Taming the Factor Zoo*

Guanhao Feng† — Booth School of Business, University of Chicago

Stefano Giglio‡ — Booth School of Business, University of Chicago

Dacheng Xiu§ — Booth School of Business, University of Chicago

This Version: April 4, 2017

Abstract

The asset pricing literature has produced hundreds of potential risk factors. Organizing this "zoo of factors" and distinguishing between useful, useless, and redundant factors require econometric techniques that can deal with the curse of dimensionality. We propose a model-selection method that allows us to systematically evaluate the contribution to asset pricing of any new factor, above and beyond what a high-dimensional set of existing factors explains. Our procedure selects the best parsimonious model out of the large set of existing factors, and uses it as the control in making statistical inference about the contribution of new factors. Our inference allows for model-selection mistakes, and is therefore more reliable in finite sample. We derive the asymptotic properties of our test and apply it to a large set of factors proposed in the literature. We show that despite the fact that hundreds of factors have been proposed in the last 30 years, some recent factors – such as profitability – have statistically significant explanatory power in addition to existing ones. We confirm the effectiveness of our procedure in discriminating factors in a recursive and out-of-sample experiment, and show our procedure results in a parsimonious model with a small number of factors and high cross-sectional explanatory power, even as the pool of candidate factors has expanded dramatically.

Key words: Factors, Risk Price, Post-Selection Inference, Regularized Two-Pass Estimation, Machine Learning, LASSO.

*We appreciate insightful comments from John Campbell, John Cochrane, and Chris Hansen. We are also grateful to helpful comments from seminar and conference participants at the City University of Hong Kong and the 2016 Financial Engineering and Risk Management Symposium in Guangzhou. We acknowledge research support by the Fama-Miller Center for Research in Finance at Chicago Booth.

†Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: guanhao.feng@chicagobooth.edu.

‡Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: stefano.giglio@chicagobooth.edu.

§Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: dacheng.xiu@chicagobooth.edu.

1 Introduction

The search for factors that explain the cross section of expected stock returns has produced hundreds of potential factor candidates, as noted by Cochrane (2011) and most recently by Harvey et al. (2015), McLean and Pontiff (2016), and Hou et al. (2016). A fundamental task facing the asset pricing field today is to bring more discipline to this "zoo" of factors. In particular, how can we discriminate truly useful pricing factors from redundant and useless factors that appear significant due to data mining? How do we judge whether a new factor adds explanatory power for asset pricing, relative to the existing set of hundreds of factors the literature has so far produced?

This paper provides a framework for systematically evaluating the contribution of individual factors relative to the myriad of existing factors the literature has proposed, and conducting appropriate statistical inference in this high-dimensional setting. In particular, we show how to estimate and test the marginal importance of any factor g_t in pricing the cross section of expected returns *beyond* what is explained by a high-dimensional set of potential factors h_t, where g_t and h_t could be tradable or non-tradable factors. We assume the true asset pricing model is approximately low-dimensional; however, in addition to relevant asset pricing factors, g_t and h_t include redundant ones that add no explanatory power to the other factors, as well as useless ones that have no explanatory power at all. Selecting the relevant factors from h_t to conduct proper inference on the contribution of g_t is the aim of this paper.

When h_t consists of a small number of factors, testing whether g_t is useful in explaining asset prices while controlling for the factors in h_t is straightforward: it simply requires estimating the loadings of the stochastic discount factor on g_t and h_t (i.e., the *price of risk* of these factors), and testing whether the price of risk of g_t is different from zero (see Cochrane (2009)). This exercise not only tells us whether g_t is useful for pricing the cross section, but it also reveals how shocks to g_t affect marginal utility (which has a direct economic interpretation).

When h_t consists of potentially hundreds of factors, however, standard statistical methods to estimate and test risk prices become infeasible or result in poor estimates and invalid inference, because of the curse of dimensionality. Although dimension-reduction techniques (e.g., least absolute shinkage and selection operator, LASSO) can be useful in (asymptotically) selecting the right model and reducing the dimensionality of h_t, they produce erroneous *inference* unless appropriate econometric methods are used to explicitly account for the model-selection mistakes that can occur in any finite sample (see Chernozhukov et al. (2015)).

The methodology we propose in this paper marries these new econometric methods (in particular, the double-selection LASSO method of Belloni et al. (2014b)) with two-pass regressions such as Fama-MacBeth to specifically estimate risk prices in a high-dimensional setting. Without relying on

prior knowledge about which factors to include as controls among the hundreds in h_t, our procedure selects factors that are *either* useful in explaining the cross section of expected returns *or* are useful in mitigating the omitted variable bias problem. We show that including both types of factors as controls is essential to conduct correct inference on the price of risk of g_t.

We apply our methodology to a large set of factors proposed in the last 30 years. We collect and construct a factor data library, containing more than 100 risk factors, that includes both tradable and nontradable factors. We perform a variety of empirical exercises to illustrate the importance of taking into account model-selection mistakes when conducting inference about risk prices and assessing the importance of new factors. We start by evaluating the marginal contribution of recent factors proposed in the last five years to the large set of factors proposed before 2010. The new factors include – among others – the two new factors introduced by Fama and French (2015) and two intermediary-based factors from He et al. (2016) and Adrian et al. (2014). Given that the set of potential control factors includes more than 100 factors, one might wonder whether, in practice, any additional factor could make any significant contribution to explaining the cross section of expected returns. We show that indeed some of these new factors (e.g., profitability) have significant marginal explanatory power for expected returns.

We also confirm the ability of our procedure to select useful factors through a recursive exercise. In each year since 1994, we consider all factors introduced in that year and test whether they add significant explanatory power for the cross section of expected returns to the set of factors available up to that point. We then verify that the factors our test deems to have been useful when they were introduced actually appear to be useful in pricing assets in later years.

Over time, the total number of factors available has increased. We can therefore use our recursive exercise to also study how the selected model evolves over time, as more and more factors are added to the set h_t (from around 20 in 1994 to over 100 in 2016). Two results emerge. First, although the number of factors is not determined ex ante, but is chosen optimally, it is remarkably stable over time. Despite the fact that the number of potential factors increased by a factor of 5 over the last 20 years, the number of factors selected by the model increases modestly (from 10 in 1994 to 18 in 2016), and, in fact, the number of *significant* factors among those is close to constant at around five for the entire period. At the same time, the cross-sectional R^2 achieved by the selected model has improved as well, an indication that new research has indeed discovered better factors over time, and our procedure successfully selects them. This result is consistent with the finding that even now, with an existing pool of hundreds of factors, we still find significant contributions from several of the newest factors.

The stability of the number of selected factors over time is also striking when compared to the number of factors that seem to command a significant risk premium in the literature (documented,

e.g., by Harvey et al. (2015)). Of the around 100 factors we examine, about half of them have a risk premium significant with a t-stat above 2, and about one sixth have a risk premium significant with a t-stat above 3. These fractions of significance over total factors have been remarkably stable over time. We would therefore expect that, going forward, as more and more potential factors are introduced, an increasing number of them will carry a significant risk premium, even using the higher significance level imposed by requiring a t-stat of 3. This highlights the importance of evaluating a factor based on its contribution relative to the existing factors (risk price, which we estimate in our paper), rather than its risk premium (which does not capture the marginal contribution of a factor to explaining asset prices, but simply the correlation with the stochastic discount factor): most of the new factors with positive risk premia are simply redundant relative to the existing factors. Our procedure only keeps, over time, the most relevant factors, and weeds out those that do not contain any additional information to explain asset prices, thus naturally bringing discipline to this growing set of factors.

Of course, the existing literature has routinely attempted to evaluate the contribution of new factors relative to some benchmark model, typically by estimating and testing the alpha of a regression of the new factor onto existing factors (e.g., Barillas and Shanken (2015) and Fama and French (2016)). Our methodology differs from the existing procedures in several ways. First, we do not select the control model in an ad-hoc way (e.g., using the three Fama-French factors), but rather we select the control model that best explains the cross section of returns. In addition, our procedure aims to minimize the potential omitted variable bias while enhancing statistical efficiency. Second, we not only test whether the factor of interest g_t is useful in explaining asset prices, but we also estimate its role in driving marginal utility (its coefficient in the stochastic discount factor, or risk price). Third, we handle both traded and non-traded factors. Lastly, our inference is valid given a large dimensional set of controls and test assets in addition to an increasing span of time series.

The methodology we propose builds on the double-selection technique of Belloni et al. (2014b) (that was proposed for linear treatment effect models), combining it with two-step cross-sectional regressions. Our procedure first uses a double-selection method to select "control" factors from h_t, and then estimates the risk price of g_t from cross-sectional regressions that include g_t and the selected factors from h_t.

As the name implies, the "double selection" of factors from h_t happens in two stages; both stages are crucial to obtain correct inference on g_t. A first set of factors is selected from h_t based on their pricing ability for the cross section of returns. Factors that appear to contribute little to pricing assets are excluded from the set of controls. This first step has the advantage of selecting factors based on their usefulness in pricing the cross section of assets, as opposed to other commonly used selection methods (e.g., principal components) that select factors based on their ability to explain

the time-series variation of returns.

This first step is, however, not sufficient to ensure valid inference on g_t, because the first selection may exclude some factors that have small risk prices in sample, but whose covariance with returns are nonetheless highly cross-sectionally correlated with that of g_t. That is, in any finite sample, we can never be sure of having selected the correct model. Any omission of relevant factors due to finite-sample model-selection errors distorts the inference on the risk price of g_t, leading to incorrect inference on the significance – and even the sign – of g_t. This issue is a well-known problem with model-selection methods; see, for example, Leeb and Pötscher (2005), which has spurred a large econometrics literature on uniformly valid inference, and has important consequences for asset pricing tests that require selecting a model in large dimensional settings.

Instead, we demonstrate that to obtain correct asymptotic inference for g_t, including a second stage of factor selection is crucial. The second step adds to the set of controls additional factors whose covariances with returns are highly correlated in the cross section with the covariance between returns and g_t. Intuitively, we want to make sure to include even factors with small in-sample risk prices, if omitting them may still induce a large omitted variable bias due to the cross-sectional correlation between their risk exposures and the risk exposures to g_t. This procedure takes explicitly into account the fact that model-selection procedures can never guarantee the model selected from h_t is the true one in any finite sample.

After selecting the set of controls from h_t (including all factors selected in either of the two selection stages), we conduct inference on g_t by estimating the coefficient of a standard two-pass regression using g_t and the selected small number of control factors from h_t. This post-selection estimation step is also useful to remove biases arising from regularization in any LASSO procedure; see, for example, Friedman et al. (2009). We then conduct asymptotic inference on the risk price of g_t using a central-limit result we derive in this paper. We show by simulation that our estimator performs well in finite samples, and substantially outperforms alternative estimators.

Our paper builds on several strands of the asset pricing and econometrics literature. First and most directly, the paper is related to the recent literature on the high dimensionality of cross-sectional asset pricing models. Green et al. (2016) test 94 firm characteristics through Fama-Macbeth regressions and find 8-12 characteristics are significant independent determinants of average returns. McLean and Pontiff (2016) use an out-of-sample approach to study the post-publication bias of 97 discovered risk anomalies. Harvey et al. (2015) adopt a multiple testing framework to re-evaluate past research and suggest a new benchmark for current and future factor fishing. Following on this multiple-testing issue, Harvey and Liu (2016) provide a bootstrap technique under factor orthogonalization. Recently, Freyberger et al. (2017) propose a group LASSO procedure to select characteristics and to estimate how they affect expected returns nonparametrically.

The paper naturally builds on a large literature that has identified a variety of pricing factors, starting with the CAPM of Sharpe (1964) and Lintner (1965). Among the factors that literature has proposed, some are based on economic theory (e.g., Breeden (1979), Chen et al. (1986), Jagannathan and Wang (1996), Lettau and Ludvigson (2001), Yogo (2006), Pástor and Stambaugh (2003), Adrian et al. (2014), He et al. (2016)); others have been constructed using firm characteristics, such as Fama and French (1993, 2015), Carhart (1997), and Hou et al. (2014). Excellent reviews of cross-sectional asset pricing include Campbell (2000), Lewellen et al. (2010), Goyal (2012), and Nagel (2013).

We also build upon the econometrics literature devoted to the estimation and testing of asset pricing models using two-pass regressions, dating back to Jensen et al. (1972) and Fama and MacBeth (1973). Over the years, the econometric methodologies have been refined and extended; see, for example, Ferson and Harvey (1991), Shanken (1992), Jagannathan and Wang (1996), Welch (2008), and Lewellen et al. (2010). These papers, along with the majority of the literature, rely on large T and fixed n asymptotic analysis for statistical inference and only deal with models in which all factors are specified and observable. Bai and Zhou (2015) and Gagliardini et al. (2016) extend the inferential theory to the large n and large T setting, which delivers better small-sample performance when n is large relative to T. Connor et al. (2012) use semiparametric methods to model time variation in the risk exposures as a function of observable characteristics, again when both n and T are large. Giglio and Xiu (2016) rely on a similar large n and large T analysis, but estimate risk premia (not risk prices as in this paper) in the case in which not all relevant pricing factors are observed. Raponi et al. (2016), on the other hand, study the ex-post risk premia using large n and fixed T asymptotics. For a review of this literature, see Shanken (1996), Jagannathan et al. (2010), and more recently, Kan and Robotti (2012).

A more recent literature has focused on various pitfalls in estimating and testing linear factor models. For instance, ignoring model misspecification and identification failure leads to an overly positive assessment of the pricing performance of spurious (Kleibergen (2009)) or even useless factors (Kan and Zhang (1999a,b); Jagannathan and Wang (1998)), and biased risk-premia estimates of true factors in the model. Therefore, the use of inference methods that are robust to model misspecification is more reliable (Shanken and Zhou (2007); Kan and Robotti (2008); Kleibergen (2009); Kan and Robotti (2009); Kan et al. (2013); Gospodinov et al. (2013); Kleibergen and Zhan (2014); Gospodinov et al. (2014b); Bryzgalova (2015); Burnside (2016)). We study a different model-misspecification form – priced factors omitted from the model, which would also bias the estimates for the observed factors.

Last but not least, our paper is related to a large statistical and machine-learning literature on variable selection and regularization using LASSO and post-selection inference. For theoretical properties of LASSO, see Bickel et al. (2009), Meinshausen and Yu (2009), Tibshirani (2011), Wainwright

(2009), Zhang and Huang (2008), Belloni and Chernozhukov (2013). For the post-selection-inference method, see, for example, Belloni et al. (2012), Belloni et al. (2014b), and review articles by Belloni et al. (2014a) and Chernozhukov et al. (2015). Our asymptotic results are new to the existing literature in two important respects. First, our setting is a large panel regression with a large number of factors (p), in which both cross-sectional and time-series dimensions (n and T) increase. Second, our procedure in fact selects covariances between factors and returns, which are contaminated by estimation errors, rather than factors themselves that are immediately observable.

The rest of the paper is organized as follows. In Section 2, we set up the model, present our methodology, and develop relevant statistical inference. Section 3 provides Monte Carlo simulations that demonstrate the finite-sample performance of our estimator. In Section 4, we show several empirical applications of the procedure. Section 5 concludes. The appendix contains technical details.

2 Methodology

2.1 Model Setup

We set up the model with a linear specification of the stochastic discount factor (SDF):

$$m_t := \gamma_0^{-1} - \gamma_0^{-1}\lambda_v^\intercal v_t := \gamma_0^{-1}(1 - \lambda_g^\intercal g_t - \lambda_h^\intercal h_t), \tag{1}$$

where γ_0 is the zero-beta rate, g_t is a $d \times 1$ vector of factors to be tested, and h_t is a $p \times 1$ vector of potentially confounding factors. Both g_t and h_t are de-meaned; that is, they are factor innovations satisfying $\mathrm{E}(g_t) = \mathrm{E}(h_t) = 0$. λ_g and λ_h are $d \times 1$ and $p \times 1$ vectors of parameters, respectively. We refer to λ_g and λ_h as the risk prices of the factors g_t and h_t.

Our goal here is to make inference on the risk prices of a small set of factors g_t while accounting for the explanatory power of a large number of existing factors, collected in h_t. These factors are not necessarily all useful factors: their corresponding risk prices may be equal to zero. This framework potentially includes completely useless factors (factors that have a risk price of zero and whose covariances with returns are uncorrelated with the covariances of returns and the SDF), as well as redundant factors (factors that have a price of zero but whose covariances with returns are correlated in the cross section with the covariance between returns and the SDF).

We want to estimate and test the risk price of g_t for two reasons. First, it directly reveals whether g_t drives the SDF after controlling for h_t, that is, whether g_t contains additional pricing information relative to h_t (Cochrane (2009)), or whether it is instead redundant or useless. Second, the coefficient λ_g indicates *how* g_t affects marginal utility. For example, a positive sign for λ_g tells us that states where g_t is low are high-marginal-utility states. The estimate of λ_g can therefore be used to test predictions of asset pricing models about how investors perceive g_t shocks.

In addition to g_t and h_t, we observe a $n \times 1$ vector of test asset returns, r_t. Because of (1), the expected return satisfies:

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + C_v \lambda_v = \iota_n \gamma_0 + C_g \lambda_g + C_h \lambda_h, \tag{2}$$

where ι_n is a $n \times 1$ vector of 1s, $C_a = \mathrm{Cov}(r_t, a_t)$, for $a = g, h$, or v. Furthermore, we assume the dynamics of r_t follow a standard linear factor model:

$$r_t = \mathrm{E}(r_t) + \beta_g g_t + \beta_h h_t + u_t, \tag{3}$$

where β_g and β_h are $n \times d$ and $n \times p$ factor-loading matrices, u_t is a $n \times 1$ vector of idiosyncratic components with $\mathrm{E}(u_t) = 0$ and $\mathrm{Cov}(u_t, v_t) = 0$.

Equation (2) represents expected returns in terms of (univariate) covariances with the factors, multiplied by risk prices λ_g and λ_h. An equivalent representation of expected returns can be obtained in terms of multivariate betas:

$$\mathrm{E}(r_t) = \iota_n \gamma_0 + \beta_g \gamma_g + \beta_h \gamma_h, \tag{4}$$

where β_g and β_h are the factor exposures (i.e., multivariate betas) and γ_g and γ_h are the *risk premia* of the factors. Risk prices λ and risk premia γ are directly related through the covariance matrix of the factors, but they differ substantially in their interpretation. In this paper, we aim to estimate the risk *prices* of the factors g_t, not their risk premia. The risk premium γ of a factor tells us whether investors are willing to pay to hedge a certain risk factor, but it does not tell us whether that factor is useful in pricing the cross section of returns. For example, a factor could command a nonzero risk premium without even appearing in the SDF, by simply being correlated with the true factors. As discussed extensively in Cochrane (2009), to understand whether a factor is useful in pricing the cross section of assets, we want to study its risk price λ, not its risk premium γ.

Because the link between risk prices and risk premia depends on the covariances among factors, it is useful to write explicitly the projection of g_t on h_t as

$$g_t = \eta h_t + z_t, \quad \text{where} \quad \mathrm{Cov}(z_t, h_t) = 0. \tag{5}$$

Finally, for the estimation of λ_g, it is essential to characterize the cross-sectional dependence between C_g and C_h, so we write the cross-sectional projection of C_g onto C_h as:

$$C_g = \iota_n \xi^\intercal + C_h \chi^\intercal + C_e, \tag{6}$$

where ξ is a $d \times 1$ vector, χ is a $d \times p$ matrix, and C_e is a $n \times d$ matrix of cross-sectional regression residuals.[1]

[1] For the sake of clarity and simplicity, we assume the set of testing assets used is not sampled randomly but deterministically, so that these covariances and loadings are treated as non-random. This is without loss of generality, because their sampling variation does not affect the first-order asymptotic inference. By contrast, Gagliardini et al. (2016) consider random loadings as a result of a random sampling scheme from a continuum of assets.

