Received: 20 September 2022 | Revised: 13 November 2023 | Accepted: 14 November 2023

DOI: 10.1002/jae.3031

WILEY Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE

Identifying factors via automatic debiased machine learning

Esfandiar Maasoumi[1] | **Jianqiu Wang**[2] | **Zhuo Wang**[3] | **Ke Wu**[3]

[1]Department of Economics, Emory University, Atlanta, Georgia, USA

[2]School of Finance, Capital University of Economics and Business, Beijing, China

[3]School of Finance, Renmin University of China, Beijing, China

Correspondence
Jianqiu Wang, School of Finance, Capital University of Economics and Business, Beijing 100070, China.
Email: jianqiu.wang@cueb.edu.cn

Ke Wu, School of Finance, Renmin University of China, Beijing 100872, China.
Email: ke.wu@ruc.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 72173127, 72192804, 72303162; Beijing Municipal Social Science Foundation, Grant/Award Number: 23JJC029; Public Computing Cloud at Renmin University of China; FinTech Lab at Capital University of Economics and Business

Summary

Identifying risk factors that have significant explanatory power for the cross-sectional asset returns is fundamental in asset pricing. We adopt a novel automatic debiased machine learning (ADML) method proposed by Chernozhukov, Newey, and Singh (2022) to robustly estimate partial pricing effect of a certain factor controlling for a large number of confounding factors under a nonlinear stochastic discount factor (SDF) assumption. The ADML resolves biased estimation, non-robustness, and overfitting issues that are common to traditional machine learning approaches. We find that the most significant factors selected by the ADML outperform the Fama–French sparse factors and factors identified via the double-selection LASSO method under a linear factor model assumption. Out of a high-dimensional zoo of US stock market factors commonly tested in the finance literature, we identify approximately 30 to 50 factors having significant but declining pricing power in explaining the cross-section of stock returns. Our findings are robust to hyperparameter settings and choices of test assets and machine learning methods.

KEYWORDS

automatic debiased machine learning, factor zoo, nonlinear SDF

1 | INTRODUCTION

A fundamental challenge in asset pricing lies in the reliable identification of factors that explain cross-sectional returns (Cochrane, 2011; Feng et al., 2020; Green et al., 2017; Kozak et al., 2020). Traditional regression-based methods, such as Fama and MacBeth (1973) regressions, come with at least two potential drawbacks. First, given the hundreds of potential factors documented by prior research, linear regressions may result in very noisy estimates in high-dimensional settings. Second, traditional factor models assume asset payoffs are linear functions of pricing factors (Bansal & Viswanathan, 1993). However, Jagannathan and Korajczyk (1986) demonstrate that a passive portfolio return of option-like stocks (highly levered firms) may exhibit a concave or a convex relation with the market return. Further evidence includes the option-like behavior of momentum factors (Daniel & Moskowitz, 2016) and nonlinear hedge fund payoffs (Agarwal & Naik, 2004; Fung & Hsieh, 2001).

In the finance literature, machine learning tools have primarily been applied to model conditional expectations of returns. Compared with traditional statistical models, they perform significantly better in terms of out-of-sample prediction (Gu et al., 2020). While machine learning with regularization yields sparsity models in high-dimensional settings, these methods often suffer from biased and non-robust estimation because their main focus lies in optimal prediction

(Chernozhukov et al., 2018, 2022). Chernozhukov et al. (2018) proposes double machine learning (DML) approaches to correct biases stemming from regularization and overfitting, using Neyman orthogonalization (NO) and sample splitting, respectively. We adopt an automatic debiased machine learning (ADML) approach proposed by Chernozhukov et al. (2022) to robustly identify a factor's average partial pricing effect under a general nonlinear stochastic discount factor (SDF) structure. The ADML method shares a common feature with the DML, in "withholding" anyone or a set of factors from biased "selection" of the LASSO or other regularization methods. In addition, the model-free debiasing and inference is carried out automatically in one single step and hence is robust to potential model or moment misspecifications. Finally, this approach employs rigorous and standard asymptotic inference tools, making it straightforward to provide confidence bands.

