, and $\hat{\Sigma}^{-1}_{\gamma}$ obtained by Tikhonov regularization.

In small samples, we can include an extra term corresponding to the impact of regularization (see the proof of Proposition 2 in Appendix):

$$\hat{h}^{extra}_t = \gamma\hat{C}^{-1}_{\beta}\left[\left(\frac{\hat{\beta}^{\prime}\widehat{G}_{\gamma}\overline{r}_t}{N}\right)\left(\frac{\overline{r}^{\prime}_t\widehat{G}_{\gamma}\hat{\alpha}}{N}\right) - \frac{\hat{\beta}^{\prime}\widehat{G}_{\gamma}\widehat{\Sigma}\widehat{G}_{\gamma}\hat{\alpha}}{N}\right]$$

where $\widehat{G}_{\gamma} = \left(\hat{\Sigma}^2 + \gamma I\right)^{-1}$. This extra term is asymptotically negligible but provides a more accurate finite-sample approximation of the variance.

For a kernel ω and a bandwidth m_T, the HAC estimator of S is defined as

$$\hat{S}_T = \sum_{|j|<T}\omega\left(\frac{j}{m_T}\right)\widehat{\Gamma}_j$$

where $\widehat{\Gamma}_j = \frac{1}{T}\sum_{t=j+1}^{T} \widehat{h}_t \widehat{h}'_{t-j}$ for $j \geq 0$ and $\widehat{\Gamma}_j = \widehat{\Gamma}'_{-j}$ for $j < 0$. The kernel ω is assumed to belong to the class $\mathcal{W}$:

$$\begin{aligned}\mathcal{W} &= \{\omega(.) : \mathbb{R} \to [-1,1], \omega(0) = 1, \omega(x) = \omega(-x), \forall x \in \mathbb{R}, \\ &\quad \int |\omega(x)|\, dx < \infty, \omega(.) \text{ is continuous at 0 and at all but a finite number of points}\Big\}.\end{aligned}$$

This class corresponds to the class $\mathcal{K}_1$ of Andrews (1991) with the extra condition $\int |\omega(x)|\, dx < \infty$.

Note that h_t depends on two types of parameters, the finite dimensional parameters denoted ξ_1 and the $N \times 1$ dimensional parameters denoted ξ_2. More specifically, $\xi_1 = \left(\theta', \mu_1', vecV_{11}', vecC_\beta'\right)'$, $\xi_2 = \left((\Sigma^{-1}\beta)', \left(\Sigma^{-1}\alpha\right)', \mu_2', \beta'\right)'$. To establish the consistency of $\hat{S}_T$, we need some extra conditions.

Assumption 5. *(i)* $\omega \in \mathcal{W}$.

(ii) $\{h_t\}$ *is* $\alpha - mixing$ *with*

$$\sum_{l=1}^{\infty} l^2 \alpha(l)^{\frac{\rho}{2+\rho}} < \infty$$

where ρ *is the same as in Assumption 2.*

(iii) $E\left[\|h_t\|^{4+2\rho}\right] < \infty$ *and* $E\left[\sup_{\xi_1 \in \Xi} \left\|\frac{\partial h_t}{\partial \xi_1}\right\|^2\right] < \infty$, *where* Ξ *is a convex neighborhood of the true value of* ξ_1.

Proposition 3. *Suppose Assumptions 1-5 hold,* T, N *go to infinity,* γ *and* m_T *are chosen such that* $\gamma T \to \infty$, $\gamma^2 T \to 0$, $m_T \to \infty$, $m_T/\sqrt{\gamma T} \to 0$, *and* $m_T\sqrt{\gamma} \to 0$ *as* $T \to \infty$. *Then, we have*

$$\left\|\widehat{S}_T - S\right\| = O_p\left(m_T\left(\frac{1}{\sqrt{\gamma T}} + \sqrt{\gamma}\right)\right) + o_p(1) = o_p(1).$$

Assumption 5(ii) is slightly stronger than Assumption 2(i). Our conditions are similar to Assumptions A and B in Andrews (1991). The proof of the consistency of $\widehat{S}_T$ is more involved than that of Andrews because of the presence of parameters whose dimension N increases. To the best of our knowledge, Proposition 3 is the first result establishing the convergence of HAC estimator based on Tikhonov regularization in high dimension. Some results on the HAC estimator in high dimension have been derived by Li and Liao (2020) for series estimators and Babii, Ghysels, and Striaukas (2024) for Lasso estimators. Proposition 3 provides an upper bound for the convergence rate of $\widehat{S}_T - S$. We observe that this rate depends on the regularization parameter γ and is slower than that obtained by Andrews (1991) which was $m_T/\sqrt{T} \to 0$. This is due to the large dimensional parameters ξ_2 which estimators converge slower than $\sqrt{T}$.

3.4 Choice of the regularization parameter

We adopt a data-driven procedure to select the regularization parameter γ. From Proposition 2, consistency requires that $\gamma T \to \infty$ and $\gamma^2 T \to 0$ which implies that γ must decrease at a rate $T^{-\kappa}$ with $\kappa \in (1/2, 1)$. We impose this rate by setting $\gamma = c\lambda_{\max} T^{-0.8}$ where $\lambda_{\max}$ is the largest eigenvalue of $\left(\widehat{V}_{22}/N\right)^2$ and c is a scalar selected from a grid such that $0 < c_{\min} \leq c \leq c_{\max} < \infty$. This parametrization separates scale (captured by $\lambda_{\max}$) from shrinkage intensity (captured by c). Scaling by $\lambda_{\max}$ ensures that the regularization is invariant to the magnitude of the covariance matrix, while restricting c to a grid allows for a simple one-dimensional search. To select c, we use a two-fold cross-validation procedure designed to approximate the out-of-sample Hansen–Jagannathan distance.

For a given sample size T, we split the data into two equal subsamples. For each candidate c, we estimate θ using the first subsample, evaluate the pricing errors on the second subsample, and compute the out-of-sample criterion:

$$HJ_{oos} = (\bar{\mu}_2^o - \widehat{V}_{21}^o \hat{\theta}_c)' \left(diag\left(\widehat{V}_{22}^o\right)\right)^{-1} (\bar{\mu}_2^o - \widehat{V}_{21}^o \hat{\theta}_c), \tag{10}$$

where quantity with ^{o} are estimated from the withheld sample. We then select the value of c that minimizes the average of the HJ_{oos} obtained on the two different subsamples.

In principle, the Hansen–Jagannathan distance involves weighting by $\left(\widehat{V}_{22}^o\right)^{-1}$. However, in high dimension, $\left(\widehat{V}_{22}^o\right)^{-1}$ is poorly estimated and its inverse is highly unstable. Using a regularized inverse with the same γ inside the cross-validation criterion would mechanically favor large values of γ, since stronger regularization reduces the criterion. To avoid this issue, we replace $\widehat{V}_{22}^{o-1}$ by $\left(diag\left(\widehat{V}_{22}^o\right)\right)^{-1}$, which provides a stable and scale-correct weighting scheme[4].

4 Tests of equality of HJ-distances of asset pricing models

The distribution of the difference in HJ distances depends on whether models are nested, share the same SDF, or are distinct. We treat these cases separately because they lead to different convergence rates and limiting distributions.

We compare two competing models (Models a and b) using their regularized HJ-distances. Their SDFs are defined as $y_{at}(\eta) = 1 - (x_{at} - E[x_{at}])'\theta_a$ and $y_{bt}(\lambda) = 1 - (x_{bt} - E[x_{bt}])'\theta_b$. $x_{at} = [f_{1t}', f_{2t}']'$ and $x_{bt} = [f_{1t}', f_{3t}']'$ are two sets of factors, that are used in Model a and Model b, respectively. f_{it} is of dimension $K_i \times 1$, $i = 1, 2, 3$.

We note $\delta_m^2, m = a, b$ the HJ-distances of the two models.

$$\delta_m^2 = \mu_2' V_{22}^{-1} \mu_2 - \mu_2' V_{22}^{-1} V_{m,21} (V_{m,12} V_{22}^{-1} V_{m,21})^{-1} V_{m,12} V_{22}^{-1} \mu_2 \tag{11}$$

[4]The diagonal weighting matrix is used only in the cross-validation step. The estimators and HJ distances involved in Propositions 2, 4-7 use the full matrix $\widehat{\Sigma}_\gamma^{-1}$.

Model m is estimated using solely factors in x_m. Let $\theta_m = (\theta'_{m1}, \theta'_{m2})'$ for $m = a, b$ where θ_{m1} corresponds to the coefficient of the common factors f_{1t} and θ_{m2} to that of the other factors.

When $K_1 = 0$, the two models do not share factors. When $K_2 = 0$ or $K_3 = 0$, one of the models nests the other one. Finally, when $K_1 > 0$, $K_2 > 0$, and $K_3 > 0$, the two models are non-nested with overlapping factors.

4.1 Comparison of nested models

In this section, we assume without loss of generality that $K_2 = 0$. When the models are nested, the equality of HJ-distances is equivalent to the equality of the SDFs as pointed out by Kan and Robotti (2009). Let $V_{a,12} = E(x_{at}r'_t) = [E(f_{1t}r'_t) \;\; E(f_{2t}r'_t)]$ and $V_{b,12} = E(x_{bt}r'_t) = [E(f_{1t}r'_t) \;\; E(f_{3t}r'_t)]$. Let us define $C_m = \left(V_{m,12}V_{22}^{-1}V_{m,21}\right)^{-1}$, $m = a, b$ and partition it as below

$$C_m = \begin{bmatrix} C_{m,11} & C_{m,12} \\ C_{m,21} & C_{m,22} \end{bmatrix}. \tag{12}$$

We assume $C_{b,22}^{-1}$ is a full rank matrix. It follows from 12 in appendix that the difference of HJ-distances $(\delta_a^2 - \delta_b^2)$ between the two models is equal to

$$\delta_a^2 - \delta_b^2 = \theta_{b2}^{'} C_{b,22}^{-1} \theta_{b2}. \tag{13}$$

The following proposition can be viewed as a generalization of Kan and Robotti (2009) Proposition 2 where N and T are allowed to go to ∞ using regularization.

Proposition 4. *Suppose Assumptions 1-4 are satisfied, $cov(f_1, f_2, f_3)$ and $C_{b,22}$ are non-singular. We have the following results:*

1. $\delta_a^2 = \delta_b^2$ if and only if $\theta_{b2} = 0_{K_3}$

2. Under the hypothesis $\theta_{b2} = 0_{K_3}$, as T, N go to infinity and if γ is chosen such that $\gamma T \to \infty$ and $\gamma^2 T \to 0$,

$$T(\hat{\delta}_{a,\gamma}^2 - \hat{\delta}_{b,\gamma}^2) \xrightarrow{d} \sum_{j=1}^{K_3} \xi_j \chi_j^2(1)$$

where $\chi_j^2(1)$ are independent chi-square random variables with 1 degree of freedom and ξ_j are the eigenvalues of $V\left(\widehat{\theta}_{b2}^{\gamma}\right)^{1/2} C_{b,22}^{-1} V\left(\widehat{\theta}_{b2}^{\gamma}\right)^{1/2}$ *and $V\left(\widehat{\theta}_{b2}^{\gamma}\right)$ is the asymptotic variance of $\widehat{\theta}_{b2}^{\gamma}$.*

Proposition 4 implies that we can perform two kinds of tests to compare Model a with factor f_1 and Model b with factors f_1 and f_3.

On the one hand, we can focus on the SDF parameter θ_{b2} (coefficient of f_3) and test $H_0 : \theta_{b2} = 0_{K_3}$ using a t- or Wald statistics based on the misspecification-robust asymptotic distribution given by Proposition 2.

On the other hand, we can compute the HJ-distance difference of the two models $T(\hat{\delta}_{a,\gamma}^2 - \hat{\delta}_{b,\gamma}^2)$ using the same level of penalization or (13) and use the asymptotic distribution given in point 2

of Proposition 4 to compare them. Note that the coefficients ξ_j are all nonnegative, hence the test presented in Proposition 4 is a one-sided test.

When comparing models, it is essential to use the same penalization value γ across models. A small value of penalization provides maximum power without compromising size, while a larger value diminishes it. This comes from the fact that large values of γ shrinks the regularized HJ-distances toward 0.

4.2 Comparison of non-nested models

In this section, we assume $K_1 > 0$, $K_2 > 0$, and $K_3 > 0$. The two models are non-nested with overlapping factors. In this case, equality of HJ-distance can be achieved in two cases. The first case corresponds to the setting where the SDFs coincide. The second is when $y_a \neq y_b$ but $\delta_a^2 = \delta_b^2$. Both cases need to be treated separately.

4.2.1 Test of SDFs equality

In this section, we test the equality of the SDFs, $y_a = y_b$. Given the models are non-nested, the equality of SDFs can be achieved only if both SDF depend on f_1 only. Consider $C_a = \left(V_{a,12}V_{22}^{-1}V_{a,21}\right)^{-1}$, partition it as in (12), and assume $C_{a,22}$ is a full rank matrix. Lemma 12 shows that the difference between the HJ-distances is

$$\delta_a^2 - \delta_b^2 = -\theta_{a2}^{'} C_{a,22}^{-1}\theta_{a2} + \theta_{b2}^{'} C_{b,22}^{-1}\theta_{b2}.$$

The following proposition outlines the asymptotic distribution of the test statistics.

Proposition 5. *Suppose Assumptions 1-4 are satisfied. Moreover, assume that* $cov(f_1, f_2, f_3)$, $C_{a,22}$, *and* $C_{b,22}$ *are non-singular. We have the following result:*

1. $y_a = y_b$ *if and only if* $\theta_{a2} = 0_{K_2}$ *and* $\theta_{b2} = 0_{K_3}$.

2. Under the hypothesis $y_a = y_b$ *, as* T, N *go to infinity and* γ *goes to zero such that* $\gamma T \to \infty$ *and* $\gamma^2 T \to 0$,

$$T(\hat{\delta}_{a,\gamma}^2 - \hat{\delta}_{b,\gamma}^2) \xrightarrow{d} \sum_{i=1}^{K_2+K_3} \xi_i \chi_i^2(1) \tag{14}$$

where ξ_i *are the eigenvalues of* $V\left(\begin{bmatrix}\hat{\theta}_{a2}\\ \hat{\theta}_{b2}\end{bmatrix}\right)^{\frac{1}{2}} \begin{bmatrix}-C_{a,22}^{-1} & 0_{K_2\times K_3}\\ 0_{K_3\times K_2} & C_{b,22}^{-1}\end{bmatrix} V\left(\begin{bmatrix}\hat{\theta}_{a2}\\ \hat{\theta}_{b2}\end{bmatrix}\right)^{\frac{1}{2}}$, *and* $\chi_i^2(1)$ *are independent* $\chi^2(1)$ *random variables.*

Note that the ξ_i in Equation (14) may be positive or negative, hence this test is a two-sided test. Proposition 5.1. shows that to compare asset pricing models with overlapping factors, one can also test the simultaneous nullity of the coefficients of the non common factors (θ_{a2} and θ_{b2}). Then, Proposition 2 can be used to construct a classic Wald test. This option does not directly test the nullity of the difference in HJ-distances, but the equality of the SDFs of the two models.

4.2.2 Comparison of non-nested models with distinct SDFs

In this section, we wish to compare two non-nested models with distinct SDFs ($y_a \neq y_b$) when both models are misspecified. Hansen, Heaton, and Luttmer (1995) and Kan and Robotti (2008) have already given the distribution of the HJ-distance (for gross returns) and the modified HJ-distance (for excess returns) respectively when models are misspecified and N is fixed.

Kan and Robotti (2008) show that the modified HJ-distance has the following distribution

$$\sqrt{T}(\hat{\delta}^2 - \delta^2) \xrightarrow{d} \mathcal{N}(0, v_1)$$

where $v_1 = var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} q_t^m\right)$ and $q_t^m = y_t^2 - (y_t - \nu'(r_t - \mu_2))^2 + 2\nu'\mu_2 - \delta^2$ with ν is the Lagrange multiplier ($\nu = V_{22}^{-1}E[r_t y_t]$) of problem (5). It is worth noticing that the distribution of the distance is not affected by the uncertainty brought forth by the estimation of the Lagrange multiplier ν.

Below, we give the distribution of the penalized HJ-distance when models are misspecified and N goes to infinity. To do so, we exploit the following expression of the penalized HJ-distance given in Equations (18) and (20) in Appendix B:

$$\delta_\gamma^2 = \min_{\theta} \max_{\nu_1 \in \mathbb{R}^N, \nu_2 \in \mathbb{R}} E\left[2y\frac{\nu_1' r}{N} - \frac{\nu_1' r r' \nu_1}{N^2} - \nu_2^2 - 2\frac{\nu_1' r \nu_2}{N} + \psi_\Sigma(\nu_1)\right], \tag{15}$$

where $\psi : \mathbb{R}^N \to \mathbb{R}$ is a concave function representing the penalty, namely $\psi_\Sigma(\nu_1) = -\gamma \left\|\nu_1\right\|_N^2$ for Ridge and $\psi_\Sigma(\nu_1) = -\gamma \left\langle \nu_1, \Sigma^\dagger \nu_1 \right\rangle_N$ for Tikhonov, where $\Sigma^\dagger$ denotes the Moore-Penrose inverse of Σ.

Let $\nu_{1,\gamma}$ and $\nu_{2,\gamma}$ be the optimal values of the Lagrange multipliers ν_1 and ν_2, and θ_γ the solution of (15) in θ. We can rewrite δ^2 as $E\left[2y\frac{\nu_{1,\gamma}' r}{N} - \frac{\nu_{1,\gamma}' r r' \nu_{1,\gamma}}{N^2} - \nu_{2,\gamma}^2 - 2\frac{\nu_{1,\gamma}' r \nu_{2,\gamma}}{N} + \psi_\Sigma(\nu_{1,\gamma})\right]$. Now, by the first-order condition, we can replace $\nu_{2,\gamma}$ by $-\nu_{1,\gamma}' E(r)/N$. This yields

$$\delta_\gamma^2 = E\left[2y\frac{\nu_{1,\gamma}' r}{N} - \frac{\nu_{1,\gamma}' \tilde{r}_t \tilde{r}_t' \nu_{1,\gamma}}{N^2} + \psi_\Sigma(\nu_{1,\gamma})\right]$$

or equivalently using $E(y_t) = 1$,

$$\begin{aligned} \delta_\gamma^2 &= E\left(q_t(\eta_\gamma)\right) \text{ with} \\ q_t(\eta_\gamma) &= 2y_t\frac{\nu_{1,\gamma}' \tilde{r}_t}{N} - \frac{\nu_{1,\gamma}' \tilde{r}_t \tilde{r}_t' \nu_{1,\gamma}}{N^2} + \frac{2\nu_{1,\gamma}' \mu_2}{N} + \psi_\Sigma(\nu_{1,\gamma}) \end{aligned}$$

and $\eta_\gamma = \left(\theta_\gamma', \nu_{1,\gamma}\right)'$.

We can also show that, for both types of regularizations,

$$\widehat{\delta}_{\gamma}^{2}=\frac{1}{T}\sum_{t=1}^{T}\left[2y_{t}\left(\widehat{\theta}_{HJ}\right)\frac{\widehat{\nu}_{1,\gamma}^{\prime}r_{t}}{N}-\frac{\widehat{\nu}_{1,\gamma}^{\prime}\overline{r}_{t}\overline{r}_{t}^{\prime}\widehat{\nu}_{1,\gamma}}{N^{2}}+\psi_{\hat{\Sigma}}(\widehat{\nu}_{1,\gamma})\right]=\frac{1}{N}\widehat{\alpha}^{\prime}\hat{\Sigma}_{\gamma}^{-1}\widehat{\alpha}$$

where $\widehat{\nu}_{1,\gamma}=\hat{\Sigma}_{\gamma}^{-1}\widehat{\alpha}$ and $\widehat{\alpha}=\sum_{t=1}^{T}y_{t}\left(\widehat{\theta}_{HJ}\right)r_{t}/T$. See Lemma 13 in Appendix.

Assumption 6. *(i) For $0<\gamma<\infty$, $q_t(\eta)$ is differentiable on an open set N of η_γ and* $E\left[\sup_{\eta\in N}\parallel\nabla q_{t}(\eta)\parallel\right]<\infty$.
(ii) Let ℓ_t be the influence function $2y_tu_t-u_t^2+\delta^2$, then

$$0<v_{2}=\lim_{N,T}\sum_{j=-\infty}^{\infty}E(\ell_{t}\ell_{t-j})<\infty.$$

Assumption 6(i) ensures the interchangeability between integration and differentiation for any $0<\gamma<\infty$, (ii) is used to establish the asymptotic normality of the centered HJ distance.

Proposition 6. *Let $\hat{\delta}_{\gamma}^{2}$ be the regularized Hansen-Jagannathan distance with Tikhonov regularization. Suppose Assumptions 1-4 and 6 are satisfied and $\delta\neq0$. As T, N go to infinity and $\gamma T\to\infty$, and $\gamma^2T\to0$,*

$$\sqrt{T}\left(\hat{\delta}_{\gamma}^{2}-\delta^{2}\right)\xrightarrow{d}\mathcal{N}(0,v_{2}),$$

where

$$v_{2}=\lim_{N,T\to\infty}\sum_{j=-\infty}^{\infty}E\left(\ell_{t}\ell_{t-j}\right),$$

and $\ell_{t}=2y_{t}\nu_{1,0}^{\prime}\frac{\tilde{r}_{t}}{N}-\frac{\nu_{1,0}^{\prime}\tilde{r}_{t}\tilde{r}_{t}^{\prime}\nu_{1,0}}{N^{2}}+\frac{2\nu_{1,0}^{\prime}\mu_{2}}{N}-\delta^{2}=2y_{t}u_{t}-u_{t}^{2}+\delta^{2}$, $u_{t}=\tilde{r}_{t}^{\prime}\Sigma^{-1}\alpha/N$, *and* $\nu_{1,0}=\Sigma^{-1}\alpha$.

Proposition 6 gives the distribution of the penalized HJ-distance when the model is misspecified. It cannot be used when the model is correctly specified. Indeed, when $\alpha=0$, $\nu_1=0$ and hence $\ell_t=0$ which yields a degenerate asymptotic distribution. The asymptotic variance v_2 can be estimated using a HAC estimator and replacing ℓ_t by $\widehat{\ell_t}$:

$$\widehat{\ell_{t}}=2y_{t}\left(\widehat{\theta}_{HJ}^{\gamma}\right)\widehat{u}_{t}-\widehat{u}_{t}^{2}+\hat{\delta}_{\gamma}^{2}+\gamma\left(\left(\frac{\overline{r}_{t}^{\prime}w_{\gamma}}{N}\right)^{2}-\frac{w_{\gamma}^{\prime}\widehat{\Sigma}w_{\gamma}}{N}\right)$$

where $\widehat{u}_t=\widehat{\nu}_1^{\prime}\overline{r}_t/N$, $\hat{\nu}_1=\widehat{\Sigma}_{\gamma}^{-1}\widehat{\alpha}$, and $w_{\gamma}=\left(\hat{\Sigma}^{2}+\gamma I\right)^{-1}\widehat{\alpha}$. The term in γ reflects the impact of the regularization and vanishes asymptotically. A formal proof of the HAC estimator consistency could be done along the line of that of Proposition 3.

Considering two competing asset pricing models a and b, we are interested in the difference of their squared HJ distances. To establish its limiting distribution, we need an extra assumption. Define $L_{N,t} = (\ell_{a,t}, \ell_{b,t})'$ where $\ell_{m,t} = 2y_{m,t}u_{m,t} - u_{m,t}^2 + \delta_m^2$, $m = a, b$ and

$$\Omega_N = \sum_{j=-\infty}^{\infty} E\left(L_{N,t}L'_{N,t-j}\right).$$

Assumption 7. *Assume that* $\sup_N E\left\|L_{N,t}\right\|^{2+\eta} < \infty$, $\Omega_N \to \Omega$ *and* $0 < v_3 = e'\Omega e < \infty$, *where* $e = (1, -1)'$.