2.2 Challenges with Standard Two-Pass Methods

Using two-pass regressions to estimate empirical asset pricing models dates back to Jensen et al. (1972) and Fama and MacBeth (1973). Partly because of its simplicity, this approach is widely used in practice. The procedure involves two steps, including one asset-by-asset time-series regression to estimate individual factor loadings βs, and one cross-sectional regression of expected returns on the estimated factor loadings to estimate risk premia γ.

Because our parameter of interest is the risk price of g_t, λ_g, instead of the risk premium, the first step needs to be modified to use covariances between returns and factors rather than factor betas. In a low-dimensional setting, this method would work smoothly for the estimation of λ_g, as pointed out by Cochrane (2009).

Nevertheless, the empirical asset pricing literature has created hundreds of factors, which can include useless and redundant factors in addition to useful factors; all and only the useful ones should be used as controls in estimating the risk price of newly proposed factors g_t and testing for their contribution to asset pricing (λ_g). Over time, the number of potential factors p discovered in the literature has increased to the same scale as, if not greater than, n or T. In such a scenario, the standard cross-sectional regression with all factor covariances included is at best highly inefficient, because the optimal convergence rate in this regression is $p^{1/2}n^{-1/2}$. Moreover, if p is smaller than n yet of the same scale, asymptotic inference fails entirely to converge. When p is larger than n, the regression approach becomes infeasible because the number of parameters exceeds the sample size.

Standard methodologies therefore do not work well if at all in a high-dimensional setting due to the curse of dimensionality, so that dimension-reduction and regularization techniques are inevitable for valid inference. The existing literature has so far employed ad hoc solutions to this dimensionality problem. To test a new factor, cherry-picking a handful of control factors, such as the prominent Fama-French three factors, is common, effectively imposing an assumption that the selected model is the true one (and is not missing any additional factors). However, this assumption is clearly unrealistic: these standard models have generally poor performance in explaining the large available cross section of expected returns, indicating omitted factors are likely to be present in the data. The stake of selecting an incorrect model is high, because it leads to model misspecification and omitted variable bias when useful factors are not included (Giglio and Xiu (2016)). Relatedly, including useless factors may also lead to incorrect inference (Kan and Zhang (1999b)).

2.3 A Regularized Two-Pass Regression Approach

This issue is not unique to asset pricing. To address it, we need to impose a certain low-dimensional structure in the model. In this paper, we impose a sparsity assumption that has a natural economic

interpretation and has recently been studied at length in the machine-learning literature. Imposing sparsity in our setting means a relatively small number of factors exist in h_t, whose linear combinations along with g_t yield the SDF m_t, and that alone are relevant for the estimation of λ_g. More specifically, sparsity in our setting means there are only s non-zero entries in λ_h, and in each row of η and χ, where s is small relative to n and T. The sparsity assumption allows us to extract the most influential factors, while making valid inference on the parameters of interest, without *prior knowledge* or *perfect recovery* of the useful factors that determine m_t.

To leverage sparsity, Tibshirani (1996) proposes the so-called LASSO estimator, which incorporates into the least-squares optimization a penalty function on the $\mathbb{L}_1$ norm of parameters, which leads to an estimator that has many zero coefficients in the parameter vector. The LASSO estimator has appealing properties in particular for prediction purposes. With respect to parameter estimation, a well-documented finite-sample bias is associated with the non-zero coefficients of the LASSO estimate because of the regularization. For these reasons, Belloni and Chernozhukov (2013) and Belloni et al. (2012) suggest the use of a "Post-LASSO" estimator, which has more desirable statistical properties. The Post-LASSO estimator runs LASSO as a model selector, and then re-fits the least-squares problem without penalty, using only variables that have non-zero coefficients in the first step.

In the asset pricing context, the LASSO and Post-LASSO procedures could theoretically be used to select the factors in h_t with non-zero risk prices as controls for g_t, therefore accounting for the possibility that h_t contains useless or redundant factors. In fact, when the number of factors is large, LASSO and Post-LASSO will asymptotically recover the true model under certain assumptions.

Unfortunately, these procedures are not appropriate when we want to conduct *inference* about risk prices (e.g., about the price of g_t as in our context), because in any finite sample, we can never be sure LASSO or Post-LASSO will select the correct model from h_t. But if the model is misspecified, that is, if important factors from h_t are excluded, inference about risk prices will be affected by an omitted variable bias. Therefore, standard LASSO or Post-LASSO regressions will generally yield erroneous inference about risk prices, as we confirm in simulations in Section 3.

This omitted variable bias due to model-selection mistakes is exacerbated if risk exposures to the omitted factors are highly correlated in the cross section with the exposures to g_t (even though these factors may have a small in-sample price of risk, which is why they may be omitted by LASSO). We will therefore need to ensure these factors are included in the set of controls *even if LASSO would suggest excluding them.* Note this problem is not unique to high-dimensional problems – see, for example, Leeb and Pötscher (2005) – but it is arguably more severe in such a scenario because model selection is inevitable.

To guard against omitted variable biases due to selection mistakes, we therefore adopt a double-selection strategy in the same spirit as what Belloni et al. (2014b) propose for estimating the treatment effect. The first selection searches for factors in h_t whose covariances with returns are useful for explaining the cross section of expected returns. A second selection is then added to search for factors in h_t potentially missed from the first step, but that, if omitted, would induce a large omitted variable bias. Factors excluded from both stages of the double-selection procedure must have small risk prices and have covariances that correlate only mildly in the cross section with the covariance between factors of interest g_t and returns – these factors can be excluded with minimal omitted variable bias. This strategy results in a parsimonious model that minimizes the omitted factor bias ex ante when estimating and testing λ_g.

The regularized two-pass estimation proceeds as follows:

(1) Variable Selection

(1.a) Run a cross-sectional LASSO regression of average returns on sample covariances between factors in h_t and returns:[2]

$$\min_{\gamma,\lambda}\left\{n^{-1}\left\|\bar{r}-\iota_n\gamma-\widehat{C}_h\lambda\right\|^2+\tau_0 n^{-1}\|\lambda\|_1\right\}, \tag{7}$$

where $\widehat{C}_h=\widehat{\text{Cov}}(r_t,h_t)=T^{-1}\bar{R}\bar{H}^\intercal$.[3] This step selects among the factors in h_t those that best explain the cross section of expected returns. Denote $\{\widehat{I}_1\}$ as the set of indices corresponding to the selected factors in this step.

(1.b) Run d LASSO regressions using the covariance between the jth factor in g_t and returns sequentially on the covariances between each factor in h_t and returns for $j=1,\cdots,d$:

$$\min_{\xi_j,\chi_{j,\cdot}}\left\{n^{-1}\left\|(\widehat{C}_{g,\cdot,j}-\iota_n\xi_j-\widehat{C}_h\chi_{j,\cdot}^\intercal)\right\|^2+\tau_j n^{-1}\|\chi_{j,\cdot}^\intercal\|_1\right\}. \tag{8}$$

This step identifies factors whose exposures are highly correlated to the exposures to g_t in the cross-section. This is the crucial second step in the double-selection algorithm, that identifies factors that may be missed by the first step but that may still induce large omitted variable bias in the estimation of λ_g if omitted, due to their covariance properties. Denote $\{\widehat{I}_{2,j}\}$ as the set of indices corresponding to the selected factors in the $j$the regression, and $\widehat{I}_2=\bigcup_{j=1}^d\widehat{I}_{2,j}$.

(2) Post-selection Estimation

[2] We use $\|A\|$ and $\|A\|_1$ to denote the operator norm and the $\mathbb{L}_1$ norm of a matrix $A=(a_{ij})$, that is, $\sqrt{\lambda_{\max}(A^\intercal A)}$, $\max_j\sum_i|a_{ij}|$, where $\lambda_{\max}(\cdot)$ denotes the largest eigenvalue of a matrix.

[3] For any matrix $A=(a_1:a_2:\ldots a_T)$, we write $\bar{a}=T^{-1}\sum_{t=1}^T a_t$, $\bar{A}=A-\iota_T^\intercal\bar{a}$.

Run an OLS cross-sectional regression using covariances between the selected factors from *both* steps and average returns:

$$(\widehat{\gamma}_0, \widehat{\lambda}_g, \widehat{\lambda}_h) = \arg \min_{\gamma_0, \lambda_g, \lambda_h} \left\{ \left\| \bar{r} - \iota_n \gamma_0 - \widehat{C}_g \lambda_g - \widehat{C}_h \lambda_h \right\|^2 : \quad \lambda_{h,j} = 0, \quad \forall j \notin \widehat{I} = \widehat{I}_1 \bigcup \widehat{I}_2 \right\}. \quad (9)$$

We refer to this procedure as a double-selection approach, as opposed to the single-selection approach which only involves (1.a) and (2).

The LASSO estimators involve only convex optimizations, so that the implementation is quite fast. Statistical software such as R and Matlab have existing packages that implement LASSO using efficient algorithms. Note that other variable-selection procedures are also applicable. For instance, the second selection (1.b) can instead adopt group LASSO, for example, Yuan and Lin (2006), which requires that the selected factors from h_t matter for all factors in g_t, which is more aggressive in terms of factor exclusion than the procedure we recommend here. Also, either (1.a) or (1.b) can be replaced by other machine-learning methods such as regression tree, random forest, boosting, and neural network, as shown in Chernozhukov et al. (2016) for treatment-effect estimation.

Our LASSO regression contains a nonnegative regularization parameter, for example, τ_j ($j = 0, 1, \ldots, d$), to control the level of penalty. A higher τ_j indicates a greater penalty and hence results in a smaller model. The optimization becomes a least-squares problem if τ_j is 0. To determine the regularization parameter, we adopt the most commonly used 10-fold cross-validation, BIC, and AIC; see, for example, Friedman et al. (2009). BIC tends to select a more parsimonious model than the cross-validation and AIC.

We can also give different weights to λ_h. Belloni et al. (2012) recommend a data-driven method for choosing a penalty that allows for non-Gaussian and heteroskedastic disturbances. We adopt a strategy in the same spirit of Bryzgalova (2015), which assigns weights to λ_h proportional to the inverse of the operator norm of the univariate betas of the corresponding factor in h_t. This strategy helps remove spurious factors in h_t because of a higher penalty assigned on those factors with smaller univariate betas.

Our definition of spurious factors differs from the existing literature, in that the weak-identification issue occurs when covariances between factors and returns are zero or small, rather than when factors have zero or small multivariate betas. Detecting spurious factors and conducting inference in the presence of such factors have been discussed by, for example, Kleibergen (2009), Gospodinov et al. (2014a), and Bryzgalova (2015). In the current framework, we recommend pre-screening to rule out such factors in g_t prior to any empirical analysis.

2.4 Statistical Inference

We derive the asymptotic distribution of the estimator for λ_g under a jointly large n and T asymptotic design. Whereas d is fixed throughout, s and p can either be fixed or increasing. In the appendix, we prove the following theorem:

Theorem 1. *Under Assumptions A.1 - A.6 in Appendix A.2, if* $s^2T^{1/2}(n^{-1}+T^{-1})\log(n\vee p\vee T)=o(1)$, *we have*

$$T^{1/2}(\widehat{\lambda}_g-\lambda_g)\xrightarrow{\mathcal{L}}\mathcal{N}_d\left(0,\Pi\right),$$

where the asymptotic variance is given by

$$\Pi=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{E}\left((1-\lambda^\intercal v_t)(1-\lambda^\intercal v_s)\Sigma_z^{-1}z_t z_s^\intercal\Sigma_z^{-1}\right),\quad \Sigma_z=\mathrm{Var}(z_t).$$

We stress this result holds even with imperfect model selection. That is, the selected models from (7) and (8) may omit certain useful factors and include redundant ones, which nonetheless has a negligible effect on the inference of λ_g. Using analysis similar to Belloni et al. (2014b), the results can be strengthened to hold uniformly over a sequence of data-generating processes that may vary with the sample size, so that our inference is valid without relying on a perfect recovery of the correct model in finite sample. Moreover, the asymptotic distribution of $\widehat{\lambda}_g$ does not rely on covariances or factor loadings of g_t and h_t, because they appear in strictly higher-order terms, which further facilitates our inference. The next theorem provides a Newey-West-type estimator of the asymptotic variance Π.

Theorem 2. *Suppose the same assumptions as in Theorem 1 hold. In addition, Assumption A.7 holds. If* $qs^{3/2}(T^{-1/2}+n^{-1/2})\left\|V\right\|_{\mathrm{MAX}}\left\|Z\right\|_{\mathrm{MAX}}=o_p(1)$,[4] *we have*

$$\widehat{\Pi}\xrightarrow{p}\Pi,$$

where $\widehat{\lambda}=(\widehat{\lambda}_g:\widehat{\lambda}_h)$ *is given by (9), and*

$$\begin{aligned}\widehat{\Pi}=&\frac{1}{T}\sum_{t=1}^{T}(1-\widehat{\lambda}^\intercal v_t)^2\widehat{\Sigma}_z^{-1}\widehat{z}_t\widehat{z}_t^\intercal\widehat{\Sigma}_z^{-1}\\&+\frac{1}{T}\sum_{k=1}^{q}\sum_{t=k+1}^{T}\left(1-\frac{k}{q+1}\right)\left((1-\widehat{\lambda}^\intercal v_t)(1-\widehat{\lambda}^\intercal v_{t-k})\widehat{\Sigma}_z^{-1}\left(\widehat{z}_t\widehat{z}_{t-k}^\intercal+\widehat{z}_{t-k}\widehat{z}_t^\intercal\right)\widehat{\Sigma}_z^{-1}\right),\\\widehat{\Sigma}_z=&\frac{1}{T}\sum_{t=1}^{T}\widehat{z}_t\widehat{z}_t^\intercal,\quad \widehat{z}_t=g_t-\widetilde{\eta}_{\widetilde{I}}h_t,\quad \widetilde{\eta}_{\widetilde{I}}=\arg\min_{\eta}\left\{\left\|G-\eta H\right\|^2:\eta_{\cdot,j}=0,\quad j\notin\widetilde{I}\right\},\end{aligned}$$

[4] We use a capital letter A to denote the matrix $(a_1:a_2:\ldots:a_T)$ and $\left\|A\right\|_{\mathrm{MAX}}$ to denote the $\mathbb{L}_\infty$-norm of A in the vector space.

and $\widetilde{I}$ is the union of selected variables using a LASSO regression of each factor in g_t on h_t:

$$\min_{\eta_j} \left\{ T^{-1} \left\| G_{j,\cdot} - \eta_j H \right\|^2 + \bar{\tau}_j T^{-1} \|\eta_j\|_1 \right\}, \quad j = 1, 2, \ldots, d. \tag{10}$$

3 Simulation Evidence

One of the central advantages of our double-selection method is that it obtains proper inference on the risk premia λ_g of a factor, taking explicitly into account the possibility that the model-selection step (based on LASSO) may mistakenly include some irrelevant factors or exclude useful factors in any finite sample.

In this section, we therefore study the finite-sample performance of our inference procedure using Monte Carlo simulations. In particular, we show that if one were to make inference on λ_g by selecting the control factors via standard LASSO (and ignoring potential mistakes in model selection), the omitted variable bias resulting from selection mistakes would yield incorrect inference about λ_g. Instead, our double-selection procedure fully corrects for this problem in finite sample, and produces valid inference.

More specifically, in our simulation, we are interested in making inference on λ_g, the vector of prices of risk of three factors in g_t. g_t includes a useful factor (denoted as g_{1t}) as well as a useless factor and a redundant factor (denoted together as a 2×1 vector g_{2t}). g_{2t} has zero risk price, that is, $\lambda_{g_2} = 0$, but the covariance of the redundant factor is correlated with the cross section of expected returns. In our simulation, h_t is a large set of factors that includes four useful factors h_{1t}, and $p - 4$ useless and redundant factors collected in h_{2t} (so the total dimension of h_t is p).

The motivation for this setup is the case where the true model has five true factors (as in the Fama-French 5-factor model), one captured by g_{1t}, and the remaining four captured by h_{1t}. Therefore, we make inference about the price of risk of one of the true factors (g_{1t}), where h_t contains the remaining four, and allow both g_t and h_t to have useless and redundant factors.

In what follows, we first give details of the simulation procedure, and then show the results of the Monte Carlo experiment.

3.1 Simulating the Data-Generating Process

The simulation proceeds as follows. Recall that for any factor or set of factors a_t, $C_a = Cov(r_t, a_t)$. Also, recall that our data-generating process (DGP) involves a cross-sectional relationship between the covariances C_{g_1} and C_{h_1}:

$$C_{g_1} = \iota_n \xi_1 + C_{h_1} \chi_1^{\intercal} + C_{e_1} \tag{11}$$

as well as a time-series projection of the true factor g_{1t} onto the remaining true factors h_{1t}:

$$g_{1t} = \eta_1 h_{1t} + z_{1t} \tag{12}$$

which in turns implies the covariance relation

$$C_{z_1} = C_{g_1} - C_{h_1}\eta_1^\intercal \tag{13}$$

We use these relations to simulate our environment. We first choose a calibration for the model parameters ($\eta_1, \xi_1, \chi_1^\intercal, \lambda_{g_1}, \lambda_{h_1}$, etc.), discussed in detail below. Given the chosen parameters, we generate the covariances of returns and factors C: first drawing C_{e_1} and C_{h_1} independently from multivariate normal distributions, then generating C_{g_1} using equation (11), and C_{z_1} using equation (13). We then generate the time series of the true factors. To do so, we simulate $h_{1t} \sim \mathcal{N}(0, \Sigma_{h_1})$, and $z_{1t} \sim \mathcal{N}(0, \Sigma_{z_1})$; g_{1t} is generated using equation (12).

Finally, we simulate returns. To do so, we need to generate the cross section of expected returns, $\mathrm{E}(r_t)$, and the time series of return innovations, $r_t - \mathrm{E}(r_t)$. Expected returns are generated according to the true model: $\mathrm{E}(r_t) = \iota_n\gamma_0 + C_{g_1}\lambda_{g_1} + C_{h_1}\lambda_{h_1}$. Return innovations are generated from the factor model $\beta_{g_1} g_{1t} + \beta_{h_1} h_{1t} + u_t$, where the betas are $\beta_{g_1} = C_{z_1}\Sigma_{z_1}^{-1}$ and $\beta_{h_1} = C_{h_1}\Sigma_{h_1}^{-1} - \beta_{g_1}\eta_1$, as implied by the DGP. u_t is simulated from a Student's t distribution with 5 degrees of freedom and a covariance matrix Σ_u.

The steps described so far simulate the *true* model, based on the factors h_{1t} and g_{1t}. Next, we add a simulation of useless and redundant factors. Both are unpriced (they have zero risk price), but *useless* factors are also uncorrelated with the true factors (g_{1t}, h_{1t}), whereas *redundant* factors are correlated with the true factors: so they will command a risk *premium* simply due to this correlation, even though they have zero risk price because they do not affect marginal utility once the true factors are controlled for. We include a total of $p-4$ factors in h_t, half useless and half redundant. We include one useless and one redundant factor in g_t.[5]

We calibrate our DGP to mimic the actual Fama-French five-factor model. In particular, we calibrate χ, η, λ, Σ_z, the mean and covariance matrices of C_e, C_{h_1}, as well as Σ_{h_1} to match the

[5] More technically, we simulate useless and redundant factors such that r_t is conditionally independent of g_{2t} and h_{2t} given g_{1t} and h_{1t}; g_{1t} is conditionally independent of h_{2t} given h_{1t}; C_{g_1} is conditionally independently of C_{h2} given C_{h1}; and $\mathrm{E}(r_t)$ is conditionally independent of C_{g_2} and C_{h_2}. For both useless and redundant factors, we calculate $(C_{g_2} : C_{h_2}) = \iota_n\theta_0 + (C_{g_1} : C_{h_1})\theta_1 + C_\epsilon$, where θ_0 is a $1 \times (p-2)$ matrix, θ_1 is a $5 \times (p-2)$ matrix, and C_ϵ is simulated independently from a multivariate normal distribution. We set $(p/2-1)$ columns of θ_1 to 0s, so that half the factors in g_{2t} and h_{2t} are useless, because in the cross section, their covariances are not correlated with $\mathrm{E}(r_t)$. The remaining half are redundant factors. The factors g_{2t} and h_{2t} are simulated as $(g_{2t}^\intercal : h_{2t}^\intercal)^\intercal = \phi^\intercal(r_t - \mathrm{E}(r_t)) + \nu_t$, where $\phi = \Sigma_r^{-1}(C_{g_2} : C_{h_2})$ is a $n \times (p-2)$ matrix, $\Sigma_r = C_{z_1}\Sigma_{z_1}^{-1}C_{z_1}^\intercal + C_{h_1}\Sigma_{h_1}^{-1}C_{h_1} + \Sigma_u$, and ν_t is simulated independently from a multivariate normal distribution.

summary statistics (times series and cross-sectional R^2, factor-return covariances, etc.) of the Fama-French five factors estimated using 202 characteristics-sorted portfolios, described in detail in the next section. We calibrate a diagonal Σ_u so that the average time series R^2 for this five-factor model is 85%. For redundant and useless factors, we calibrate the parameters using all the other factors in our data library, again described in detail in the next section.