Following Feng et al. (2020), we consider useful factors (those with non-zero SDF loadings), redundant factors (those with zero SDF loadings but are correlated with the useful factors), and useless factors (those with zero SDF loadings and zero correlation with the useful factors). Additionally, we allow for nonlinearity and interactions among factors in the return generating process. With simulated data, we demonstrate that standard machine learners, such as the LASSO, yield significant biases, and estimates that deviate from normality, resulting in highly unreliable and misleading confidence bands. By contrast, the ADML produces accurate estimates with asymptotically normal distributions, benefiting from the double-robust property. Furthermore, simulation shows that the ADML can consistently identify useful factors from the factor zoo mixed with redundant and useless factors with the probability close to 1 under parameter settings mimicking the real-world data.

By utilizing all the other factors as controls, we identify each factor's incremental pricing effect. To make the empirical results comparable, we use the same data and factors as in Feng et al. (2020), which includes 150 risk factors from 1972 to 2017. We identify more than 30 factors with significant explanatory power for the cross-sectional returns. Specifically, we find that net debt finance, operating leverage, 36-month momentum, composite equity issuance, percentage change sales-to-inventory, and industry-adjusted change in asset turnover are among the most significant factors. On the other hand, factors such as current ratio, order backlog, dividend initiation, R&D-to-sales, growth in advertising expense, and book-asset liquidity show minimal significance.

Further, we investigate whether these identified factors perform well in the classical spanning test of pricing factors (see, e.g., Hou et al., 2018). We construct three-, five-, and six-factor models based on the most significant factors identified by the ADML, denoting them as ADML3, ADML5, and ADML6, respectively. The spanning test indicates that ADML3, ADML5, and ADML6 outperform the Fama–French three-, five-, and six-factor models in the number of factors explained and the average magnitude of the absolute factor alphas. Especially for the factors in the intangible category, the ADML3, ADML5, and ADML6 models explain all the other factors in this category at the 1% level, while the FF3, FF5, and FF6 models explain significantly less. These results confirm that the factors identified by the ADML are considerably more useful in explaining the other risk factors.

We also report time-series variation in the partial pricing power of each factor and in the number of significant factors for each category. We find well-known factors, such as SMB, HML, RMW, and CMA in Fama and French (2015) five-factor model and IA and ROE factors in Hou et al. (2015) four-factor model, show persistent pricing power. More importantly, overall factor significance declines over time, consistent with the findings of Green et al. (2017). The ADML approach identifies around 30 to 50 significant factors, much more than that selected by linear models, such as the OLS (less than 10) and the double-selection LASSO method proposed by Feng et al. (2020) (around 15). It indicates that factors with important nonlinear pricing information are more prevalent than those previously identified by the linear factor models.

Finally, to validate the nonlinear SDF framework and the ADML estimation work beyond the US market, we conduct an international robustness analysis using pricing factors in the Chinese stock market, the second-largest stock market in the world. More sentiment-based factors are identified through the ADML method, consistent with prior research suggesting that the Chinese stock market is predominantly influenced by individual investors. In addition, we conduct more robustness checks by varying the values of hyperparameters, such as the number of units in the hidden layer in the neural network and the number of sample splits, using alternative nonlinear machine learning methods, such as the random forest, and using alternative test assets.

Our finding of a large number of significant factors is potentially compatible with prior studies that highlight only a few predominant factors. This phenomenon is typically observed during market booming period (like the post-2008 era), where few factors can efficiently encapsulate others and deliver equivalently good return predictions. While many correlated factors may appear redundant for the return prediction purpose, they are nonetheless crucial for understanding pricing mechanisms and informing investment decisions.

1.1 | Relation to the literature

This paper contributes to a strand of literature that estimates high-dimensional factor models. For instance, Lewellen (2015) employs the Fama–Macbeth regression to jointly investigate the capacity of multiple characteristics to predict expected stock returns and concludes that using multiple predictive variables simultaneously vastly outperforms the use of a single variable. Green et al. (2017) also use Fama–Macbeth regressions to simultaneously test 94 characteristic variables in the US stock market, finding that only eight to 12 variables are significant predictors for stock returns. In more recent studies, Kozak et al. (2020) argue that a sparse number of firm characteristics are insufficient to fully explain the cross-sectional variations in stock returns. However, several principal components of characteristic-based factor returns can reasonably estimate the SDF. Using the adaptive group LASSO method, Freyberger et al. (2020) select 13 out of 62 cross-sectional return predictors and stress the importance of nonlinear factor relationship. Gu et al. (2020) apply machine learning methods to predict cross-sectional stock returns and emphasize the substantial improvements with nonlinear methods, such as the neural networks. While most of these studies focus on predicting cross-sectional returns, a more fundamental question is which factors affect the investor's marginal utility and hence constitute the SDF. As pointed out by Feng et al. (2020), a direct estimation of the SDF via machine learning methods with regularization, such as the LASSO, leads to biased estimates. To address this issue, they employ a two-step double-selection LASSO method.