Proposition 7. *Let* $\hat{\delta}^2_{m\gamma}$ *be the regularized Hansen-Jagannathan distance with Tikhonov regularization for model* m, $m = a, b$. *Assumptions 1-4 and 6-7 are satisfied for both models,* $y_a \neq y_b$, *and* $\delta_a^2, \delta_b^2 \neq 0$. *As* T, N *go to infinity and if* γ *is chosen such that* $\gamma T \to \infty$ *and* $\gamma^2 T \to 0$,

$$\sqrt{T}\left((\hat{\delta}^2_{a\gamma} - \hat{\delta}^2_{b\gamma}) - (\delta_a^2 - \delta_b^2)\right) \xrightarrow{d} \mathcal{N}(0, v_3),$$

where

$$v_3 = \lim_{N,T\to\infty} var\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\tilde{\ell}_{at} - \tilde{\ell}_{bt}\right)\right],$$

where $\tilde{\ell}_{m,t} = 2y_{m,t}u_{m,t} - u_{m,t}^2 + \delta_m^2$ *for* $m = a, b$, $y_{m,t}$ *is the SDF of model* m *and* $\nu_{m,1} = \Sigma^{-1}\alpha_m$ *where* α_m *represents the pricing errors of model* m. *Moreover, under* $H_0 : \delta_a^2 = \delta_b^2$, *we have the following simplification*

$$\sqrt{T}\left(\hat{\delta}^2_{a\gamma} - \hat{\delta}^2_{b\gamma}\right) \xrightarrow{d} \mathcal{N}(0, v_3).$$

When using Proposition 7 to compare two asset pricing models, one should use the same value of the regularization parameter. We can use Proposition 7 to construct a so-called Normal test of $H_0 : \delta_a^2 = \delta_b^2$ of the form $\sqrt{T}\left(\hat{\delta}^2_{a\gamma} - \hat{\delta}^2_{b\gamma}\right)\widehat{v}_4^{-1/2}$, where $\widehat{v}_4$ is a HAC estimator of v_4. This statistics converges asymptotically to a standard normal distribution. Note the difference of rates of $\left(\hat{\delta}^2_{a\gamma} - \hat{\delta}^2_{b\gamma}\right)$ which is $\sqrt{T}$ here instead of T in the previous sections. Why such a difference? The $\sqrt{T}$ rate comes from the rate of $\hat{\delta}^2_{\gamma}$ obtained in Proposition 6. It would be expected to have the same rate for the difference $\left(\hat{\delta}^2_{a\gamma} - \hat{\delta}^2_{b\gamma}\right)$ in all scenarios. However, when $y_{at} = y_{bt}$, we have $\alpha_a = \alpha_b$ and $\nu_{a,1} = \nu_{b,1}$ so that $\tilde{\ell}_{at} - \tilde{\ell}_{bt} = 0$ and $v_3 = 0$. Because of this simplification, the rate of convergence of $\left(\hat{\delta}^2_{a\gamma} - \hat{\delta}^2_{b\gamma}\right)$ is faster than the standard $\sqrt{T}$ rate when the SDFs coincide.

Given $\hat{\delta}^2_{a\gamma} - \hat{\delta}^2_{b\gamma}$ has a different asymptotic distribution depending on whether the model is correctly specified or not, it is not obvious to know which test to use. Kan and Robotti (2009) propose to apply a sequential test inspired from Vuong (1989). However, Vuong's test has been shown to exhibit large size distortions. Shi (2015) and Schennach and Wilhelm (2017) propose one-step tests which have good size and power. The first approach relies on simulated critical

values and the second on sample splitting. Implementing these procedures in our setting would be interesting but goes beyond the scope of the present paper.

4.3 Multiple comparison

In this section, we present a comparison test of multiple models. The test is based on the work of Wolak (1963) and Wolak (1989), see also Gospodinov, Kan, and Robotti (2013). Suppose we have $p+1$ models with HJ-distances given by $\delta_i^2, \quad i = 1, \ldots, p+1$. We are interested in testing whether a benchmark model (Model 1) has an aggregate pricing error as low as the other p models. Let $d_i = \delta_1^2 - \delta_i^2$, $i = 2, \ldots, p+1$ be the difference between the HJ-distance of the benchmark and those of the remaining models and $d = (d_2, \cdots, d_{p+1})'$. The null hypothesis of the test is $H_0 : d \leq 0_p$. The test uses the sample counterpart of d, $\hat{d}_\gamma = \left(\hat{d}_{2\gamma}, \cdots, \hat{d}_{p+1\gamma}\right)'$ for a given value of γ. To have the same framework as Wolak (1989), we rely on the fact that under the conditions of Proposition 7,

$$\sqrt{T}(\hat{d}_\gamma - d) \xrightarrow{d} \mathcal{N}(0_p, \Omega_d).$$

The latter is valid only when the models have distinct SDFs and they are misspecified, that is $\delta_i > 0$. This seems to be a reasonable assumption given asset pricing models can be viewed as approximations of the reality and usually include different factors. Let $\tilde{d}_\gamma$ be the optimal solution of the following quadratic programming problem

$$\min_d (\hat{d}_\gamma - d)' \hat{\Omega}_{d,\gamma}^{-1} (\hat{d}_\gamma - d) \quad s.t.\ d \leq 0_p,$$

where $\hat{\Omega}_{d,\gamma}$ is a consistent estimator of Ω_d when $N, T \to \infty$ and $\gamma \to 0$. The likelihood ratio statistic of the null hypothesis is

$$LR_\gamma = T(\hat{d}_\gamma - \tilde{d}_\gamma)' \hat{\Omega}_{d,\gamma}^{-1} (\hat{d}_\gamma - \tilde{d}_\gamma). \tag{16}$$

The distribution of the previous statistics is obtained under the least favorable value, i.e. $d = 0_p$. We have $LR_\gamma \xrightarrow{d} \sum_{i=0}^{p} w_{p-i}(\Omega_d)\chi^2(i)$, where the weights w_i sum up to one[5] and $\chi^2(i)$ are independent Chi-square random variables with i degrees of freedom and $\chi^2(0)$ is a dirac at 0. See Kudo (1987) and Wolak (1963) for more details.

5 Monte Carlo Simulations

In this section, we run several Monte Carlo simulations to showcase the advantage of the regularization schemes described previously. The parameters for the simulations are calibrated on real

[5] Appendix C of Gospodinov, Kan, and Robotti (2013) gives the procedure to compute $w_i(\Omega_d)$ and the p-value of the test.

data.

5.1 SDF parameter estimates

In this subsection, we analyze the small sample properties of the t-test of $H_0 : \theta_{HJj} = 0$. This corresponds to testing whether the factor j has a zero coefficient in the HJ minimizing SDF. We compare the small-samples size properties of our test with that of Kan and Robotti (2008).

We consider the three factor model of Fama and French (1993) (FF3), where the risk factors are the market excess return (MKT, f_1), the return difference between portfolios of small and large stocks (SMB, f_2), and the return difference between portfolios of high and low book-to-market ratios (HML, f_3). Our simulation strategy follows closely that of Fan, Fan, and Lv (2008).

First, we estimate the factor model

$$r_t = \mu_2 + \beta_1 \tilde{f}_{1t} + \beta_2 \tilde{f}_{2t} + \beta_3 \tilde{f}_{3t} + \varepsilon_t \tag{17}$$

using for r_t the excess returns of 202 combined portfolios going from July 1963 to January 2026 extracted from the Kenneth French's Website; these portfolios are the same equity portfolios used in Giglio and Xiu (2021). The portfolios list is presented in Table 7 in Appendix C. The factors are also from Kenneth French's website. So we obtain a 202×3 matrix B of estimators of $(\beta_1, \beta_2, \beta_3)$, we compute their 3×1 mean vector μ_b and 3×3 variance matrix Σ_b. We also estimate the mean μ_1 and variance V_{11} of the factors. We compute the standard deviation (std) of the residuals of the regression (17) and obtain their sample mean and std. In the sequel, we will draw the standard deviations from a Gamma distribution. To calibrate the two parameters of the Gamma distribution, we use the method of moments in order to match the sample mean and sample std of the residual std with the theoretical ones.

For each simulation, we proceed as follows:

(a) We first generate a random sample of $f = (f_1, f_2, f_3)'$ of size T from the normal distribution $N(\mu_1, V_{11})$.

(b) Then, we generate a random sample of N factor loadings $(b_1, b_2, ..., b_N)$ from $N(\mu_b, \Sigma_b)$.

(c) We generate a random sample of N standard deviations $\sigma_1, ..., \sigma_N$ from a Gamma distribution whose two parameters have been estimated as previously explained.

(d) Once we have N standard deviations of the errors, we generate a random sample of $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)'$ of size T from $N\left(0, diag\left(\sigma_1^2, ..., \sigma_N^2\right)\right)$.

(e) Then, we get a random sample $r = (r_1, ..., r_T)$ from (17). If $N > 202$, we construct μ_2 by concatenating several μ_2.

When the mean μ_2 is not constrained, the SDF is typically misspecified. For our sample with $N = 202$ test assets and $T = 751$, the (non-regularized) Hansen-Jagannathan distance equals 1.25.

We consider $N = 50$, 300, and 1000 and $T = 150$, 300, and 1000. We run 10000 simulations to evaluate the empirical level and power of the tests. In all cases, the regularization parameter γ is selected using the procedure described in Section 3.4 where c is chosen in a log-scale grid going

from 0.00001 to 0.001. Standard errors and covariance matrices are robust to misspecification but not to autocorrelation.

Table 1 reports the empirical size of the individual t-tests for the SDF coefficients. To evaluate the test of the jth coefficient, we generate r_t as a three-factor model but we choose the mean μ so that the corresponding $\theta_{HJj} = 0$ (see Appendix C for details). We compare the Tikhonov test with the test of Kan and Robotti (2008). The latter is not directly applicable when $N > T$ because $\widehat{V}_{22}$ is singular; in those cases we replace its inverse by the Moore–Penrose inverse. This implementation is equivalent to setting $\gamma = 0$ (with the convention $0/0 = 0$) in our procedure. Table 1 shows that Tikhonov regularization substantially reduces the size distortions of the Moore–Penrose procedure when $N = T$. At $N = T = 300$ and $N = T = 1000$, the Moore–Penrose tests exhibit severe overrejection. Near the interpolation boundary $N/T = 1$, small eigenvalues of the sample return covariance matrix are inverted without regularization. The resulting instability inflates the standard errors of the estimators. The size improves again once N/T moves away from one. The nonmonotonic pattern across (N, T) is reminiscent of the double-descent phenomenon discussed by Hastie, Montanari, Rosset, and Tibshirani (2022) and Liao, Ma, Neuhierl, and Shi (2026). On the other hand, the regularized rejection frequencies remain much closer to the nominal 5% level when $N = T$. Regularization does not eliminate all finite-sample distortions: the HML test, for example, rejects in 8.3% of replications when $T = 1000$ and $N = 50$. Away from $N = T$, the two procedures generally perform similarly.

We now turn our attention to the empirical power of our t-test. Table 2 reports rejection rates under $H_1 : \theta_{HJj} \neq 0$. In that case, we generate r_t following a FF3 calibrated on the empirical observations. Away from $N = T$, the Tikhonov and Moore–Penrose procedures generally have similar power. At $N = T$, comparisons of their rejection frequencies must be interpreted cautiously because the Moore–Penrose procedure severely overrejects under the null. The regularized tests gain substantial power as T increases. The effect of increasing N is less uniform and varies across coefficients: it is particularly pronounced for HML in the largest time-series sample, while SMB power changes relatively little. Overall, the principal advantage of regularization in these experiments is improved size control near $N = T$, while retaining power comparable to the unregularized procedure in the other designs.

The rejection rates of Tikhonov test (Panel B) generally rise with T and, for a given T, with N. Power is highest for the market coefficient and lower for SMB and HML, reflecting differences in signal strength across the three factors. Appendix C reports the cross-validated values of γ. They decline with T, as expected when covariance estimation becomes more precise, and vary very little with N.

In summary, Tikhonov regularization is crucial for reliable inference when the sample covariance matrix is ill-conditioned. A data-driven regularization parameter therefore permits to take advantage of a large cross-section of test assets.

Table 1: Empirical size of t-tests under misspecification

N	50	300	1000	50	300	1000	50	300	1000
Panel A: Test using Moore-Penrose inverse									
	θ_{MKT}			θ_{SMB}			θ_{HML}		
$T = 150$	0.048	0.046	0.047	0.049	0.049	0.044	0.049	0.051	0.042
$T = 300$	0.047	0.351	0.047	0.047	0.503	0.047	0.050	0.663	0.048
$T = 1000$	0.049	0.047	0.526	0.051	0.052	0.659	0.051	0.054	0.795
Panel B: Test using Tikhonov regularized inverse									
$T = 150$	0.047	0.045	0.046	0.050	0.048	0.044	0.047	0.043	0.041
$T = 300$	0.048	0.047	0.046	0.051	0.048	0.046	0.058	0.051	0.047
$T = 1000$	0.053	0.051	0.048	0.064	0.060	0.055	0.083	0.070	0.058

Note: The theoretical size is 5%.

Table 2: Empirical power of t-tests under misspecification

N	50	300	1000	50	300	1000	50	300	1000
Panel A: Test using Moore-Penrose inverse									
	θ_{MKT}			θ_{SMB}			θ_{HML}		
$T = 150$	0.484	0.531	0.533	0.102	0.095	0.088	0.168	0.165	0.148
$T = 300$	0.779	0.687	0.839	0.149	0.527	0.136	0.289	0.687	0.297
$T = 1000$	0.997	1.000	0.875	0.364	0.361	0.697	0.595	0.684	0.815
Panel B: Test using Tikhonov regularized inverse									
$T = 150$	0.483	0.529	0.532	0.097	0.090	0.088	0.151	0.149	0.143
$T = 300$	0.781	0.832	0.840	0.141	0.134	0.133	0.274	0.294	0.290
$T = 1000$	0.998	1.000	1.000	0.362	0.367	0.365	0.606	0.716	0.756

Note: Tests are conducted at the 5% significance level.

5.2 Model comparison tests

This section studies the finite-sample behavior of the pairwise and multiple model-comparison tests. All models within a comparison are estimated with the same regularization parameter; otherwise, differences in penalization could be confounded with differences in pricing performance.

We compare the Fama–French three-factor model (FF3) with a three-factor model denoted q3. The q3 model contains the market factor and the first two factors of q5 from Hou, Mo, Xue, and Zhang (2021): the size factor (ME) and the investment factor (IA). The calibration data, obtained from Lu Zhang's webpage[6], cover January 1967 through December 2024.

Panel A of Table 3 considers the equality of two nonnested SDFs using Proposition 5. The competing models are FF3 and q3. To evaluate the size, we generate a five factor model encompassing the factors of the two competing models and set the return mean μ so that the SDF coefficients are all zero except for the market coefficient; both candidate models nevertheless remain misspecified. The Wald procedure tests the joint significance of the nonmarket coefficients

[6] https://global-q.org/factors.html

in the encompassing five-factor model. The weighted χ^2 procedure instead estimates FF3 and q3 separately and applies (14). For the power experiment, the data-generating process contains all factors from both models without constraint, making the coefficients of the nonoverlapping factors nonzero. In both experiments, γ is selected by two-fold cross-validation applied to the encompassing model. Panel A reveals appreciable finite-sample size distortions even after regularization. The tests are generally conservative at $T = 150$, but overreject at $T = 1000$, particularly the Wald test. The weighted chi-square test exhibits smaller size distortions in these largest-sample designs, although its rejection frequencies also exceed the nominal level. Both procedures gain substantial power with T. The Wald test generally has higher rejection frequencies under the alternative, but part of this difference may reflect its greater tendency to reject under the null. Results for the Moore-Penrose versions of the tests are reported in Table 11 in Appendix C and show severe overrejections for $T = N$.

Panel B examines equality of HJ distances when the competing models imply different SDFs, $y_a \neq y_b$, using the normal pairwise statistic of Proposition 7. For the size experiment, the two misspecified three-factor models share SMB and HML. Their respective market factors equal MKT plus independent Gaussian measurement errors (independent from the returns), each with variance equal to 20% of the market variance. This guarantees that the two models have the same population HJ distance while their SDFs differ. Under the null, the DGP is the uncontaminated FF3 model.

The Tikhonov test is markedly conservative under this null, particularly for smaller T. Its rejection frequencies generally move toward the nominal level as T increases, but remain below 5% in the reported size experiments. Despite this conservativeness, power increases substantially with the time-series sample size and approaches one in several larger-sample designs. The Moore–Penrose test is also conservative in several designs except when $N = T$ where it overrejects (see Table 11 in Appendix C).

For the power experiment, the competitors are a one-factor model based on squared MKT and a two-factor model containing SMB and HML, while the data-generating process is a three-factor model encompassing all factors of the two competing models and with return mean, μ, unconstrained (obtained by calibration). As a result, the SDF coefficients of all the factors are non zero under the alternative. Power rises substantially with T and in most cases with N.

Panel C studies the finite-sample behavior of the comparison test using the Wolak statistic in (16). The size experiments follow the symmetric construction of Panel B, with two FF3 candidate models when $p = 1$ and three FF3 models when $p = 2$. In the power experiments, the data-generating process is again the uncontaminated FF3 model. The benchmark uses SMB and HML. For $p = 1$ it is compared with the squared MKT model; for $p = 2$ the additional competitor is IA. All models are estimated using the regularization parameter selected using the joint model including all factors, ensuring that their HJ distances are comparable. For illustration, when $N = 50$ and $T = 300$, the average regularized squared HJ distances are 0.25 for the benchmark, 0.09 for the square MKT model, and 0.23 for IA.

Like the normal pairwise procedure, the Wolak test is conservative under the null, particularly

in small samples, but it retains substantial power. For $p = 1$ its power exceeds that of the two-sided normal comparison in Panel B, consistent with the one-sided alternative embodied in the Wolak statistic. Increasing the number of comparisons to $p = 2$ generally lowers power, as the procedure must account jointly for two inequalities. A comparison with the Wolak test based on Moore-Penrose inverse in Table 11 shows that the Moore-Penrose version of the test has lower power than the Tikhonov version when $N = T$, and similar power otherwise.

Table 3. Model comparison tests under misspecification (Tikhonov)

N	50	300	1000	50	300	1000	50	300	1000	50	300	1000
	Panel A: Pairwise tests of equality of two SDFs											
	Wald test						Weighted χ^2 test					
T	Size			Power			Size			Power		
150	0.030	0.017	0.010	0.156	0.143	0.120	0.035	0.023	0.016	0.133	0.121	0.101
300	0.058	0.049	0.029	0.421	0.496	0.523	0.056	0.047	0.032	0.313	0.371	0.388
1000	0.147	0.197	0.114	0.888	0.971	0.995	0.107	0.121	0.075	0.697	0.815	0.884
	Panel B: Normal pairwise test of equality of two HJ-distances (with different SDFs)											
T	Size			Power								
150	0.001	0.001	0.000	0.023	0.056	0.013						
300	0.004	0.003	0.001	0.377	0.955	0.981						
1000	0.020	0.021	0.010	0.991	1.000	1.000						
	Panel C: Multiple comparison test (Wolak test)											
	Wolak test ($p = 1$)						Wolak test ($p = 2$)					
T	Size			Power			Size			Power		
150	0.004	0.003	0.000	0.079	0.281	0.230	0.002	0.001	0.000	0.030	0.101	0.085
300	0.011	0.008	0.003	0.546	0.991	0.998	0.006	0.004	0.002	0.319	0.811	0.905
1000	0.033	0.032	0.017	0.996	1.000	1.000	0.026	0.024	0.013	0.977	1.000	1.000

Note: The theoretical size is 5%.

6 Empirical application

For the empirical application, we compare four popular models, namely the Fama and French (1993) model (FF3), the q5 of Hou, Mo, Xue, and Zhang (2021), the nonlinear model of Dittmar (2002), and the Fama and French (2015) model (FF5). The latter adds two new factors to the FF3: the profitability (RMW) and investment (CMA) factors. The q5 model has the following factors: the market, the size factor (ME) and the investment factor (IA), the return-on equity factor (ROE), and the expected growth factor (EG). We also consider the polynomial specification of Dittmar (2002), whose SDF is

$$y = 1 - \theta_{MKT}\left(r_{MKT} - E\left[r_{MKT}\right]\right) - \theta_{mkt,2}\left(r_{MKT}^2 - E\left[r_{MKT}^2\right]\right) - \theta_{mkt,3}\left(r_{MKT}^3 - E\left[r_{MKT}^3\right]\right)$$

The quadratic and cubic market terms are supposed to reflect the preference of agents for higher moments. Without being fully nonparametric, these terms permit to incorporate nonlinearity in the SDF.

For this analysis, we use the large set of test assets put together by Giglio and Xiu (2021)[7]. This set of $N = 647$ portfolios includes 202 equity portfolios from Kenneth French's website, treasury bonds, corporate bonds, currencies, and US equities sorted by a large number of characteristics, see Table 8 in Appendix C. The data span the period from March 1976 to January 2010, so that we have $T = 407$. In this sample, T is quite a bit smaller than N so that regularization is necessary.

Table 4 reports SDF coefficient estimates using a regularization parameter selected separately for each model. In FF3, HML is significant at the 5% level, while the market coefficient is significant at 10% but not at 5%; SMB is not individually significant. In FF5, the HML coefficient is not statistically distinguishable from zero, whereas the other four coefficients are significant at conventional levels. This outcome is in line with the results of Fama and French (2015), who argue that the value factor is redundant as the model with the five factors does not improve upon the model with just four factors without HML. None of the three coefficients in the polynomial market model is individually significant. In q5, the market, size, and expected-growth coefficients are significant, while investment and return-on-equity are not. These are individual coefficient tests and do not establish joint insignificance of groups of factors. For this table, the value of the regularization parameter has been optimized separately for each model, therefore the squared HJ distances $\hat{\delta}^2_\gamma$ are not comparable across the four models because they use different values of γ.

[7]The data was downloaded https://stefanogiglio.org/#data. We thank Dacheng Xiu and Stefano Giglio for letting us use this dataset.

Table 4: SDF parameter estimates

	FF3 model				Nonlinear model		
Factors	θ_{MKT}	θ_{SMB}	θ_{HML}		θ_{MKT}	$\theta_{MKT\hat{}2}$	$\theta_{MKT\hat{}3}$
SDF	0.037*	0.030	0.066**		-0.001	0.003	0.000
t-ratio	1.895	1.518	2.103		-0.042	0.740	0.985
p-value	0.058	0.129	0.036		0.966	0.459	0.325
$\hat{\delta}^2_\gamma$	0.1552	γ	0.00347		1.0378	γ	0.00004
	FF5 model						
Factors	θ_{MKT}	θ_{SMB}	θ_{HML}	θ_{RMW}	θ_{CMA}		
SDF	0.060***	0.063***	-0.025	0.119***	0.158***		
t-ratio	3.394	2.955	-0.615	3.418	3.537		
p-value	0.001	0.003	0.539	0.001	0.000		
$\hat{\delta}^2_\gamma$	0.5361	γ	0.00015				
	q5 model						
Factors	θ_{MKT}	θ_{ME}	θ_{IA}	θ_{ROE}	θ_{EG}		
SDF	0.079***	0.084***	0.033	-0.052	0.411***		
t-ratio	3.996	3.853	0.610	-1.486	4.391		
p-value	0.000	0.000	0.542	0.137	0.000		
$\hat{\delta}^2_\gamma$	0.066	γ	0.00347				

***,**,* indicate that the null hypothesis of zero SDF parameter is rejected at the 1%, 5%, and 10% levels. The standard errors are robust to both misspecification and autocorrelation (they use Newey-West HAC estimator).