Our calibrated data-generating process by construction achieves the desired sparsity of $\chi, \lambda_h^\intercal$, and η, because their last $(p-4)$ columns are 0s. At the same time, it produces non-zero unconditional correlations among *all* factors in the time series and among their covariances in the cross section. The total number of Monte Carlo trials is 2000. Because we assume non-random selection of assets, we simulate only once C_g, C_h, and hence β_g, β_h, so that they are constant throughout the rest of the Monte Carlo trials.

3.2 Simulation Results

We report here the results of various simulations from the model. We consider various settings with number of total factors $p = 25, 100$, number of assets $n = 50, 100, 200$, and length of time series $T = 360, 480, 600$. Also, we report simulation results using different choices of regularization parameters for robustness, including BIC, AIC, and 10-fold cross-validation.

Figure 1 compares the asymptotic distributions of the proposed double-selection estimator with that of the single-selection estimator for the case $p = 100$, $n = 200$, and $T = 600$. The right side of the figure shows the distribution of the t-test for the price of risk λ_g of the three factors (useful in the first row, redundant in the second row, and useless in the third row) when using the controls selected by standard LASSO (i.e., a single-selection-based estimator). The panels show that inference without double-selection adjustment displays substantial biases and distortion from normality. The left side of the figure shows instead that our double-selection procedure produces an unbiased and asymptotically normal test, as predicted in Theorem 1.

Figure 2 plots the histograms of the select variables for each selection step. Note the LASSO procedure does include many useless factors as controls: if the model selection were able to perfectly identify the correct controls in h_t, exactly the four factors h_{1t} would be selected – yet the LASSO often selects more than 20 factors. The key to correct inference is that the two-step selection procedure minimizes the potential omitted factor bias.

Tables 1, 2, and 3 compare the biases and root-mean-squared errors (RMSEs) for double-selection, single-selection, and the OLS estimators of each entry of λ_g, respectively. We only report results based on BIC and 10-fold cross-validation for selecting regularization parameters. The results based on AIC is similar to that of the cross-validation.

Both double- and single-selection estimators outperform OLS in terms of the RMSEs, particularly when p is large relative to n. When p is greater than n, OLS becomes infeasible. This result confirms the efficiency benefits of dimension-reduction techniques. In addition, the double-selection estimator has a smaller bias and a smaller RMSE than the single-selection estimator. But the main advantage of double-selection relative to single-selection is in removing the distortions to inference, visible from the distribution of standardized statistics in Figure 1.

Note the biases and RMSEs become smaller as n and T increase. Instead, when p is larger, the results exacerbate slightly. Overall, the simulation results confirm our econometric analysis that the double-selection estimator outperforms the benchmarks.

4 Empirical Analysis

In this section, we apply our methodology to the data library of hundreds of factors. We first show how our estimation procedure can be used to evaluate whether a newly proposed factor actually provides useful pricing information compared to the myriad of existing factors. We document that indeed some of the recently proposed factors (e.g., profitability) do contribute significantly to explaining asset prices, even controlling for the hundreds of factors the literature had proposed previously. We also evaluate our methodology recursively and out of sample, verifying that factors that appear significant when they were introduced indeed tend to be selected in the best parsimonious asset pricing model in later years.

In addition, we show our model-selection procedure selects a parsimonious model even as the number of proposed factors increases over time; yet the fit of the selected model keeps improving as new potential factors are added. This finding suggests the asset pricing literature has indeed uncovered better factors over time, and that our procedure allows us to bring discipline to this ever-growing list of factors, identifying the most useful ones.

4.1 The Zoo of Factors

Our factor library contains 114 factors (both tradable and non-tradable) at the monthly frequency from July 1980 to December 2015 from multiple sources. For factors introduced in 2016, we use their data up to 2015. First, we downloaded all workhorse factors in the U.S. market from Ken French's data library. Then we added several published factors directly from authors' websites, such as liquidity of Pástor and Stambaugh (2003), the q-factor model of Hou et al. (2014),[6] the intermediary asset pricing model of He et al. (2016), the Betting-Against-Beta, and Quality-Minus-Junk factors from AQR. Finally, in addition to those publicly available factors, for the firm characteristics reported

[6]We are grateful to Lu Zhang for sharing the factors data.

in Green et al. (2016),[7] we follow the Fama-French portfolio-sorting rule and construct the long-short portfolio spreads (top 30% - bottom 30% or 1-0 dummy difference) based on the security sorting on the previous June.

In Tables 1, 2, and 3 of the Supplemental Appendix, we report descriptive statistics for the set of 114 factors (monthly average returns, standard deviations, annualized Sharpe ratios) as well as the academic sources. We only use 105 of them in the empirical analysis because of missing values. We follow Hou et al. (2014) and provide six main categories for the factor classification: Momentum, Value-versus-Growth, Investment, Profitability, Intangibles, and Trading Frictions. Each category contains more than 10 factors.

As a brief summary of the factors created by long-short portfolio spreads, 95 factors have annualized Sharpe ratios greater than 0.1, and 27 of them greater than 0.5. For the time-series test of the risk premium (expected excess return) of tradable factors, 51 factors have t-stat greater than 2, and 26 of them have t-stat greater than 3. We also test the time-series alphas of these tradable factors relative to the Fama-French three-factor model. Seventy-four of them have t-stat greater than 2, and 37 of them have t-stat greater than 3.

4.2 Test Portfolios

We conduct our empirical analysis on a large set of standard portfolios of U.S. equities. We target U.S. equities because of their better data quality and because they are available for a long time period; however, our methodology could be applied to any set of countries or asset classes. We include in our analysis 202 portfolios: 25 portfolios sorted by size and book-to-market ratio, 17 industry portfolios, 25 portfolios sorted by operating profitability and investment, 25 portfolios sorted by size and variance, 35 portfolios sorted by size and net issuance, 25 portfolios sorted by size and accruals, 25 portfolios sorted by size and momentum, and 25 portfolios sorted by size and beta. This set of portfolios captures a vast cross section of anomalies and exposures to different factors; at the same time, they are easily available on Kenneth French's website, and therefore represent a natural starting point to illustrate our methodology.[8] We conduct our analysis on the period from July of 1980 to December of 2015 (426 months), for which all the returns and factors are available. We perform the analysis at the monthly frequency, and work with factors that are available at the monthly frequency.

[7] We are grateful to Jeremiah Green for sharing the firm-characteristics data.

[8] See the description of all portfolio construction on Kenneth French's website: http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html.

4.3 Are New Factors Useful?

One of the motivations for our methodology is to bring discipline to the large number of factors the literature has proposed, teasing out which ones truly contain new useful asset pricing information (that helps explain the cross section of prices), and which ones instead are redundant or useless in pricing the panel of returns.

In this section, we apply our methodology to factors that have been proposed in the last five years, drawing the "control" factors from the set of more than 100 factors that were proposed before 2011. That is, we ask whether the recently introduced factors add any new pricing information to the existing factors, or are redundant or outright useless in pricing the panel of returns. We have no ex-ante reason to expect the results to go in either direction. On the one hand, given the set of potential control factors is already extremely large, one might think new factors are unlikely to contribute much to pricing the cross section of returns. On the other hand, we expect new research to potentially uncover better factors over time, yielding factors that improve over the existing ones.

Table 4 reports the results for many factors proposed in the last five years, among which we find quality-minus-junk (QMJ), betting-against-beta (BAB), two investment factors (CMA from Fama-French and IA from HXZ), two profitability factors (RMW from Fama-French and ROE from HXZ), and the intermediary capital factor from He et al. (2016).

The table contains three panels. The left panel reports the results of single- and double-selection methods. The middle panel shows the results when the controls are not selected optimally, but are simply the three Fama-French factors; if the true model has additional factors, this approach would suffer from omitted variable bias. The last panel contains information about the risk *premium* associated with each factor (when the factor is traded): simply the time-series average excess return of each factor.

The first column of the table shows the slope of the cross-sectional regression of returns on (univariate) betas for each factor, controlling for the factors selected by our two-stage procedure. The number in this column represents the average excess return in basis points per month of a portfolio with unit univariate beta with respect to that factor. This number is equal to λ_g but scaled to excess return units for ease of interpretation. A positive number in the first column indicates that high values of the factor capture states of low marginal utility (good states of the world). The second column (labeled "t-stat DS") reports the t-statistic for the test that this coefficient is different from zero, using our double-selection methodology.

Although many recent factors do not appear to contain useful new information, we find significant exceptions. In particular, the profitability factors (ROE and, to a lesser extent, RMW) seem to provide new useful information for pricing the cross section of returns.

The last column of the double-selection panel reports the cross-sectional R^2 achieved by the selected model. Although the R^2 varies from model to model, it ranges between 71% and 77%, a high value for such a large cross section of returns.

The middle panel of the table shows the results where the "control" factors in the cross-sectional regression are assumed to be the three Fama-French factors. Two observations are noteworthy. First, the R^2 of the three-factor model (plus the additional factor considered) is low in the panel of 202 portfolios, mostly lower than 30%. Clearly, more factors are needed to explain the large cross section of returns we study. This finding suggests testimates of the risk price of a factor will be affected by omitted variable bias. In fact, comparing the middle and left panels of the table, one immediately sees that many of the results change significantly when using our model-selection procedure instead of the three Fama-French factors as controls. For example, the betting-against-beta factor has a positive and significant price of risk when controlling for the FF3 factors, but it is insignificant when controlling for all other factors. Profitability instead appears statistically more significant when controlling for existing factors.

The rightmost panel of the table shows the average excess return of the factor, when tradable, that is, its risk premium. This number represents the compensation investors obtain from bearing exposure to that factor, holding all other risk factors constant. As discussed, for example, in Cochrane (2009), the risk premium of a factor does *not* correspond to its ability to price other assets, that is, its coefficient in the SDF. Using the risk premium to assess the importance of a factor in a pricing model can therefore be misleading. For example, consider two factors that are both equally exposed to the same underlying risk, plus some noise. Both factors will command an identical risk premium. Yet those factors are not *both* useful to price other assets—regardless of their level of statistical significance. The most promising way to reduce the proliferation of factors is not to look at their risk premium (no matter how significant it is), but to evaluate whether they add any pricing information to the existing factors. Our paper proposes a way to make this feasible even in a context of high dimensionality, when the set of potential control factors is large.

Finally, the table also illustrates the importance of *both* steps in the double-selection methodology. The column "tstat-SS" in the left panel shows the inference that would be made if one were to select the controls via standard LASSO and then performing statistical inference on g_t, that is, ignoring the second stage in the double-selection procedure. Comparing the single-selection ("tstat-SS") and the double-selection ("tstat-DS") columns shows that inference about the price of risk of many factors would change. For example, QMJ and the maximal negative return (maxret) would appear useful under single-selection, but double-selection reveals their risk price to be smaller and insignificant. This, combined with our simulation evidence in the previous section, emphasizes the importance of selecting the controls appropriately. In the context of this paper, that means using

double-selection to minimize the ex-ante omitted variable bias.

Given that the core of this paper is a procedure to select controls from h_t in two steps, looking at what controls are chosen in this case among those introduced before 2011 is interesting. We start by looking at the first stage of the double-selection, which, in this case, selects 14 factors. Table 5 reports them together with their risk price and a test of significance (itself based on double-selection); it also reports their risk premia.

Two points are noteworthy about the first selection stage. First, several – but not all – of the selected factors are statistically significant. This is typically the case with LASSO-based procedures, which eliminate some factors entirely (setting their coefficients to 0), but that also tend to include some non-significant factors. Second, the factors selected are not the standard factors the literature often used as controls (in particular, the Fama-French factors). This indicates that in the 20 years since they were introduced, newer factors have improved pricing relative to those early factors.

To the 14 factors selected in the first stage and reported in Table 5, our double-selection procedure adds additional control factors in a second stage; these 14 factors are those whose risk exposures are cross-sectionally correlated with those of the target factor g_t, which are crucial to minimize the omitted variable bias in risk prices. Due to space constraints, we do not report the additional factors for all the g_t of Table 4: each factor g_t induces a different second-stage selection. The typical second stage adds around three to five factors; for example, the profitability factor from HXZ includes four additional factors that were not already in the first-stage selection.

The results have so far described the application of our double-selection methodology to the factors introduced in the period 2011-2016, highlighting how some new factor appears significant even relative to the large set of existing factors. Next, we extend this exercise in a fully recursive way, where each year we test new factors relative to the ones existing up to then, and then verify recursively whether factors that are deemed significant indeed are selected as part of the best model in future years.

4.4 A Recursive and Out-of-Sample Evaluation

One of the advantages of our procedure is that even as new potential factors are added to the pool of controls, the procedure always selects a low-dimensional model for the SDF when evaluating new factors. Over time, as new factors are proposed, our procedure will retain the ones with the best explanatory power for the cross section of returns and evaluate new factors against the best of the existing factors.

To illustrate this point, we perform the following recursive testing exercise. In each year starting in 1994, we consider the factors introduced during that year, and use our double-selection procedure

to test whether they are useful or redundant relative to factors existing up to then. We can then follow over time which factors appear useful when they are introduced, and which ones do not. Note that in this exercise, we update the pool of factors recursively but use the entire time series to construct our tests; thus, the only thing changing over time is the set of factors available in the pool of potential controls h_t. This approach allows us to focus on how the set of selected and significant factors evolved as new factors were added in the literature, without being contaminated by the fact that, over time, the time series of returns also changed.[9]

Table 6 reports the factors introduced each year starting in 1994, identified by their id, and underlines the ones that appear to be statistically significant according to our test. Of the 80 factors introduced between 1994 and 2016, our procedure found only 11 of them to be useful at the time they were introduced. This finding is a sign that – according to our estimates – many of the factors in the "zoo of factors" are redundant or useless.

To evaluate the usefulness of our methodology, we next construct a recursive evaluation exercise in the following way. Every time a new factor is introduced, we test whether its risk price is nonzero, controlling for all factors existing up to that point, as described above. We then look at whether in future years – when this factor will now belong to the set of potential controls h_t – this factor will be selected as part of the best model. This analysis tells us whether factors our double-selection test determined to be significant tend to actually replace older factors and be selected in future selections, and whether factors our test determines to be insignificant are not included in the best model in future years.

We start by looking in Table 7 at the evolution of the model selected using recursively the factors available up to each year. For ease of reading, we report only the significant factors. The table shows the "best" control model selected out of the (recursively expanding) h_t is remarkably stable over time. Some factors (e.g., No.63, corporate investment) tend to appear throughout the 20 years considered, being selected almost all the time. Others appear at the beginning but are substituted with more modern factors (e.g., No.41, introduced in 1998, disappears from the optimal model after 2001).

Among the factors the model-selection procedure selected as part of the model – reported in Table 7 – how many of them did our test deem to be useful when they were introduced? The table underlines factors that were deemed to be significant at introduction and greys out the ones that were never tested, because they were introduced before 1994. The table shows that in fact all but a couple of the factors we see being selected as part of the parsimonious asset pricing models were originally deemed to be useful, thus confirming our inference (applied at the introduction of the factor) is able to identify factors that contain strong pricing information.

[9] We discuss a full out-of-sample extension of this exercise below.

To summarize the results of the recursive evaluation, Table 8 shows how many factors are later selected in the parsimonious model, depending on whether our test established they were significant for explaining asset pricing when they were introduced. The table shows that 80% of the factors deemed significant are then selected in the model later; 70% of those not determined to be significant in fact do not appear important ex post and do not get selected (here we consider a factor to be selected if it is selected at least three times in future years).

Results are similar when we update not only the pool of factors, but also the time series of the data recursively, in a fully out-of-sample recursive evaluation. As expected, results are weaker in this case due to the shorter time series now available for estimation and model selection. Still, 70% of significant factors are selected in later years at least three times, and 55% of the insignificant ones are not selected later three times or more.

Overall, the recursive analysis confirms our procedure is able to uncover factors that represent useful additions to the asset pricing model, and to discard factors that are redundant or useless.

4.5 Risk Prices, Risk Premia, and Student t-Stats of 3

Recent research (McLean and Pontiff (2016), Harvey et al. (2015)) has pointed out the vast number of anomalies and factors the literature has found. In particular, these papers note the large number of anomalies and risk factors that appear significant at standard significance levels has increased dramatically, leading to concerns of data mining. In response to this concern, Harvey et al. (2015) propose adopting a stricter requirement for significance, such as using a threshold for the t-stat of 3, motivated from the multiple-testing literature.

Although motivated by this literature, in this paper we propose a different solution to the expanding myriad of factors and anomalies: to evaluate each new potential factor's contribution to explaining asset prices relative to the existing factors, and to use model-selection techniques to select the best model out of potentially hundreds of factors.

Our approach differs from the approaches proposed in the existing literature in three substantial ways. First, it directly takes into account the correlation among factors, rather than considering factors individually and using Bonferroni-type bounds to assess their joint significance. We provide a statistical test of a factor's contribution with desirable asymptotic properties, as demonstrated in the previous sections.

Second, it directly handles hundreds of factors, exploiting machine-learning advances to reduce the dimensionality of the factor set. As discussed more below, this approach yields a parsimonious model whose size is stable even when the set of potential factors has been growing rapidly.

Third, the criterion we employ for selecting factors is based on the risk price, not the risk

premium, of the factors; it therefore captures the contribution of a factor to explaining asset prices. As discussed in Cochrane (2009) as well as, more recently, in Fama and French (2016) and in Barillas and Shanken (2015), to test whether a factor contributes to explaining the cross section of expected return, one should do inference on its risk *price*: the loading of the SDF on that factor or, equivalently, the slope of a cross-sectional regression onto that factor and the other factors in the SDF using *univariate* betas. The risk premium (or equivalently the slope of a cross-sectional regression that uses *multivariate* betas) does not capture the pricing ability of a factor for the cross section of assets, and therefore is not the right criterion to select factors in a model.

Consider, for example, a factor A with a strongly significant excess return (e.g., with a t-stat above 3). Construct now a new factor B, equal to A plus a small amount of orthogonal noise. Both A and B will yield the same expected return (risk premium), and if the noise has low-enough variance, both will appear (3 standard-deviations) statistically significant. Yet they contain the same information. Our procedure—based on risk prices— would correctly reveal B to be redundant relative to A.

To illustrate this point, consider Figure 3. The solid line in the figure plots the total (cumulative) number of factors we collected, as they were added in the literature over time. Of these, about half have a risk premium (in our data) significant with a t-stat above 2, and about a sixth have a t-stat above 3 (blue and black dashed lines). The figure shows that irrespective of the statistical criterion used for significance, the number of significant factors has increased linearly over time together with the number of total factors produced. Consider in particular the number of factors with t-stat above 3. This number went from 3 in 1994 to 26 in 2016 — an eight-fold increase in 20 years.

This finding should not be surprising: any new portfolios that are simply correlated with existing factors or anomalies necessarily command a risk premium, including factors that are entirely redundant. We have no reason to expect the set of statistically significant anomalies to stop expanding as we add more factors over time, regardless of the hurdle chosen for significance, and indeed, it does not in our data.