Our work is mostly related to Feng et al. (2020). By relaxing their assumption of a linear SDF structure, our approach is able to take into account potential interaction and nonlinear relations among factors. Specifically, we apply the ADML method proposed by Chernozhukov et al. (2022) to empirically estimate the average coefficient of a given factor after controlling for a large number of potential factors in a general nonlinear SDF framework.

The remainder of this paper is organized as follows. Section 2 introduces the nonlinear factor model and the econometric methodology. Section 3 outlines the simulation process and its resultant findings. Section 4 presents several empirical applications and robustness tests of these findings. Lastly, Section 5 concludes.

2 | METHODOLOGY

This section introduces the model setup and discusses the potential regularization and overfitting issues when utilizing complex machine learning algorithms. We then introduce the ADML method, discuss how to estimate the SDF loadings, and present the statistical tools for inference.

2.1 | Model setup

As discussed by Cochrane (2009), in classical factor pricing models, the ability of a given factor to explain the asset returns is reflected by its loading in the SDF. Bansal and Viswanathan (1993) considers a general form of nonlinear SDF or a pricing kernel as follows:

$$m_t := H(f_t), \tag{1}$$

where f_t is a set of factors and $H(\cdot)$ is a general nonlinear function. The intuition is that nonlinear payoffs do not lie in the linear span of the factor payoffs in an incomplete market, and thus, linear representation is not feasible for all the payoffs. By contrast, the general pricing kernel $H(f_t)$ does not suffer from these shortcomings and is able to price any security.

The nonlinear pricing kernel exists under several assumptions. First, there exists an M-dimensional process f_t such that for all nonlinear functions $L(\cdot)$, $\mathbf{E}[L(f_\tau)|I_t, f_t] = \mathbf{E}[L(f_\tau)|f_t]$ for $\tau > t$. This assumption implies that f_t is a sufficient statistic for predicting any nonlinear function of f_τ at time t. Second, there exists an agent j whose intertemporal marginal rate of substitution is a function of f_t, that is, $H(f_{t+1})$. This assumption implies that there exists an individual whose marginal rate of substitution is a function of a few factors. Besides, the assumption of incomplete markets is crucial for the necessity of a nonlinear SDF. In a complete market, any nonlinear payoff can be represented by a linear combination of a set of other payoffs. However, in incomplete markets, there does not exist enough assets to hedge risks in all states of the market, and the SDF is not unique (Cochrane, 2009). A nonlinear SDF is more flexible to reflect the lack of opportunities to perfectly hedge risks and may better price payoffs in incomplete markets. In reality, due to various market frictions, such as transaction costs, short-sale constraints, and information asymmetry, the stock market can hardly be a complete market, which motivates us to consider a nonlinear SDF.

More supporting evidence is documented in the literature. For instance, abundant evidence for nonlinearities in expected returns has been provided by prior research, such as Bansal and Viswanathan (1993), Chapman (1997), Harvey and Siddique (2000), Dittmar (2002), and Adrian et al. (2019). Bansal and Viswanathan (1993) and Chapman (1997) show that the pricing kernel is not a linear function of the market return. Under a representative agent model with a utility function that is third-order differentiable, Harvey and Siddique (2000) and Dittmar (2002) argue that the square and cubic terms of the market return enter the SDF. Their empirical results suggest that the nonlinear SDFs outperform the linear ones in explaining the cross-sectional variation of expected returns. Our methodology allows for general nonlinearity in the pricing kernel and thereby provides a more robust explanation of the behavior of asset returns in the real-world markets.