We also examine whether the models exhibit different explanatory power. To compare models, it is essential to keep the same level of penalization. We use a penalization level $\gamma = 0.000035$ which is optimal for a model including the factors of FF5 and q5 together. For this value of γ, the regularized squared HJ distances of the different models are given by 0.808 (q5), 0.925 (FF5), 1.009 (FF3), and 1.038 (nonlinear).

First, we perform pairwise comparison tests. Table 5 reports the results. For the nested models FF3 and FF5, we use a Wald test of the joint significance of the additional factors RMW and CMA. The null hypothesis is strongly rejected, indicating that FF5 provides a better fit than FF3.

For all other pairs of models, we apply the asymptotic normal test[8] of Proposition 7, based on differences in regularized squared HJ distances. This test requires that the competing models have distinct SDFs and are misspecified. The first condition is satisfied since the models differ in their factor structures; in particular, equality of SDFs would require that all models reduce to a function of the market factor alone, which is not supported by the data as evidenced in Table 4. The second condition is also plausible in this setting, as all models considered are parsimonious and do not fully span the cross-section of returns.

The results can be summarized as follows. The pairwise tests show that both FF5 and q5 outperform FF3 and the polynomial model at the 5% level. The difference between q5 and FF5

[8]The table reports the squared statistics, so that the p-values are computed using the Chi-square(1) distribution.

is significant at 10%, but not at 5% ($p = 0.068$). The difference between FF3 and the polynomial model is not statistically significant ($p = 0.348$).

Table 5: Pairwise HJ-distance comparison tests

	FF3	Nonlinear	q5
FF5	20.692	8.468	3.329
	(0.000)	(0.004)	(0.068)
FF3		0.883	9.304
		(0.348)	(0.002)
Nonlinear			16.242
			(0.000)

P-values are in parentheses. Covariance matrices use Newey-West HAC estimator.

Finally, to determine which model is best, we implement the multiple model comparison test of Wolak (1989). The null hypothesis of this test is that the squared HJ-distance of a benchmark model is smaller than those of two or more alternative models. In turn, we consider each model as a benchmark and compare it against the others. For each test, we remove alternative models nested by the benchmark model as H_0 is verified by construction (the benchmark has already lower pricing errors (HJ)). Within the remaining alternatives, we also remove models nested by others. Finally, we remove alternative models that nest the benchmark as the asymptotic normality assumption on d_i does not hold under the null of $d_i = 0$. For example, to compare the FF3 against the other models, we remove FF5 from the alternatives.

Table 6 tests whether each benchmark has a regularized squared HJ distance no greater than those of its listed alternatives. The null is rejected for FF3 and the polynomial market model at the 5% level, indicating that at least one alternative has a smaller population distance. For FF5, the null is rejected at 10% but not at 5% ($p = 0.066$). Finally, the null hypothesis cannot be rejected for q5 at conventional significance levels. In conclusion, q5 has the smallest pricing error, significantly outperforms FF3 and the nonlinear model at 5%, while the q5-FF5 difference is significant only at 10%.

Table 6: Multiple model comparison tests

Benchmark	Alternatives	$\hat{\delta}^2_\gamma$	LR_γ	p-value
FF3	q5 and nonlinear	1.009	9.304	0.003
FF5	q5 and nonlinear	0.925	3.329	0.066
Nonlinear	FF5 and q5	1.038	21.784	0.000
q5	FF5 and nonlinear	0.808	0.000	0.602

Note: Covariance matrices use Newey-West HAC estimator.

7 Conclusion

This paper develops a framework for evaluating and comparing asset pricing models in settings with a large number of test assets. We introduce regularized versions of the Hansen–Jagannathan

distance based on Tikhonov and ridge schemes to stabilize the inversion of the return covariance matrix. This allows the SDF parameters to be estimated even when the number of assets exceeds the number of time observations. In addition to improving numerical stability, these regularizations admit an economic interpretation in terms of allowing controlled pricing errors, which can be viewed as arising from transaction costs or market frictions. The regularization parameter is selected using a data-driven procedure based on out-of-sample performance.

Building on this framework, we propose a set of tests for comparing competing asset pricing models. Our approach focuses on differences in regularized HJ distances, providing a natural way to rank models under misspecification. For linear SDF specifications, the resulting statistics admit simple analytical forms and are straightforward to implement. Monte Carlo simulations show that regularization stabilizes inference near $N = T$ while retaining comparable or higher power than the Moore-Penrose inverse in many designs.

Several directions for future research remain. First, it would be useful to develop inference procedures that are robust to weak factors, i.e., factors with low correlation with returns. Second, our analysis is restricted to linear SDFs; extending the framework to nonlinear models, for example along the lines of Gospodinov, Kan, and Robotti (2013), would be of clear interest. Third, since the HJ distance is based on second moments, alternative discrepancy measures incorporating higher moments, as in Almeida and Garcia (2012), could provide additional insights. Finally, our empirical analysis relies on a set of test portfolios. Expanding the cross-section using transformations of firm characteristics, as in Didisheim, Ke, Kelly, and Malamud (2024), would generate a much richer set of moment conditions and provide a natural setting where regularization becomes even more critical.

8 Appendix A: Short review on regularization

In this section, we introduce some notations and results on norms and operators. Then, we explain how Ridge and Tikhonov regularizations act on eigenvalues to stabilize the inverse of the covariance matrix. Finally, we compare the two regularizations.

Definition 2.

1. *For a vector* $v \in \mathbb{R}^N$, $\|v\| = \sqrt{v'v}$ *is the euclidean norm and* $\|v\|_N = \sqrt{v'v/N}$.
2. *For a* $N \times N$ *matrix* A, *the operator norm of* A *is* $\|A\|_{op,N} = \sup_{\|\phi\|_N=1} \|A\phi\|_N$. *Therefore, for any vector* $u \in \mathbb{R}^N$, $\|Au\|_N \leq \|A\|_{op,N} \parallel \|u\|_N$.
3. *For an arbitrary* $N \times K$ *matrix* V, *the operator norm of* V *is* $\|V\|_{op,N} = \sup_{\|\phi\|=1} \|V\phi\|_N$ *where* $\parallel \phi \parallel$ *is the euclidean norm of* ϕ.
4. *We define the Frobenius norm as* $\|V\|_F = \left(tr(V'V)\right)^{\frac{1}{2}}$. *We have* $\|V\| \leq \|V\|_F$ *and for any vector* $u \in \mathbb{R}^N$, $\|Vu\| \leq \|V\|_F \|u\|$.
5. *If* $\|v\|_N < \infty$ *when* $N \to \infty$, *we note* $\|v\|_\infty$ *its limit value.*

Operators and spectral decomposition

Recall that $\mathbb{R}^N$ is endowed by the norm $\parallel \phi \parallel_N^2 = \phi'\phi/N$. Let $\mathbb{R}^T$ be endowed with norm $\parallel v \parallel_T^2 = \frac{v'v}{T}$ induced by inner product $\langle v_1, v_2 \rangle_T = \frac{v_1'v_2}{T}$. Let H be the operator from $\mathbb{R}^N$ to $\mathbb{R}^T$ defined by $H\phi = \frac{\bar{R}\phi}{N}$ and its adjoint H^* is the operator from $\mathbb{R}^T$ to $\mathbb{R}^N$ such that $H^*v = \frac{\bar{R}'v}{T}$. With that, we have the operator $H^*H\phi = \frac{\bar{R}'\bar{R}}{NT}\phi = \hat{\Sigma}\phi$ which goes from $\mathbb{R}^N$ to $\mathbb{R}^N$. Let $\left\{\sqrt{\hat{\lambda}_j}, \hat{\phi}_j, \hat{v}_j\right\}$, $j = 1, 2, ...$ be the singular value decomposition of H such that $H\widehat{\phi}_j = \sqrt{\hat{\lambda}_j}\hat{v}_j$ and $H^*\widehat{v}_j = \sqrt{\hat{\lambda}_j}\hat{\phi}_j$. Note that $\left\{\hat{\lambda}_j, \hat{\phi}_j\right\}$, $j = 1, 2, ..., min(N, T-1)$ are the non zero eigenvalues and eigenvectors of $\hat{\Sigma}$.

Regularity space

Notice that as $\hat{\beta} = \frac{\bar{R}'P_F}{T}$, where $P_F = \bar{F}(\frac{\bar{F}'\bar{F}}{T})^{-1} = \begin{bmatrix} P_F^1 & \cdots & P_F^K \end{bmatrix}$. Then, $\hat{\beta}$ can be rewritten as $\hat{\beta} = \begin{bmatrix} H^*P_F^1 & \cdots & H^*P_F^K \end{bmatrix}$. Therefore, $\hat{\beta}_k \in \mathcal{R}(H^*), k = 1, \cdots, K$. From Proposition 6.2 of Carrasco, Florens, and Renault (2007), $\mathcal{R}(H^*) = \mathcal{H}(\hat{\Sigma}) = \mathcal{R}(\hat{\Sigma}^{\frac{1}{2}})$ where $\mathcal{H}(\hat{\Sigma})$ is the Reproducing Kernel Hilbert Space (RKHS) of $\hat{\Sigma}$. In Assumption 3(i), we impose the stronger assumption $\beta \in \mathcal{R}\left(\Sigma^2\right)$.

Relationship between Ridge and Tikhonov

For Ridge regularization, $\widehat{\Sigma}^{-1}$ is replaced by $\widehat{\Sigma}_R^{-1} = \left(\widehat{\Sigma} + \gamma I_N\right)^{-1}$. So this regularization method replaces the explosive eigenvalues of $\widehat{\Sigma}^{-1}$, $\frac{1}{\widehat{\lambda}_j}$, $j = 1, \cdots, N$ by $\frac{1}{\widehat{\lambda}_j + \gamma}$ which is bounded by $1/\gamma$. On the other hand, for Tikhonov regularization, $\widehat{\Sigma}^{-1}$ is replaced by $\widehat{\Sigma}_K^{-1} = \left(\widehat{\Sigma}^2 + \gamma I_N\right)^{-1} \widehat{\Sigma}$. Therefore, the eigenvalues $\frac{1}{\widehat{\lambda}_j}$ are replaced by $\frac{\widehat{\lambda}_j}{\widehat{\lambda}_j^2 + \gamma}$ which is bounded by $\widehat{\lambda}_j/\gamma$ and is exactly zero when $\widehat{\lambda}_j = 0$. It may seem like quite different regularization techniques, however the difference may be minor in some relevant cases as discussed in the online Appendix.

9 Appendix B: Penalization on the Lagrange multiplier

In Section 3.1, we provided an interpretation of the regularization in terms of a penalization on the norm of the pricing error. Here, we show that it involves a penalization on the Lagrange multipliers. To get insights on the mechanism behind the penalization in (6), we consider the dual of the optimization problem (6) (see the proof of Proposition 1 in the online appendix):

$$\delta_R^2 = \underset{\nu_1 \in \mathbb{R}^N, \nu_2 \in \mathbb{R}}{max} E\left\{2y\nu_1^{'}\frac{r}{N} - \frac{\nu_1^{'} r r^{'} \nu_1}{N^2} - \nu_2^2 - 2\frac{\nu_1^{'} r \nu_2}{N} - \frac{\gamma}{N} \parallel \nu_1 \parallel^2\right\}. \tag{18}$$

Remark that Equation (18) is the penalized version of the dual of (5) with a penalization applied to ν_1, the Lagrange multiplier associated with $E(mr)$.

The first order condition with respect to ν_1 gives

$$\alpha - \frac{E(rr')}{N}\nu_1 - \nu_2\mu_2 - \gamma\nu_1 = 0. \tag{19}$$

The first order condition with respect to ν_2 gives $\nu_2 = -\nu_1'\mu_2/N$. Replacing ν_2 in Equation (19) gives

$$\nu_1 = (\Sigma + \gamma I_N)^{-1}\alpha.$$

And the solution is the ridge regularized Hansen-Jagannathan distance

$$\delta_R^2 = \frac{\alpha'(\Sigma + \gamma I_N)^{-1}\alpha}{N}.$$

For Tikhonov regularization, the dual is given by

$$\delta_K^2 = \underset{\nu_1 \in \mathbb{R}^N, \nu_2 \in \mathbb{R}}{max} E\left\{2y\nu_1^{'}\frac{r}{N} - \frac{\nu_1^{'} r r^{'} \nu_1}{N^2} - \nu_2^2 - 2\frac{\nu_1^{'} r \nu_2}{N} - \frac{\gamma}{N}\nu_1'\Sigma^{\dagger}\nu_1\right\} \tag{20}$$

where $\Sigma^{\dagger}$ is the Moore-Penrose inverse of Σ when Σ is singular, and is Σ^{-1} when Σ is positive

definite. Solving in ν_1 and ν_2 yields $\nu_1 = [\Sigma^2 + \gamma I_N]^{-1}\Sigma E[ry] = [\Sigma^2 + \gamma I_N]^{-1}\Sigma\alpha$, $\nu_2 = -\nu_1'\mu_2/N$, and

$$\delta_K^2 = \frac{\alpha'[\Sigma^2 + \gamma I_N]^{-1}\Sigma\alpha}{N}.$$

So for both regularizations, the penalization on $\|E(mr)\|_N^2$ translates into a penalization on the Lagrange multiplier ν_1 and hence relaxes the pricing condition.

10 Appendix C: Extra information for the simulations and empirical application

Algorithm for the simulations

To obtain the simulations of Tables 1 and 3 (panel A), we must generate a factor model, so that the model is misspecified and the jth SDF coefficient $\theta_{HJj} = 0$ in population.

We follow the steps:

1. Calibrate V_{11}, β, μ_2 and Σ_ϵ from the data and define

$$V_{21} = \beta V_{11},\ V_{22} = \beta V_{11}\beta' + \Sigma_\varepsilon.$$

2. The pseudo true value is

$$\theta^* = \left(V_{12}V_{22}^{-1}V_{21}\right)^{-1} V_{12}V_{22}^{-1}\mu_2.$$

   and the pricing error is

$$\alpha^* = \mu_2 - V_{21}\theta^*.$$

   The first order condition implies that $V_{21}'V_{22}^{-1}\alpha^* = 0$.

   Now set only the tested component to the target parameter to zero

$$\theta^{null} = (\theta_1^*, ..., 0, ...\theta_K^*)'.$$

3. Construct the returns mean as

$$\mu_2^{null} = \alpha^* + V_{21}\theta^{null}.$$

   Because $V_{21}'V_{22}^{-1}\alpha^* = 0$, the pseudo-true value associated with μ_2^{null} is exactly θ^{null}.

4. Generate

$$r_t = \mu_2^{null} + \beta\widetilde{f}_t + \varepsilon_t.$$

This construction keeps the original β, preserves identification, preserves the original misspecification error vector α^*, and changes only the tested pseudo-true value by replacing μ_2 by μ_2^{null}.

Table 7: List of test portfolios for simulations

25 portfolios formed on size and book-to-market
17 industry portfolios
25 portfolios formed on operating profitability and investment
25 portfolios formed on size and variance
35 portfolios formed on size and net share issues
25 portfolios formed on size and accruals
25 portfolios formed on size and market beta
25 portfolios formed on size and momentum

Table 8: List of test portfolios for the empirical application

202 portfolios from Table 7
10 maturity-sorted government bond portfolios
10 corporate bond portfolios sorted on yield spread
6 currency portfolios sorted on interest rate differentials
6 currency portfolios sorted on currency momentum
413 anomaly portfolios based on 103 characteristics

For more details, see Section III.1. of the online appendix of Giglio and Xiu (2021).

Table 9: Regularization parameter γ used for Table 1

N	50	300	1000
		θ_{MKT}	
$T = 150$	0.006 (0.006)	0.006 (0.006)	0.006 (0.006)
$T = 300$	0.004 (0.004)	0.003 (0.004)	0.003 (0.004)
$T = 1000$	0.002 (0.001)	0.001 (0.001)	0.001 (0.001)
		θ_{SMB}	
$T = 150$	0.006 (0.006)	0.006 (0.006)	0.006 (0.006)
$T = 300$	0.003 (0.004)	0.003 (0.004)	0.003 (0.004)
$T = 1000$	0.002 (0.001)	0.001 (0.001)	0.001 (0.001)
		θ_{HML}	
$T = 150$	0.006 (0.006)	0.006 (0.006)	0.006 (0.006)
$T = 300$	0.003 (0.004)	0.003 (0.004)	0.003 (0.004)
$T = 1000$	0.002 (0.001)	0.001 (0.001)	0.001 (0.001)

Note: The values in the cells correspond to the mean of γ over simulations and its standard deviation in parentheses.

Table 10: Regularization parameter used for Table 2

	N=50	N=300	N=1000
T=150	0.0059 (0.0064)	0.0061 (0.0064)	0.0063 (0.0065)
T=300	0.0035 (0.0036)	0.0032 (0.0035)	0.0034 (0.0036)
T=1000	0.0015 (0.0013)	0.0012 (0.0013)	0.0012 (0.0013)

Notes: The values in the cells correspond to the mean of γ over simulations and its standard deviation in parentheses.

Table 11. Model comparison tests under misspecification (Moore-Penrose)

N	50	300	1000	50	300	1000	50	300	1000	50	300	1000
	Panel A: Pairwise tests of equality of two SDFs											
	Wald test						Weighted χ^2 test					
T	Size			Power			Size			Power		
150	0.044	0.072	0.012	0.248	0.301	0.145	0.045	0.054	0.020	0.192	0.208	0.115
300	0.043	0.989	0.045	0.497	0.990	0.609	0.047	0.903	0.041	0.359	0.916	0.381
1000	0.049	0.074	0.998	0.898	0.950	0.999	0.049	0.063	0.968	0.708	0.740	0.971
	Panel B: Normal pairwise test of equality of two HJ-distances (with different SDFs)											
T	Size			Power								
150	0.004	0.045	0.000	0.022	0.058	0.009						
300	0.007	0.794	0.047	0.375	0.719	0.976						
1000	0.026	0.051	0.928	0.991	1.000	0.904						
	Panel C: Multiple comparison test (Wolak test)											
	Wolak test ($p = 1$)						Wolak test ($p = 2$)					
T	Size			Power			Size			Power		
150	0.009	0.055	0.001	0.078	0.279	0.210	0.005	0.044	0.000	0.030	0.110	0.083
300	0.017	0.416	0.055	0.545	0.585	0.997	0.010	0.584	0.047	0.318	0.635	0.878
1000	0.037	0.058	0.480	0.996	1.000	0.777	0.028	0.057	0.630	0.977	1.000	0.819

Note: The theoretical size is 5%.

References

Almeida, C., and R. Garcia (2012): "Assessing misspecified asset pricing models with empirical likelihood estimators," *Journal of Econometrics*, 170(2), 519–537.

——— (2017): "Economic Implications of Nonlinear Pricing Kernels," *Management Science*, 63, 3361–3380.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

Antoine, B., K. Proulx, and E. Renault (2020): "Pseudo-True SDFs in Conditional Asset Pricing Models," *Journal of Financial Econometrics*, 18(4), 656–714.

Babii, A., E. Ghysels, and J. Striaukas (2024): "High-Dimensional Granger Causality Tests with an Application to VIX and News," *Journal of Financial Econometrics*, 22(3), 605–635.

Barillas, F., and J. Shanken (2018): "Comparing Asset Pricing Models," *The Journal of Finance*, 73(2), 715–754.

Barras, L. (2019): "A large-scale approach for evaluating asset pricing models," *Journal of Financial Economics*, 134, 549–569.

Bauschke, H. H., and P. L. Combettes (2017): *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books in Mathematics. Springer.

Bernstein, D. S. (2009): *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press.

Borwein, J. M. (1993): "On the failure of maximum entropy reconstruction of Fredholm equations and other infinite systems," *Mathematical Programming*, 61, 251–261.

Borwein, J. M., and A. S. Lewis (1992): "Partially finite convex programming, Part I: Quasi relative interiors and duality theory," *Mathematical Programming*, 57(1-3), 15–48.

Brodie, J., I. Daubechies, C. De Mol, and I. Loris (2009): "Sparse and Stable Markowitz Portfolios," *Proceedings of the National Academy of Science*, 106, 12267–12272.

Callot, L., M. Caner, O. Onder, and E. Ulasan (2021): "A Nodewise Regression Approach to Estimating Large Portfolios," *Journal of Business and Economic Statistics*, 39, 520–531.

Carrasco, M. (2012): "A regularization approach to the many instruments problem," *Journal of Econometrics*, 170(2), 383–398.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Chapter 77 Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, vol. 6, pp. 5633–5751. Elsevier.

Carrasco, M., N. Kone, and N. Noumon (2019): "Optimal Portfolio Selection using Regularization," Working paper.

Carrasco, M., and B. Rossi (2016): "In-Sample Inference and Forecasting in Misspecified Factor Models," *Journal of Business & Economic Statistics*, 34(3), 313–338.

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

Chang, J., S. X. Chen, and X. Chen (2015): "High dimensional generalized empirical likelihood for moment restrictions with dependent data," *Journal of Econometrics*, 185, 283–304.

Chernov, M., B. T. Kelly, S. Malamud, and J. Schwab (2025): "A Test of the Efficiency of a Given Portfolio in High Dimensions," Working Paper 33565, National Bureau of Economic Research.

Cochrane, J. H. (1996): "A cross-sectional test of an investment-based asset pricing," *Journal of Political Economy*, 104(3), 572.

——— (2005): *Asset pricing*. Princeton University Press, Princeton, N.J.

Daniel, K., and S. Titman (2012): "Testing Factor-Model Explanations of Market Anomalies," *Critical Finance Review*, 1(1), 103–139, Publisher: Now Publishers, Inc.

Davydov, Y. A. (1968): "Convergence of Distributions Generated by Stationary Stochastic Processes," *Theory of Probability & Its Applications*, 13(4), 691–696.

Didisheim, A., S. B. Ke, B. T. Kelly, and S. Malamud (2024): "APT or AIPT? The Surprising Dominance of Large Factor Models," Working Paper 33012, National Bureau of Economic Research.

Dittmar, R. F. (2002): "Nonlinear Pricing Kernels, Kurtosis Preference, and Evidence from the Cross Section of Equity Returns," *The Journal of Finance*, 57(1), 369–403.

D'Hondt, C., R. De Winne, E. Ghysels, and S. Raymond (2020): "Artificial Intelligence Alter Egos: Who might benefit from robo-investing?," *Journal of Empirical Finance*, 59, 278–299.

Fama, E. F., and K. R. French (1992): "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, 47(2), 427–465.

——— (1993): "Common risk factors in the returns on stocks and bonds," *Journal of Financial Economics*, 33(1), 3–56.

——— (2015): "A five-factor asset pricing model," *Journal of Financial Economics*, 116(1), 1–22.

Fan, J., Y. Fan, and J. Lv (2008): "High dimensional covariance matrix estimation using a factor model," *Journal of Econometrics*, 147(1), 186–197.

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

Ferson, W. (2019): "Empirical asset pricing: Models and methods," Publisher: MIT Press.

Francq, C., and J.-M. Zakoïan (2005): "A Central Limit Theorem for Mixing Triangular Arrays of Variables Whose Dependence Is Allowed to Grow with the Sample Size," *Econometric Theory*, 21(6), 1165–1171.

Gagliardini, P., E. Ossola, and O. Scaillet (2016): "Time-varying risk premium in large cross-sectional equity data sets," *Econometrica*, 84, 985–1046.

Gagliardini, P., and D. Ronchetti (2020): "Comparing Asset Pricing Models by the Conditional Hansen-Jagannathan Distance," *Journal of Financial Econometrics*, 18(2), 333–394.

Giglio, S., and D. Xiu (2021): "Asset Pricing with Omitted Factors," *Journal of Political Economy*, 129(7), 1947–1990.

Gospodinov, N., R. Kan, and C. Robotti (2013): "Chi-squared tests for evaluation and comparison of asset pricing models," *Journal of Econometrics*, 173(1), 108–125.