Now contrast this with the model selected by our procedure. Figure 3 also reports the dimension of our recursively selected model (the same one used for Table 7). In particular, the red dotted line plots the total number of factors selected (both significant and nonsignificant), whereas the dashed blue line reports just the number of significant factors among those selected. Both lines are extremely stable during this period, resulting in a parsimonious model that only retains the most relevant factors for asset pricing.

Interestingly, while the number of factors of our selected model has not increased dramatically over time, the cross-sectional R^2 achieved by the model has increased from around 50% in 1994 to

around 75% in 2016, indicating that indeed the asset pricing literature has proposed better factors over time.

5 Conclusion

In this paper we propose a regularized two-pass cross-sectional regression approach to establish the contribution to asset pricing of each factor relative to a set of control factors h_t, where the potential control set can have high dimensionality and include useless or redundant factors. Our procedure uses machine-learning techniques (specifically the double-selection procedure of Belloni et al. (2014b)) to systematically select the best control model out of the large set of factors, while explicitly taking into account that in any finite sample we cannot be sure to have selected the correct model.

We apply this methodology to a large set of factors that the literature has proposed in the last 30 years. We uncover several interesting empirical findings. First, several newly proposed factors (for example, profitability) are useful in explaining asset prices, even after accounting for the large set of existing factors proposed up to 2011. Second, factors that are deemed significant by our test when they are introduced in the literature tend to be selected as part of the best "control" model in later years, confirming that our procedure does select useful pricing factors. Third, the best parsimonious model selected recursively (adding over time new factors as they are proposed) achieves an increasingly higher cross-sectional R^2 over time, even as its dimension stays stable. This confirms that asset pricing research has indeed been producing better factors over time. Fourth, we demonstrate how our results differ starkly from the conclusions one would obtain simply by using the risk premia of the factors or the standard Fama-French three factor model as control (as opposed to the model selection procedure we advocate).

Taken together, our results are quite encouraging about the continuing progress of asset pricing research, and suggest that studying the marginal contribution of new factors relative to the vast set of existing ones is a conservative and productive way to screen new factors as they are proposed, as well as to organize the current "zoo of factors".

References

Adrian, T., Etula, E., and Muir, T. (2014). Financial intermediaries and the cross-section of asset returns. *The Journal of Finance*, 69(6):2557–2596.

Bai, J. and Zhou, G. (2015). Fama–macbeth two-pass regressions: Improving risk premia estimates. *Finance Research Letters*, 15:31–40.

Barillas, F. and Shanken, J. (2015). Comparing asset pricing models. Technical report, National Bureau of Economic Research.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Belloni, A. and Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *The Journal of Economic Perspectives*, 28(2):29–50.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732.

Breeden, D. T. (1979). An intertemporal asset pricing model with stochastic consumption and investment opportunities. *Journal of Financial Economics*, 7(3):265–296.

Bryzgalova, S. (2015). Spurious factors in linear asset pricing models. Technical report, Stanford University.

Burnside, C. (2016). Identification and inference in linear stochastic discount factor models with excess returns. *Journal of Financial Econometrics*, 14(2):295–330.

Campbell, J. Y. (2000). Asset pricing at the millennium. *The Journal of Finance*, 55(4):1515–1567.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance*, 52(1):57–82.

Chen, N.-F., Roll, R., and Ross, S. A. (1986). Economic forces and the stock market. *Journal of Business*, pages 383–403.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. K. (2016). Double machine learning for treatment and causal parameters. Technical report, MIT.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, 7(1):649–688.

Cochrane, J. H. (2009). *Asset Pricing:(Revised Edition)*. Princeton university press.

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of Finance*, 66(4):1047–1108.

Connor, G., Hagmann, M., and Linton, O. (2012). Efficient semiparametric estimation of the fama–french model and extensions. *Econometrica*, 80(2):713–754.

Fama, E. F. and French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, 33(1):3–56.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116(1):1–22.

Fama, E. F. and French, K. R. (2016). Choosing factors. Technical report, University of Chicago.

Fama, E. F. and MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, 81(3):607–636.

Fan, J., Liao, Y., and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, 39:3320–3356.

Ferson, W. E. and Harvey, C. R. (1991). The variation of economic risk premiums. *Journal of Political Economy*, 99(2):385–415.

Freyberger, J., Neuhierl, A., and Weber, M. (2017). Dissecting characteristics nonparametrically. Technical report, University of Wisconsin-Madison.

Friedman, J., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning - Data Mining, Inference, and Prediction, Second Edition*. New York, NY: Springer-Verlag New York.

Gagliardini, P., Ossola, E., and Scaillet, O. (2016). Time-varying risk premium in large cross-sectional equity data sets. *Econometrica*, 84(3):985–1046.

Giglio, S. W. and Xiu, D. (2016). Inference on risk premia in the presence of omitted factors. Technical report, University of Chicago.

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

Gospodinov, N., Kan, R., and Robotti, C. (2014a). Misspecification-robust inference in linear asset-pricing models with irrelevant risk factors. *Review of Financial Studies*, 27(7):2139–2170.

Gospodinov, N., Kan, R., and Robotti, C. (2014b). Spurious inference in unidentified asset-pricing models. Technical report, Federal Reserve Bank of Atlanta.

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

Green, J., Hand, J. R., and Zhang, F. (2016). The characteristics that provide independent information about average u.s. monthly stock returns. Technical report, Penn State University.

Harvey, C. R. and Liu, Y. (2016). Lucky factors. Technical report, Duke University.

Harvey, C. R., Liu, Y., and Zhu, H. (2015). ... and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68.

He, Z., Kelly, B., and Manela, A. (2016). Intermediary asset pricing: New evidence from many asset classes. Technical report, National Bureau of Economic Research.

Hou, K., Xue, C., and Zhang, L. (2014). Digesting anomalies: An investment approach. *Review of Financial Studies*, pages 650–705.

Hou, K., Xue, C., and Zhang, L. (2016). A comparison of new factor models. Technical report, National Bureau of Economic Research.

Jagannathan, R., Skoulakis, G., and Wang, Z. (2010). The analysis of the cross section of security returns. *Handbook of financial econometrics*, 2:73–134.

Jagannathan, R. and Wang, Z. (1996). The conditional capm and the cross-section of expected returns. *The Journal of finance*, 51(1):3–53.

Jagannathan, R. and Wang, Z. (1998). An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *The Journal of Finance*, 53(4):1285–1309.

Jensen, M. C., Black, F., and Scholes, M. S. (1972). The capital asset pricing model: Some empirical tests. In *Studies in the theory of capital markets*. New York: Praeger.

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

Kan, R. and Robotti, C. (2009). Model comparison using the hansen-jagannathan distance. *Review of Financial Studies*, 22(9):3449–3490.

Kan, R. and Robotti, C. (2012). Evaluation of asset pricing models using two-pass cross-sectional regressions. In *Handbook of computational finance*, pages 223–251. Springer.

Kan, R., Robotti, C., and Shanken, J. (2013). Pricing model performance and the two-pass cross-sectional regression methodology. *The Journal of Finance*, 68(6):2617–2649.

Kan, R. and Zhang, C. (1999a). Gmm tests of stochastic discount factor models with useless factors. *Journal of Financial Economics*, 54(1):103–127.

Kan, R. and Zhang, C. (1999b). Two-pass tests of asset pricing models with useless factors. *the Journal of Finance*, 54(1):203–235.

Kleibergen, F. (2009). Tests of risk premia in linear factor models. *Journal of econometrics*, 149(2):149–173.

Kleibergen, F. and Zhan, Z. (2014). Mimicking portfolios of macroeconomic factors. Technical report, Brown University Working Paper.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21(01):21–59.

Lettau, M. and Ludvigson, S. (2001). Resurrecting the (c) capm: A cross-sectional test when risk premia are time-varying. *Journal of Political Economy*, 109(6):1238–1287.

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

Lintner, J. (1965). Security prices, risk, and maximal gains from diversification. *The Journal of Finance*, 20(4):587–615.

McLean, R. D. and Pontiff, J. (2016). Does academic research destroy stock return predictability? *The Journal of Finance*, 71(1):5–32.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 37(1):246–270.

Nagel, S. (2013). Empirical cross-sectional asset pricing. *Annual Review of Financial Economics*, 5(1):167–199.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55:703–708.

Pástor, L. and Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, 111(3):642–685.

Raponi, V., Robotti, C., and Zaffaroni, P. (2016). Testing beta-pricing models using large cross-sections. Technical report, Imperial College London.

Shanken, J. (1992). On the estimation of beta-pricing models. *Review of Financial studies*, 5(1):1–33.

Shanken, J. (1996). Statistical methods in tests of portfolio efficiency: A synthesis. *Handbook of statistics*, 14:693–711.

Shanken, J. and Zhou, G. (2007). Estimating and testing beta pricing models: Alternative methods and their performance in simulations. *Journal of Financial Economics*, 84(1):40–86.

Sharpe, W. F. (1964). Capital asset prices: A theory of market equilibrium under conditions of risk. *The Journal of Finance*, 19(3):425–442.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202.

Welch, I. (2008). The link between fama-french time-series tests and fama-macbeth cross-sectional tests. Technical report, UCLA.

White, H. (2000). *Asymptotic Theory for Econometricians: Revised Edition*. Emerald Group Publishing Limited.

Yogo, M. (2006). A consumption-based explanation of the cross-section of expected stock returns. *The Journal of Finance*, 61(2):539–580.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, 36(4):1567–1594.

Table 1: Asymptotic Approximation Performance for λ_{useful}

	Bias						RMSE					
	$p = 25$			$p = 100$			$p = 25$			$p = 100$		
n	DS	SS	OLS	DS	SS	OLS	DS	SS	OLS	DS	SS	OLS
	Panel A: BIC											
	$T = 360$											
50	-0.37	-4.47	1.18	-8.68	-12.26	NaN	5.97	7.28	7.17	17.46	14.17	NaN
100	0.08	-2.74	1.08	-1.27	-5.34	NaN	5.21	6.38	5.76	6.55	7.54	NaN
200	0.33	-0.57	1.08	-0.22	-3.45	3.93	4.74	5.05	4.99	5.48	6.06	10.97
	$T = 480$											
50	-1.48	-2.49	0.76	-7.32	-11.70	NaN	5.46	5.68	6.43	10.56	12.72	NaN
100	0.03	-1.74	0.88	-1.47	-5.70	NaN	4.40	4.77	4.76	5.66	7.35	NaN
200	0.26	-0.39	0.90	-0.18	-2.27	3.47	4.18	4.23	4.36	4.74	5.23	8.24
	$T = 600$											
50	-1.26	-3.15	0.40	-7.93	-12.22	NaN	4.80	5.27	5.40	12.02	13.44	NaN
100	0.02	-1.51	0.74	-0.98	-4.03	NaN	3.91	4.23	4.11	5.04	6.02	NaN
200	0.21	-0.21	0.63	-0.51	-2.55	2.49	3.52	3.56	3.62	4.36	4.79	7.13
	Panel B: Cross-Validation											
	$T = 360$											
50	-0.56	-3.12	1.22	-5.41	-10.21	NaN	5.73	7.02	6.69	12.06	12.17	NaN
100	0.13	-1.74	1.15	-1.53	-5.46	NaN	4.62	5.64	5.09	6.53	7.62	NaN
200	0.22	-0.56	0.91	-0.02	-3.73	4.20	4.69	4.97	4.88	5.69	6.04	10.85
	$T = 480$											
50	-3.32	-5.61	0.48	-5.40	-10.52	NaN	6.63	7.47	8.04	8.84	11.66	NaN
100	-0.12	-1.12	0.69	-1.57	-5.82	NaN	4.25	5.16	4.48	5.81	7.43	NaN
200	0.33	-0.69	0.89	0.03	-2.45	3.17	4.20	4.21	4.40	5.05	5.32	8.24
	$T = 600$											
50	-1.60	-4.88	0.76	-5.39	-10.20	NaN	4.86	6.73	6.52	8.69	11.45	NaN
100	0.22	-2.36	0.87	-1.25	-4.18	NaN	3.85	4.73	4.14	5.31	6.18	NaN
200	0.11	-0.62	0.60	-0.91	-3.10	2.39	3.56	3.81	3.72	4.42	5.04	7.06

Note. This table provides the biases and root-mean-squared errors (RMSE) of the estimates of the price of risk λ of the useful factor from Monte Carlo simulations. DS is the double-selection estimator, SS is the single-selection estimator, and OLS is the ordinary least squares without selection. The regularization parameters in the LASSO are selected either by minimizing BIC or using 10-fold Cross-Validation. The true value λ_{useful} is 16.76.

Table 2: Asymptotic Approximation Performance for $\lambda_{\text{redundant}}$

	Bias						RMSE					
	$p = 25$			$p = 100$			$p = 25$			$p = 100$		
n	DS	SS	OLS	DS	SS	OLS	DS	SS	OLS	DS	SS	OLS
						Panel A: BIC						
						$T = 360$						
50	-0.07	-0.96	-0.12	-0.51	-0.50	NaN	1.51	1.54	1.88	2.66	1.16	NaN
100	-0.18	-0.20	-0.19	-0.13	-0.63	NaN	1.01	1.01	1.19	1.32	1.12	NaN
200	-0.09	-0.22	-0.10	-0.09	-0.72	-0.12	0.71	0.84	0.77	0.72	1.02	1.45
						$T = 480$						
50	-0.26	-0.57	-0.19	-0.11	-0.09	NaN	1.51	1.42	1.83	1.93	1.17	NaN
100	-0.15	-0.90	-0.20	-0.26	-0.64	NaN	1.08	1.52	1.26	1.20	1.02	NaN
200	-0.10	-0.35	-0.10	-0.15	-0.37	-0.17	0.67	0.83	0.73	0.79	0.88	1.29
						$T = 600$						
50	-0.06	-0.97	-0.11	-0.40	-0.33	NaN	1.22	1.37	1.57	1.96	1.22	NaN
100	-0.17	-0.42	-0.15	-0.11	-0.82	NaN	0.93	1.05	1.02	1.06	1.23	NaN
200	-0.04	-0.11	-0.05	-0.09	-0.74	-0.19	0.61	0.72	0.66	0.67	1.07	1.28
						Panel B: Cross-Validation						
						$T = 360$						
50	-0.30	-0.18	-0.42	-0.53	-0.68	NaN	1.67	1.62	2.23	2.22	1.07	NaN
100	-0.25	-0.29	-0.31	-0.11	-0.60	NaN	1.24	1.24	1.46	1.35	1.13	NaN
200	-0.08	-0.15	-0.07	-0.07	-0.58	-0.11	0.75	0.78	0.83	0.79	0.94	1.48
						$T = 480$						
50	0.14	-0.59	-0.12	-0.16	-0.12	NaN	1.40	1.33	1.89	1.82	1.19	NaN
100	-0.13	-0.16	-0.12	-0.20	-0.61	NaN	0.93	0.94	1.03	1.26	1.00	NaN
200	-0.06	-0.32	-0.06	-0.13	-0.33	-0.16	0.61	0.82	0.65	0.81	0.83	1.28
						$T = 600$						
50	-0.32	-0.93	-0.38	-0.52	-0.49	NaN	1.62	1.94	2.24	1.73	1.39	NaN
100	-0.12	-0.41	-0.13	-0.15	-0.69	NaN	0.89	1.05	1.04	1.05	1.13	NaN
200	-0.05	-0.24	-0.06	-0.06	-0.54	-0.13	0.64	0.79	0.71	0.67	0.92	1.25

Note. This table provides the biases and root-mean-squared errors (RMSE) of the estimates of the price of risk λ of the redundant factor from Monte Carlo simulations. DS is the double-selection estimator, SS is the single-selection estimator, and OLS is the ordinary least squares without selection. The regularization parameters in the LASSO are selected either by minimizing BIC or using 10-fold Cross-Validation. The true value $\lambda_{\text{redundant}}$ is 0.

Table 3: Asymptotic Approximation Performance for λ_{useless}

	Bias						RMSE					
	$p = 25$			$p = 100$			$p = 25$			$p = 100$		
n	DS	SS	OLS	DS	SS	OLS	DS	SS	OLS	DS	SS	OLS
						Panel A: BIC						
						$T = 360$						
50	0.01	-0.07	-0.02	0.14	0.12	NaN	0.41	0.39	0.59	0.94	0.49	NaN
100	0.02	-0.07	0.01	0.02	0.00	NaN	0.31	0.33	0.36	0.34	0.31	NaN
200	0.00	0.01	0.00	0.01	0.02	0.00	0.18	0.18	0.20	0.19	0.18	0.35
						$T = 480$						
50	0.00	0.01	0.00	-0.18	-0.19	NaN	0.53	0.51	0.72	0.55	0.45	NaN
100	0.00	0.05	0.00	0.02	0.03	NaN	0.27	0.28	0.30	0.28	0.26	NaN
200	0.01	0.03	0.01	0.00	-0.01	0.00	0.17	0.18	0.18	0.18	0.17	0.35
						$T = 600$						
50	-0.01	-0.05	0.00	0.10	0.01	NaN	0.33	0.33	0.42	0.52	0.40	NaN
100	0.00	-0.03	0.01	0.01	0.04	NaN	0.28	0.28	0.31	0.24	0.24	NaN
200	0.01	0.01	0.01	0.00	0.00	-0.02	0.15	0.15	0.15	0.16	0.16	0.30
						Panel B: Cross-Validation						
						$T = 360$						
50	-0.04	-0.10	-0.05	0.10	0.12	NaN	0.49	0.48	0.65	0.69	0.50	NaN
100	0.01	0.04	0.00	0.02	0.01	NaN	0.29	0.30	0.33	0.35	0.31	NaN
200	0.01	0.02	0.01	0.00	0.02	-0.01	0.19	0.19	0.21	0.21	0.18	0.34
						$T = 480$						
50	-0.01	-0.01	0.00	-0.18	-0.22	NaN	0.44	0.43	0.63	0.50	0.47	NaN
100	0.00	-0.01	0.00	0.02	0.03	NaN	0.27	0.27	0.31	0.32	0.25	NaN
200	-0.01	-0.01	-0.01	0.00	-0.02	0.01	0.17	0.17	0.18	0.22	0.17	0.34
						$T = 600$						
50	-0.02	0.02	0.02	0.14	0.04	NaN	0.36	0.37	0.47	0.49	0.40	NaN
100	0.01	0.10	0.00	0.02	0.03	NaN	0.22	0.27	0.25	0.28	0.24	NaN
200	0.01	0.00	0.01	0.00	0.00	0.00	0.16	0.17	0.17	0.17	0.15	0.30

Note. This table provides the biases and root-mean-squared errors (RMSE) of the estimates of the price of risk λ of the useless factor from Monte Carlo simulations. DS is the double-selection estimator, SS is the single-selection estimator, and OLS is the ordinary least squares without selection. The regularization parameters in the LASSO are selected either by minimizing BIC or using 10-fold Cross-Validation. The true value λ_{useless} is 0.

Table 4: Testing for factors introduced in 2011-2016

		Regularized two-pass				Fama-MacBeth		Average excess ret	
id	Factor	λ_s (bp)	tstat (DS)	tstat (SS)	R^2	tstat (FF3)	R^2 (FF3)	avg.ret. (bp)	tstat
98	Maximum daily return	-61	-0.16	1.90*	71.7%	-1.53	25.0%	15	0.61
99	Cash holdings	88	0.53	-0.50	71.7%	-0.10	22.2%	30*	1.65
100	Quality Minus Junk	54	1.14	1.88*	77.6%	0.32	22.3%	49***	3.99
101	Gross profitability	40	0.65	-0.33	76.7%	2.58***	28.6%	29***	2.63
102	Organizational capital	22	0.33	-0.43	71.7%	2.05**	29.1%	51***	3.52
103	AEM Leverage	-29	-0.53	0.57	70.0%	3.42***	44.7%		
104	HXZ Investment	-34	-1.63	0.45	71.7%	1.27	25.6%	38***	4.04
105	HXZ Profitability	66	2.09**	1.26	71.9%	1.88*	29.5%	59***	4.63
106	Betting Against Beta	-40	-0.77	-0.28	71.7%	2.76***	41.1%	96***	5.45
107	Employee growth	27	0.76	1.40	71.7%	2.12**	31.5%	24***	2.73
108	RMW	49	1.54	2.20**	74.2%	0.91	23.0%	38***	3.34
109	CMA	-25	-0.95	0.65	71.7%	1.58	27.3%	30***	3.07
110	Intermediary Capital	-33	-0.50	0.83	71.7%	-0.26	22.3%		
111	Intermediary Investment	-62	-0.63	0.83	71.7%	-0.35	22.4%	114***	3.48
112	Convertible Debt	29	1.65*	2.11**	74.8%	2.56***	34.7%	27***	3.77

Note. The table reports tests for the contribution of factors introduced in 2011-2016 relative to the set of factors introduced up to 2010. The left panel shows the estimate of risk price λ_g for each factor, together with the t-statistic obtained using our double-selection procedure and standard errors (t-stat DS), the t-statistic obtained using a single-selection procedure (t-stat SS), and the cross-sectional R^2 achieved by the model. λ_s is expressed in basis points per month. The middle panel shows the corresponding risk price estimates where the three Fama-French factors (RmRf, SMB, HML) are used as controls instead of the optimally-selected factors. We report the t-stat for the risk price as well as the cross-sectional R^2 achieved by the model that includes g_t and the three Fama-French factors. The right panel shows average excess returns (risk premia) for tradable factors, with corresponding test of significance.