With the nonlinear pricing kernel, the expected return can be written as

$$E(r_t) - \tau_n r_f = cov(H(f_t), r_t) = g(cov(f_t, r_t)), \tag{2}$$

where E (r_t) is a $n \times 1$ vector of expected returns, r_f is the risk free rate, and τ_n is a $n \times 1$ vector of ones.

Equation (2) can be derived by Jacobian transforming the joint density of $H(f_t)$ and r_t to the joint density of f_t and r_t. Specifically, assume $(f_{1t}, f_{2t}, \cdots, f_{nt}, r_t)$ be a random vector with pdf $f(f_{1t}, f_{2t}, \cdots, f_{nt}, r_t)$. Define $y_i = H_i(f_{1t}, \cdots, f_{nt})$ for $i = 1, \cdots, n-1$ and $y_n = H(f_{1t}, \cdots, f_{nt})$. Consider the new random vector $(y_1, \cdots, y_n, r_t)$. Assume the following inverse exists, denoted by $f_{it} = h_i(y_1, \cdots, y_n)$ for $i = 1, \cdots, n$. Then, for the joint density of $(y_1, \cdots, y_n, r_t)$, we have

$$p(y_1, \cdots, y_n, r_t) = f(f_{1t}, \cdots, f_{nt}, r_t)\,|J|, \tag{3}$$

where $|J|$ is the Jacobian of the transformation, defined as

$$|J| = \begin{vmatrix} \frac{\partial h_1}{\partial y_1} & \cdots & \frac{\partial h_1}{\partial y_n} & 0 \\ \vdots & & \vdots & \vdots \\ \frac{\partial h_n}{\partial y_1} & \cdots & \frac{\partial h_n}{\partial y_n} & 0 \\ 0 & \cdots & 0 & 1 \end{vmatrix}. \tag{4}$$

Assume the joint distribution of f_t and r_t follows an elliptical distribution,[1] then we can define an one-to-one mapping $g(\cdot)$ from the covariance space of f_t and r_t to the covariance space of $H(f_t)$ and r_t. Therefore, the expected return can be expressed by Equation (2).

Equation (2) is nonparametric, and the functional form $g(\cdot)$ is determined by machine learning methods. Equation (2) separates the risks contained in these factors, and their relationships with the expected returns are reflected in $g(\cdot)$, which includes the higher order moments of the risk factors and the interactions of these factors. Following Feng et al. (2020), f_t contains a large set of factors including useful, redundant, and useless factors. The problem is how to identify useful factors in the high-dimensional setting.

Based on Equation (2), we use the following moment function to identify the average partial effect of each factor λ_k:

$$\lambda_k = E[m(W, \gamma_0)] = E\left[\frac{\partial \gamma_0(X)}{\partial C_k}\right], \tag{5}$$

where W stands for a data observation, $X = C_f$ is a covariance vector for all factors, $\gamma_0(X) = E[r_t|X]$ is the conditional expected return (conditional on covariances C_f) in Equation (2), and $C_k = cov(f_{kt}, r_t)$ is the covariance term for factor k. The identification equation (5) can be considered as a generalization of the linear SDF loading[2] and gives the average partial pricing effect of each factor after controlling for all the other factors. The nonlinear SDF allows the partial pricing effect to be varying with the covariance term C_k taking on different values.

[1] An elliptical distribution is a generalization of the normal distribution, such that the iso-density contours of the probability density are ellipsoids, centered around the mean. Typical examples of elliptical distributions include the multivariate normal distribution, multivariate Student's t-distribution, and others.

[2] See eq. (9) in Feng et al. (2020) for comparison.

2.2 | ADML

Here, we assume that $E[m(W,\gamma)]$ is a mean-square continuous function of the function γ, where γ is a possible conditional expectation function of returns. By Riesz representation theorem, there exists a function $\alpha_0(X)$ with $E\left[\alpha_0(X)^2\right] < \infty$ and

$$E[m(W,\gamma)] = E\left[\alpha_0(X)\gamma(X)\right] \quad \text{for all} \quad \gamma \quad \text{such that} \quad E\left[\gamma(X)^2\right] < \infty, \tag{6}$$

where α_0 is the Riesz representor. If we directly apply machine learning methods to identify the moment condition $\mathrm{E}\left[m\left(W,\gamma_0\right) - \lambda_k\right] = 0$, the estimator could be biased due to the regularization condition in the machine learning methods as in Chernozhukov et al. (2018, 2022). The debiased estimator can be obtained by adding the influence function $\alpha[\gamma_0(X) - \gamma(X)]$ as in Chernozhukov et al. (2022), and the bias-corrected moment function is

$$\psi\left(w, \lambda_k, \gamma, \alpha\right) = m(w,\gamma) - \lambda_k + \alpha[\gamma_0(X) - \gamma(X)]. \tag{7}$$