——— (2016): "On the properties of the constrained Hansen-Jagannathan distance," *Journal of Empirical Finance*, 36, 121–150.

Goyal, A. (2012): "Empirical cross-sectional asset pricing: a survey," *Financial Markets and Portfolio Management*, 26(1), 3–38.

Hall, A., and D. Pelletier (2011): "Nonnested testing in models estimated via generalized method of moments," *Econometric Theory*, 27, 443–456.

Hall, A. R., and A. Inoue (2003): "The large sample behaviour of the generalized method of moments estimator in misspecified models," *Journal of Econometrics*, 114(2), 361–394.

Hansen, L. P., J. Heaton, and E. G. J. Luttmer (1995): "Econometric Evaluation of Asset Pricing Models," *The Review of Financial Studies*, 8(2), 237–274.

Hansen, L. P., and R. Jagannathan (1997): "Assessing Specification Errors in Stochastic Discount Factor Models," *The Journal of Finance*, 52(2), 557–590.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2022): "Surprises in high-dimensional ridgeless least squares interpolation," *The Annals of Statistics*, 50, 949Ů986.

Hodrick, R. J., and X. Zhang (2001): "Evaluating the specification errors of asset pricing models," *Journal of Financial Economics*, 62(2), 327–376.

Hou, K., H. Mo, C. Xue, and L. Zhang (2021): "An Augmented q-Factor Model with Expected Growth*," *Review of Finance*, 25(1), 1–41.

Kan, R., and C. Robotti (2008): "Specification tests of asset pricing models using excess returns," *Journal of Empirical Finance*, 15(5), 816–838.

——— (2009): "Model Comparison Using the Hansen-Jagannathan Distance," *The Review of Financial Studies*, 22(9), 3449–3490.

Kleibergen, F., and Z. Zhan (2020): "Robust inference for consumption-based asset pricing," *Journal of Finance*, 75, 507–5502.

Korsaye, S. A., A. Quaini, and F. Trojani (2025): "Smart Stochastic Discount Factors," *Management Science*, 72, 6580–6602.

Kozak, S., S. Nagel, and S. Santosh (2020): "Shrinking the cross-section," *Journal of Financial Economics*, 135, 271–292.

Kress, R. (2014): *Linear Integral Equations*, Applied Mathematical Sciences. Springer-Verlag, New York.

Kudo, A. (1987): "An exact test for multiple inequality and equality constraints in the linear regression model," *Journal of the American Statistical Association*, 82, 782–793.

Ledoit, O., and M. Wolf (2003): "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, 10(5), 603–621.

——— (2017): "Nonlinear Shrinkage of the Covariance Matrix for Portfolio Selection: Markowitz Meets Goldilocks," *The Review of Financial Studies*, 30, 4349–4388.

Lewellen, J., S. Nagel, and J. Shanken (2010): "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, 96(2), 175–194.

Li, J., and Z. Liao (2020): "Uniform nonparametric inference for time series," *Journal of Econometrics*, 219(1), 38–51.

Liao, Y., X. Ma, A. Neuhierl, and Z. Shi (2026): "Benign Overfitting in Economic Forecasting via Noise Regularization," Working paper.

Ludvigson, S. C. (2013): "Advances in Consumption-Based Asset Pricing: Empirical Tests," in *Handbook of the Economics of Finance*, vol. 2, pp. 799–906. Elsevier.

Luenberger, D. G. (1969): *Optimization by vector space methods*. N.Y., Wiley, N.Y.

Lönn, R., and P. Schotman (2024): "Empirical Asset Pricing with Many Test Assets," *Journal of Financial Econometrics*, 22, 1236–1263.

Manresa, E., F. Panaranda, and E. Sentana (2023): "Empirical evaluation of overspecified asset pricing models," *Journal of Financial Economics*, 147, 338–351.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219–255.

Pesaran, M. H., and R. Smith (2021): "Factor Strengths, Pricing Errors, and Estimation of Risk Premia," *CESifo working paper*.

——— (2023): "Arbitrage pricing theory, the stochastic discount factor and estimation of risk premia from portfolios," *Econometrics and Statistics*, 26, 17–30.

Phillips, P. C. B., and H. R. Moon (1999): "Linear Regression Limit Theory for Nonstationary Panel Data," *Econometrica*, 67(5), 1057–1111.

Qiu, C., and T. Otsu (2022): "Information theoretic approach to high-dimensional multiplicative models: Stochastic discount factor and treatment effect," *Quantitative Economics*, 13, 63–94.

Raponi, V., C. Robotti, and P. Zaffaroni (2020): "Testing Beta-Pricing Models Using Large Cross-Sections," *The Review of Financial Studies*, 33(6), 2796–2842.

Ren, Y., and K. Shimotsu (2009): "Improvement in finite sample properties of the HansenŰ-Jagannathan distance test," *Journal of Empirical Finance*, 16(3), 483–506.

Rio, E. (1993): "Covariance inequalities for strongly mixing processes," in *Annales de l'IHP Probabilités et statistiques*, vol. 29, pp. 587–597. Issue: 4.

Schennach, S., and D. Wilhelm (2017): "A Simple Parametric Model Selection Test," *Journal of the American Statistical Association*, 112, 1663–1674.

Shi, X. (2015): "A nondegenerate Vuong test," *Quantitative Economics*, 6, 85–121.

Vuong, Q. (1989): "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57, 307–333.

WOLAK, F. A. (1963): "A multivariate analogue of the one-sided test," *Biometrika*, 50, 403–418.

——— (1989): "Testing inequality constraints in linear econometric models," *Journal of Econometrics*, 41(2), 205–235.

ZELLNER, A. (1971): *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York.

11 Online Appendix for "Hansen-Jagannathan distance with many assets" by Marine Carrasco and Cheikh Nokho

11.1 Comparison between Ridge and Tikhonov

In this section, we show that Ridge and Tikhonov are very close regularization techniques when applied to an element of the RKHS of $\hat{\Sigma}$. In the formulas of $\widehat{\delta}_{\gamma}^{2}$ and $\hat{\theta}$, $\widehat{\Sigma}^{-1}$ is applied to either $\hat{\beta} = \dfrac{\overline{R}'\overline{F}}{T}\left(\dfrac{\overline{F}'\overline{F}}{T}\right)^{-1}$ or $\widehat{\mu}_2$. $\hat{\beta}$ is an element of the RKHS of $\hat{\Sigma}$. This yields simplification as explained below.

Let the $N \times 1$ vector v be an element of the RKHS of $\hat{\Sigma}$, i.e. there exists a $T \times 1$ vector w such that $v = \overline{R}' w/T$.

1. **Ridge regularization**

$$\begin{aligned}\widehat{\Sigma}_R^{-1} v &= \sum_j \frac{1}{\hat{\lambda}_j + \gamma} \left\langle v, \widehat{\phi}_j \right\rangle_N \widehat{\phi}_j = \sum_j \frac{1}{\hat{\lambda}_j + \gamma} \left\langle \overline{R}' w, \widehat{\phi}_j \right\rangle_N \widehat{\phi}_j \\ &= \sum_j \frac{\sqrt{\hat{\lambda}_j}}{\hat{\lambda}_j + \gamma} \langle w, \widehat{v}_j \rangle_T \widehat{\phi}_j = \sum_j \frac{q(\gamma, \widehat{\lambda}_j)}{\sqrt{\hat{\lambda}_j}} \langle w, \widehat{v}_j \rangle_T \widehat{\phi}_j\end{aligned}$$

where $q(\gamma, \mu) = \dfrac{\mu}{\gamma + \mu}$.

2. **Tikhonov regularization**

$$\begin{aligned}\widehat{\Sigma}_K^{-1} v &= \sum_j \frac{\hat{\lambda}_j}{\hat{\lambda}_j^2 + \gamma} \left\langle v, \widehat{\phi}_j \right\rangle_N \widehat{\phi}_j = \sum_j \frac{\hat{\lambda}_j}{\hat{\lambda}_j^2 + \gamma} \left\langle \overline{R}' w, \widehat{\phi}_j \right\rangle_N \widehat{\phi}_j \\ &= \sum_j \frac{\sqrt{\hat{\lambda}_j}\hat{\lambda}_j}{\hat{\lambda}_j^2 + \gamma} \langle w, \widehat{v}_j \rangle_T \widehat{\phi}_j = \sum_j \frac{q(\gamma, \widehat{\lambda}_j^2)}{\sqrt{\hat{\lambda}_j}} \langle w, \widehat{v}_j \rangle_T \widehat{\phi}_j.\end{aligned}$$

Therefore, Tikhonov and Ridge regularizations have the same formula except for $q(\gamma, \widehat{\lambda}_j^2)$ replaced by $q(\gamma, \widehat{\lambda}_j)$. So both regularizations give basically the same results when applied to an element of the RKHS of $\hat{\Sigma}$.

11.2 Proof of lemma 1

1. We have

$$\begin{aligned} V_{22} &= E\left[(r_t-\mu_2)(r_t-\mu_2)'\right] \\ &= E\left[\left(\beta\tilde{f}_t+\epsilon_t\right)\left(\beta\tilde{f}_t+\epsilon_t\right)'\right] \\ &= \beta E\left(\tilde{f}_t\tilde{f}_t'\right)\beta'+E\left(\epsilon_t\epsilon_t'\right). \end{aligned}$$

Therefore,

$$tr(\Sigma)=tr\left(\frac{\beta V_{11}\beta'}{N}+\frac{\Sigma_\epsilon}{N}\right).$$

From Assumption 1, $tr(\Sigma_\epsilon)/N=O(1)$, $tr(\beta'\beta)/N=tr(\beta\beta')/N=\sum_i \beta_i'\beta_i/N=O(1)$. Hence,

$$\frac{tr(\beta V_{11}\beta')}{N}=\frac{tr(V_{11}\beta'\beta)}{N}\leq\frac{|trV_{11}|\,|tr(\beta'\beta)|}{N}$$

where we used the fact $|tr(AB)|\leq trAtrB$ when A and B are positive semi-definite matrices (see Bernstein (2009)). As a result $tr(\Sigma)=O(1)$.

2. We have

$$E\|r_t\|_N^2=E(r_t'r_t)/N=tr(\Sigma)+\mu_2'\mu_2/N.$$

Using $\mu_2=\alpha+\beta\lambda$, we obtain

$$\frac{\mu_2'\mu_2}{N}=\frac{\alpha'\alpha}{N}+2\frac{(\beta\lambda)'\alpha}{N}+\frac{\lambda'\beta'\beta\lambda}{N}.$$

$$\frac{\lambda'\beta'\beta\lambda}{N}=\frac{tr(\beta'\beta\lambda\lambda')}{N}\leq\frac{tr(\beta'\beta)}{N}tr(\lambda\lambda')=O(1)$$

by Assumption 1. Moreover, using Cauchy-Schwarz

$$\begin{aligned} \frac{\left|(\beta\lambda)'\alpha\right|}{N} &\leq \sqrt{\frac{(\beta\lambda)'(\beta\lambda)}{N}}\sqrt{\frac{\alpha'\alpha}{N}} \\ &= \sqrt{\frac{tr(\beta'\beta\lambda\lambda')}{N}}\sqrt{\frac{\alpha'\alpha}{N}} \\ &\leq \sqrt{\frac{tr(\beta'\beta)}{N}}\sqrt{tr(\lambda\lambda')}\sqrt{\frac{\alpha'\alpha}{N}} \\ &= O(1). \end{aligned}$$

Therefore, $\mu_2'\mu_2/N=O(1)$ and hence $E\|r_t\|_N^2=O(1)$.

11.3 Proof of Proposition 1

We transform the primal problem to be able to use the Fenchel-Rockafellar Duality, see Chapter 15 of Bauschke and Combettes (2017) or Borwein and Lewis (1992) as well as Korsaye, Quaini, and Trojani (2025).

Let X be the $(N+1)$ vector defined as $X = \begin{bmatrix} \frac{2r'}{N} & 2 \end{bmatrix}'$.

Let $f_y : L^2 \to \mathbb{R}$ be the function defined by $f_y(x) = E[(x-y)^2]$ and $A : L^2 \to \mathbb{R}^{N+1}$ be the operator such that $A(m) = E[mX]$.

Let $g : \mathbb{R}^{N+1} \to (-\infty, +\infty]$ be defined by $g(x) = h(x_1) + \chi_{\{2\}}(x_{-1})$ where $x = (x_1', x_{-1})' \in \mathbb{R}^N \times \mathbb{R}$, $\chi_{\{2\}}$ is

$$\chi_{\{2\}}(x) = \begin{cases} 0 & \text{if } x = 2, \\ \infty & \text{otherwise,} \end{cases}$$

and $h(x) = \dfrac{N}{4\gamma} \|x\|^2$.

First, we prove the result for Ridge regularization. Problem (6) can be rewritten as

$$\delta_R^2 = \inf_{m \in L^2} \{ f_y(m) + g(A(m)) \}.$$

The functions f_y and g are convex as $x \mapsto x^2$ is convex. Moreover, A is a linear bounded function. From Theorem 4.2 of Borwein and Lewis (1992), strong duality holds if $(ri\, dom(g)) \cap (ri\, A(dom(f_y))) \neq \emptyset$.[9]

To check this condition, it suffices to find an element of $(ri\, dom(g)) \cap (ri\, A(dom(f_y)))$. Let $m_0 \in L^2$ be a candidate SDF[10] with finite pricing error, i.e. $E(m_0) = 1$ and $\| E[m_0 r] \|_N^2 < \infty$. As $m_0 \in L^2$ and $E(m_0 - y)^2 < \infty$, then $m_0 \in dom(f_y)$ and $A(m_0) \in ri\, A(dom(f_y))$. In addition, because $E[2m_0] = 2$ and $\| E[m_0 r] \|_N^2 < \infty$, $g(A(m_0)) = \dfrac{N}{4\gamma} \| E[m_0 \dfrac{2r}{N}] \|^2 = \dfrac{1}{\gamma} \| E[m_0 r] \|_N^2 < \infty$, therefore $A(m_0) \in ri\, dom(g)$. In conclusion, $A(m_0)$ is an element of $(ri\, dom(g)) \cap (ri\, A(dom(f_y)))$ and hence $(ri\, dom(g)) \cap (ri\, A(dom(f_y))) \neq \emptyset$.

Then by Theorem 4.2 of Borwein and Lewis (1992), we have[11]

$$\delta_R^2 = \max_{\nu \in \mathbb{R}^{N+1}} \left\{ -f_y^*(A^*(-\nu)) - g^*(\nu) \right\},$$

where f_y^* and g^* are the conjugate functions of f_y and g respectively and A^* is the adjoint of A.

Let us determine the relevant conjugate functions.

[9] For a convex set $S \subseteq \mathbb{R}^N$, $ri\, S$ is its relative interior. The latter is the interior with respect to the affine hull of S, $aff\, S$. Specifically, $ri\, S = \{x \in S : B_\epsilon(x) \cap aff\, S \subseteq S\}$, where $aff\, S = \{\theta_1 x_1 + \cdots + \theta_k x_k : x_1, \ldots, x_k \in S, \theta_1 + \cdots + \theta_k = 1\}$ and $B_\epsilon(x) = \{y \in \mathbb{R}^N : \| y - x \| < \epsilon\}$.

[10] Such a random variable exists, for instance $m_0 = 1$ satisfies the conditions.

[11] For convenience, we replaced ν by $-\nu$ in Borwein and Lewis' formula. This does not affect the final result.

$f_y^*(z) = E\left\{\sup_{w\in L^2} zw-(w-y)^2\right\} = E\left\{zy+\frac{1}{4}z^2\right\}$[12], $A^*:\mathbb{R}^{N+1}\to L^2$ and $A^*(\theta)=X'\theta$. Finally $g^*(\nu)=h^*(\nu_1)+\chi^*_{\{2\}}(\nu_2)$ where $\nu=(\nu_1',\nu_2)\in\mathbb{R}^{N+1}$. The conjugates[13] are given by $\chi^*_{\{2\}}(\nu_2)=\sup_{x\in\{2\}} x\nu_2=2\nu_2$ and $h^*(\nu_1)=\frac{\gamma}{N}\|\nu_1\|^2$. So, $g^*(\nu)=2\nu_2+\frac{\gamma}{N}\|\nu_1\|^2$.

Therefore

$$\delta_R^2=\max_{\nu_1\in\mathbb{R}^N,\nu_2\in\mathbb{R}} E\left\{2y\nu_1'\frac{r}{N}+2y\nu_2-\frac{\nu_1' rr' \nu_1}{N^2}-\nu_2^2-2\frac{\nu_1' r\nu_2}{N}-2\nu_2-\frac{\gamma}{N}\parallel\nu_1\parallel^2\right\}.$$

Now, using the fact that $E[y]=1$, we obtain

$$\delta_R^2=\max_{\nu_1\in\mathbb{R}^N,\nu_2\in\mathbb{R}} E\left\{2y\nu_1'\frac{r}{N}-\frac{\nu_1' rr' \nu_1}{N^2}-\nu_2^2-2\frac{\nu_1' r\nu_2}{N}-\frac{\gamma}{N}\parallel\nu_1\parallel^2\right\}.$$

Solving in ν_1 and ν_2 gives

$$\nu_1=(\Sigma+\gamma I)^{-1}\alpha \text{ and } \nu_2=-\nu_1'\mu_2/N,$$

and hence $\delta_R^2=\alpha'(\Sigma+\gamma I)^{-1}\alpha/N$.

We can do the same for Tikhonov by setting $h(x)=\frac{N}{4\gamma}\|x\|_\Sigma^2$. The latter can be rewritten as $h(x)=\frac{N}{2\gamma}n(x)$ with $n(x)=\frac{1}{2}\|x\|_\Sigma^2$. Therefore the convex conjugate of h is $h^*(z)=\frac{N}{2\gamma}n^*\left(\frac{z}{\frac{N}{2\gamma}}\right)$ with

$$n^*(z)=\sup_{w\in R^N}\left\{w'z-\frac{w'\Sigma w}{2}\right\}.$$

When Σ is non singular, the expression in brackets is maximized at $w=\Sigma^{-1}z$. Therefore, $n^*(z)=\frac{z'\Sigma^{-1}z}{2}$. As a result, $h^*(z)=\frac{\gamma}{N}\|z\|_{\Sigma^{-1}}^2$. When Σ is singular,

$$n^*(z)=\begin{cases}\frac{z'\Sigma^\dagger z}{2},\ z\in\mathcal{R}(\Sigma),\\ +\infty,\ z\notin\mathcal{R}(\Sigma)\end{cases}$$

where $\Sigma^\dagger$ is the Moore-Penrose inverse of Σ. So

$$h^*(z)=\frac{\gamma}{N}z'\Sigma^\dagger z+i_{\mathcal{R}(\Sigma)}z$$

[12] This comes from the definition of the functional conjugate of a convex function p.196 of Luenberger (1969) and the use of Riesz Theorem in the L^2 space equipped with the usual inner product.

[13] To determine the conjugate of h, note that the conjugate of $\frac{1}{2}\|x\|^2$ is again $\frac{1}{2}\|x\|^2$. In addition, if $f(x)=ag(x)+b$, then $f^*(x)=ag^*\left(\frac{x}{a}\right)+b$.

where $i_{\mathcal{R}(\Sigma)} = 0$ on the range of Σ and $+\infty$ elsewhere. Substituting this conjugate into the dual problem and using $E[y] = 1$ gives

$$\max_{\nu_1,\nu_2} E\left[2y\frac{\nu_1' r}{N} - \frac{\nu_1' r r' \nu_1}{N^2} - \nu_2^2 - 2\frac{\nu_1' r \nu_2}{N} - \frac{\gamma}{N}\nu_1'\Sigma^\dagger \nu_1\right].$$

The first-order condition with respect to ν_2 gives

$$\nu_2 = -\frac{\nu_1'\mu_2}{N}.$$

Using $E[rr'] = V_{22} + \mu_2\mu_2'$, $V_{22} = N\Sigma$, and $E[yr] = \alpha$, the terms involving $\mu_2\mu_2'$ cancel. The dual problem therefore reduces to

$$\max_{\nu_1 \in \mathcal{R}(\Sigma)} \frac{1}{N}\left[2\nu_1'\alpha - \nu_1'\Sigma\nu_1 - \gamma\nu_1'\Sigma^\dagger\nu_1\right].$$

Its maximizer satisfies

$$(\Sigma + \gamma\Sigma^\dagger)\nu_1 = P_\Sigma \alpha,$$

where $P_\Sigma = \Sigma\Sigma^\dagger$ and hence

$$\nu_1 = (\Sigma + \gamma\Sigma^\dagger)^\dagger \alpha = (\Sigma^2 + \gamma I_N)^{-1}\Sigma\alpha.$$

Substituting the maximizer into the dual objective yields

$$\frac{1}{N}\alpha'\nu_1 = \frac{1}{N}\alpha'(\Sigma^2 + \gamma I_N)^{-1}\Sigma\alpha = \delta_K^2,$$

which establishes the Tikhonov result.

11.4 Proof of Proposition 2

The proof of Proposition 2 uses the following lemmas.

Lemma 2. *Suppose Assumptions 1 and 2 are satisfied. Then,* $\left\|\epsilon'\bar{F}\right\|_F^2 = O_p(NT)$, *where* ϵ is a $T \times N$ *matrix with (t,i) element* ϵ_{ti}.

Proof of Lemma 2.
Let us denote $Y_{\epsilon f,t} = \epsilon_t \vec{f}_t'$.

First,

$$\begin{aligned} E\left[\left\|\frac{\epsilon'\bar{F}}{T}\right\|_F^2\right] &= E\left[tr\left\{\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\bar{f}_t'\right)'\left(\frac{1}{T}\sum_{t=1}^{T}\epsilon_t\bar{f}_t'\right)\right\}\right] \\ &= E\left[tr\left\{\left(\frac{1}{T^2}\sum_{t=1}^{T}Y_{\epsilon f,t}'Y_{\epsilon f,t}+\frac{1}{T^2}\sum_{t\neq s}^{T}Y_{\epsilon f,t}'Y_{\epsilon f,s}\right)\right\}\right] \\ &= \frac{1}{T}E\left[tr(Y_{\epsilon f,1}'Y_{\epsilon f,1})\right]+\frac{2}{T}\sum_{l=1}^{T}(1-\frac{l}{T})E\left[tr(Y_{\epsilon f,1}'Y_{\epsilon f,1+l})\right] \end{aligned}$$

We have

$$\begin{aligned} trE[Y_{\epsilon f,t}'Y_{\epsilon f,t}] &= E[tr(\bar{f}_t\epsilon_t'\epsilon_t\bar{f}_t')] \\ &= trE[\bar{f}_t'\bar{f}_t\epsilon_t'\epsilon_t] \end{aligned}$$

From Cauchy-Schwarz, $\left|E[\bar{f}_t'\bar{f}_t\epsilon_t'\epsilon_t]\right| \leq \sqrt{E[\left\|\bar{f}_t\right\|^4]E[\|\epsilon_t\|^4]} = O(N)$. Therefore, $\frac{1}{T}E\left[tr(Y_{\epsilon f,1}'Y_{\epsilon f,1})\right] = O(\frac{N}{T})$.
Using Davydov's inequality (Davydov (1968), Rio (1993)) [14] (with $q = r = 2+\rho$),

$$\begin{aligned} trE\left[(Y_{\epsilon f,1}'Y_{\epsilon f,1+l})\right] &= \sum_{i=1}^{N}\sum_{k=1}^{K}E\left[\left(\bar{f}_{k1}\epsilon_{i1}\right)\left(\bar{f}_{k1+l}\epsilon_{i1+l}\right)\right] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K}cov(\bar{f}_{k1}\epsilon_{i1},\bar{f}_{k1+l}\epsilon_{i1+l}) \\ &\leq 12\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_x(l)^{\frac{\rho}{2+\rho}}E[(\bar{f}_{kt}\epsilon_{it})^{2+\rho}]^{\frac{2}{2+\rho}}. \end{aligned}$$

As a result,

$$\begin{aligned} \frac{2}{T}\sum_{l=1}^{T}(1-\frac{l}{T})E\left[tr(Y_{\epsilon f,1}'Y_{\epsilon f,1+l})\right] &\leq \frac{24}{T}\sum_{i=1}^{N}\sum_{k=1}^{K}E[(\bar{f}_{kt}\epsilon_{it})^{2+\rho}]^{\frac{2}{2+\rho}}\sum_{l=1}^{T}(1-\frac{l}{T})\alpha_x(l)^{\frac{\rho}{2+\rho}} \\ &\leq \frac{24}{T}\sum_{i=1}^{N}\sum_{k=1}^{K}E[(\bar{f}_{kt}\epsilon_{it})^{2+\rho}]^{\frac{2}{2+\rho}}\sum_{l=1}^{T}\alpha_x(l)^{\frac{\rho}{2+\rho}}. \end{aligned}$$

[14] For any positive real numbers p, q, r such that $\frac{1}{p}+\frac{1}{q}+\frac{1}{r}=1$, the covariance between two r.vs X and Y is bounded as follows: $cov(X,Y) \leq 12\alpha(\sigma(X),\sigma(Y))^{\frac{1}{p}}E\left[\mid X\mid^q\right]^{\frac{1}{q}}E\left[\mid Y\mid^r\right]^{\frac{1}{r}}$, where $\sigma(X)$ is the sigma algebra generated by X and α is the strong mixing coefficient.