Table 5: Testing for factors selected in 2010

		Regularized two-pass		Average excess ret	
id	Factor	λ_s (bp)	tstat (DS)	avg.ret. (bp)	tstat
37	Industry adjusted % change in capital expenditures	12	0.87	3	0.47
39	% change in sales - % change in inventory	-16	-0.63	8*	1.70
40	% change in sales - % change in A/R	5	0.69	11***	2.40
44	Number of earnings increases	20	2.17**	11***	2.65
53	Volatility of liquidity (dollar trading volume)	70	0.84	34***	2.51
55	Change in inventory	-10	-0.59	20***	3.20
59	Liquidity (Pastor and Stambaugh)	-83	-2.22**	32	1.13
63	Corporate investment	33	2.71***	14***	2.57
69	Growth in long-term debt	30	1.41	38***	6.35
71	Price delay	55	2.20**	5	0.66
72	Change in 6-month momentum	6	0.17	31***	3.08
82	Abnormal earnings announcement volume	7	0.48	1	0.14
85	Change in shares outstanding	-31	-0.73	41***	3.82
93	Capital expenditures and inventory	-32	-0.79	37***	4.30

Note. The table reports the list of factors selected as controls for the tests of Table 4 (first step of the double-selection procedure). In addition to reporting the list of factors, we also report their significance (itself obtained using our double-selection asymptotic results). The right part of the table shows average excess returns (risk premia) for the control factors, with corresponding test of significance.

Table 6: Testing Factors by Published Year

Year	Factor id						
1994	30						
1995	31	32	33				
1996	34	35					
1997	36						
1998	37	38	39	40	41	42	
1999	43	44					
2000	45	46	47	48	49		
2001	50	53	54				
2002	55	56	58				
2003	59	60	61				
2004	62	63	64	65	66		
2005	67	68	69	70	71		
2006	72	73	74	75	76	77	78
2007	80	81					
2008	82	83	84	85	86	87	
2009	88	89	90	91			
2010	92	93	94	96	97		
2011	98						
2012	99						
2013	100	101	102				
2014	104	105	106	107			
2015	108	109					
2016	110	111	112				

Note. The table reports factor testing results recursively from 1994 to 2016. For each year t, we report the id of the factors that were introduced during that year. We underline a factor if its risk price is statistically significant according to our double-selection test, controlling for all factors introduced up to year $t-1$.

Table 7: Recursive Test for Selected Models

Year	Factor id																	
1994		8	12	15	16		27											
1995	1			15	16		27		33									
1996	1				16				33									
1997	1								33									
1998	1				16				33	39	40							
1999					16					39	40	41	44					
2000					16				33	39	40	41						
2001								30	33	39	40	41						
2002										39	40		44		55			
2003										39	40		44		55		63	
2004										39	40		44		55		63	
2005						26				39			44		55		63	
2006						26							44	53			63	
2007						26							44				63	71
2008						26							44				63	71
2009						26							44			59	63	71
2010													44			59	63	71
2011										39						59	63	71
2012										39						59	63	71
2013						26				39							63	71
2014						26							44				63	71
2015						26							44				63	
2016						26							44				63	71

Year	Factor id																	
1994																		
1995																		
1996																		
1997																		
1998																		
1999																		
2000																		
2001																		
2002																		
2003																		
2004																		
2005																		
2006																		
2007																		
2008																		
2009																		
2010																		
2011																		
2012																		
2013																		
2014																		
2015																		
2016																		

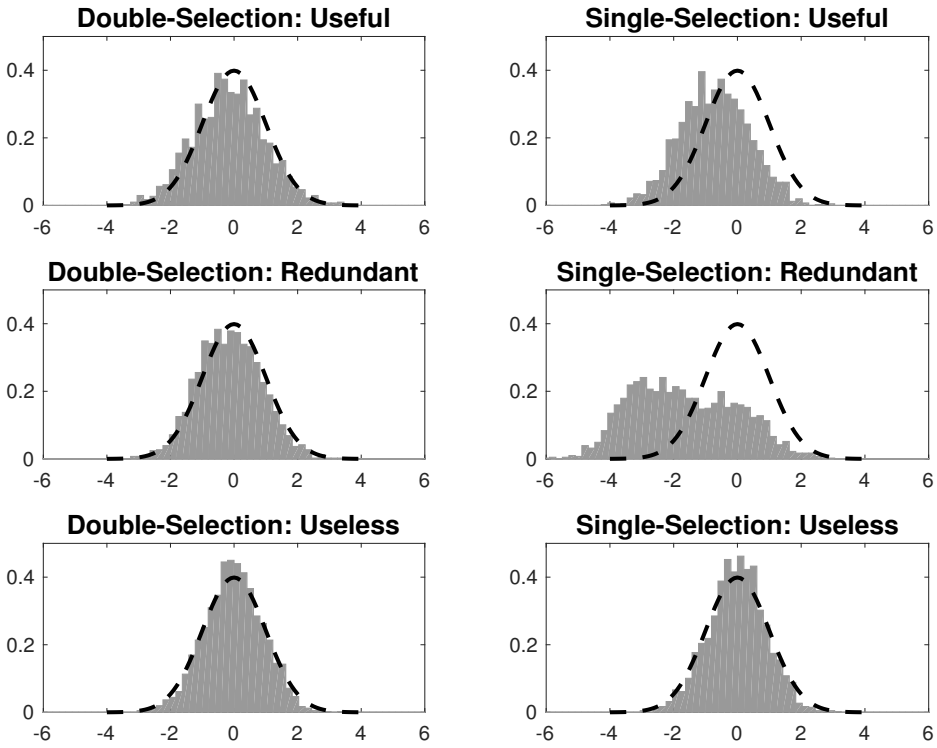
Note. The table reports in each year t the factors that are selected by our procedure among all the ones available up to that year. The selection is operated via LASSO and corresponds to the first, model-selection step of our two-step procedure. Underlined factors are those that had been deemed significant using our statistical test at the time of their introduction. Non-underlined factors are those that are selected in the model in year t yet were insignificant when first introduced. Shaded factors are those that were never tested when introduced (since they were introduced in 1994 and we start testing factors in 1994, as shown in Table 6.)

Table 8: Recursive Test Summary

	Later Selected	Later Not Selected
Significant at introduction	81.8%	18.2%
Insignificant at introduction	30.3%	69.7%

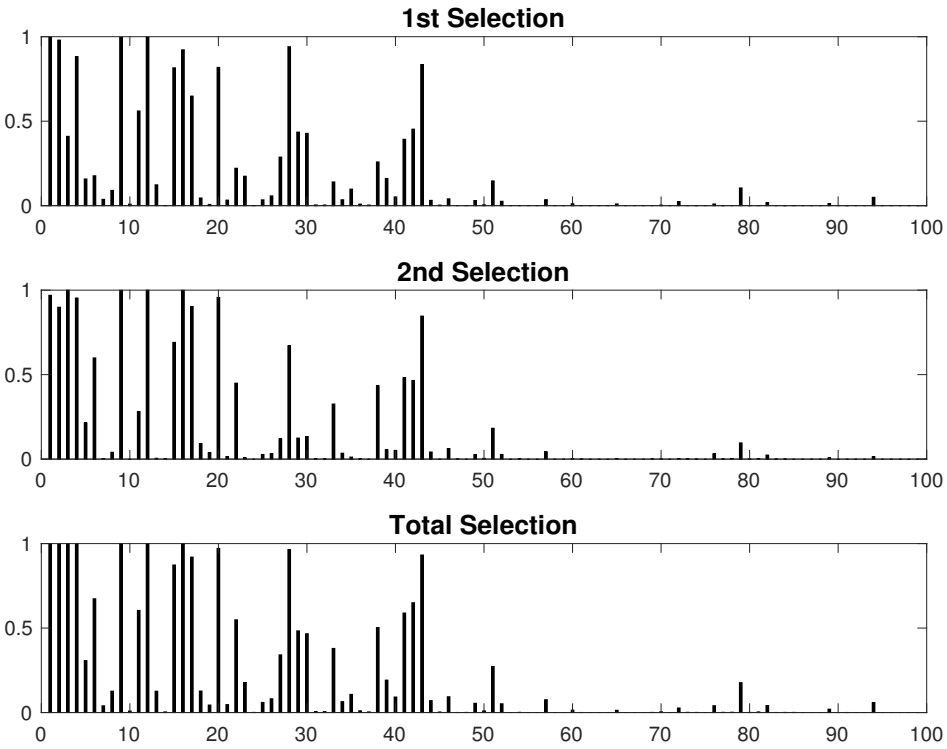
Note. The table reports the recursive performance of our statistical test. For each factor introduced since 1994, the table notes whether it was statistically significant when it was introduced (established using our test), and counts whether that same factor was later on selected to be part of the best parsimonious model in future years, for at least 3 years. The table then summarizes how often significant or insignificant factors get included in the model subsequently.

Figure 1: Histograms of the Standardized Estimates in Simulations



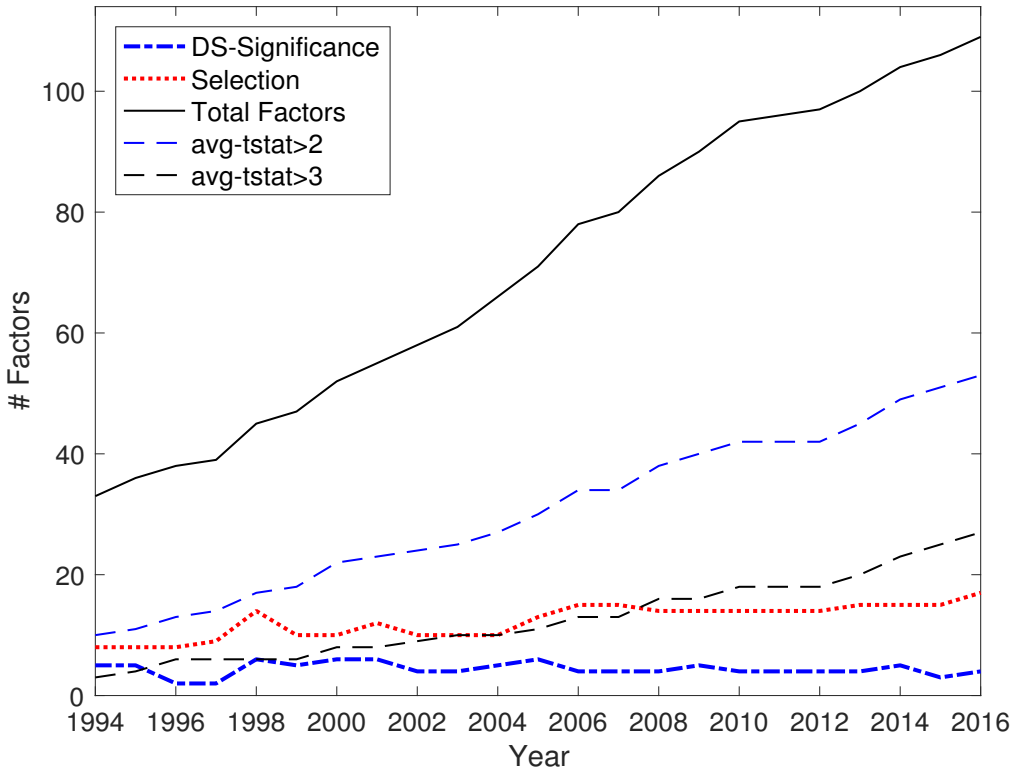
Note. This figure presents the histograms of the standardized double- and single-selection estimates using estimated standard errors, compared with the standard normal density in solid dash lines. We set $T = 600$, $N = 200$, and $p = 100$. The regularization parameters in each selection are chosen by minimizing their BICs.

Figure 2: Histograms of the Selection Variables



Note. The figure reports how often each factor is selected in each step of our double selection (and their union in the bottom panel) in Monte Carlo simulations. Each factor corresponds to a number on the X axis. Factors 1 - 4 are part of the true factors in the DGP. Factors 5 - 52 are redundant, whereas factors 53 - 100 are useless. We set $T = 600$, $N = 200$, and $p = 100$. The regularization parameters in each selection are chosen by minimizing their BICs.

Figure 3: Cumulative #Factor Discovery



Note. The figure provides time series plots for the cumulative number of factors introduced in the literature and chosen by the model between 1994 to 2016. In particular, the solid line reports the cumulative number of factors in our data library, based on the year in which they were introduced. The dashed blue and black lines report the cumulative number of significant factors by the risk premium, using a critical value for the t-stat of 2 and 3, respectively. The dotted line represents the total number of factors selected by our model-selection procedure, and the blue dotted line reports the selected factors that are statistically significant.

Appendix A Technical Details

A.1 Notation

We summarize the notation used throughout. Let e_i be a vector with 1 in the ith entry and 0 elsewhere, whose dimension depends on the context. Let ι_k denote a k-dimensional vector with all entries being 1. We use $a \vee b$ to denote the max of a and b, and $a \wedge b$ as their min for any scalars a and b. We also use the notation $a \lesssim b$ to denote $a \leq Kb$ for some constant $K > 0$; and $a \lesssim_p b$ to denote $a = O_p(b)$. For any time series of vectors $\{a_t\}_{t=1}^T$, we denote $\bar{a} = T^{-1} \sum_{t=1}^T a_t$. In addition, we write $\bar{a}_t = a_t - \bar{a}$. We use the capital letter A to denote the matrix $(a_1 : a_2 : \ldots : a_T)$, and write $\bar{A} = A - \iota_T^\intercal \bar{a}$ correspondingly. We use $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of A. We use $\|A\|_1$, $\|A\|_\infty$, $\|A\|$, and $\|A\|_{\mathrm{F}}$ to denote the $\mathbb{L}_1$ norm, the $\mathbb{L}_\infty$ norm, the operator norm (or $\mathbb{L}_2$ norm), and the Frobenius norm of a matrix $A = (a_{ij})$, that is, $\max_j \sum_i |a_{ij}|$, $\max_i \sum_j |a_{ij}|$, $\sqrt{\lambda_{\max}(A^\intercal A)}$, and $\sqrt{\mathrm{Tr}(A^\intercal A)}$, respectively. We also use $\|A\|_{\mathrm{MAX}} = \max_{i,j} |a_{ij}|$ to denote the $\mathbb{L}_\infty$ norm of A on the vector space. When a is a vector, both $\|a\|$ and $\|a\|_{\mathrm{F}}$ are equal to its Euclidean norm. We use $\|a\|_0$ to denote $\sum_i 1_{\{a_i \neq 0\}}$. We also denote $Supp(a) = \{i : a_i \neq 0\}$. We write the projection operator with respect to a matrix A as $\mathbb{P}_A = A(A^\intercal A)^{-1} A^\intercal$, and the corresponding annihilator as $\mathbb{M}_A = \mathbb{I} - \mathbb{P}_A$, where $\mathbb{I}$ is the identity matrix whose size depends on the context. For a set of indices I, let $A[I]$ denote a sub-matrix of A, which contains all columns indexed in I.

A.2 Technical Assumptions

Assumption A.1 (Sparsity)**.** $\|\lambda_h\|_0 \leq s$, $\|\chi_{j,\cdot}\|_0 \leq s$, $\|\eta_{j,\cdot}\|_0 \leq s$, $1 \leq j \leq d$, *for some* s *such that* $sn^{-1} \to 0$.

Definition 1 (LASSO and Post-LASSO Estimators)**.** *We consider a generic linear regression problem with sparse coefficients:*

$$Y = X\beta + \varepsilon, \quad \textit{subject to} \quad \|\beta\|_0 \leq s,$$

where Y is a $n \times 1$ vector, X is a $n \times p$ matrix, β is $p \times 1$ vector of parameters. We define the LASSO estimator as

$$\overline{\beta} = \arg\min_\beta \left\{ n^{-1} \|Y - X\beta\|^2 + n^{-1}\tau \|\beta\|_1 \right\}.$$

We define the Post-LASSO estimator $\widetilde{\beta}_{\widehat{I}}$ as

$$\widetilde{\beta}_{\widehat{I}} = \arg\min_\beta \left\{ n^{-1} \|Y - X\beta\|^2 : \quad \beta_j = 0, \quad j \notin \widehat{I} \right\},$$

where $\widehat{I}$ is the set of indices of variables selected by a first-step LASSO, that is, $\widehat{I} = Supp(\overline{\beta})$.

We adopt a high-level assumption on the model selection properties of LASSO and the prediction error bounds of the Post-LASSO estimators in (7) and (8). Belloni and Chernozhukov (2013) provide more primitive conditions for these bounds to hold.

Assumption A.2 (Properties of Post-LASSO Estimators)**.** *The Post-LASSO estimators in (7) and (8) satisfy the following properties:*

1. $\widehat{s} = |\widehat{I}_1 \bigcup \widehat{I}_2| \lesssim_p s$.
2. *Moreover, if* $\tau_0 \geq 2c \left\| \lambda_g^\intercal C_e^\intercal (\iota_n : \widehat{C}_h) \right\|_1$, *for some* $c > 1$, *then*

$$n^{-1/2} \left\| \iota_n (\widetilde{\gamma}_{\widehat{I}_1} - \breve{\gamma}_0) + \widehat{C}_h (\widetilde{\lambda}_{\widehat{I}_1} - \breve{\lambda}_h) \right\| \lesssim_p s T^{-1/2} (\log(n \vee p \vee T))^{1/2} + \tau_0 s^{1/2} n^{-1}, \tag{A.1}$$

where $\breve{\gamma}_0 = \gamma_0 + \xi^\intercal \lambda_g$ *and* $\breve{\lambda}_h = \chi^\intercal \lambda_g + \lambda_h$ *are the true parameter values given in (2) and (6). If* $\tau_j \geq 2c_j \left\| e_j^\intercal C_e^\intercal (\iota_n : \widehat{C}_h) \right\|_1$, *for some* $c_j > 1$ *and* $j = 1, 2, \ldots, d$, *then*

$$n^{-1/2} \left\| \iota_n (\widetilde{\xi}_{\widehat{I}_2} - \xi)^\intercal + \widehat{C}_h (\widetilde{\chi}_{\widehat{I}_2} - \chi)^\intercal \right\| \lesssim_p s T^{-1/2} (\log(n \vee p \vee T))^{1/2} + \|\tau\|_{\text{MAX}} s^{1/2} n^{-1}, \tag{A.2}$$

where $\tau = (\tau_1, \tau_2, \ldots, \tau_d)^\intercal$, ξ *and* χ *are the true parameter values given in (6).*

Assumption A.2 gives a probabilistic upper bound on $\widehat{s}$. The prediction error bounds in (A.1) and (A.2) are non-standard, because the regressors here are estimated. We provide a sketch of the proof for (A.1) in Appendix A.4, for which we need the following sparse eigenvalues assumption. The proof of (A.2) is similar and simpler. Our theoretical result below would also hold if other model selection procedures are employed, provided that they obey similar properties in Assumption A.2.

Assumption A.3 (Sparse Eigenvalues)**.** *There exist* $K_1, K_2 > 0$ *and a sequence* $l_n \to \infty$, *such that with probability approaching 1,*

$$K_1 \leq \phi_{\min}(l_n s) \left[n^{-1} (\iota_n : \widehat{C}_h)^\intercal (\iota_n : \widehat{C}_h) \right] \leq \phi_{\max}(l_n s) \left[n^{-1} (\iota_n : \widehat{C}_h)^\intercal (\iota_n : \widehat{C}_h) \right] \leq K_2,$$

where we denote

$$\phi_{\min}(k)[A] = \min_{1 \leq \|v\|_0 \leq k} \frac{v^\intercal A v}{\|v\|^2}, \quad \text{and} \quad \phi_{\max}(k)[A] = \max_{1 \leq \|v\|_0 \leq k} \frac{v^\intercal A v}{\|v\|^2}.$$

Assumption A.3 resembles one of the sufficient conditions that lead to desirable statistical properties of LASSO, which has been adopted by, e.g., Belloni et al. (2014b). It implies the restricted eigenvalue condition proposed by Bickel et al. (2009).

Assumption A.4 (Large Deviation Bounds)**.** *The stochastic discount factor, the returns, and the factors satisfy*

$$\|\bar{a}\|_{\text{MAX}} \lesssim_p T^{-1/2} (\log(n \vee p \vee T))^{1/2}, \text{ where } a \in \{m, v, z, u\}. \tag{A.3}$$

$$\left\| T^{-1} \bar{A} \bar{B}^\intercal - \text{Cov}(a_t, b_t) \right\|_{\text{MAX}} \lesssim_p T^{-1/2} (\log(n \vee p \vee T))^{1/2}, \text{ where } A, B \in \{M, V, Z, U\}. \tag{A.4}$$

Assumption A.4 imposes high-level assumptions on the large deviation type bounds, which can be verified using the same arguments as in Fan et al. (2011) under stationarity, ergodicity, strong mixing, and exponential-type tail conditions.

Next, we impose additional uniform bounds that impose restrictions on the cross-sectional dependence of the "residuals" in the covariance projection (6). Similar assumptions on factor loadings are employed by Giglio and Xiu (2016).

Assumption A.5 ("Moment" Conditions)**.** *The following restrictions on the $n \times d$ matrix C_e hold:*

$$\|C_e\|_{\text{MAX}} \lesssim 1, \quad \|C_e^\intercal \iota_n\|_{\text{MAX}} \lesssim n^{1/2}, \quad \|C_e^\intercal C_h\|_{\text{MAX}} \lesssim n^{1/2}, \tag{A.5}$$

$$\|C_e^\intercal \bar{u}\|_{\text{MAX}} \lesssim_p n^{1/2}T^{-1/2}, \quad \left\|C_e^\intercal \bar{U}\bar{V}^\intercal\right\|_{\text{MAX}} \lesssim_p n^{1/2}T^{1/2}, \tag{A.6}$$

$$\lambda_{\min}(n^{-1}C_e^\intercal C_e) \geq K, \quad \|C_e^\intercal(\beta_g \eta + \beta_h)\|_\infty \lesssim sn^{1/2}, \quad \|\beta_h\|_\infty \lesssim s. \tag{A.7}$$

In addition, for $a \in \{m, v, z, u\}$, it holds that

$$\|\Sigma_a\|_{\text{MAX}} \lesssim 1, \quad \|C_a\|_{\text{MAX}} \lesssim 1. \tag{A.8}$$

Finally, we impose a joint central limit theorem for $(z_t, \lambda^\intercal v_t z_t) = (z_t, (1 - \gamma_0 m_t) z_t)$. This can be verified by the standard central limit theory for dependent stochastic processes, if more primitive assumptions are satisfied, see, e.g., White (2000).

Assumption A.6 (CLT)**.** *The following results hold as $T \to \infty$:*

$$T^{1/2}\begin{pmatrix} \bar{z} \\ -T^{-1}\gamma_0 \bar{M}\bar{Z} - \Sigma_z \lambda_g \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{12}^\intercal & \Pi_{22} \end{pmatrix}\right),$$

where Π_{11}, Π_{12}, and Π_{22} are given by

$$\Pi_{11} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{E}(z_s z_t^\intercal),$$

$$\Pi_{12} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{E}\left(\lambda^\intercal v_s z_s z_t^\intercal\right),$$

$$\Pi_{22} = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{E}\left(\lambda^\intercal v_s \lambda^\intercal v_t z_s z_t^\intercal\right).$$

Assumption A.7 (Selection for the Asymptotic Variance Estimator)**.** *The Post-LASSO estimator $\widetilde{\eta}_{\widetilde{I}}$ satisfies the usual bounds. That is, if $\bar{\tau}_j \geq 2\bar{c}_j \|HZ\|_\infty$, for some $\bar{c}_j > 1$, $j = 1, 2, \ldots, d$, then we have*

$$\left\|(\widetilde{\eta}_{\widetilde{I}} - \eta)H\right\| \lesssim_p s^{1/2}(\log(p \vee T))^{1/2}, \quad \textit{and} \quad \left\|\widetilde{\eta}_{\widetilde{I}} - \eta\right\| \lesssim_p s^{1/2}T^{-1/2}(\log(p \vee T))^{1/2}.$$

A.3 Proof of Main Theorems

Proof of Theorem 1. The estimator of λ_g can be written in closed-form as

$$\widehat{\lambda}_g = \left(\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])} \widehat{C}_g\right)^{-1} \left(\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])} \bar{r}\right). \tag{A.9}$$

Moreover, by (2) and (5), we can relate C_g and C_h to β_g and β_h:

$$C_g = C_h \eta^\intercal + C_z, \quad \text{where} \quad C_h = (\beta_g \eta + \beta_h)\Sigma_h, \quad C_z = \beta_g \Sigma_z. \tag{A.10}$$

Using (3), (5), (A.10), and the fact that

$$\begin{aligned}
\widehat{C}_g - C_g &= (\widehat{C}_h - C_h)\eta^\intercal + (\widehat{C}_z - C_z), \\
\widehat{C}_z - C_z &= \beta_g \left(T^{-1}\bar{Z}\bar{Z}^\intercal - \Sigma_z\right) + T^{-1}\bar{U}\bar{Z}^\intercal + T^{-1}\left(\beta_g\eta + \beta_h\right)\bar{H}\bar{Z}^\intercal, \\
\widehat{C}_h - C_h &= (\beta_g\eta + \beta_h)\left(T^{-1}\bar{H}\bar{H}^\intercal - \Sigma_h\right) + T^{-1}\bar{U}\bar{H}^\intercal + T^{-1}\beta_g\bar{Z}\bar{H}^\intercal,
\end{aligned}$$

we obtain the following decomposition:

$$\begin{aligned}
&T^{1/2}(\widehat{\lambda}_g - \lambda_g) \\
=&\left(n^{-1}\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_g\right)^{-1} n^{-1}T^{1/2}\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left((C_g - \widehat{C}_g)\lambda_g + C_h\lambda_h + \beta_g\bar{z} + ((\beta_g\eta+\beta_h)\bar{h} + \bar{u})\right) \\
=&T^{1/2}\Sigma_z^{-1}\left(\bar{z} - \left(T^{-1}\bar{Z}\bar{V}^\intercal\lambda - \Sigma_z\lambda_g\right)\right) \\
&+\left(n^{-1}\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_g\right)^{-1}\left(n^{-1}T^{1/2}\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u} - T^{-1}\bar{U}\bar{V}^\intercal\lambda\right)\right. \\
&+n^{-1}T^{1/2}\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g - \widehat{C}_g\Sigma_z^{-1}) \times \left(\bar{z} - \left(T^{-1}\bar{Z}\bar{V}^\intercal\lambda - \Sigma_z\lambda_g\right)\right) \\
&-n^{-1}T^{1/2}\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta + \beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda - \Sigma_h(\eta^\intercal\lambda_g + \lambda_h) - \bar{h}\right) \\
&\left.+n^{-1}T^{1/2}\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right).
\end{aligned}$$

We first analyze the leading term. Note that $\gamma_0\bar{M} = -\bar{V}^\intercal\lambda$, by Assumption A.6 and applying the Delta method, we have

$$\begin{aligned}
&T^{1/2}\left(\Sigma_z^{-1}\bar{z} - \Sigma_z^{-1}\left(-T^{-1}\gamma_0\bar{Z}\bar{M} - \Sigma_z\lambda_g\right)\right) \\
&\xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\mathrm{E}\left((1-\lambda^\intercal v_t)(1-\lambda^\intercal v_s)\Sigma_z^{-1}z_t z_s^\intercal\Sigma_z^{-1}\right)\right).
\end{aligned} \tag{A.11}$$

Next, we show that the reminder terms are of a smaller order. By (A.42), we have

$$n^{-1}T^{1/2}\left\|\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u} - T^{-1}\bar{U}\bar{V}^\intercal\lambda\right)\right\| \lesssim_p s(n^{-1/2} + T^{-1/2})\log(n\vee p\vee T).$$

By (A.27), we have

$$n^{-1}T^{1/2}\left\|\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \lesssim_p s^2(n^{-1}T^{1/2} + T^{-1/2})\log(n\vee p\vee T).$$

By (A.40), we have

$$
\begin{aligned}
&n^{-1}T^{1/2}\left\|\widehat{C}_g^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^{\intercal}\lambda-\Sigma_h(\eta^{\intercal}\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
\lesssim_p& s^2(n^{-1/2}+T^{-1/2})\log(n\vee p\vee T).
\end{aligned}
$$

By Assumption A.4, (A.11), and (A.35), we have

$$
\begin{aligned}
&n^{-1}T^{1/2}\left\|\widehat{C}_g^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g-\widehat{C}_z\Sigma_z^{-1})\left(\bar{z}-\left(T^{-1}\bar{Z}\bar{V}^{\intercal}\lambda-\Sigma_z\lambda_g\right)\right)\right\| \\
\leq& n^{-1}T^{1/2}\left\|\widehat{C}_g^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g-\widehat{C}_z\Sigma_z^{-1})\right\|\left\|\bar{z}-\left(T^{-1}\bar{Z}\bar{V}^{\intercal}\lambda-\Sigma_z\lambda_g\right)\right\| \\
\lesssim_p& s(n^{-1/2}+T^{-1/2})\log(n\vee p\vee T).
\end{aligned}
$$

This concludes the proof. □

Proof of Theorem 2. By the identical argument in the proof of Theorem 2 of Newey and West (1987), we have

$$
\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T}Q_{tr}(1-\lambda^{\intercal}v_t)(1-\lambda^{\intercal}v_r)\left(z_tz_r^{\intercal}+z_rz_t^{\intercal}\right)\xrightarrow{p}\Sigma_z\Pi\Sigma_z.
$$

So applying the continuous mapping theorem, it is sufficient to show that

$$
\widehat{\Sigma}_z\xrightarrow{p}\Sigma_z, \tag{A.12}
$$

$$
\widetilde{\Pi}-\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T}Q_{tr}(1-\lambda^{\intercal}v_t)(1-\lambda^{\intercal}v_r)\left(z_tz_r^{\intercal}+z_rz_t^{\intercal}\right)\xrightarrow{p}0, \tag{A.13}
$$

where

$$
Q_{tr}=\left(1-\frac{|r-t|}{q+1}\right)1_{\{|t-r|\leq q\}},\quad \widetilde{\Pi}=\widehat{\Sigma}_z\widehat{\Pi}\widehat{\Sigma}_z.
$$

To prove (A.15), we note that by Assumptions A.4 and A.7, we have

$$
\begin{aligned}
&\left\|\widehat{\Sigma}_z-\Sigma_z\right\|_{\mathrm{MAX}} \\
\lesssim& T^{-1/2}\left\|(\widetilde{\eta}_{\widetilde{I}}-\eta)H\right\|\|Z\|_{\mathrm{MAX}}+T^{-1}\left\|(\widetilde{\eta}_{\widetilde{I}}-\eta)H\right\|^2+\left\|T^{-1}ZZ^{\intercal}-\Sigma_z\right\|_{\mathrm{MAX}} \\
\lesssim_p& s^{1/2}T^{-1/2}(\log(p\vee T))^{1/2}\|Z\|_{\mathrm{MAX}}+sT^{-1}\log(p\vee T)+T^{-1/2}(\log(n\vee p\vee T))^{1/2} \\
=& o_p(1).
\end{aligned} \tag{A.14}
$$

As to (A.13), we can decompose its left-hand side as

$$
\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T}Q_{tr}(\widehat{\lambda}-\lambda)^{\intercal}v_t(1-\widehat{\lambda}^{\intercal}v_r)\left(\widehat{z}_t\widehat{z}_r^{\intercal}+\widehat{z}_r\widehat{z}_t^{\intercal}\right) \tag{A.15}
$$

$$+\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T} Q_{tr}(1-\lambda^\intercal v_t)(\widehat{\lambda}-\lambda)^\intercal v_r\left(\widehat{z}_t\widehat{z}_r^\intercal+\widehat{z}_r\widehat{z}_t^\intercal\right) \tag{A.16}$$

$$+\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T} Q_{tr}(1-\lambda^\intercal v_t)(1-\lambda^\intercal v_r)\left((\widehat{z}_t-z_t)\,\widehat{z}_r^\intercal+(\widehat{z}_r-z_r)\,\widehat{z}_t^\intercal\right) \tag{A.17}$$

$$+\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T} Q_{tr}(1-\lambda^\intercal v_t)(1-\lambda^\intercal v_r)\left(z_t\,(\widehat{z}_r-z_r)^\intercal+z_r\,(\widehat{z}_t-z_t)^\intercal\right). \tag{A.18}$$

Analyzing each of these terms, we can obtain that

$$\begin{aligned}
&\left\|\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T} Q_{tr}(\widehat{\lambda}-\lambda)^\intercal v_t(1-\widehat{\lambda}^\intercal v_r)\left(\widehat{z}_t\widehat{z}_r^\intercal+\widehat{z}_r\widehat{z}_t^\intercal\right)\right\|_{\text{MAX}}\\
\lesssim& qT^{-1}\left\|\widehat{Z}\right\|\left\|\iota_T^\intercal-\widehat{\lambda}^\intercal V\right\|\left\|(\widehat{\lambda}-\lambda)^\intercal V\right\|_{\text{MAX}}\left\|\widehat{Z}\right\|_{\text{MAX}}\lesssim_p qs^{1/2}(T^{-1/2}+n^{-1/2})\left\|V\right\|_{\text{MAX}}\left\|Z\right\|_{\text{MAX}},\\
&\left\|\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T} Q_{tr}(1-\lambda^\intercal v_t)(\widehat{\lambda}-\lambda)^\intercal v_r\left(\widehat{z}_t\widehat{z}_r^\intercal+\widehat{z}_r\widehat{z}_t^\intercal\right)\right\|_{\text{MAX}}\\
\lesssim& qT^{-1}\left\|\iota_T^\intercal-\lambda^\intercal V\right\|\left\|\widehat{Z}\right\|\left\|(\widehat{\lambda}-\lambda)^\intercal V\right\|_{\text{MAX}}\left\|\widehat{Z}\right\|_{\text{MAX}}\lesssim_p qs^{1/2}(T^{-1/2}+n^{-1/2})\left\|V\right\|_{\text{MAX}}\left\|Z\right\|_{\text{MAX}},\\
&\left\|\frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{T} Q_{tr}(1-\lambda^\intercal v_t)(1-\lambda^\intercal v_r)\left((\widehat{z}_t-z_t)\,\widehat{z}_r^\intercal+(\widehat{z}_r-z_r)\,\widehat{z}_t^\intercal\right)\right\|_{\text{MAX}}\\
\lesssim& qT^{-1}\left\|\iota_T^\intercal-\lambda^\intercal V\right\|\left\|(\widehat{\eta}-\eta)H\right\|\left\|\widehat{Z}\right\|_{\text{MAX}}\left\|\iota_T^\intercal-\lambda^\intercal V\right\|_{\text{MAX}}\\
\lesssim_p& qs^{3/2}(T^{-1/2}+n^{-1/2})\left\|V\right\|_{\text{MAX}}\left\|Z\right\|_{\text{MAX}},
\end{aligned}$$

where we use

$$\begin{aligned}
&\left\|\iota_T^\intercal-\lambda^\intercal V\right\|\lesssim T^{1/2}+\left\|\bar{M}\right\|+\left\|\lambda^\intercal\bar{v}\right\|\lesssim_p T^{1/2},\\
&\left\|\iota_T^\intercal-\lambda^\intercal V\right\|_{\text{MAX}}\lesssim 1+\left\|\lambda^\intercal V\right\|_{\text{MAX}}\lesssim s\left\|V\right\|_{\text{MAX}},\\
&\left\|\iota_T^\intercal-\widehat{\lambda}^\intercal V\right\|\le\left\|\iota_T^\intercal-\lambda^\intercal V\right\|+\left\|(\widehat{\lambda}-\lambda)^\intercal V\right\|\lesssim_p T^{1/2}+\left\|\widehat{\lambda}-\lambda\right\|\left\|V\right\|\lesssim_p T^{1/2},\\
&\left\|\widehat{Z}\right\|\lesssim T^{1/2}\left\|\widehat{\Sigma}_z\right\|^{1/2}\lesssim_p T^{1/2}\left\|\Sigma_z\right\|^{1/2}\lesssim T^{1/2},\\
&\left\|(\widehat{\lambda}-\lambda)^\intercal V\right\|_{\text{MAX}}\le\left\|\widehat{\lambda}-\lambda\right\|_{\infty}\left\|V\right\|_{\text{MAX}}\le\left\|\widehat{\lambda}-\lambda\right\|\left\|V\right\|_{\text{MAX}}\lesssim_p s^{1/2}(T^{-1/2}+n^{-1/2})\left\|V\right\|_{\text{MAX}},\\
&\left\|\widehat{Z}\right\|_{\text{MAX}}\le\left\|(\widehat{\eta}-\eta)H\right\|+\left\|Z\right\|_{\text{MAX}}\lesssim_p\left\|Z\right\|_{\text{MAX}},
\end{aligned}$$

which hold by (A.14), Assumption A.4, and Lemma 7. This concludes the proof. □

A.4 Proof of Lemmas

Proof of (A.1). We provide a sketch of the proof, as they are identical to Belloni and Chernozhukov (2013). With respect to the optimization problem (7), we define

$$Q(\gamma,\lambda)=n^{-1}\left\|\bar{r}-\iota_n\gamma-\widehat{C}_h\lambda\right\|^2.$$

We denote the solution to this problem as $\widetilde{\gamma}$ and $\widetilde{\lambda}$. Let $\delta = \widetilde{\lambda} - \breve{\lambda}_h$. Note by (5) and (2), we have

$$\mathrm{E}(r_t) = \iota_n \breve{\gamma}_0 + C_h \breve{\lambda}_h + C_e \lambda_g, \quad \text{and} \quad \bar{r} = \mathrm{E}(r_t) + \beta_g \bar{g} + \beta_h \bar{h} + \bar{u}.$$

By direct calculations, we have

$$\begin{aligned}
& Q(\widetilde{\gamma}, \widetilde{\lambda}) - Q(\breve{\gamma}_0, \breve{\lambda}_h) - n^{-1} \left\| \iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right\|^2 \\
= & -2n^{-1} \left(\bar{r} - \iota_n \breve{\gamma}_0 - \widehat{C}_h \breve{\lambda}_h \right)^{\intercal} \left(\iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right) \\
= & -2n^{-1} \left(\beta_g \bar{g} + \beta_h \bar{h} + \bar{u} + (C_h - \widehat{C}_h) \breve{\lambda}_h + C_e \lambda_g \right)^{\intercal} \left(\iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right) \\
\geq & -2n^{-1} \left\| \beta_g \bar{g} + \beta_h \bar{h} + \bar{u} + (C_h - \widehat{C}_h) \breve{\lambda}_h \right\| \left\| \iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right\| \\
& -2n^{-1} \left\| (C_e \lambda_g)^{\intercal} (\iota_n : \widehat{C}_h) \right\|_1 \| (\widetilde{\gamma} - \breve{\gamma}_0 : \delta^{\intercal})^{\intercal} \|_1 \\
\geq & -2n^{-1} \left\| \beta_g \bar{g} + \beta_h \bar{h} + \bar{u} + (C_h - \widehat{C}_h) \breve{\lambda}_h \right\| \left\| \iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right\| \\
& - \tau_0 K^{-1} n^{-1} \left(|\widetilde{\gamma} - \breve{\gamma}_0| + \|\delta_I\|_1 + \|\delta_{I^c}\|_1 \right),
\end{aligned}$$

where I is the set of non-zeros in $\breve{\lambda}_h$, I^c is its complement, and δ_I is a sub-vector of δ with all entries taken from I.