From Assumption 2(iii) and Cauchy-Schwarz, $\left|E[(\bar{f}_{kt}\epsilon_{it})^{2+\rho}]\right| \leq E[\bar{f}_{kt}^{4+2\rho}]^{\frac{1}{2}} E[\epsilon_{it}^{4+2\rho}]^{\frac{1}{2}} \leq c^{\frac{1}{2}} E[\bar{f}_{kt}^{4+2\rho}]^{\frac{1}{2}}$. So,

$$\frac{2}{T}\sum_{l=1}^{T}(1-\frac{l}{T})E\left[tr(Y_{\epsilon f,1}^{'}Y_{\epsilon f,1+l})\right] = O(\frac{N}{T}).$$

Hence, $E\left[\left\|\frac{\epsilon'\bar{F}}{T}\right\|_F^2\right] = O(\frac{N}{T})$. In conclusion,

$$\left\|\epsilon'\bar{F}\right\|_F^2 = O_p(NT).$$

Lemma 3. *Suppose Assumptions 1 and 2 are satisfied. Then,* $\left\|\hat{\beta}-\beta\right\|_F^2 = O_p\left(\frac{N}{T}\right)$.

Proof of Lemma 3. Using the fact that $\left\|\epsilon'\bar{F}\right\|_F^2 = O_p(NT)$, we have

$$\begin{aligned}
\left\|\hat{\beta}-\beta\right\|_F^2 &= \left\|\frac{\epsilon^{'}\bar{F}}{T}\hat{V}_{11}^{-1}\right\|_F^2 \\
&= \left\|\frac{\epsilon^{'}\bar{F}}{T}V_{11}^{-1} + \frac{\epsilon^{'}\bar{F}}{T}(\hat{V}_{11}^{-1}-V_{11}^{-1})\right\|_F^2 \\
&\leq 2\left\|\frac{\epsilon^{'}\bar{F}}{T}V_{11}^{-1}\right\|_F^2 + 2\left\|\frac{\epsilon^{'}\bar{F}}{T}(\hat{V}_{11}^{-1}-V_{11}^{-1})\right\|_F^2 \\
&\leq 2\left\|\frac{\epsilon^{'}\bar{F}}{T}V_{11}^{-1}\right\|_F^2 + 2\left\|\frac{\epsilon^{'}\bar{F}}{T}\right\|_F^2 . \left\|\hat{V}_{11}^{-1}-V_{11}^{-1}\right\|_F^2 \\
&= O_p\left(\frac{NT}{T^2}\right) + O_p\left(\frac{NT}{T^2}.\frac{1}{T}\right) \\
&= O\left(\frac{N}{T}\right).
\end{aligned}$$

Therefore,

$$\left\|\hat{\beta}-\beta\right\|_F^2 = O_p\left(\frac{N}{T}\right)$$

Lemma 4. *Suppose Assumptions 1 and 2 are satisfied. For* $k = 1, \cdots, K$, $\left\|\hat{\beta}_k-\beta_k\right\|_N = O_p\left(\frac{1}{\sqrt{T}}\right)$.

Proof of Lemma 4. As $\left\|\hat{\beta}-\beta\right\|_F^2 = O_p\left(\frac{N}{T}\right)$, we have $\frac{1}{N}tr\left[(\hat{\beta}-\beta)^{'}(\hat{\beta}-\beta)\right] = O_p\left(\frac{1}{T}\right)$.

$$(\hat{\beta}-\beta)(\hat{\beta}-\beta)' = \sum_{k=1}^{K}((\hat{\beta}_k-\beta_k)(\hat{\beta}_k-\beta_k)^{'}).$$

As a result,

$$\frac{1}{N}tr\left[\sum_{k=1}^{K}((\hat{\beta}_k-\beta_k)(\hat{\beta}_k-\beta_k)')\right]=\sum_{k=1}^{K}\parallel\hat{\beta}_k-\beta_k\parallel_N^2=O_p(\frac{1}{T})$$

In conclusion, $\|\hat{\beta}_k-\beta_k\|_N=O_p\left(\frac{1}{\sqrt{T}}\right)$.

Lemma 5. *Suppose Assumptions 1, 2, and 4(i), we have*

$$\left\|\widehat{\Sigma}-\Sigma\right\|_{op}\leq\left\|\widehat{\Sigma}-\Sigma\right\|_{HS}=O_p\left(\frac{1}{\sqrt{T}}\right).$$

Proof of Lemma 5. $\widehat{\Sigma}$ and Σ are elements of the Hilbert space of Hilbert-Schmidt operators with inner product $\langle A,B\rangle=tr\left(A'B\right)$ and norm $\|A\|_{HS}^2=tr\left(A'A\right)$.

Using the notations $\tilde{r}_t=r_t-\mu_2$ and $X_t=\tilde{r}_t\tilde{r}_t'/N-\Sigma$, we have by Davydov's inequality and the mixing propriety of r_t

$$\begin{aligned}E\left[\left\|\frac{1}{T}\sum_{t=1}^{T}X_t\right\|_{HS}^2\right] &= \frac{1}{T}E\left\|X_1\right\|_{HS}^2+\frac{2}{T}\sum_{l=1}^{T}E\left\langle X_1,X_{1+l}\right\rangle\\ &\leq \frac{1}{T}E\left\|X_1\right\|_{HS}^2+\frac{24}{T}\left(E\left(\left\|X_1\right\|_{HS}^{2+\rho}\right)\right)^{\frac{2}{2+\rho}}\sum_{l=1}^{T}\left(1-\frac{l}{T}\right)\alpha\left(l\right)^{\frac{\rho}{2+\rho}}.\end{aligned}$$

Under the mixing property $\sum_l\alpha\left(l\right)^{\frac{\rho}{2+\rho}}<\infty$ and under the assumption 4(i), we have $\sup_N E\left(\left\|X_1\right\|_{HS}^{2+\rho}\right)<\infty$. Moreover,

$$\begin{aligned}\widehat{\Sigma}-\Sigma &= \frac{\sum_{t=1}^{T}X_t}{T}-\frac{\left(\overline{r}-\mu_2\right)\left(\overline{r}-\mu_2\right)'}{N}\\ &= \frac{\sum_{t=1}^{T}X_t}{T}+O_p\left(\frac{1}{T}\right)\end{aligned}$$

because $\left\|\frac{\left(\overline{r}-\mu_2\right)\left(\overline{r}-\mu_2\right)'}{N}\right\|=\left\|\overline{r}-\mu_2\right\|_N^2=O_p\left(1/T\right)$. It follows that $\left\|\widehat{\Sigma}-\Sigma\right\|_{HS}^2=O_p\left(1/T\right)$ by Markov's inequality.

Lemma 6. *Suppose that Assumptions 1 to 3. If $\beta_k\in\mathcal{R}\left(\Sigma^{1+\omega}\right)$ for some $1\geq\omega>0$ (instead of the stronger condition $\beta_k\in\mathcal{R}\left(\Sigma^2\right)$), we have the following result:*

$$\left\|\hat{\Sigma}_\gamma^{-1}\hat{\beta}_k-\Sigma^{-1}\beta_k\right\|_N^2=O_p\left(\frac{1}{\gamma T}\right)+O(\gamma^\omega).$$

Proof of Lemma 6. We consider Tikhonov only. We follow the proof of Lemma 3 of Carrasco

(2012). We have the following decomposition.

$$\begin{aligned}\left\|\hat{\Sigma}_\gamma^{-1}\hat{\beta}_k - \Sigma^{-1}\beta_k\right\|_N^2 &\leq 3\left\|\hat{\Sigma}_\gamma^{-1}\left(\hat{\beta}_k - \beta_k\right)\right\|_N^2 && (21)\\ &\quad +3\left\|\left(\hat{\Sigma}_\gamma^{-1} - \Sigma_\gamma^{-1}\right)\beta_k\right\|_N^2 && (22)\\ &\quad +3\left\|\left(\Sigma_\gamma^{-1} - \Sigma^{-1}\right)\beta_k\right\|_N^2 && (23)\end{aligned}$$

First, consider (21). Let $v = \hat{\beta}_k - \beta_k$. So,

$$\begin{aligned}\left\|\widehat{\Sigma}_\gamma^{-1}v\right\|_N^2 &= \left\|\left(\widehat{\Sigma}^2 + \gamma I\right)^{-1}\widehat{\Sigma}v\right\|_N^2\\ &= \sum_{j=1}^{\infty}\frac{\widehat{\lambda}_j^2}{\left(\widehat{\lambda}_j^2 + \gamma\right)^2}\left\langle\hat{\phi}_j, v\right\rangle_N^2\\ &\leq \frac{1}{\gamma}\|v\|_N^2 = O_p\left(\frac{1}{\gamma T}\right)\end{aligned}$$

where we used $\|v\|_N^2 = \left\|\hat{\beta}_k - \beta_k\right\|_N^2 = O_p\left(1/T\right)$ by Lemma 4. Hence (21) is $O_p\left(\frac{1}{\gamma T}\right)$.

For (22), we will use the relation

$$\begin{aligned}&\left(\hat{\Sigma}^2 + \gamma I\right)^{-1}\hat{\Sigma} - \left(\Sigma^2 + \gamma I\right)^{-1}\Sigma\\ =\ &\left(\hat{\Sigma}^2 + \gamma I\right)^{-1}\left[\hat{\Sigma}\left(\Sigma - \hat{\Sigma}\right)\Sigma + \gamma\left(\hat{\Sigma} - \Sigma\right)\right]\left(\Sigma^2 + \gamma I\right)^{-1}\end{aligned}$$

As $\beta_k \in \mathcal{R}\left(\Sigma^{1+\omega}\right)$, there exists ϕ such that $\|\phi\|_N < \infty$ and $\beta_k = \Sigma^{1+\omega}\phi$. We have

$$\begin{aligned}\left\|\left(\hat{\Sigma}_\gamma^{-1} - \Sigma_\gamma^{-1}\right)\beta_k\right\|_N^2 \leq\ &\\ &\left\|\left(\hat{\Sigma}^2 + \gamma I\right)^{-1}\hat{\Sigma}\left(\Sigma - \hat{\Sigma}\right)\Sigma\left(\Sigma^2 + \gamma I\right)^{-1}\beta_k\right\|_N^2 && (24)\\ &+\left\|\left(\hat{\Sigma}^2 + \gamma I\right)^{-1}\gamma\left(\hat{\Sigma} - \Sigma\right)\left(\Sigma^2 + \gamma I\right)^{-1}\beta_k\right\|_N^2 && (25)\end{aligned}$$

Regarding (24), we have

$$\begin{aligned} & \left\| \left(\hat{\Sigma}^2 + \gamma I \right)^{-1} \hat{\Sigma} \left(\Sigma - \hat{\Sigma} \right) \Sigma \left(\Sigma^2 + \gamma I \right)^{-1} \beta_k \right\|_N^2 \\ \leq & \left\| \left(\hat{\Sigma}^2 + \gamma I \right)^{-1} \hat{\Sigma} \right\|_{op,N}^2 \left\| \Sigma - \hat{\Sigma} \right\|_{op,N}^2 \left\| \Sigma \left(\Sigma^2 + \gamma I \right)^{-1} \beta_k \right\|_N^2 \\ = & O_p \left(\frac{1}{\gamma T} \right) \end{aligned}$$

because $\left\| \Sigma - \hat{\Sigma} \right\|_{op,N}^2 = O_p \left(1/T \right)$ by Lemma 5.

Regarding (25), we have

$$\begin{aligned} \left\| \left(\hat{\Sigma}^2 + \gamma I \right)^{-1} \gamma \left(\hat{\Sigma} - \Sigma \right) \left(\Sigma^2 + \gamma I \right)^{-1} \beta_k \right\|_N^2 & \leq \left\| \left(\hat{\Sigma} - \Sigma \right) \left(\Sigma^2 + \gamma I \right)^{-1} \beta_k \right\|_N^2 \\ & = O_p \left(\frac{1}{\gamma T} \right). \end{aligned}$$

In summary, the rate of (22) is $O_p \left(\frac{1}{\gamma T} \right)$.

Finally, the term (23) satisfies

$$\begin{aligned} & \left\| \left(\Sigma_\gamma^{-1} - \Sigma^{-1} \right) \beta_k \right\|_N^2 \\ = & \left\| \left(\Sigma_\gamma^{-1} - \Sigma^{-1} \right) \Sigma^{1+\omega} \phi \right\|_N^2 \\ = & \left\| \left(\left(\Sigma^2 + \gamma I \right)^{-1} \Sigma - \Sigma^{-1} \right) \Sigma \Sigma^\omega \phi \right\|_N^2 \\ = & \left\| \left(\left(\Sigma^2 + \gamma I \right)^{-1} \Sigma^2 - I \right) \Sigma^\omega \phi \right\|_N^2 \\ = & \left\| \left(\Sigma^2 + \gamma I \right)^{-1} \left(\Sigma^2 - \Sigma^2 - \gamma \right) \Sigma^\omega \phi \right\|_N^2 \\ = & \gamma^2 \left\| \left(\Sigma^2 + \gamma I \right)^{-1} \Sigma^\omega \phi \right\|_N^2 \\ = & \gamma^2 \left\| \left(\Sigma^2 + \gamma I \right)^{-1+\omega/2} \left(\Sigma^2 + \gamma I \right)^{-\omega/2} \Sigma^\omega \phi \right\|_N^2 \\ = & O \left(\gamma^\omega \right). \end{aligned}$$

This completes the proof of Lemma 6.

Lemma 7. *Let*

$$X_{T,N} = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} < \tilde{r}_t, u >_N \equiv \frac{1}{\sqrt{T}} \sum_{t=1}^{T} Y_{t,T,N},$$

where $u \in \mathbb{R}^N$ *is not random and* $\|u\|_N = O(1)$.

If Assumptions 2(i) and 4(i) hold, then $Var\left(X_{T,N}\right) = O\left(1\right)$ *and* $X_{T,N} = O_p\left(1\right)$.
If moreover, $\lim_{N,T\to\infty} var(X_{T,N}) = \omega^2 \in (0,\infty)$, *then*

$$X_{T,N} \xrightarrow{d} \mathcal{N}(0,\omega^2)$$

when T, N *go simultaneously to* ∞.

Proof of Lemma 7.

First, consider the case when $\{r_t\}_{t=1,\cdots,T}$ are independent. From Assumptions 4(i), $E\left[\parallel r_t \parallel_N^{2+\rho}\right] = O(1)$ when N goes to infinity for $\rho > 0$. To establish the central limit theorem, we need to verify the Lindeberg condition for a double indexed process of Phillips and Moon (1999)(see their Theorem 2). In our setting, this condition can be rewritten as

$$\lim_{N,T\to\infty} \frac{1}{\sigma_{T,N}^2} \sum_{t=1}^{T} E\left[Y_{t,T,N}^2 1_{|Y_{t,T,N}|>\sigma_{T,N}\varepsilon}\right] \to 0$$

for every $\varepsilon > 0$, with $\sigma_{T,N}^2 = T.var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_{t,T,N}\right) = T.\frac{u'\Sigma u}{N} \equiv T\sigma_N^2 > 0$.

To see that this condition is satisfied, observe that when $\left|\frac{Y_{t,T,N}}{\sigma_{T,N}\varepsilon}\right| > 1$,

$$\varepsilon^{\rho}\frac{Y_{t,T,N}^2}{\sigma_{T,N}^2} \leq \frac{Y_{t,T,N}^{2+\rho}}{\sigma_{T,N}^{2+\rho}}.$$

Therefore,

$$\varepsilon^{\rho} E\left[\frac{Y_{t,T,N}^2}{\sigma_{T,N}^2} 1_{|Y_{t,T,N}|>\sigma_{T,N}\varepsilon}\right] \leq E\left[\frac{Y_{t,T,N}^{2+\rho}}{\sigma_{T,N}^{2+\rho}} 1_{|Y_{t,T,N}|>\sigma_{T,N}\varepsilon}\right] \leq E\left[\frac{Y_{t,T,N}^{2+\rho}}{\sigma_{T,N}^{2+\rho}}\right].$$

Moreover

$$\lim_{N,T\to\infty} \sum_{t=1}^{T} E\left[\frac{Y_{t,T,N}^{2+\rho}}{\sigma_{T,N}^{2+\rho}}\right] = \lim_{N,T\to\infty} \frac{1}{T^{\rho/2}} \frac{1}{\sigma_N^{2+\rho}} E\left[< \tilde{r}_t, u >_N{}^{2+\rho}\right] = 0$$

as $\sigma_N^2 \to \omega^2 > 0$ by assumption, $E\left[\mid \langle \tilde{r}_t, u\rangle_N \mid^{2+\rho}\right] \leq E\left[\parallel \tilde{r}_t \parallel_N^{2+\rho}\right] \parallel u \parallel_N^{2+\rho} = O(1)$ by Assumption 4(i). Hence, the Lindeberg condition is satisfied and the central limit theorem follows.

For the dependent case**,** by Davydov's inequality (Davydov (1968), Rio (1993)) (with $q = r =$

$2+\rho$), we have

$$\begin{aligned} var(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} Y_{t,T,N}) &= E[Y_{1,T,N}^2] + 2\sum_{l=1}^{T}(1-\frac{l}{T})E\left[Y_{1,T,N}Y_{1+l,T,N}\right] \\ &\leq E[Y_{1,T,N}^2] + 24\left(E[\mid Y_{t,T,N}\mid^{2+\rho}]\right)^{\frac{2}{2+\rho}}\sum_{l=1}^{T}(1-\frac{l}{T})a(l)^{\frac{\rho}{2+\rho}} \\ &\leq E[Y_{1,T,N}^2] + 24\left(E[\mid Y_{t,T,N}\mid^{2+\rho}]\right)^{\frac{2}{2+\rho}}\sum_{l=1}^{T}a(l)^{\frac{\rho}{2+\rho}} \\ &\leq C\left[1+\sum_{l=1}^{T}a(l)^{\frac{\rho}{2+\rho}}\right] \\ &= O\left(1\right) \end{aligned}$$

uniformly in N, T. Hence, $X_{T,N} = O_p\left(1\right)$.

In addition, the central limit theorem of Francq and Zakoïan (2005) applies because the Lindeberg condition (iii) of page 1168 still applies when N goes to infinity, see also Chang, Chen, and Chen (2015) p.288 for a similar application of this result. Hence, $X_{T,N}$ asymptotically converges to a normal distribution when T, N go to ∞.

Lemma 8. *Suppose Assumptions 2(i) and 4 are satisfied. For any $u, v \in \mathbb{R}^N$ with $\|u\|_N = O\left(1\right)$ and $\|v\|_N = O\left(1\right)$, as N and T go to ∞, if*

$$\lim_{N,T\to\infty} var\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\left\langle\tilde{r}_t, v\right\rangle_N\left\langle\tilde{r}_t, u\right\rangle_N - E\left(\left\langle\tilde{r}_t, v\right\rangle_N\left\langle\tilde{r}_t, u\right\rangle_N\right)\right)\right] \equiv \sigma_{u,v}^2 > 0,$$

then $\sqrt{T}\left\langle\left(\hat{\Sigma}-\Sigma\right)v, u\right\rangle_N$ converges to a Gaussian distribution of mean 0 and variance $\sigma_{u,v}^2$.

Proof of Lemma 8. We have the following decomposition of $\hat{\Sigma}-\Sigma$:

$$\begin{aligned} \hat{\Sigma}-\Sigma &= \frac{1}{NT}\sum_{t=1}^{T}\left(r_t-\hat{\mu}_2\right)\left(r_t-\hat{\mu}_2\right)' - \frac{1}{N}E[\tilde{r}_t\tilde{r}_t'] \\ &= \frac{1}{NT}\sum_{t=1}^{T}\left(r_t-\mu_2+\mu_2-\hat{\mu}_2\right)\left(r_t-\mu_2+\mu_2-\hat{\mu}_2\right)' - \frac{1}{N}E[\tilde{r}_t\tilde{r}_t'] \\ &= \frac{1}{NT}\sum_{t=1}^{T}\left(\tilde{r}_t\tilde{r}_t' - E[\tilde{r}_t\tilde{r}_t']\right) + \left(\mu_2-\hat{\mu}_2\right)\frac{1}{NT}\sum_{t=1}^{T}\left(r_t-\mu_2\right)' \\ &\quad + \frac{1}{NT}\sum_{t=1}^{T}\left(r_t-\mu_2\right)\left(\mu_2-\hat{\mu}_2\right)' \\ &\quad + \frac{1}{N}\left(\mu_2-\hat{\mu}_2\right)\left(\mu_2-\hat{\mu}_2\right)'. \end{aligned}$$

Therefore,

$$\begin{aligned}\left\langle \sqrt{T}\left[\hat{\Sigma}-\Sigma\right]v,u\right\rangle_N &= \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\left\langle \tilde{r}_t,v\right\rangle_N\left\langle \tilde{r}_t,u\right\rangle_N - E\left(\left\langle \tilde{r}_t,v\right\rangle_N\left\langle \tilde{r}_t,u\right\rangle_N\right)\right\} \\ &\quad + \left\langle v,(\mu_2-\hat{\mu}_2)\right\rangle_N\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\langle \tilde{r}_t,u\right\rangle_N \\ &\quad + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\langle v,\tilde{r}_t\right\rangle_N\left\langle (\mu_2-\hat{\mu}_2),u\right\rangle_N \\ &\quad + \sqrt{T}\left\langle v,(\mu_2-\hat{\mu}_2)\right\rangle_N\left\langle (\mu_2-\hat{\mu}_2),u\right\rangle_N .\end{aligned}$$

Following Lemma 7, $\left\langle v,(\mu_2-\hat{\mu}_2)\right\rangle_N\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\langle \tilde{r}_t,u\right\rangle_N = O_p\left(\frac{1}{\sqrt{T}}\right)$, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\langle v,\tilde{r}_t\right\rangle_N\left\langle (\mu_2-\hat{\mu}_2),u\right\rangle_N = O_p\left(\frac{1}{\sqrt{T}}\right)$, and $\sqrt{T}\left\langle v,(\mu_2-\hat{\mu}_2)\right\rangle_N\left\langle (\mu_2-\hat{\mu}_2),u\right\rangle_N = O_p\left(\frac{1}{\sqrt{T}}\right)$. As a result,

$$\left\langle \sqrt{T}\left[\hat{\Sigma}-\Sigma\right]v,u\right\rangle_N = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\{\left\langle \tilde{r}_t,v\right\rangle_N\left\langle \tilde{r}_t,u\right\rangle_N - E\left(\left\langle \tilde{r}_t,v\right\rangle_N\left\langle \tilde{r}_t,u\right\rangle_N\right)\right\} + O_p\left(\frac{1}{\sqrt{T}}\right).$$

From here on, the proof is similar to that of Lemma 7. To apply the central limit theorem of Francq and Zakoïan (2005), we need

$$\lim_{N\to\infty}\sup_t E[(\left\langle v,\tilde{r}_t\right\rangle_N\left\langle \tilde{r}_t,u\right\rangle_N)^{2+\rho}] < \infty$$

for some $\rho > 0$. This condition is met because of Assumption 4(i).