On the other hand, by definition of $\widetilde{\gamma}$ and $\widetilde{\lambda}$, we have

$$\begin{aligned}
Q(\widetilde{\gamma}, \widetilde{\lambda}) - Q(\breve{\gamma}_0, \breve{\lambda}_h) \leq & \tau_0 n^{-1} \left(\left\| (\breve{\gamma}_0 : \breve{\lambda}_h^{\intercal})^{\intercal} \right\|_1 - \left\| (\widetilde{\gamma} : \widetilde{\lambda}^{\intercal})^{\intercal} \right\|_1 \right) \\
\leq & \tau_0 n^{-1} (|\widetilde{\gamma} - \breve{\gamma}_0| + \|\delta_I\|_1 - \|\delta_{I^c}\|_1).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
& n^{-1} \left\| \iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right\|^2 - \tau_0 c^{-1} n^{-1} \left(|\widetilde{\gamma} - \breve{\gamma}_0| + \|\delta_I\|_1 + \|\delta_{I^c}\|_1 \right) \\
& \qquad - 2n^{-1} \left\| \beta_g \bar{g} + \beta_h \bar{h} + \bar{u} + (C_h - \widehat{C}_h) \breve{\lambda}_h \right\| \left\| \iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right\| \\
& \qquad \leq \tau_0 n^{-1} (|\widetilde{\gamma} - \breve{\gamma}_0| + \|\delta_I\|_1 - \|\delta_{I^c}\|_1), \qquad \text{(A.19)}
\end{aligned}$$

where we use the fact that

$$\tau_0 \geq 2c \left\| \lambda_g^{\intercal} C_e^{\intercal} (\iota_n : \widehat{C}_h) \right\|_1.$$

If it holds that

$$n^{-1} \left\| \iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right\| - 2n^{-1} \left\| \beta_g \bar{g} + \beta_h \bar{h} + \bar{u} + (C_h - \widehat{C}_h) \breve{\lambda}_h \right\| < 0,$$

we can establish that

$$n^{-1/2} \left\| \iota_n(\widetilde{\gamma} - \breve{\gamma}_0) + \widehat{C}_h \delta \right\| \lesssim_p s T^{-1/2} (\log(n \vee p \vee T))^{1/2},$$

where we use the fact that

$$n^{-1/2}\left\|\beta_g\bar{g}\right\| \lesssim \left\|\beta_g\right\|_{\mathrm{MAX}}\left\|\bar{g}\right\|_{\mathrm{MAX}} \lesssim_p T^{-1/2}, \tag{A.20}$$

$$n^{-1/2}\left\|\bar{u}\right\| \lesssim \left\|\bar{u}\right\|_{\mathrm{MAX}} \lesssim_p T^{-1/2}(\log(n\vee p\vee T))^{1/2}, \tag{A.21}$$

$$n^{-1/2}\left\|\beta_h\bar{h}\right\| \leq \left\|\beta_h\right\|_{\infty}\left\|\bar{h}\right\|_{\mathrm{MAX}} \lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2}, \tag{A.22}$$

$$n^{-1/2}\left\|(C_h-\widehat{C}_h)\check{\lambda}_h\right\| \lesssim \left\|C_h-\widehat{C}_h\right\|_{\mathrm{MAX}}\left\|\check{\lambda}_h\right\|_1 \lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2}. \tag{A.23}$$

Otherwise, from (A.19) it follows that

$$-c^{-1}\left(|\widetilde{\gamma}-\check{\gamma}_0|+\|\delta_I\|_1+\|\delta_{I^c}\|_1\right) \leq |\widetilde{\gamma}-\check{\gamma}_0|+\|\delta_I\|_1-\|\delta_{I^c}\|_1,$$

which leads to, writing $\bar{c}=(c+1)(c-1)^{-1}$,

$$\|\delta_{I^c}\| \leq \bar{c}(|\widetilde{\gamma}-\check{\gamma}_0|+\|\delta_I\|_1).$$

Then by (A.19) again as well as the restricted eigenvalue condition in Belloni and Chernozhukov (2013), we obtain

$$\begin{aligned}\left\|\iota_n(\widetilde{\gamma}-\check{\gamma}_0)+\widehat{C}_h\delta\right\|^2 - 2\left\|\beta_g\bar{g}+\beta_h\bar{h}+\bar{u}+(C_h-\widehat{C}_h)\check{\lambda}_h\right\|\left\|\iota_n(\widetilde{\gamma}-\check{\gamma}_0)+\widehat{C}_h\delta\right\| \\ \leq (1+c^{-1})\tau_0(|\widetilde{\gamma}-\check{\gamma}_0|+\|\delta_I\|_1) \lesssim \tau_0 s^{1/2}n^{-1/2}\left\|\iota_n(\widetilde{\gamma}-\check{\gamma}_0)+\widehat{C}_h\delta\right\|.\end{aligned}$$

Therefore, we have

$$\begin{aligned}n^{-1/2}\left\|\iota_n(\widetilde{\gamma}-\check{\gamma}_0)+\widehat{C}_h\delta\right\| &\lesssim n^{-1/2}\left\|\beta_g\bar{g}+\beta_h\bar{h}+\bar{u}+(C_h-\widehat{C}_h)\check{\lambda}_h\right\|+\tau_0 s^{1/2}n^{-1} \\ &\lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2}+\tau_0 s^{1/2}n^{-1}.\end{aligned}$$

The Post-LASSO estimator converges at the same rate following the same arguments as in Belloni and Chernozhukov (2013).

□

Lemma 1. *Under Assumptions A.1, A.2, A.4, A.5, we have*

$$n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\chi^{\intercal}\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}. \tag{A.24}$$

$$n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}. \tag{A.25}$$

Proof of Lemma 1. Using the fact that $\widehat{I}_2\subseteq\widehat{I}$ and by (A.2), we have

$$\begin{aligned}n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\chi^{\intercal}\right\| &= n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\widehat{C}_h\chi^{\intercal}+\iota_n\xi^{\intercal})\right\| \leq n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_2])}(\widehat{C}_h\chi^{\intercal}+\iota_n\xi^{\intercal}))\right\| \\ &\leq n^{-1/2}\left\|\iota_n(\xi-\widetilde{\xi}_{\widehat{I}_2})^{\intercal}+\widehat{C}_h\chi^{\intercal}-\widehat{C}_h\widetilde{\chi}^{\intercal}_{\widehat{I}_2}\right\| \\ &\lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2}+\|\tau\|_{\mathrm{MAX}}\, s^{1/2}n^{-1}.\end{aligned}$$

Since by Assumptions A.4 and A.5, our choice of τ satisfies:

$$\begin{aligned} n^{-1}\left\|\tau\right\|_{\mathrm{MAX}} \lesssim & n^{-1}\max_{1\le j\le d}\left\|e_j^\intercal C_e^\intercal \widehat{C}_h\right\|_1 \lesssim n^{-1}\left\|C_e^\intercal C_h\right\|_{\mathrm{MAX}} + n^{-1}\left\|C_e^\intercal(\widehat{C}_h - C_h)\right\|_{\mathrm{MAX}} \\ \lesssim_p & (n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}. \end{aligned} \tag{A.26}$$

This concludes the proof of (A.24).

Similarly, to prove (A.25), by (A.1) we have

$$\begin{aligned} & n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_1])}\left(\widehat{C}_h\breve{\lambda}_h + \iota_n\breve{\gamma}_0\right)\right\| \\ \le & n^{-1/2}\left\|(\iota_n:\widehat{C}_h)(\widetilde{\gamma}_{\widehat{I}_1}-\breve{\gamma}_0:(\widetilde{\lambda}_{\widehat{I}_1}-\breve{\lambda}_h)^\intercal)^\intercal\right\| \lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2} + \tau_0 s^{1/2}n^{-1}. \end{aligned}$$

Because we can select τ_0 that satisfies

$$\begin{aligned} n^{-1}\tau_0 \le & n^{-1}\left\|\lambda_g^\intercal C_e^\intercal(\iota_n:\widehat{C}_h)\right\|_1 \le n^{-1}\left|\lambda_g^\intercal C_e^\intercal\iota_n\right| + n^{-1}\left\|\lambda_g^\intercal C_e^\intercal\widehat{C}_h\right\|_{\mathrm{MAX}} \\ \lesssim & n^{-1}\left\|C_e\iota_n\right\|_{\mathrm{MAX}} + \left\|C_e\right\|_{\mathrm{MAX}}\left\|\widehat{C}_h - C_h\right\|_{\mathrm{MAX}} + n^{-1}\left\|C_e^\intercal C_h\right\|_{\mathrm{MAX}} \\ \lesssim_p & (n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}, \end{aligned}$$

hence it follows that

$$n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_1])}\left(\widehat{C}_h(\lambda_h+\chi^\intercal\lambda_g)+\iota_n\gamma_0\right)\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}.$$

By the triangle inequality and $\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_1])}\iota_n = 0$, we have

$$\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_1])}\widehat{C}_h\lambda_h\right\| \le \left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_1])}\left(\widehat{C}_h(\lambda_h+\chi^\intercal\lambda_g)+\iota_n\gamma_0\right)\right\| + \left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_1])}\widehat{C}_h\chi^\intercal\right\|\left\|\lambda_g\right\|,$$

which, combined with (A.24) and $\left\|\lambda_g\right\| \lesssim 1$, lead to the conclusion. □

Lemma 2. *Under Assumptions A.1, A.2, A.3, A.4, A.5, we have*

$$n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \lesssim_p s^2(n^{-1}+T^{-1})\log(n\vee p\vee T). \tag{A.27}$$

Proof of Lemma 2 . We note by (6) that

$$\widehat{C}_g = \widehat{C}_h\chi^\intercal + \widehat{C}_g - C_g + \iota_n\xi^\intercal + (C_h-\widehat{C}_h)\chi^\intercal + C_e, \tag{A.28}$$

thereby it follows

$$\begin{aligned} n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \le & n^{-1}\left\|\chi\widehat{C}_h^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| + n^{-1}\left\|C_e^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \\ & + n^{-1}\left\|(\widehat{C}_g - C_g + (C_h-\widehat{C}_h)\chi^\intercal)^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\|. \end{aligned}$$

On the one hand, by Lemma 1, we have

$$n^{-1}\left\|\chi\widehat{C}_h^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \le n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\chi^\intercal\right\| n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\|$$

$$\lesssim_p s^2(n^{-1}+T^{-1})\log(n\vee p\vee T). \tag{A.29}$$

On the other hand, note that

$$\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h = (\iota_n\gamma_0+\widehat{C}_h\lambda_h)-(\iota_n:\widehat{C}_h)(\hat{\gamma}_0:\hat{\lambda}_h^\intercal)^\intercal = (\iota_n:\widehat{C}_h)(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)^\intercal,$$

where $(\hat{\gamma}_0:\hat{\lambda}_h^\intercal)^\intercal = \arg\min_{\gamma,\lambda}\{\iota_n\gamma_0+\widehat{C}_h\lambda_h-\iota_n\gamma-\widehat{C}_h\lambda:\lambda_j=0, j\in\widehat{I}^c\}$. By Assumption A.3, we have

$$\begin{aligned}
n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| =& n^{-1/2}\left\|(\iota_n:\widehat{C}_h)(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)^\intercal\right\| \\
\geq& \phi_{\min}^{1/2}(s+\hat{s}+1)\left[n^{-1}(\iota_n:\widehat{C}_h)^\intercal(\iota_n:\widehat{C}_h)\right]\left\|(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)\right\| \\
\gtrsim& \left\|(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)\right\|,
\end{aligned}$$

hence it follows from (A.25) that

$$\left\|(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}. \tag{A.30}$$

Using this, we have

$$\begin{aligned}
n^{-1}\left\|C_e^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| =& n^{-1}\left\|C_e^\intercal(\iota_n:\widehat{C}_h)(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)^\intercal\right\| \\
\lesssim& n^{-1}\left\|C_e^\intercal(\iota_n:\widehat{C}_h)\right\|_{\mathrm{MAX}}\left\|(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)^\intercal\right\|_1.
\end{aligned} \tag{A.31}$$

Using (A.5) and Assumption A.4, it follows that

$$\begin{aligned}
n^{-1}\left\|C_e^\intercal(\iota_n:\widehat{C}_h)\right\|_{\mathrm{MAX}} \leq& n^{-1}\left\|C_e^\intercal(\widehat{C}_h-C_h)\right\|_{\mathrm{MAX}}+n^{-1}\left\|C_e^\intercal C_h\right\|_{\mathrm{MAX}}+n^{-1}\left\|C_e^\intercal\iota_n\right\|_{\mathrm{MAX}} \\
\lesssim& \|C_e\|_{\mathrm{MAX}}\left\|\widehat{C}_h-C_h\right\|_{\mathrm{MAX}}+n^{-1}\left\|C_e^\intercal C_h\right\|_{\mathrm{MAX}}+n^{-1}\left\|C_e^\intercal\iota_n\right\|_{\mathrm{MAX}} \\
\lesssim_p& (n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}.
\end{aligned} \tag{A.32}$$

Moreover, since by sparsity of λ_h and $\hat{\lambda}_h$, we have

$$\left\|(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)^\intercal\right\|_1 \leq (s+\hat{s}+1)^{1/2}\left\|(\gamma_0-\hat{\gamma}_0:\lambda_h^\intercal-\hat{\lambda}_h^\intercal)^\intercal\right\|.$$

Combining (A.30), (A.31), and (A.32), we obtain

$$n^{-1}\left\|C_e^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \lesssim_p s^{3/2}(n^{-1}+T^{-1})\log(n\vee p\vee T). \tag{A.33}$$

Finally, by (A.25) we have

$$\begin{aligned}
& n^{-1}\left\|(\widehat{C}_g-C_g+(C_h-\widehat{C}_h)\chi^\intercal)^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \\
\lesssim& \left\|\widehat{C}_g-C_g+(C_h-\widehat{C}_h)\chi^\intercal\right\|_{\mathrm{MAX}} n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\lambda_h\right\| \\
\lesssim_p& s^2(n^{-1/2}T^{-1/2}+T^{-1})(\log(n\vee p\vee T))^{1/2}.
\end{aligned}$$

The above estimate, along with (A.33) and (A.29), conclude the proof of (A.27).

□

Lemma 3. *Under Assumptions A.1, A.2, A.3, A.4, A.5, we have*

$$n^{-1}\left\|\widehat{C}_g^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}. \tag{A.34}$$

$$n^{-1}\left\|\widehat{C}_g^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g-\widehat{C}_g\Sigma_z^{-1})\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}. \tag{A.35}$$

Proof of Lemma 3. (i) By (6), we have

$$\begin{aligned}n^{-1}\left\|\widehat{C}_g^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\| \leq& n^{-1}\left\|C_e^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\| + n^{-1}\left\|\chi\widehat{C}_h^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\| \\ &+ n^{-1}\left\|\left((\widehat{C}_g-C_g)^{\intercal}+\chi(C_h-\widehat{C}_h)^{\intercal}\right)\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\|.\end{aligned}$$

Moreover, by (A.24), we obtain

$$\begin{aligned}n^{-1}\left\|\chi\widehat{C}_h^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\| \leq& n^{-1/2}\left\|\chi\widehat{C}_h^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\right\| n^{-1/2}\left\|C_h\eta^{\intercal}\right\| \\ \lesssim_p& s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2},\end{aligned} \tag{A.36}$$

where we use the fact that $C_g=C_h\eta^{\intercal}+C_z$, and that

$$n^{-1/2}\left\|C_h\eta^{\intercal}\right\| \lesssim \left\|C_h\eta^{\intercal}\right\|_{\text{MAX}} \lesssim \left\|C_g\right\|_{\text{MAX}}+\left\|C_z\right\|_{\text{MAX}} \lesssim 1.$$

In addition, we have

$$n^{-1}\left\|C_e^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\| \leq n^{-1}\left\|C_e^{\intercal}C_h\eta^{\intercal}\right\| + n^{-1}\left\|C_e^{\intercal}\mathbb{P}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\|.$$

To bound the first term, we have

$$n^{-1}\left\|C_e^{\intercal}C_h\eta^{\intercal}\right\| \lesssim n^{-1}\left\|C_e^{\intercal}C_h\right\|_{\text{MAX}}\left\|\eta\right\|_{\infty} \lesssim_p sn^{-1/2}(\log(n\vee p\vee T))^{1/2}.$$

As to the second term, using (A.32) we obtain

$$\begin{aligned}&n^{-1}\left\|C_e^{\intercal}\mathbb{P}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^{\intercal}\right\| \\ =&n^{-1}\left\|C_e^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\left((\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\right)^{-1}(\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}C_h\eta^{\intercal}\right\| \\ \leq&n^{-1}\left\|C_e^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\right\|\left\|\left((\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\right)^{-1}\right\|\left\|(\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}C_h\eta^{\intercal}\right\| \\ \lesssim&(1+\widehat{s})\phi_{\min}^{-1}(\widehat{s}+1)\left[n^{-1}(\iota_n:\widehat{C}_h)^{\intercal}(\iota_n:\widehat{C}_h)\right]n^{-1}\left\|C_e^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\right\|_{\text{MAX}}n^{-1}\left\|(\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}C_h\eta^{\intercal}\right\|_{\text{MAX}} \\ \lesssim_p& s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2},\end{aligned}$$

where we also use $\left\|C_h\eta\right\|_{\text{MAX}} \leq \left\|C_g\right\|_{\text{MAX}}+\left\|C_z\right\|_{\text{MAX}} \lesssim 1$, and

$$\begin{aligned}n^{-1}\left\|(\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}C_h\eta\right\|_{\text{MAX}} \leq& n^{-1}\left\|(\iota_n:\widehat{C}_h)^{\intercal}C_h\eta\right\|_{\text{MAX}} \lesssim \left\|(\iota_n:\widehat{C}_h)\right\|_{\text{MAX}}\left\|C_h\eta\right\|_{\text{MAX}} \\ \lesssim&\left(\left\|(\iota_n:C_h)\right\|_{\text{MAX}}+\left\|\widehat{C}_h-C_h\right\|_{\text{MAX}}\right)\left\|C_h\eta\right\|_{\text{MAX}} \lesssim_p 1.\end{aligned}$$

Therefore, we have

$$n^{-1}\left\|C_e^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])} C_h \eta^\intercal\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2}. \tag{A.37}$$

Similarly, because we have

$$\begin{aligned}
&n^{-1}\left\|\left((\widehat{C}_g-C_g)^\intercal+\chi(C_h-\widehat{C}_h)^\intercal\right)C_h\eta^\intercal\right\| \\
\lesssim&\left\|(\widehat{C}_g-C_g)^\intercal+\chi(C_h-\widehat{C}_h)^\intercal\right\|_{\mathrm{MAX}}\|C_h\eta^\intercal\|_{\mathrm{MAX}} \lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2}. \\
&n^{-1}\left\|\left((\widehat{C}_g-C_g)^\intercal+\chi(C_h-\widehat{C}_h)^\intercal\right)(\iota_n:\widehat{C}_h[\widehat{I}])\right\|_{\mathrm{MAX}} \\
\leq& K\left\|(\widehat{C}_g-C_g)^\intercal+\chi(C_h-\widehat{C}_h)^\intercal\right\|_{\mathrm{MAX}}\left\|(\iota_n:\widehat{C}_h[\widehat{I}])\right\|_{\mathrm{MAX}} \lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2},
\end{aligned}$$

it follows that

$$n^{-1}\left\|\left((\widehat{C}_g-C_g)^\intercal+\chi(C_h-\widehat{C}_h)^\intercal\right)\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^\intercal\right\| \lesssim_p s(n^{-1/2}+T^{-1/2})(\log(n\vee p\vee T))^{1/2},$$

which, along with (A.36) and (A.37), establish the first claim.