Lemma 9. *Suppose the assumptions of Proposition 2 hold. Then*

$$\left\|\widehat{\alpha}-\alpha\right\|_N = O_p\left(T^{-1/2}\right).$$

Proof of Lemma 9. Since

$$\widehat{\alpha} = \widehat{\mu}_2 - \widehat{V}_{21}\widehat{\theta}^{\gamma}_{HJ} \quad \text{and} \quad \alpha = \mu_2 - V_{21}\theta_{HJ},$$

we have

$$\widehat{\alpha}-\alpha = (\widehat{\mu}_2-\mu_2) - (\widehat{V}_{21}-V_{21})\theta_{HJ} - \widehat{V}_{21}(\widehat{\theta}^{\gamma}_{HJ}-\theta_{HJ}).$$

Because K is fixed,

$$\|\widehat{\mu}_2-\mu_2\|_N = O_p(T^{-1/2}), \qquad \|\widehat{V}_{21}-V_{21}\|_{op} = O_p(T^{-1/2}),$$

and Proposition 2 gives

$$\|\widehat{\theta}^{\gamma}_{HJ} - \theta_{HJ}\| = O_p(T^{-1/2}).$$

Moreover, $\|\widehat{V}_{21}\|_{op} = O_p(1)$. The result follows from the triangle inequality.

Proof of Proposition 2

Consistency:

Recall that by Equation (9), we have

$$\begin{aligned}\hat{\theta}^{\gamma}_{HJ} - \theta_{HJ} = &(\hat{V}_{11}^{-1} - V_{11}^{-1})\lambda \\ &+ \hat{V}_{11}^{-1}(\hat{\mu}_1 - \mu_1) \\ &+ \hat{V}_{11}^{-1}\left(\frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\hat{\beta}}{N}\right)^{-1}\left[\frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}(\beta - \hat{\beta})}{N}(\lambda + \hat{\mu}_1 - \mu_1)\right. \\ &\left.+\frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\alpha}{N} + \frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\bar{\epsilon}}{N}\right].\end{aligned} \tag{26}$$

For the first two rows, $(\hat{V}_{11}^{-1} - V_{11}^{-1})\lambda$ and $\hat{V}_{11}^{-1}(\hat{\mu}_1 - \mu_1)$ converge to 0 in probability by the law of large numbers and Assumption 2 (i).

$$\begin{aligned}\frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\hat{\beta}}{N} &= \left[\frac{\hat{\beta}'_{k_1}\hat{\Sigma}_{\gamma}^{-1}\hat{\beta_{k_2}}}{N}\right]_{k_1,k_2=1,\cdots,K} \\ &= \left\langle \hat{\beta}_{k_1}, \hat{\Sigma}_{\gamma}^{-1}\hat{\beta}_{k_2}\right\rangle_{N;k_1,k_2=1,\cdots,K}.\end{aligned}$$

We have

$$\begin{aligned}\left\langle \hat{\beta}_{k_1}, \hat{\Sigma}_{\gamma}^{-1}\hat{\beta}_{k_2}\right\rangle_N &= \left\langle \hat{\beta}_{k_1} - \beta_{k_1}, \hat{\Sigma}_{\gamma}^{-1}\hat{\beta}_{k_2} - \Sigma^{-1}\beta_{k_2}\right\rangle_N + && (27)\\ &\quad \left\langle \hat{\beta}_{k_1} - \beta_{k_1}, \Sigma^{-1}\beta_{k_2}\right\rangle_N + && (28)\\ &\quad \left\langle \beta_{k_1}, \hat{\Sigma}_{\gamma}^{-1}\hat{\beta}_{k_2} - \Sigma^{-1}\beta_{k_2}\right\rangle_N + && (29)\\ &\quad \left\langle \beta_{k_1}, \Sigma^{-1}\beta_{k_2}\right\rangle_N - C_{k_1,k_2} && (30)\\ &\quad + C_{k_1,k_2}. && (31)\end{aligned}$$

where C_{k_1,k_2} is the (k_1, k_2) element of the matrix C defined in Assumption 3.

$|(27)| \leq \left\|\hat{\beta}_{k_1} - \beta_{k_1}\right\|_N \left\|\hat{\Sigma}_{\gamma}^{-1}\hat{\beta}_{k_2} - \Sigma^{-1}\beta_{k_2}\right\|_N \to 0$ as $N, T \to \infty$, and $\gamma T \to \infty$ using Lemma 6.

For (28), we have

$$\left|\left\langle \hat{\beta}_{k_1} - \beta_{k_1}, \Sigma^{-1}\beta_{k_2}\right\rangle_N\right| \leq \left\|\hat{\beta}_{k_1} - \beta_{k_1}\right\|_N \left\|\Sigma^{-1}\beta_{k_2}\right\|_N \to 0$$

as $N, T \to \infty$, by Assumption 3(i) and Lemma 4.

By Lemma 6, $\|(29)\| = o_p(1)$.

Finally, using Assumption 3, (30) goes to 0 as N goes to ∞.

In conclusion, $\dfrac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\hat{\beta}}{N} \to C$ as $N, T \to \infty$, $\gamma T \to \infty$, and $\gamma \to 0$.

Using the same argument as before, we have $\dfrac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}(\beta - \hat{\beta})}{N} = \dfrac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\beta}{N} - \dfrac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\hat{\beta}}{N} \xrightarrow{P} 0$ as $N, T \to \infty$, $\gamma T \to \infty$, and $\gamma \to 0$.

For $\dfrac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\alpha}{N}$, we have $\hat{\Sigma}_\gamma^{-1}\hat{\beta} \xrightarrow{P} \Sigma^{-1}\beta$ when as $N, T \to \infty$, and $\gamma T \to \infty$. Moreover, $\beta'\Sigma^{-1}\alpha = 0$ by the first order condition of (2). Therefore when as $N, T \to \infty$, and $\gamma T \to \infty$, $\dfrac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\alpha}{N} \xrightarrow{P} 0$.

The same is true for $\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\bar{\epsilon}$, which converges in probability to 0 as $N, T \to \infty$, and $\gamma T \to \infty$. The consistency of $\hat{\theta}_{HJ}^{\gamma}$ follows.

Distribution:

We detail the proof of the asymptotic normality for Tikhonov estimator. The result for ridge could be shown similarly. We analyze the decomposition (26) using the following results:

- $\hat{\beta} - \beta = \dfrac{1}{T}\sum_{t=1}^{T} \epsilon_t \bar{f}_t' \hat{V}_{11}^{-1}$
- Note that $\hat{C}_\beta = \dfrac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\hat{\beta}}{N}$ and we already have shown that $\hat{C}_\beta - C_\beta \xrightarrow{P} 0_{K,K}$.
- $\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\alpha = (\hat{\beta} - \beta)'\hat{\Sigma}_\gamma^{-1}\alpha + \beta'(\hat{\Sigma}_\gamma^{-1} - \Sigma^{-1})\alpha$ using the fact that $\beta'\Sigma^{-1}\alpha = 0$ by the population first-order condition of (2).
- We have

$$\begin{aligned} \beta'(\hat{\Sigma}_\gamma^{-1} - \Sigma^{-1})\alpha &= \beta'(\hat{\Sigma}_\gamma^{-1} - \Sigma_\gamma^{-1} + \Sigma_\gamma^{-1} - \Sigma^{-1})\alpha \\ &= \beta'(\hat{\Sigma}_\gamma^{-1} - \Sigma_\gamma^{-1})\alpha + \beta'(\Sigma_\gamma^{-1} - \Sigma^{-1})\alpha \end{aligned}$$

and

$$\begin{aligned} \beta'(\hat{\Sigma}_\gamma^{-1} - \Sigma_\gamma^{-1})\alpha &= \beta'\left(\hat{\Sigma}^2 + \gamma I\right)^{-1}\left[\hat{\Sigma}\left(\Sigma - \hat{\Sigma}\right)\Sigma + \gamma\left(\hat{\Sigma} - \Sigma\right)\right]\left(\Sigma^2 + \gamma I\right)^{-1}\alpha \\ &= \beta'\hat{\Sigma}_\gamma^{-1}\left(\Sigma - \hat{\Sigma}\right)\Sigma_\gamma^{-1}\alpha + \gamma\beta'\left(\hat{\Sigma}^2 + \gamma I\right)^{-1}\left(\hat{\Sigma} - \Sigma\right)\left(\Sigma^2 + \gamma I\right)^{-1}\alpha \end{aligned}$$

Therefore,

$$\begin{aligned}
&\sqrt{T}\left(\hat{\theta}_{HJ}^{\gamma}-\theta_{HJ}\right)\\
&-\sqrt{T}\hat{V}_{11}^{-1}\hat{C}_{\beta}^{-1}\left\{\gamma\beta'\left(\hat{\Sigma}^{2}+\gamma I\right)^{-1}\left(\hat{\Sigma}-\Sigma\right)\left(\Sigma^{2}+\gamma I\right)^{-1}\frac{\alpha}{N}+\beta'(\Sigma_{\gamma}^{-1}-\Sigma^{-1})\frac{\alpha}{N}\right\} && (32)\\
&=\hat{V}_{11}^{-1}\Bigg\{-\sqrt{T}(\hat{V}_{11}-V_{11})V_{11}^{-1}\lambda && (33)\\
&+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(f_{t}-\mu_{1}) && (34)\\
&+\hat{C}_{\beta}^{-1}\Bigg[-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\epsilon_{t}\bar{f}_{t}'}{N}\hat{V}_{11}^{-1}(\lambda+\hat{\mu}_{1}-\mu_{1}) && (35)\\
&-\sqrt{T}\beta'\hat{\Sigma}_{\gamma}^{-1}(\hat{\Sigma}-\Sigma)\Sigma_{\gamma}^{-1}\frac{\alpha}{N} && (36)\\
&+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{V}_{11}^{-1}\frac{\bar{f}_{t}\epsilon_{t}'\hat{\Sigma}_{\gamma}^{-1}\alpha}{N} && (37)\\
&+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\epsilon_{t}}{N}\Bigg]\Bigg\}. && (38)
\end{aligned}$$

First, we simplify the terms in the expansion.

- Regarding (33), note that

$$\begin{aligned}
\sqrt{T}(\hat{V}_{11}-V_{11})V_{11}^{-1}\lambda &= \sqrt{T}\hat{V}_{11}\theta_{HJ}-\sqrt{T}\lambda\\
\sqrt{T}\hat{V}_{11} &= \frac{1}{\sqrt{T}}\sum_{t}\left(f_{t}-\widehat{\mu}_{1}\right)\left(f_{t}-\widehat{\mu}_{1}\right)'\\
&= \frac{1}{\sqrt{T}}\sum_{t}\tilde{f}_{t}\tilde{f}_{t}'+\frac{1}{\sqrt{T}}\sum_{t}\tilde{f}_{t}\left(\mu_{1}-\widehat{\mu}_{1}\right)'\\
&\quad+\frac{1}{\sqrt{T}}\sum_{t}\left(\mu_{1}-\widehat{\mu}_{1}\right)\tilde{f}_{t}'+\sqrt{T}\left(\mu_{1}-\widehat{\mu}_{1}\right)\left(\mu_{1}-\widehat{\mu}_{1}\right)'\\
&= \frac{1}{\sqrt{T}}\sum_{t}\tilde{f}_{t}\tilde{f}_{t}'+o_{p}\left(1\right).
\end{aligned}$$

because $\sqrt{T}\left(\mu_{1}-\widehat{\mu}_{1}\right)=O_{p}\left(1\right)$ and $\frac{1}{\sqrt{T}}\sum_{t}\tilde{f}_{t}=O_{p}\left(1\right)$. Hence

$$\sqrt{T}(\hat{V}_{11}-V_{11})V_{11}^{-1}\lambda=\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\tilde{f}_{t}\tilde{f}_{t}'\theta_{HJ}-\lambda\right)+o_{p}(1).$$

- Consider term (35):

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\epsilon_t\bar{f}_t'}{N}\hat{V}_{11}^{-1}(\lambda+\hat{\mu}_1-\mu_1) = \frac{\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}}{N}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\epsilon_t\tilde{f}_t'V_{11}^{-1}\lambda\right. \quad (39)$$

$$+ \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\epsilon_t\bar{f}_t'(\hat{V}_{11}^{-1}-V_{11}^{-1})\lambda \quad (40)$$

$$- \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\epsilon_t(\hat{\mu}_1-\mu_1)'V_{11}^{-1}\lambda \quad (41)$$

$$\left. + \frac{1}{\sqrt{T}}\sum_{t=1}^{T}\epsilon_t\bar{f}_t'\hat{V}_{11}^{-1}(\hat{\mu}_1-\mu_1)\right]. \quad (42)$$

The term (40) can be rewritten as $T^{-\frac{1}{2}}\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}\frac{\epsilon'\bar{F}}{N}(\hat{V}_{11}^{-1}-V_{11}^{-1})\lambda$. From Lemma 2, $\left\|\epsilon'\bar{F}\right\|_F^2 = O_p(NT)$. So $\left\|\epsilon'\bar{F}\right\|^2 = O_p(NT)$. From Lemma 6, $\left\|\hat{\Sigma}_{\gamma}^{-1}\hat{\beta}_k-\Sigma^{-1}\beta_k\right\|_N \to 0,$ as $N,T\to\infty,$ and $\gamma T\to\infty$. In addition, $\left\|(\hat{V}_{11}^{-1}-V_{11}^{-1})\lambda\right\|^2 = O_p\left(\frac{1}{T}\right)$. Therefore, (40) is $o_p(1)$ when $N,T\to\infty$ and $\gamma T\to\infty$.

Using the fact that $\theta_{HJ}=V_{11}^{-1}\lambda$, we can rewrite (41) as

$$-(\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}-\beta'\Sigma^{-1})\frac{1}{N\sqrt{T}}\sum_t\epsilon_t(\hat{\mu}_1-\mu_1)'\theta_{HJ}-\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\langle\Sigma^{-1}\beta,\epsilon_t\right\rangle_N\right)(\hat{\mu}_1-\mu_1)'\theta_{HJ}.$$

Using a proof similar to that of Lemma 2, we can show that $\left\|\sum_t\epsilon_t\right\|^2 = \|\epsilon'1\|^2 = O_p(NT)$. Moreover, $\|\hat{\mu}_1-\mu_1\|^2 = O_P\left(\frac{1}{T}\right)$. As a result, $(\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}-\beta'\Sigma^{-1})\frac{1}{N\sqrt{T}}\sum_t\epsilon_t(\hat{\mu}_1-\mu_1)'\theta_{HJ}$ is $o_p(1)$ when $N,T\to\infty,$ and $\gamma T\to\infty$.

Using Lemma 7, $\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\langle\Sigma^{-1}\beta,\epsilon_t\right\rangle_N$ is $O_p(1)$ as $\left\|\Sigma^{-1}\beta\right\|_N<\infty$ and ϵ_t has the same characteristics as $\tilde{r}_t$. Then the second term is $o_p(1)$ when $N,T\to\infty$. Therefore, (41) is $o_p(1)$.

Regarding (42), we can rewrite it as

$$\frac{\sqrt{T}}{N}(\hat{\beta}'\hat{\Sigma}_{\gamma}^{-1}-\beta'\Sigma^{-1})(\hat{\beta}-\beta)(\hat{\mu}_1-\mu_1)+\frac{\sqrt{T}}{N}\beta'\Sigma^{-1}(\hat{\beta}-\beta)(\hat{\mu}_1-\mu_1).$$

From Lemmas 3 and 6, this expression is $o_p(1)$ as $N,T\to\infty,$ and $\gamma T\to\infty$.

Finally, (39) is equal to

$$\frac{1}{N\sqrt{T}}\left(\hat{\beta}'\hat{\Sigma}_\gamma^{-1}-\beta'\Sigma^{-1}\right)\epsilon'\tilde{F}V_{11}^{-1}\lambda+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left\langle\Sigma^{-1}\beta,\epsilon_t\tilde{f}_t'V_{11}^{-1}\lambda\right\rangle_N.$$

The first part is $o_P(1)$ when $N,T\to\infty$, and $\gamma T\to\infty$.

- Consider (36). We have

$$\sqrt{T}\beta'\hat{\Sigma}_\gamma^{-1}(\hat{\Sigma}-\Sigma)\Sigma_\gamma^{-1}\frac{\alpha}{N}=\sqrt{T}\beta'\Sigma^{-1}(\hat{\Sigma}-\Sigma)\frac{\Sigma^{-1}\alpha}{N}+o_p(1).$$

- Consider (37). We have the following decomposition:

$$\begin{aligned}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\hat{V}_{11}^{-1}\frac{\bar{f}_t\epsilon_t'\hat{\Sigma}_\gamma^{-1}\alpha}{N} &= V_{11}^{-1}\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\tilde{f}_t\epsilon_t'\hat{\Sigma}_\gamma^{-1}\alpha}{N} &(43)\\ &+ \frac{1}{N\sqrt{T}}\left(\hat{V}_{11}^{-1}-V_{11}^{-1}\right)\bar{F}'\epsilon\hat{\Sigma}_\gamma^{-1}\alpha &(44)\\ &+ \frac{1}{N\sqrt{T}}V_{11}^{-1}(\mu_1-\hat{\mu}_1)\sum_{t=1}^{T}\epsilon_t'\hat{\Sigma}_\gamma^{-1}\alpha. &(45)\end{aligned}$$

Using Lemmas 2 and 6, (44) and (45) are $o_p(1)$ when $N,T\to\infty$, and $\gamma T\to\infty$.

- The term in (32) can be dealt with using the same approach as in the proof of Lemma 6 with $\omega=1$. First, consider

$$\begin{aligned}\beta'(\Sigma_\gamma^{-1}-\Sigma^{-1})\frac{\alpha}{N} &= \beta'(\left(\Sigma^2+\gamma I\right)^{-1}\Sigma-\Sigma^{-1})\frac{\alpha}{N}\\ &= \beta'(\left(\Sigma^2+\gamma I\right)^{-1}\Sigma^2-I)\Sigma^{-1}\alpha/N\\ &= -\gamma\beta'\left(\Sigma^2+\gamma I\right)^{-1}\Sigma^{-1}\alpha/N\\ &= O(\gamma)\end{aligned}$$

  under the assumptions $\alpha,\beta\in\mathcal{R}\left(\Sigma^2\right)$. So $\sqrt{T}\widehat{V}_{11}^{-1}\widehat{C}_\beta^{-1}\beta'(\Sigma_\gamma^{-1}-\Sigma^{-1})\frac{\alpha}{N}$ is $O_p\left(\sqrt{T}\gamma\right)$. Therefore, this term is $o_p(1)$ as $N,T,\gamma T\to\infty$ and $\gamma^2T\to0$. It corresponds to a regularization bias which vanishes asymptotically.

  Second, using $A^{-1}-B^{-1}=A^{-1}(B-A)B^{-1}$, we have

$$\left(\left(\hat{\Sigma}^2+\gamma I\right)^{-1}-\Sigma^{-2}\right)\beta_k=\left(\hat{\Sigma}^2+\gamma I\right)^{-1}\left[\Sigma^2-\widehat{\Sigma}^2-\gamma I\right]\Sigma^{-2}\beta_k.$$

Hence,

$$\begin{aligned}
&\sqrt{T}\gamma\widehat{V}_{11}^{-1}\widehat{C}_{\beta}^{-1}\beta_k\left(\hat{\Sigma}^2+\gamma I\right)^{-1}\left(\hat{\Sigma}-\Sigma\right)\left(\Sigma^2+\gamma I\right)^{-1}\frac{\alpha}{N}\\
= \;& \sqrt{T}\gamma\widehat{V}_{11}^{-1}\widehat{C}_{\beta}^{-1}\left\langle\left(\hat{\Sigma}^2+\gamma I\right)^{-1}\beta_k,\left(\hat{\Sigma}-\Sigma\right)\left(\Sigma^2+\gamma I\right)^{-1}\alpha\right\rangle_N = O_p\left(\gamma+T^{-1/2}\right)=o_p\left(1\right).
\end{aligned}$$

This term also vanishes asymptotically but could be kept in the influence function to obtain more accurate estimates of the standard errors of the estimators. Now, we explain how to incorporate this term in the influence function. Let $G_\gamma=\left(\Sigma^2+\gamma I\right)^{-1}$. The leading term is

$$\frac{\sqrt{T}}{N}\gamma V_{11}^{-1}C_{\beta}^{-1}\beta' G_\gamma\left(\hat{\Sigma}-\Sigma\right)G_\gamma\alpha.$$

From the expansion of the sample covariance matrix,

$$\sqrt{T}\left(\hat{\Sigma}-\Sigma\right)=\frac{1}{\sqrt{T}}\sum_{t=1}^{n}\left(\frac{\widetilde{r}_t\widetilde{r}_t'}{N}-\Sigma\right)+o_p\left(1\right).$$

The additional contribution of Tikhonov regularization to the influence function is

$$h_t^{extra}=\gamma V_{11}^{-1}C_{\beta}^{-1}\left[\left(\frac{\beta' G_\gamma\widetilde{r}_t}{N}\right)\left(\frac{\widetilde{r}_t' G_\gamma\alpha}{N}\right)-\frac{\beta' G_\gamma\Sigma G_\gamma\alpha}{N}\right].$$

Using the previous results, we have

$$\sqrt{T}\left(\hat{\theta}_{HJ}^{\gamma}-\theta_{HJ}\right)=\hat{V}_{11}^{-1}.A+o_p(1),$$

where

$$\begin{aligned}
A \;=\;& \left(-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{f}_t\tilde{f}_t'\theta_{HJ}+\sqrt{T}\lambda\right)+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\tilde{f}_t+C_{\beta}^{-1}\left[-\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\beta'\Sigma^{-1}\frac{\epsilon_t\tilde{f}_t'\theta_{HJ}}{N}\right.\\
&\left.+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}V_{11}^{-1}\tilde{f}_t\frac{\epsilon_t'\Sigma^{-1}\alpha}{N}-\beta'\Sigma^{-1}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\left(\frac{\tilde{r}_t\tilde{r}_t'}{N^2}-\frac{\Sigma}{N}\right)\right)\Sigma^{-1}\alpha+\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\frac{\beta'\Sigma^{-1}\epsilon_t}{N}\right]\\
\equiv\;& \frac{1}{\sqrt{T}}\sum_{t=1}^{T}h_t
\end{aligned}$$

and

$$\begin{aligned}
h_t \;=\;& -\tilde{f}_t\tilde{f}_t'\theta_{HJ}+\lambda+\tilde{f}_t-C_{\beta}^{-1}\beta'\Sigma^{-1}\frac{\epsilon_t\tilde{f}_t'\theta_{HJ}}{N}+C_{\beta}^{-1}V_{11}^{-1}\tilde{f}_t\frac{\epsilon_t'\Sigma^{-1}\alpha}{N}\\
&-C_{\beta}^{-1}\beta'\Sigma^{-1}\frac{\tilde{r}_t\tilde{r}_t'}{N^2}\Sigma^{-1}\alpha+C_{\beta}^{-1}\frac{\beta'\Sigma^{-1}\epsilon_t}{N}.
\end{aligned}$$

Using the fact that $y_t = 1 - \tilde{f}_t'\theta_{HJ}$ and noting $\tilde{u}_t = \dfrac{\tilde{r}_t'\Sigma^{-1}\alpha}{N}$, h_t can be rewritten as follows

$$h_t = \tilde{f}_t y_t + \lambda + \frac{C_\beta^{-1}\beta'\Sigma^{-1}}{N}(\epsilon_t y_t - \tilde{r}_t\tilde{u}_t) + C_\beta^{-1}V_{11}^{-1}\widetilde{f}_t\frac{\epsilon_t'\Sigma^{-1}\alpha}{N}.$$

We have for a constant C independent of N,

$$\|h_t\| \leq C\left(1 + \left\|\widetilde{f}_t\right\|^2 + \|\epsilon_t\|_N^2 + \|\tilde{r}_t\|_N^2\right).$$

Assumptions 2(ii)-(iii) and 4(i) imply

$$\sup_N E\left(\|h_t\|^{2+\rho}\right) < \infty.$$

For the error term, the needed bound follows directly from

$$E\left[\|\epsilon_t\|_N^{4+2\rho}\right] = E\left[\left(\frac{1}{N}\sum_i \epsilon_{it}^2\right)^{2+\rho}\right] \leq \frac{1}{N}\sum_i E\left|\epsilon_{it}\right|^{4+2\rho}.$$

Moreover, $Eh_t = 0$ and h_t is a measurable function of $x_{N,t}$, so it inherits its strong mixing bound. Let $\Gamma_N(j) = E\left(h_t h_{t-j}'\right)$. Davydov's inequality gives

$$\|\Gamma_N(j)\| \leq Ca\left(|j|\right)^{\rho/(2+\rho)}, \; j \neq 0.$$

The bound is summable uniformly in N. By Assumption 4(ii), the variance converges to S :

$$Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^T h_t\right) = \sum_{|j|<T}\left(1 - \frac{|j|}{T}\right)\Gamma_N(j) \to S.$$

For any fixed a, the centered scalar array $a'h_t$ satisfies Francq and Zakoïan (2005)'s conditions. Their theorem yields

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T a'h_t \xrightarrow{d} \mathcal{N}(0, a'Sa).$$

By Cramer-Wold device, we have

$$\sqrt{T}\left(\hat{\theta}_{HJ}^{\gamma} - \theta_{HJ}\right) \xrightarrow{d} \mathcal{N}(0_K, V_{11}^{-1}SV_{11}^{-1}),$$

where $S = \lim_{N,T\to\infty} var\left[\dfrac{1}{\sqrt{T}}\sum_{t=1}^T h_t\right]$.