(ii) Next, by (5) we have

$$\widehat{C}_g=\widehat{C}_h\eta^\intercal+\widehat{C}_z.$$

And recall that $\beta_g=C_z\Sigma_z^{-1}$, so we have

$$\begin{aligned}
&n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g-\widehat{C}_g\Sigma_z^{-1})\right\| \\
\leq& n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(C_z-\widehat{C}_z)\Sigma_z^{-1}\right\|+n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\widehat{C}_h-C_h)\eta^\intercal\Sigma_z^{-1}\right\| \\
&+n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}C_h\eta^\intercal\Sigma_z^{-1}\right\|.
\end{aligned}$$

Using Assumption A.4 and $\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\right\|\leq 1$, we have

$$\begin{aligned}
&n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(C_z-\widehat{C}_z)\Sigma_z^{-1}\right\| \\
\lesssim&\left\|\widehat{C}_g\right\|_{\mathrm{MAX}}\left\|C_z-\widehat{C}_z\right\|_{\mathrm{MAX}}\|\Sigma_z^{-1}\| \lesssim_p T^{-1/2}(\log(n\vee p\vee T))^{1/2},
\end{aligned} \tag{A.38}$$

where we also use the fact that

$$\|\Sigma_z^{-1}\|\leq\lambda_{\min}^{-1}(\Sigma_z)\lesssim 1,\quad \left\|\widehat{C}_g\right\|_{\mathrm{MAX}}\leq\left\|\widehat{C}_g-C_g\right\|_{\mathrm{MAX}}+\|C_g\|_{\mathrm{MAX}}\lesssim 1.$$

Similarly, we obtain

$$\begin{aligned}
n^{-1}\left\|\widehat{C}_g^\intercal\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\widehat{C}_h-C_h)\eta^\intercal\Sigma_z^{-1}\right\| &\lesssim\left\|\widehat{C}_g\right\|_{\mathrm{MAX}}\left\|\widehat{C}_h-C_h\right\|_{\mathrm{MAX}}\|\eta\|_\infty\|\Sigma_z^{-1}\| \\
&\lesssim_p sT^{-1/2}(\log(n\vee p\vee T))^{1/2}.
\end{aligned} \tag{A.39}$$

Combining (A.38), (A.39), and (A.34) concludes the proof.

□

Lemma 4. *Under Assumptions A.1, A.2, A.3, A.4, A.5, we have*

$$\begin{aligned}
&n^{-1}\left\|\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
\lesssim_p & s^2(n^{-1/2}T^{-1/2}+T^{-1})\log(n\vee p\vee T). \qquad \text{(A.40)}
\end{aligned}$$

Proof of Lemma 4. From (A.24) and Assumption A.4, it follows that

$$\begin{aligned}
&n^{-1}\left\|\chi\widehat{C}_h^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
\leq & n^{-1/2}\left\|\chi\widehat{C}_h^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\right\|\|\beta_g\eta+\beta_h\|_\infty\left(\left\|T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)\right\|_{\text{MAX}}+\|\bar{h}\|_{\text{MAX}}\right) \\
\lesssim_p & s^2(n^{-1/2}T^{-1/2}+T^{-1})\log(n\vee p\vee T). \qquad \text{(A.41)}
\end{aligned}$$

Next, by triangle inequality, we have

$$\begin{aligned}
&n^{-1}\left\|C_e^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{Z}^\intercal\lambda_g+\left(T^{-1}\bar{H}\bar{H}^\intercal-\Sigma_h\right)(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
\leq & n^{-1}\left\|C_e^\intercal(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
& + n^{-1}\left\|C_e^\intercal \mathbb{P}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\|.
\end{aligned}$$

For the first term, by Assumption A.5 we have

$$\begin{aligned}
&n^{-1}\left\|C_e^\intercal(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
\leq & n^{-1}\|C_e^\intercal(\beta_g\eta+\beta_h)\|_\infty\left\|\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\|_{\text{MAX}} \\
\lesssim_p & sn^{-1/2}T^{-1/2}(\log(n\vee p\vee T))^{1/2}.
\end{aligned}$$

For the second term, we use Assumptions A.1, A.3, A.4, and (A.32),

$$\begin{aligned}
&n^{-1}\left\|C_e^\intercal \mathbb{P}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
\lesssim & (1+\widehat{s})\phi_{\min}^{-1}(\widehat{s}+1)\left[n^{-1}(\iota_n:\widehat{C}_h)^\intercal(\iota_n:\widehat{C}_h)\right]n^{-1}\left\|C_e^\intercal(\iota_n:\widehat{C}_h[\widehat{I}])\right\|_{\text{MAX}} \\
& \times\left\|(\iota_n:\widehat{C}_h[\widehat{I}])^\intercal\right\|_{\text{MAX}}\|\beta_g\eta+\beta_h\|_\infty\left\|T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right\|_{\text{MAX}} \\
\lesssim_p & s^2(n^{-1/2}T^{-1/2}+T^{-1})\log(n\vee p\vee T).
\end{aligned}$$

Finally, by Assumptions A.1 and A.4, we have

$$\begin{aligned}
&n^{-1}\left\|(\widehat{C}_g-C_g+(C_h-\widehat{C}_h)\chi^\intercal)^\intercal \mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}(\beta_g\eta+\beta_h)\left(T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right)\right\| \\
\lesssim & \left\|(\widehat{C}_g-C_g+(C_h-\widehat{C}_h)\chi^\intercal)^\intercal\right\|_{\text{MAX}}\|\beta_g\eta+\beta_h\|_\infty\left\|T^{-1}\bar{H}\bar{V}^\intercal\lambda-\Sigma_h(\eta^\intercal\lambda_g+\lambda_h)-\bar{h}\right\|_{\text{MAX}} \\
\lesssim_p & s^2T^{-1}\log(n\vee p\vee T).
\end{aligned}$$

The conclusion then follows from (A.28). □

Lemma 5. *Under Assumptions A.1, A.2, A.3, A.4, we have*

$$n^{-1}\left\|\widehat{C}_g^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \lesssim_p s(n^{-1/2}T^{-1/2}+T^{-1})\log(n\vee p\vee T). \tag{A.42}$$

Proof of Lemma 5. Note that by (A.24), we have

$$\begin{aligned} n^{-1}\left\|\chi\widehat{C}_h^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \leq & n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\widehat{C}_h\chi^{\intercal}\right\| n^{-1/2}\left\|\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right\| \\ \lesssim_p & s(n^{-1/2}T^{-1/2}+T^{-1})\log(n\vee p\vee T), \end{aligned}$$

where we use the following estimates as a result of Assumptions A.1 and A.4:

$$\begin{aligned} n^{-1/2}\left\|\bar{u}\right\| \lesssim \left\|\bar{u}\right\|_{\mathrm{MAX}} \lesssim_p T^{-1/2}(\log n\vee p\vee T)^{1/2}, \\ n^{-1/2}\left\|T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right\| \lesssim \left\|T^{-1}\bar{U}\bar{M}^{\intercal}\gamma_0\right\|_{\mathrm{MAX}} \lesssim_p T^{-1/2}(\log(n\vee p\vee T))^{1/2}. \end{aligned}$$

Moreover, by triangle inequality, we have

$$\begin{aligned} & n^{-1}\left\|C_e^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \\ \leq & n^{-1}\left\|C_e^{\intercal}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| + n^{-1}\left\|C_e^{\intercal}\mathbb{P}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \end{aligned}$$

For the first term, we have

$$n^{-1}\left\|C_e^{\intercal}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \leq n^{-1}\left\|C_e^{\intercal}\bar{u}\right\| + T^{-1}n^{-1}\left\|C_e^{\intercal}\bar{U}\bar{V}^{\intercal}\lambda\right\| \lesssim_p sn^{-1/2}T^{-1/2}.$$

As to the second term, using Assumption A.3 and (A.32) we have

$$\begin{aligned} & n^{-1}\left\|C_e^{\intercal}\mathbb{P}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \\ = & n^{-1}\left\|C_e^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\left((\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\right)^{-1}(\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \\ \lesssim & sn^{-1}\left\|C_e^{\intercal}(\iota_n:\widehat{C}_h[\widehat{I}])\right\|_{\mathrm{MAX}} n^{-1}\left\|(\iota_n:\widehat{C}_h[\widehat{I}])^{\intercal}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\|_{\mathrm{MAX}} \\ \lesssim_p & s(n^{-1/2}T^{-1/2}+T^{-1})\log(n\vee p\vee T), \end{aligned}$$

where we also use the following

$$\begin{aligned} n^{-1}\left\|(\iota_n:\widehat{C}_h)^{\intercal}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\|_{\mathrm{MAX}} \leq & \left(\left\|\widehat{C}_h-C_h\right\|_{\mathrm{MAX}}+\left\|(\iota_n:C_h)\right\|_{\mathrm{MAX}}\right)\left\|\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right\|_{\mathrm{MAX}} \\ \lesssim_p & T^{-1/2}(\log(n\vee p\vee T))^{1/2}. \end{aligned}$$

Finally, we note that

$$\begin{aligned} & n^{-1}\left\|\left(\widehat{C}_g-C_g+(C_h-\widehat{C}_h)\chi^{\intercal}\right)^{\intercal}\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\left(\bar{u}-T^{-1}\bar{U}\bar{V}^{\intercal}\lambda\right)\right\| \\ \lesssim & \left\|\widehat{C}_g-C_g+(C_h-\widehat{C}_h)\chi^{\intercal}\right\|_{\mathrm{MAX}}\left\|\bar{u}-T^{-1}\bar{U}\bar{M}^{\intercal}\right\|_{\mathrm{MAX}} \lesssim_p sT^{-1}\log(n\vee p\vee T). \end{aligned}$$

This concludes the proof. □

Lemma 6. *Under Assumptions A.1, A.2, A.3, A.4, A.5, we have*

$$\left\| n \left(\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_g \right)^{-1} \right\| \lesssim_p 1.$$

Proof of Lemma 6. Note that by (A.28), we have

$$\begin{aligned}
&\widehat{C}_g^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_g \\
=&C_e^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} C_e + C_e^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal + \chi \widehat{C}_h^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} C_e + \chi \widehat{C}_h^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal \\
&+ C_e^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right) + \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right)^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} C_e \\
&+ \chi \widehat{C}_h^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right) + \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right)^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal \\
&+ \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right)^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right)
\end{aligned}$$

There are 9 terms in total on the right-hand side. By (A.24), we have

$$\begin{aligned}
n^{-1} \left\| \chi \widehat{C}_h^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} C_e \right\| =& n^{-1} \left\| C_e^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal \right\| \lesssim \|C_e\|_{\mathrm{MAX}} \, n^{-1/2} \left\| \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal \right\| \\
\lesssim_p & s(n^{-1/2} + T^{-1/2})(\log(n \vee p \vee T))^{1/2}, \\
n^{-1} \left\| \chi \widehat{C}_h^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal \right\| \leq & n^{-1} \left\| \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal \right\|^2 \lesssim_p s^2 (n^{-1} + T^{-1}) \log(n \vee p \vee T).
\end{aligned}$$

Also, we have

$$\begin{aligned}
&n^{-1} \left\| C_e^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right) \right\| = n^{-1} \left\| \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right) \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} C_e \right\| \\
\lesssim & \|C_e\|_{\mathrm{MAX}} \left\| \widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right\|_{\mathrm{MAX}} \lesssim_p s T^{-1/2} (\log(n \vee p \vee T))^{1/2}, \\
&n^{-1} \left\| \chi \widehat{C}_h^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right) \right\| = n^{-1} \left\| \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right) \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_h \chi^\intercal \right\| \\
\lesssim & n^{-1/2} \left\| \chi \widehat{C}_h^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \right\| \left\| \widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right\|_{\mathrm{MAX}} \\
\lesssim_p & s^2 (n^{-1/2} T^{-1/2} + T^{-1}) \log(n \vee p \vee T), \\
&n^{-1} \left\| \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right)^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \left(\widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right) \right\| \\
\lesssim & \left\| \widehat{C}_g - C_g + (C_h - \widehat{C}_h)\chi^\intercal \right\|_{\mathrm{MAX}}^2 \lesssim_p s^2 T^{-1} \log(n \vee p \vee T).
\end{aligned}$$

Finally, by (A.32) and Assumptions A.2 and A.3, we have

$$\begin{aligned}
n^{-1} \left\| C_e^\intercal \mathbb{P}_{(\iota_n : \widehat{C}_h[\widehat{I}])} C_e \right\| =& n^{-1} \left\| C_e^\intercal (\iota_n : \widehat{C}_h[\widehat{I}]) \left((\iota_n : \widehat{C}_h[\widehat{I}])^\intercal (\iota_n : \widehat{C}_h[\widehat{I}]) \right)^{-1} (\iota_n : \widehat{C}_h[\widehat{I}])^\intercal C_e \right\| \\
\lesssim & s n^{-2} \left\| C_e^\intercal (\iota_n : \widehat{C}_h[\widehat{I}]) \right\|_{\mathrm{MAX}}^2 \lesssim_p s(n^{-1} + T^{-1}) \log(n \vee p \vee T).
\end{aligned}$$

Hence, we obtain

$$n^{-1} \widehat{C}_g^\intercal \mathbb{M}_{(\iota_n : \widehat{C}_h[\widehat{I}])} \widehat{C}_g = n^{-1} C_e^\intercal C_e + o_p(1).$$

The conclusion follows from (A.5) and Weyl inequalities. □

Lemma 7. *Under Assumptions A.1, A.2, A.3, A.4, A.5, A.6, we have*

$$\left\|(\widehat{\gamma}_0 : \widehat{\lambda}_h^\intercal) - (\gamma_0 : \lambda_h^\intercal)\right\| \lesssim_p s(n^{-1/2} + T^{-1/2})(\log(n \vee p \vee T))^{1/2}.$$

Proof. It follows from (9) that

$$(\widehat{\gamma}_0 : \widehat{\lambda}_h[\widehat{I}]^\intercal)^\intercal = \left(\left(\iota_n : \widehat{C}_h[\widehat{I}]\right)^\intercal \left(\iota_n : \widehat{C}_h[\widehat{I}]\right)\right)^{-1} \left(\iota_n : \widehat{C}_h[\widehat{I}]\right)^\intercal \left(\bar{r} - \widehat{C}_g\widehat{\lambda}_g\right),$$

which implies that

$$\left\|(\widehat{\gamma}_0 : \widehat{\lambda}_h^\intercal)^\intercal - (\gamma_0 : \lambda_h^\intercal)^\intercal\right\| \leq \left\|(\widetilde{\gamma}_0 : \widetilde{\lambda}_h^\intercal)^\intercal - (\breve{\gamma}_0 : \breve{\lambda}_h^\intercal)^\intercal\right\| + \left\|(\widetilde{\xi} : \widetilde{\chi})^\intercal\widehat{\lambda}_g - (\xi : \chi)^\intercal\lambda_g\right\|,$$

where

$$\begin{aligned}
(\widetilde{\gamma}_0 : \widetilde{\lambda}_h^\intercal)^\intercal &= \arg\min_{\gamma,\lambda}\left\{\left\|\bar{r} - \iota_n\gamma - \widehat{C}_h\lambda\right\| : \quad \lambda_j = 0, \quad j \notin \widehat{I}\right\}, \\
(\widetilde{\xi}_j : \widetilde{\chi}_{j,\cdot})^\intercal &= \arg\min_{\xi_j,\chi_{j,\cdot}}\left\{\left\|\widehat{C}_{g,\cdot,j} - \iota_n\xi_j - \widehat{C}_h\chi_{j,\cdot}^\intercal\right\| : \quad \chi_{j,k} = 0, \quad k \notin \widehat{I}\right\}, \quad j = 1, 2, \ldots, d.
\end{aligned}$$

Moreover, because

$$\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}])}\bar{r} = \iota_n\breve{\gamma}_0 + \widehat{C}_h\breve{\lambda}_h - \iota_n\widetilde{\gamma}_0 - \widehat{C}_h\widetilde{\lambda}_h + (C_h - \widehat{C}_h)\breve{\lambda}_h + C_e\lambda_g + \beta_g\bar{g} + \beta_h\bar{h} + \bar{u}$$

we obtain, using $\widehat{I}_1 \subseteq \widehat{I}$, (A.1), (A.5), (A.20) - (A.23), (A.26),

$$\begin{aligned}
&n^{-1/2}\left\|\left(\iota_n : \widehat{C}_h\right)\left(\widetilde{\gamma}_0 - \breve{\gamma}_0 : (\widetilde{\lambda}_h - \breve{\lambda}_h)^\intercal\right)^\intercal\right\| \\
\leq &n^{-1/2}\left\|\mathbb{M}_{(\iota_n:\widehat{C}_h[\widehat{I}_1])}\bar{r}\right\| + n^{-1/2}\left\|(C_h - \widehat{C}_h)\breve{\lambda}_h + C_e\lambda_g + \beta_g\bar{g} + \beta_h\bar{h} + \bar{u}\right\| \\
\leq &n^{-1/2}\left\|\left(\iota_n : \widehat{C}_h\right)\left(\widetilde{\gamma}_{\widehat{I}_1} - \breve{\gamma}_0 : (\widetilde{\lambda}_{\widehat{I}_1} - \breve{\lambda}_h)^\intercal\right)^\intercal\right\| + 2n^{-1/2}\left\|(C_h - \widehat{C}_h)\breve{\lambda}_h + C_e\lambda_g + \beta_g\bar{g} + \beta_h\bar{h} + \bar{u}\right\| \\
\lesssim_p &s(n^{-1/2} + T^{-1/2})(\log(n \vee p \vee T))^{1/2}.
\end{aligned}$$

Since we have

$$\begin{aligned}
&n^{-1/2}\left\|\left(\iota_n : \widehat{C}_h\right)\left(\widetilde{\gamma}_0 - \breve{\gamma}_0 : (\widetilde{\lambda}_h - \breve{\lambda}_h)^\intercal\right)^\intercal\right\| \\
\geq &\phi_{\min}^{1/2}(1 + \widehat{s})\left[n^{-1}(\iota_n : \widehat{C}_h)^\intercal(\iota_n : \widehat{C}_h)\right]\left\|\left(\widetilde{\gamma}_0 - \breve{\gamma}_0 : (\widetilde{\lambda}_h - \breve{\lambda}_h)^\intercal\right)\right\|,
\end{aligned}$$

it follows that

$$\left\|\left(\widetilde{\gamma}_0 - \breve{\gamma}_0 : (\widetilde{\lambda}_h - \breve{\lambda}_h)^\intercal\right)\right\| \lesssim_p s(n^{-1/2} + T^{-1/2})(\log(n \vee p \vee T))^{1/2}.$$

Similarly, we can obtain

$$\left\|\left(\widetilde{\xi} - \xi : \widetilde{\chi} - \chi\right)\right\| \lesssim_p s(n^{-1/2} + T^{-1/2})(\log(n \vee p \vee T))^{1/2}.$$

Therefore, using this, as well as Assumption A.1 and Theorem 1, we obtain

$$\begin{aligned}
\left\|(\widetilde{\xi} : \widetilde{\chi})^\intercal\widehat{\lambda}_g - (\xi : \chi)^\intercal\lambda_g\right\| &\leq \left\|(\widetilde{\xi} - \xi : \widetilde{\chi} - \chi)\right\|\left\|\widehat{\lambda}_g\right\| + \|(\xi : \chi)\|\left\|\widehat{\lambda}_g - \lambda_g\right\| \\
&\lesssim_p s(n^{-1/2} + T^{-1/2})(\log(n \vee p \vee T))^{1/2}.
\end{aligned}$$

This concludes the proof. □