Using $\epsilon_t = \tilde{r}_t - \beta\tilde{f}_t$, $V_{21} = \beta V_{11}$, and the population first-order condition $\beta'\Sigma^{-1}\alpha = 0$, we obtain the simplified expression

$$V_{11}^{-1}h_t = \theta_{HJ} + H\left[\tilde{f}_t\tilde{u}_t + (y_t - \tilde{u}_t)\frac{V_{21}'\Sigma^{-1}\tilde{r}_t}{N}\right],$$

where

$$H = V_{11}^{-1}C_\beta^{-1}V_{11}^{-1}.$$

11.5 Proof of Proposition 3

Let

$$S_T = \sum_{|j|<T}\Gamma_j$$

where $\Gamma_j = \dfrac{1}{T}\displaystyle\sum_{t=j+1}^{T} E\left(h_t h_{t-j}'\right)$ for $j \geq 0$ and $\Gamma_j = \Gamma_{-j}'$. Note that $S = \lim_{T\to\infty} S_T$.

Let

$$\widetilde{S}_T = \sum_{|j|<T}\omega\left(\frac{j}{m_T}\right)\widetilde{\Gamma}_j$$

where $\widetilde{\Gamma}_j = \dfrac{1}{T}\displaystyle\sum_{t=j+1}^{T} h_t h_{t-j}'$ for $j \geq 0$ and $\widetilde{\Gamma}_j = \widetilde{\Gamma}_{-j}'$ and $h_t = h_t(\xi_1, \xi_2)$ depends on the (pseudo) true values of the parameters. On the other hand, $\widehat{h}_t = h_t\left(\hat{\xi}_1, \hat{\xi}_2\right)$ depends on the estimators.

We have

$$\hat{S}_T - S = \hat{S}_T - \widetilde{S}_T + \widetilde{S}_T - S_T + S_T - S$$

where $S_T - S \xrightarrow{P} 0$ by construction.

Note that under Assumption 5, the process $\{h_t\}$ satisfies the conditions of Theorem 1(a) of Andrews (1991). It follows from this theorem that $\widetilde{S}_T - S_T \to 0$ provided $m_T/\sqrt{T} \to 0$.

Now we focus on the term $\hat{S}_T - \widetilde{S}_T$. To simplify the notations, we can assume that there is only one factor, so that h_t and S are scalar, see the argument of Andrews (1991) (p.852) to justify that this can be done without loss of generality. The set of finite parameters ξ_1 includes parameters which estimators converge at the usual $\sqrt{T}$ rate, namely μ_1, V_{11}, and θ, and one parameter, C_β, which estimator converges at a slower rate. Indeed, we have

$$\begin{aligned}\left|\widehat{C}_\beta - C_\beta\right| &= \left|\frac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}\hat{\beta}}{N} - \frac{\beta'\Sigma^{-1}\beta}{N}\right| \\ &= \left|\frac{\left(\hat{\beta}-\beta\right)'\Sigma^{-1}\beta}{N} + \frac{\beta'\left(\hat{\Sigma}_\gamma^{-1}\hat{\beta} - \Sigma^{-1}\beta\right)}{N} + \frac{\left(\hat{\beta}-\beta\right)'\left(\hat{\Sigma}_\gamma^{-1}\hat{\beta} - \Sigma^{-1}\beta\right)}{N}\right|\end{aligned}$$

where we used the fact that $\widehat{a}\widehat{b}-ab=(\widehat{a}-a)\,b+a\left(\widehat{b}-b\right)+(\widehat{a}-a)\left(\widehat{b}-b\right)$. By Cauchy-Schwarz, we have

$$\begin{aligned}
\left|\frac{\left(\hat{\beta}-\beta\right)'\Sigma^{-1}\beta}{N}\right| &\leq \left\|\hat{\beta}-\beta\right\|_N\left\|\Sigma^{-1}\beta\right\|_N = O_p\left(\frac{1}{\sqrt{T}}\right),\\
\left|\frac{\beta'\left(\hat{\Sigma}_\gamma^{-1}\hat{\beta}-\Sigma^{-1}\beta\right)}{N}\right| &\leq \|\beta\|_N\left\|\hat{\Sigma}_\gamma^{-1}\hat{\beta}-\Sigma^{-1}\beta\right\|_N = O_p\left(\frac{1}{\sqrt{\gamma T}}+\sqrt{\gamma}\right),\\
\left|\frac{\left(\hat{\beta}-\beta\right)'\left(\hat{\Sigma}_\gamma^{-1}\hat{\beta}-\Sigma^{-1}\beta\right)}{N}\right| &\leq \left\|\hat{\beta}-\beta\right\|_N\left\|\hat{\Sigma}_\gamma^{-1}\hat{\beta}-\Sigma^{-1}\beta\right\|_N = O_p\left(\frac{1}{T\sqrt{\gamma}}+\frac{\sqrt{\gamma}}{\sqrt{T}}\right)
\end{aligned}$$

where we used Lemma 6. It follows that

$$\left|\widehat{C}_\beta - C_\beta\right| = O_p\left(\frac{1}{\sqrt{\gamma T}}+\sqrt{\gamma}\right).$$

Regarding the estimators of the elements of ξ_2, we have the following rates

$$\begin{aligned}
\left\|\hat{\Sigma}_\gamma^{-1}\hat{\beta}-\Sigma^{-1}\beta\right\|_N &= O_p\left(\frac{1}{\sqrt{\gamma T}}+\sqrt{\gamma}\right),\\
\left\|\hat{\Sigma}_\gamma^{-1}\widehat{\alpha}-\Sigma^{-1}\alpha\right\|_N &= O_p\left(\frac{1}{\sqrt{\gamma T}}+\sqrt{\gamma}\right),\\
\left\|\widehat{\mu}_2-\mu_2\right\|_N &= O_p\left(\frac{1}{\sqrt{T}}\right),\\
\left\|\widehat{\beta}-\beta\right\|_N &= O_p\left(\frac{1}{\sqrt{T}}\right)
\end{aligned}$$

where the first rate follows again from Lemma 6 and the second rate can be established using a similar proof as that of Lemma 6 and uses Lemma 9.

We have

$$\hat{S}_T-\widetilde{S}_T=\sum_{|j|<T}\omega\left(\frac{j}{m_T}\right)\left(\tilde{\Gamma}_j\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-\tilde{\Gamma}_j\left(\xi_1,\xi_2\right)\right)$$

and

$$\begin{aligned}
\tilde{\Gamma}_j\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-\tilde{\Gamma}_j\left(\xi_1,\xi_2\right) &= \tilde{\Gamma}_j\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-\tilde{\Gamma}_j\left(\hat{\xi}_1,\xi_2\right)\\
&\quad+\tilde{\Gamma}_j\left(\hat{\xi}_1,\xi_2\right)-\tilde{\Gamma}_j\left(\xi_1,\xi_2\right).
\end{aligned}$$

Using the mean-value theorem, we have

$$\tilde{\Gamma}_j\left(\hat{\xi}_1,\xi_2\right)-\tilde{\Gamma}_j\left(\xi_1,\xi_2\right)=\frac{\partial\tilde{\Gamma}_j\left(\overline{\xi}_1,\xi_2\right)}{\partial\xi_1}\left(\hat{\xi}_1-\xi_1\right)$$

where $\overline{\xi}_1$ is between $\widehat{\xi}_1$ and ξ_1. In addition,

$$\begin{aligned}\sup_{j\geq1}\left\|\frac{\partial\tilde{\Gamma}_j\left(\overline{\xi}_1,\xi_2\right)}{\partial\xi_1}\right\| &= \sup_{j\geq1}\left\|\frac{1}{T}\sum_{t=j+1}^{T}h_t\left(\overline{\xi}_1,\xi_2\right)\frac{\partial h_{t-j}\left(\overline{\xi}_1,\xi_2\right)}{\partial\xi_1}+h_{t-j}\left(\overline{\xi}_1,\xi_2\right)\frac{\partial h_t\left(\overline{\xi}_1,\xi_2\right)}{\partial\xi_1}\right\| \\ &\leq 2\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\xi_1\in E}h_t\left(\xi_1,\xi_2\right)^2\right)^{1/2}\left(\frac{1}{T}\sum_{t=1}^{T}\sup_{\xi_1\in E}\left\|\frac{\partial h_t\left(\xi_1,\xi_2\right)}{\partial\xi_1}\right\|^2\right)^{1/2} \\ &= O_p\left(1\right)\end{aligned}$$

by Assumption 5 and Markov's inequality.

Then, we obtain

$$\begin{aligned}\left|\sum_{|j|<T}\omega\left(\frac{j}{m_T}\right)\left(\tilde{\Gamma}_j\left(\hat{\xi}_1,\xi_2\right)-\tilde{\Gamma}_j\left(\xi_1,\xi_2\right)\right)\right| &\leq \sum_{|j|<T}\left|\omega\left(\frac{j}{m_T}\right)\right|\left\|\frac{\partial\tilde{\Gamma}_j\left(\overline{\xi}_1,\xi_2\right)}{\partial\xi_1}\right\|\left\|\hat{\xi}_1-\xi_1\right\| \\ &\leq \frac{1}{m_T}\sum_{|j|<T}\left|\omega\left(\frac{j}{m_T}\right)\right|\sup_j\left\|\frac{\partial\tilde{\Gamma}_j\left(\overline{\xi}_1,\xi_2\right)}{\partial\xi_1}\right\|m_T\left\|\hat{\xi}_1-\xi_1\right\| \\ &= O_p\left(\frac{m_T}{\sqrt{\gamma T}}+m_T\sqrt{\gamma}\right)\end{aligned}$$

where we used the fact that $\frac{1}{m_T}\sum_{|j|<T}\left|\omega\left(\frac{j}{m_T}\right)\right|\to\int\left|\omega\left(x\right)\right|dx<\infty$ (see Andrews (1991), p.852).

Now we turn our attention on the term $\tilde{\Gamma}_j\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-\tilde{\Gamma}_j\left(\hat{\xi}_1,\xi_2\right)$. We have

$$\begin{aligned}
&\tilde{\Gamma}_j\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-\tilde{\Gamma}_j\left(\hat{\xi}_1,\xi_2\right)\\
=\;&\frac{1}{T}\sum_{t=j+1}^{T}h_t\left(\widehat{\xi}_1,\widehat{\xi}_2\right)h_{t-j}\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-h_t\left(\widehat{\xi}_1,\xi_2\right)h_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\\
=\;&\frac{1}{T}\sum_{t=j+1}^{T}\left(h_t\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-h_t\left(\widehat{\xi}_1,\xi_2\right)\right)h_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\\
&+\frac{1}{T}\sum_{t=j+1}^{T}h_t\left(\widehat{\xi}_1,\xi_2\right)\left(h_{t-j}\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-h_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\right)\\
&+\frac{1}{T}\sum_{t=j+1}^{T}\left(h_t\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-h_t\left(\widehat{\xi}_1,\xi_2\right)\right)\left(h_{t-j}\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-h_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\right).
\end{aligned}$$

Observe that $h_t\left(\xi_1,\xi_2\right)$ is linear in $\Sigma^{-1}\beta$, $\Sigma^{-1}\alpha$, β and is quadratic in μ_2. The most complicated term is the one involving $\Sigma^{-1}\alpha$. We will focus on this one, the rest of the proof is similar. Let us denote $\Sigma^{-1}\alpha$ by ν and let $\xi_2=\left(\nu',\xi_{2\nu}'\right)'$.

$$\begin{aligned}
&\frac{1}{T}\sum_{t=j+1}^{T}\left(h_t\left(\widehat{\xi}_1,\widehat{\xi}_2\right)-h_t\left(\widehat{\xi}_1,\xi_2\right)\right)h_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\\
=\;&\frac{1}{T}\sum_{t=j+1}^{T}\left(h_t\left(\widehat{\xi}_1,\hat{\nu},\widehat{\xi}_{2\nu}\right)-h_t\left(\widehat{\xi}_1,\nu,\widehat{\xi}_{2\nu}\right)+h_t\left(\widehat{\xi}_1,\nu,\widehat{\xi}_{2\nu}\right)-h_t\left(\widehat{\xi}_1,\xi_2\right)\right)h_{t-j}\left(\widehat{\xi}_1,\xi_2\right).
\end{aligned}$$

We have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=j+1}^{T}\left(h_t\left(\widehat{\xi}_1,\hat{\nu},\widehat{\xi}_{2\nu}\right)-h_t\left(\widehat{\xi}_1,\nu,\widehat{\xi}_{2\nu}\right)\right)h_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\\
=\;&-\hat{C}_\beta^{-1}\frac{\hat{\beta}'\hat{\Sigma}_\gamma^{-1}}{N}\left(\frac{1}{T}\sum_{t=j+1}^{T}\overline{r}_t\overline{r}_t'h_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\right)\frac{\hat{\nu}-\nu}{N} \qquad (46)\\
&+\hat{C}_\beta^{-1}\widehat{V}_{11}^{-1}\left(\frac{1}{T}\sum_{t=j+1}^{T}\overline{f}_th_{t-j}\left(\widehat{\xi}_1,\xi_2\right)\widehat{\epsilon}_t'\right)\frac{\hat{\nu}-\nu}{N}. \qquad (47)
\end{aligned}$$

The rate of convergence is not affected by the fact that ξ_1 and $\xi_{2\nu}$ are estimated, hence to simplify

notations, we drop the hats on $\widehat{\xi}_1$ and $\widehat{\xi}_{2\nu}$. Consider first (47).

$$\begin{aligned}
& \left| C_\beta^{-1} V_{11}^{-1} \left(\frac{1}{T} \sum_{t=j+1}^{T} \widetilde{f}_t h_{t-j} \epsilon_t' \right) \frac{\hat{\nu} - \nu}{N} \right| \\
= & \left| C_\beta^{-1} V_{11}^{-1} \left\langle \frac{1}{T} \sum_{t=j+1}^{T} \widetilde{f}_t h_{t-j} \epsilon_t, \hat{\nu} - \nu \right\rangle_N \right| \\
\leq & C_\beta^{-1} V_{11}^{-1} \left\| \frac{1}{T} \sum_{t=j+1}^{T} \widetilde{f}_t h_{t-j} \epsilon_t \right\|_N \left\| \hat{\nu} - \nu \right\|_N \\
\leq & C_\beta^{-1} V_{11}^{-1} \left(\frac{1}{T} \sum_{t=j+1}^{T} h_{t-j}^2 \right)^{1/2} \left(\frac{1}{T} \sum_{t=j+1}^{T} \left\| \widetilde{f}_t \epsilon_t \right\|_N^2 \right)^{1/2} \left\| \hat{\nu} - \nu \right\|_N \\
= & O_p \left(\frac{1}{\sqrt{\gamma T}} + \sqrt{\gamma} \right).
\end{aligned}$$

Now consider (46). To simplify notations, denote $\Sigma_\gamma^{-1} \beta$ by τ. We have

$$\begin{aligned}
& C_\beta^{-1} \frac{\tau'}{N} \left(\frac{1}{T} \sum_{t=j+1}^{T} \widetilde{r}_t \widetilde{r}_t' h_{t-j} \right) \frac{\hat{\nu} - \nu}{N} \\
= & C_\beta^{-1} \left\langle \tau, \left(\frac{1}{T} \sum_{t=j+1}^{T} \widetilde{r}_t \widetilde{r}_t' h_{t-j} \right) \frac{\hat{\nu} - \nu}{N} \right\rangle_N \\
\leq & C_\beta^{-1} \left\| \tau \right\|_N \left\| \left(\frac{1}{T} \sum_{t=j+1}^{T} \widetilde{r}_t \widetilde{r}_t' h_{t-j} \right) \frac{\hat{\nu} - \nu}{N} \right\|_N \\
\leq & C_\beta^{-1} \left\| \tau \right\|_N \left\| \left(\frac{1}{T} \sum_{t=j+1}^{T} \widetilde{r}_t \widetilde{r}_t' h_{t-j} \right) \right\|_{op} \frac{1}{N} \left\| \hat{\nu} - \nu \right\|_N \\
\leq & C_\beta^{-1} \left\| \tau \right\|_N \left(\frac{1}{N^2 T} \sum_{t=j+1}^{T} \left\| \widetilde{r}_t \right\|^4 \right)^{1/2} \left(\frac{1}{T} \sum_{t=j+1}^{T} h_{t-j}^2 \right)^{1/2} \left\| \hat{\nu} - \nu \right\|_N \\
= & O_p \left(\frac{1}{\sqrt{\gamma T}} + \sqrt{\gamma} \right)
\end{aligned}$$

where we used the fact that $E \left\| \widetilde{r}_t \right\|^4 / N^2 < \infty$ by Assumption 4. The other terms can be dealt with similarly and details are omitted.

11.6 Proof of Propositions 4 and 5

Assuming f_1, f_2, and f_3 are vectors of factors such that $cov\left(f_1, f_2, f_3\right)$ is nonsingular, the proof will use the following lemmas.

Lemma 10. *In the case where Model b nests Model a (case $K_2 = 0$), $\delta_a^2 - \delta_b^2 = 0 \Leftrightarrow \theta_{b2} = 0_{K_3}$ where θ_{b2} is the coefficient of f_{3t} in Model b.*

Lemma 11. *In the case of non-nested models with overlapping factors f_{1t}, $y_a = y_b \Leftrightarrow \theta_{a2} = 0_{K_2}$ and $\theta_{b2} = 0_{K_3}$ where θ_{a2} is the coefficient of f_{2t} in Model a.*

Proof of Lemmas 10 and 11.

The proof is similar to that of Lemmas 2 and 3 of Kan and Robotti (2009). The main difference is that they work with gross returns while we use excess returns.

In the case of nested models, $\delta_a^2 - \delta_b^2 = 0 \Leftrightarrow y_a = y_b$. So we can see Lemma 10 as a special case of Lemma 11 when $K_2 = 0$. Therefore, we focus on proving Lemma 11.

It is easy to see that $y_a = y_b \Leftrightarrow \theta_{a1} = \theta_{b1}$, $\theta_{a2} = 0_{K_2}$ and $\theta_{b2} = 0_{K_3}$. So it suffices to show that $\theta_{a2} = 0_{K_2}$ and $\theta_{b2} = 0_{K_3}$ imply $\theta_{a1} = \theta_{b1}$.

We have

$$\theta_m = \left(V_{m,12}V_{22}^{-1}V_{m,21}\right)^{-1} V_{m,12}V_{22}^{-1}\mu_2,\ m = a, b.$$

It follows that

$$\left(V_{m,12}V_{22}^{-1}V_{m,21}\right)\theta_m = V_{m,12}V_{22}^{-1}\mu_2. \tag{48}$$

Using the decomposition $V_{m,21} = \left[V_{m,21}^1\ V_{m,21}^2\right]$ where $V_{m,21}^1$ is a $N \times K_1$ matrix and $V_{m,21}^2$ is a $N \times K_m$ matrix with $K_m = K_2$ for Model a and $K_m = K_3$ for Model b, we obtain

$$\begin{bmatrix} V_{m,12}^1V_{22}^{-1}V_{m,21}^1 & V_{m,12}^1V_{22}^{-1}V_{m,21}^2 \\ V_{m,12}^2V_{22}^{-1}V_{m,21}^1 & V_{m,12}^2V_{22}^{-1}V_{m,21}^2 \end{bmatrix} \begin{bmatrix} \theta_{m1} \\ \theta_{m2} \end{bmatrix} = \begin{bmatrix} V_{m,12}^1V_{22}^{-1}\mu_2 \\ V_{m,12}^2V_{22}^{-1}\mu_2 \end{bmatrix}.$$

The first row block gives

$$V_{m,12}^1V_{22}^{-1}V_{m,21}^1\theta_{m1} + V_{m,12}^1V_{22}^{-1}V_{m,21}^2\theta_{m2} = V_{m,12}^1V_{22}^{-1}\mu_2.$$

When $\theta_{a2} = 0_{K_2}$ and $\theta_{b2} = 0_{K_3}$, we have

$$\theta_{m1} = \left(V_{m,12}^1V_{22}^{-1}V_{m,21}^1\right)^{-1} V_{m,12}^1V_{22}^{-1}\mu_2,\ m = a, b.$$

It follows from $V_{a,12}^1 = E\left(\widetilde{r}_t f_{1t}'\right) = V_{b,12}^1$, that $\theta_{a1} = \theta_{b1}$.

Lemma 12. *Consider two non-nested models a,b with overlapping factors (f_1, f_2) and (f_1, f_3) as described in Section 4.2, then we have*

$$\delta_a^2 - \delta_b^2 = [\theta_{a2}'\ \theta_{b2}'] \begin{bmatrix} -C_{a,22}^{-1} & 0 \\ 0 & C_{b,22}^{-1} \end{bmatrix} \begin{bmatrix} \theta_{a2} \\ \theta_{b2} \end{bmatrix}$$

where C_m, $m = a, b$ is defined in (12).

Proof of Lemma 12.

Consider the Model 0 which includes only the factor 1: $y_t = 1 - f_{1t}'\theta_0$. Using Equation (11), we have for $m = a, b$,

$$\begin{aligned}\delta_m^2 - \delta_0^2 &= \mu_2' V_{22}^{-1}\mu_2 - \mu_2' V_{22}^{-1} V_{m,21}(V_{m,12}V_{22}^{-1}V_{m,21})^{-1}V_{m,12}V_{22}^{-1}\mu_2 \\ &\quad -\mu_2' V_{22}^{-1}\mu_2 + \mu_2' V_{22}^{-1} V_{0,21}(V_{0,12}V_{22}^{-1}V_{0,21})^{-1}V_{0,12}V_{22}^{-1}\mu_2 \\ &= \mu_2' V_{22}^{-1} V_{0,21}(V_{0,12}V_{22}^{-1}V_{0,21})^{-1}V_{0,12}V_{22}^{-1}\mu_2 \\ &\quad -\mu_2' V_{22}^{-1} V_{m,21}(V_{m,12}V_{22}^{-1}V_{m,21})^{-1}V_{m,12}V_{22}^{-1}\mu_2.\end{aligned}$$

Now using the fact that $V_{0,21} = V_{m,21}^1 = E\left(\widetilde{r}_t f_{1t}'\right)$ and $V_{m,12} = \begin{bmatrix} V_{m,12}^1 \\ V_{m,12}^2 \end{bmatrix}$, we have

$$V_{0,21}(V_{0,12}V_{22}^{-1}V_{0,21})^{-1}V_{0,12} = V_{m,21}\begin{bmatrix} (V_{m,12}^1 V_{22}^{-1} V_{m,21}^1)^{-1} & 0 \\ 0 & 0 \end{bmatrix} V_{m,12}.$$

Exploiting (48), we have

$$\begin{aligned}\delta_m^2 - \delta_0^2 &= \theta_0'(V_{0,12}V_{22}^{-1}V_{0,21})\theta_0 \\ &\quad -\theta_m'(V_{m,12}V_{22}^{-1}V_{m,21})\theta_m.\end{aligned}$$

Using

$$\begin{aligned}(V_{m,12}V_{22}^{-1}V_{m,21}) &= \begin{bmatrix} V_{m,12}^1 \\ V_{m,12}^2 \end{bmatrix} V_{22}^{-1}\left[V_{m,21}^1 \; V_{m,21}^2\right] \\ &= \begin{bmatrix} V_{m,12}^1 V_{22}^{-1} V_{m,21}^1 & V_{m,12}^1 V_{22}^{-1} V_{m,21}^2 \\ V_{m,12}^2 V_{22}^{-1} V_{m,21}^1 & V_{m,12}^2 V_{22}^{-1} V_{m,21}^2 \end{bmatrix}\end{aligned}$$

and combining the different expressions, we obtain

$$\delta_m^2 - \delta_0^2 = -\theta_{m2}' C_{m,22}^{-1}\theta_{m2}.$$

Therefore, $\delta_a^2 - \delta_b^2 = \delta_a^2 - \delta_0^2 - \delta_b^2 + \delta_0^2 = \theta_{b2}' C_{b,22}^{-1}\theta_{b2} - \theta_{a2}' C_{a,22}^{-1}\theta_{a2}$. This concludes the proof of Lemma 12.

Proof of Proposition 4.

When the models are nested, $\theta_{a2} = 0$ and by Lemma 12, we have $\delta_a^2 - \delta_b^2 = \theta_{b2}' C_{b,22}^{-1}\theta_{b2}$.

Given $C_{b,22}^{-1}$ is full rank, $\delta_a^2 - \delta_b^2 = 0$ if and only if $\theta_{b2} = 0$. This is the first result of Proposition 4.

For Tikhonov, under the hypothesis $\theta_{b2} = 0$, $z = \sqrt{T}V(\hat{\theta}_{b2}^{\gamma})^{-\frac{1}{2}}\hat{\theta}_{b2}^{\gamma} \xrightarrow{d} \mathcal{N}(0, I_{K_3})$ as T, N and γT go to infinity and $\gamma^2 T$ goes to zero, by Proposition 2.

$T(\hat{\delta}^2_{a,\gamma} - \hat{\delta}^2_{b,\gamma}) = T\hat{\theta}^{\gamma'}_{b2}\hat{C}^{-1}_{b,22,\gamma}\hat{\theta}^{\gamma}_{b2} = z'V(\hat{\theta}^{\gamma}_{b2})^{\frac{1}{2}}\hat{C}^{-1}_{b,22,\gamma}V(\hat{\theta}^{\gamma}_{b2})^{\frac{1}{2}}z$. $\hat{C}^{-1}_{b,22,\gamma}$ converges in probability to $C^{-1}_{b,22}$ as $\hat{C}^{-1}_{b,22}$ is a continuous function of $\hat{C}_b$ which converges to C_b when N, T and $\gamma T \to \infty$ and $\gamma \to 0$ (see the proof of the consistency of $\hat{\theta}^{\gamma}_{HJ}$). Therefore

$$T(\widehat{\delta}^2_{a,\gamma} - \widehat{\delta}^2_{b,\gamma}) \xrightarrow{d} z'V(\hat{\theta}^{\gamma}_{b2})^{\frac{1}{2}}C^{-1}_{b,22}V(\hat{\theta}^{\gamma}_{b2})^{\frac{1}{2}}z.$$

The result follows from the eigendecomposition of $V(\hat{\theta}^{\gamma}_{b2})^{\frac{1}{2}}C^{-1}_{b,22}V(\hat{\theta}^{\gamma}_{b2})^{\frac{1}{2}}$.

Proof of Proposition 5

The proof of Point 1 follows from Lemma 11.

Now, consider Point 2.

For Tikhonov, under the hypothesis $\begin{bmatrix}\theta_{a2}\\ \theta_{b2}\end{bmatrix} = 0$, we have

$$z = \sqrt{T}V\left(\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix}\right)^{-\frac{1}{2}}\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix} \xrightarrow{d} \mathcal{N}(0, I_{K_2+K_3})$$

as T, N and $\gamma T \to \infty$, $\gamma^2 T \to 0$, and γ goes to zero, where joint convergence follows by applying the Cramér–Wold argument of Proposition 2 to the stacked influence functions of the two models.

Moreover, by Lemma 12,

$$\begin{aligned} T(\hat{\delta}^2_{a,\gamma} - \hat{\delta}^2_{b,\gamma}) &= T\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix}'\begin{bmatrix}-\hat{C}^{-1}_{a,22,\gamma} & 0_{K_2\times K_3}\\ 0_{K_3\times K_2} & \hat{C}^{-1}_{b,22,\gamma}\end{bmatrix}\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix} \\ &= z'V\left(\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix}\right)^{\frac{1}{2}}\begin{bmatrix}-\hat{C}^{-1}_{a,22,\gamma} & 0_{K_2\times K_3}\\ 0_{K_3\times K_2} & \hat{C}^{-1}_{b,22,\gamma}\end{bmatrix}V\left(\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix}\right)^{\frac{1}{2}}z. \end{aligned}$$

The result follows from the eigendecomposition of

$$V\left(\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix}\right)^{\frac{1}{2}}\begin{bmatrix}-\hat{C}^{-1}_{a,22,\gamma} & 0_{K_2\times K_3}\\ 0_{K_3\times K_2} & \hat{C}^{-1}_{b,22,\gamma}\end{bmatrix}V\left(\begin{bmatrix}\hat{\theta}^{\gamma}_{a2}\\ \hat{\theta}^{\gamma}_{b2}\end{bmatrix}\right)^{\frac{1}{2}}.$$

11.7 Proof of Proposition 6

Let A be a self-adjoint positive semi-definite operator, and let $A^{\dagger}$ be its Moore–Penrose inverse. Throughout, we use the notations

$$\langle x, z\rangle_N = \frac{x'z}{N} \text{ and } R_\gamma(A) = \left(A^2 + \gamma I\right)^{-1}A.$$

The following lemma will be used in the proof of Proposition 6.

Lemma 13. *(duality for Tikhonov) For every $a \in \mathbb{R}^N$,*

$$\begin{aligned}\langle a, R_\gamma(A)\, a\rangle &= \max_{\nu\in\mathcal{R}(A)} \left\{2\langle \nu, a\rangle_N - \langle \nu, A\nu\rangle_N - \gamma \left\langle \nu, A^\dagger \nu\right\rangle_N\right\} \\ &= \max_{w\in\mathbb{R}^N} \left\{2\langle w, Aa\rangle_N - \left\langle w, A\left(A^2+\gamma I\right) w\right\rangle_N\right\}.\end{aligned}$$

The maximizer in the first representation is

$$\nu_\gamma = R_\gamma(A)a.$$

A (non unique) maximizer in the second is

$$w_\gamma = (A^2+\gamma I)^{-1}a, \qquad Aw_\gamma = \nu_\gamma.$$

Proof of Lemma 13. Let λ_j and ϕ_j be the eigenvalues and orthonormal eigenvectors of A. For $\nu \in \mathcal{R}(A)$, $\left\langle \nu, \phi_j\right\rangle_N \equiv \nu_j = 0$ whenever $\lambda_j = 0$. Denote $\left\langle a, \phi_j\right\rangle_N \equiv a_j$. For $\lambda_j > 0$, the j-th component of the first criterion is

$$2a_j\nu_j - \left(\lambda_j + \frac{\gamma}{\lambda_j}\right)\nu_j^2.$$

It is maximized at

$$\nu_j = \frac{\lambda_j}{\lambda_j^2+\gamma}a_j,$$

and the maximum is

$$\frac{\lambda_j}{\lambda_j^2+\gamma}a_j^2.$$

Summing over j proves the first equality. Writing $\nu = Aw$ and using $AA^\dagger A = A$ gives the second. □

Applying this lemma to $A = \widehat{\Sigma}$ and $a = \widehat{\alpha}(\theta)$, we obtain

$$\widehat{\delta}_\gamma^2 = \min_\theta \max_{\nu\in\mathcal{R}(\widehat{\Sigma})} \left\{2\langle \nu, \widehat{\alpha}(\theta)\rangle_N - \langle \nu, \widehat{\Sigma}\nu\rangle_N - \gamma\langle \nu, \widehat{\Sigma}^\dagger\nu\rangle_N\right\}. \tag{2}$$

Thus:

- The restriction $\nu \in \mathcal{R}(\widehat{\Sigma})$ is essential when $N > T$.

- The w-representation is preferable for the proof because its domain does not change with the sample.

The Ridge dual is given by:

$$\langle a, (A+\gamma I)^{-1}a\rangle_N = \max_\nu \left\{2\langle \nu, a\rangle_N - \langle \nu, A\nu\rangle_N - \gamma\|\nu\|_N^2\right\}.$$

Proof of Proposition 6

For Ridge, $q_t(\eta_\gamma) = 2y_t\frac{\nu_{1\gamma}'\tilde{r}_t}{N} - \frac{\nu_{1\gamma}'\tilde{r}_t\tilde{r}_t'\nu_{1\gamma}}{N^2} + 2\frac{\nu_{1\gamma}'\mu_2}{N} - \gamma\frac{\nu_{1\gamma}'\nu_{1\gamma}}{N}$.

For Tikhonov, $q_t(\eta_\gamma) = 2y_t\frac{\nu_{1\gamma}'\tilde{r}_t}{N} - \frac{\nu_{1\gamma}'\tilde{r}_t\tilde{r}_t'\nu_{1\gamma}}{N^2} + 2\frac{\nu_{1\gamma}'\mu_2}{N} - \gamma\frac{\nu_{1\gamma}'\Sigma^\dagger\nu_{1\gamma}}{N}$.

We give a detailed proof for Tikhonov.

Define

$$\widehat{\delta}_\gamma^2(\theta) = \left\langle \widehat{\alpha}(\theta), R_\gamma(\widehat{\Sigma})\widehat{\alpha}(\theta) \right\rangle_N,$$

and let $\widehat{\theta}_\gamma$ minimize this criterion.

First, we prove that $\sqrt{T}\left|\widehat{\delta}_\gamma^2(\widehat{\theta}_\gamma) - \widehat{\delta}_\gamma^2(\theta_0)\right| = o_p(1)$.

Let

$$\widehat{C}_\gamma = \frac{\widehat{\beta}' R_\gamma(\widehat{\Sigma})\widehat{\beta}}{N}.$$

We have

$$\begin{aligned}
& \widehat{\delta}_\gamma^2(\theta) - \widehat{\delta}_\gamma^2(\widehat{\theta}_\gamma) \\
= & \left\langle \hat{\mu}_2 - \hat{V}_{21}\theta, R_\gamma(\widehat{\Sigma})\left(\hat{\mu}_2 - \hat{V}_{21}\theta\right)\right\rangle_N \\
& - \left\langle \hat{\mu}_2 - \hat{V}_{21}(\widehat{\theta}_\gamma - \theta) - \hat{V}_{21}\theta, R_\gamma(\widehat{\Sigma})\left(\hat{\mu}_2 - \hat{V}_{21}(\widehat{\theta}_\gamma - \theta) - \hat{V}_{21}\theta\right)\right\rangle_N \\
= & (\theta - \widehat{\theta}_\gamma)\widehat{V}_{11}\widehat{\beta}' R_\gamma(\widehat{\Sigma})\widehat{\beta}\widehat{V}_{11}(\theta - \widehat{\theta}_\gamma)/N.
\end{aligned}$$

Therefore,

$$\widehat{\delta}_\gamma^2(\widehat{\theta}_\gamma) - \widehat{\delta}_\gamma^2(\theta_0) = -(\widehat{\theta}_\gamma - \theta_0)'\widehat{V}_{11}\widehat{C}_\gamma\widehat{V}_{11}(\widehat{\theta}_\gamma - \theta_0). \tag{49}$$

By Proposition 2,

$$\widehat{\theta}_\gamma - \theta_0 = O_p(T^{-1/2})$$

and $\|\widehat{C}_\gamma\|_{op} = O_p(1)$, then

$$\sqrt{T}\left|\widehat{\delta}_\gamma^2(\widehat{\theta}_\gamma) - \widehat{\delta}_\gamma^2(\theta_0)\right| = O_p(T^{-1/2}) = o_p(1). \tag{50}$$

Now, we give an expansion of the Tikhonov quadratic form.

Let $\alpha_0 = \alpha(\theta_0)$, $\widehat{\alpha}(\theta_0) = \sum_t y_t(\theta_0) r_t/T$. Using the fact that $\alpha_0 \in \mathcal{R}\left(\Sigma^2\right)$ by Assumption 3(i), we can write $\alpha_0 = \Sigma^2\phi$ for some $\phi \in \mathbb{R}^N$ such that $\|\phi\|_N < \infty$. Let

$$\widehat{\Sigma} = \Sigma + H, \quad \widehat{\alpha} = \alpha_0 + g,$$

where $H = \widehat{\Sigma} - \Sigma$ and $g = \widehat{\alpha}(\theta_0) - \alpha_0$. Put

$$w_\gamma = (\Sigma^2 + \gamma I)^{-1}\alpha_0, \qquad \nu_\gamma = \Sigma w_\gamma.$$

The source condition gives $\|w_\gamma\|_N \le \|\phi\|_N = O(1)$.

For

$$Q_{\Sigma,\alpha_0}(w) = 2\langle w, \Sigma\alpha_0\rangle_N - \langle w, \Sigma(\Sigma^2+\gamma I)w\rangle_N,$$

Lemma 13 implies

$$\langle \alpha_0, R_\gamma(\Sigma)\alpha_0\rangle_N = \max_w Q_{\Sigma,\alpha_0}(w).$$

Set

$$e = \widehat{\alpha}(\theta_0) - (\widehat{\Sigma}^2+\gamma I)w_\gamma = g - (\Sigma H + H\Sigma + H^2)w_\gamma.$$

We have

$$0 \le \langle \widehat{\alpha}(\theta_0), R_\gamma(\widehat{\Sigma})\widehat{\alpha}(\theta_0)\rangle_N - Q_{\widehat{\Sigma},\widehat{\alpha}(\theta_0)}(w_\gamma) = \langle e, R_\gamma(\widehat{\Sigma})e\rangle_N. \tag{51}$$

Since

$$\|R_\gamma(\widehat{\Sigma})\|_{op} \le \frac{1}{2\sqrt{\gamma}}, \qquad \|e\|_N = O_p(T^{-1/2}),$$

we obtain

$$\langle e, R_\gamma(\widehat{\Sigma})e\rangle_N = O_p\left(\frac{1}{T\sqrt{\gamma}}\right) = o_p(T^{-1/2}) \tag{52}$$

under $\gamma T \to \infty$.

Direct expansion gives

$$\begin{aligned} Q_{\widehat{\Sigma},\widehat{\alpha}(\theta_0)}(w_\gamma) - Q_{\Sigma,\alpha_0}(w_\gamma) &= 2\left\langle \nu_\gamma, g\right\rangle_N - \left\langle \nu_\gamma, H\nu_\gamma\right\rangle_N \\ &\quad +\gamma\left\langle w_\gamma, Hw_\gamma\right\rangle_N + \rho_T \end{aligned}$$

where

$$\rho_T = 2\langle w_\gamma, Hg\rangle_N - \langle w_\gamma, (\Sigma H^2 + H\Sigma H + H^2\Sigma + H^3)w_\gamma\rangle_N.$$

Using Lemmas 5 and 9,

$$|\rho_T| = O_p(T^{-1}) = o_p(T^{-1/2}). \tag{53}$$

Moreover,

$$\begin{aligned} \|\nu_\gamma - \nu_0\|_N &= \left\|\gamma\Sigma(\Sigma^2+\gamma I)^{-1}\phi\right\|_N \\ &= \left\|\gamma\Sigma(\Sigma^2+\gamma I)^{-1/2}(\Sigma^2+\gamma I)^{-1/2}\phi\right\|_N \le \sqrt{\gamma}\|\phi\|_N, \end{aligned} \tag{54}$$

where $\nu_0 = \Sigma^{-1}\alpha_0$. Also,

$$\sqrt{T}\,\gamma\langle w_\gamma, Hw_\gamma\rangle_N = O_p(\gamma) = o_p(1). \tag{55}$$

Equations (51)–(55) therefore imply

$$\sqrt{T}(\widehat{\delta}_\gamma^2 - Q_{\Sigma,\alpha_0}(w_\gamma)) = 2\sqrt{T}\left\langle \nu_0, g\right\rangle_N - \sqrt{T}\left\langle \nu_0, H\nu_0\right\rangle_N + o_p\left(1\right).$$

Finally, the population regularization bias is

$$\begin{aligned} Q_{\Sigma,\alpha_0}(w_\gamma) - \delta^2 &= \left\langle \alpha_0, \left(\Sigma^2 + \gamma I\right)^{-1} \Sigma\alpha_0 \right\rangle_N - \left\langle \alpha_0, \Sigma^{-1}\alpha_0 \right\rangle_N \\ &= \left\langle \alpha_0, \left(\Sigma^2 + \gamma I\right)^{-1} \Sigma^3\phi \right\rangle_N - \left\langle \alpha_0, \Sigma\phi \right\rangle_N \\ &= \left\langle \alpha_0, \left(\Sigma^2 + \gamma I\right)^{-1} \left[\Sigma^3 - \left(\Sigma^2 + \gamma I\right)\Sigma\right] \phi \right\rangle_N \\ &= -\gamma \left\langle \alpha_0, \left(\Sigma^2 + \gamma I\right)^{-1} \Sigma\phi \right\rangle_N \\ &= -\gamma \left\langle \phi, \Sigma^3 \left(\Sigma^2 + \gamma I\right)^{-1} \phi \right\rangle_N \end{aligned}$$

and hence

$$|Q_{\Sigma,\alpha_0}(w_\gamma) - \delta^2| \leq \gamma \|\Sigma\|_{op} \|\phi\|_N^2 = O(\gamma). \tag{56}$$

Thus $\sqrt{T}[Q_{\Sigma,\alpha_0}(w_\gamma) - \delta^2] = o(1)$ under $\gamma^2 T \to 0$.

Combining (50) and (56),

$$\begin{aligned} \sqrt{T}(\widehat{\delta}_\gamma^2 - \delta^2) &= \sqrt{T}(\widehat{\delta}_\gamma^2 - Q_{\Sigma,\alpha_0}(w_\gamma) + Q_{\Sigma,\alpha_0}(w_\gamma) - \delta^2) \\ &= 2\sqrt{T}\langle \nu_0, g \rangle_N - \sqrt{T}\langle \nu_0, H\nu_0 \rangle_N + o_p(1). \end{aligned} \tag{57}$$

We now derive the expression of the Influence function.

Let

$$y_t = 1 - \widetilde{f}_t'\theta_0, \qquad u_t = \frac{\widetilde{r}_t'\nu_0}{N}.$$

Sample centering gives

$$g = \frac{1}{T}\sum_{t=1}^{T} [\widetilde{r}_t y_t + V_{21}\theta_0] + o_p(T^{-1/2}).$$

The population first-order condition is

$$V_{12}\nu_0 = 0,$$

so the $V_{21}\theta_0$ term vanishes after multiplication by ν_0. Similarly,

$$\langle \nu_0, H\nu_0 \rangle_N = \frac{1}{T}\sum_{t=1}^{T} (u_t^2 - Eu_t^2) + o_p(T^{-1/2}),$$

and

$$Eu_t^2 = \langle \nu_0, \Sigma\nu_0 \rangle_N = \delta^2.$$

Consequently,

$$\sqrt{T}\left(\widehat{\delta}_\gamma^2 - \delta^2\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^{T} \ell_t + o_p(1), \ \ell_t = 2y_t u_t - u_t^2 + \delta^2. \tag{58}$$

Note that the variable ℓ_t is centered. Under the strong mixing and moment assumptions and Assumption 6, the triangular-array mixing CLT applied to (58) proves the result of Proposition 6.

For finite-sample standard errors, a more accurate finite-γ influence function contains the centered operator term

$$\gamma\left(v_{\gamma t}^2 - Ev_{\gamma t}^2\right), \qquad \text{with } v_{\gamma t} = \frac{\widetilde{r}_t' w_\gamma}{N}.$$

To see this, note that

$$\begin{aligned}\frac{1}{T}\sum_t \left(\frac{\overline{r}_t' w_\gamma}{N}\right)^2 &= \frac{1}{T}\sum_t \frac{w_\gamma' \overline{r}_t \overline{r}_t' w_\gamma}{N^2} \\ &= \left\langle w_\gamma, \widehat{\Sigma} w_\gamma \right\rangle_N\end{aligned}$$

and hence

$$\gamma \left\langle w_\gamma, H w_\gamma \right\rangle_N = \frac{1}{T}\sum_t \gamma\left(v_{\gamma t}^2 - Ev_{\gamma t}^2\right) + o_p\left(T^{-1/2}\right)$$

where the $o_p\left(T^{-1/2}\right)$ comes from replacing $\widetilde{r}_t$ by $\overline{r}_t$. This term disappears asymptotically but may matter when γ is relatively large.

11.8 Proof of Proposition 7

For $m = a, b$, Proposition 6 gives jointly

$$\sqrt{T}\left(\widehat{\delta}_{m,\gamma}^2 - \delta_m^2\right) = \frac{1}{\sqrt{T}}\sum_{t=1}^T \ell_{m,t} + o_p(1). \tag{59}$$

Let $L_{N,t} = (\ell_{a,t}, \ell_{b,t})'$. For every fixed $c \in \mathbb{R}^2$, $c'L_{N,t}$ is centered, inherits the mixing coefficients of the common data process, and has a uniformly bounded $2+\eta$ moment. Therefore the scalar triangular-array mixing CLT gives

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T c' L_{N,t} \xrightarrow{d} \mathcal{N}(0, c'\Omega c).$$

By Cramér–Wold,

$$\frac{1}{\sqrt{T}}\sum_{t=1}^T L_{N,t} \xrightarrow{d} \mathcal{N}(0, \Omega).$$

Subtracting the two expansions in (59) gives

$$\sqrt{T}\left[\left(\widehat{\delta}_{a,\gamma}^2 - \widehat{\delta}_{b,\gamma}^2\right) - \left(\delta_a^2 - \delta_b^2\right)\right] \xrightarrow{d} \mathcal{N}(0, v_3).$$

Under $H_0 : \delta_a^2 = \delta_b^2$,

$$\sqrt{T}(\widehat{\delta}_{a,\gamma}^2 - \widehat{\delta}_{b,\gamma}^2) \xrightarrow{d} \mathcal{N}(0, v_3).$$