Given the data observation W and true SDF loading λ_k, the expected moment function is

$$\begin{aligned} E\left[\psi\left(W, \lambda_k, \gamma, \alpha\right)\right] &= E[m(W,\gamma)] - \lambda_k + E[\alpha(X)\{\gamma_0(X) - \gamma(X)\}] \\ &= E\left[\alpha_0(X)\left\{\gamma(X) - \gamma_0(X)\right\}\right] + E\left[\alpha(X)\left\{\gamma_0(X) - \gamma(X)\right\}\right] \\ &= -E\left[\left\{\alpha(X) - \alpha_0(X)\right\}\left\{\gamma(X) - \gamma_0(X)\right\}\right]. \end{aligned} \tag{8}$$

Without the influence function, the expected moment function is $E\left[\alpha_0(X)\left\{\gamma(X) - \gamma_0(X)\right\}\right]$, which cannot converge to zero if the Machine Learning estimator of γ is inconsistent. By contrast, with the influence function, the moment function $\psi\left(w, \lambda, \gamma, \alpha\right)$ is double robust with zero expectation at $\lambda = \lambda_k$ when either $\gamma = \gamma_0$ or $\alpha = \alpha_0$. Here, the role of α is to adjust the bias of the Machine Learning estimator to obtain a consistent estimate.

We need to use proper machine learning methods to estimate α_0 and γ_0. For α_0, we apply the LASSO learner $\hat{\alpha}$ as in Chernozhukov et al. (2022). We assume $\hat{\alpha}$ takes the linear form:

$$\hat{\alpha}(x) = b(x)'\hat{\rho}, \tag{9}$$

where $b(x) = \left(1, C_f\right)$ is a dictionary function and $\hat{\rho}$ is a vector of estimated coefficients. Replacing the $\gamma(X)$ function in Equation (6) with $b(X)$, we can define an M function as follows:

$$M \equiv E[m(W,b)] = E\left[\alpha_0(X)b(X)\right], \tag{10}$$

where $m(w,b) = (m(w,b_1), \ldots, m(w,b_p))'$.

Then, the LASSO estimator can be

$$\widehat{\rho}_L = \arg\min_{\rho}\left\{-2\widehat{M}\rho + \rho'\widehat{G}\rho + 2r_L|\rho|_1\right\}, \ |\rho|_1 = \sum_{j=1}^{p}\left|\rho_j\right|, \tag{11}$$

where $\widehat{M} = \frac{1}{n}\sum_{i=1}^{n} m\left(W_i, b\right)$ is an unbias estimator of $E\left[\alpha_0(X)b(X)\right]$, $\widehat{G} = \frac{1}{n}\sum_{i=1}^{n} b\left(X_i\right)b(X_i)'$ is an unbias estimator of $E\left[b(X)b(X)'\right]$, and $2r_L|\rho|_1$ is the penalty. The objective function (11) can be seen as the LASSO objective with $\frac{1}{n}\sum_{i=1}^{n}\alpha_0(X_i)b(X_i)$ replaced by $\widehat{M}$ and $\frac{1}{n}\sum_{i=1}^{n}\alpha_0(X_i)^2$ dropped.

For γ_0, a variety of machine learning methods can be used, such as neural networks and random forests. After constructing $\hat{\alpha}$ and $\hat{\gamma}$, plug in these into Equation (7) to solve for λ_k. To avoid finite sample bias, we also use sample-splitting methods (Chernozhukov et al., 2017). Specifically, split the data observation into L parts. Let $I_\ell, (\ell = 1, \ldots, L)$ be a partition of the observation set of test assets $\{1, \ldots, n\}$ with equal size. Using observations that are not in I_ℓ to construct $\hat{\alpha}_\ell$ and $\hat{\gamma}_\ell$ estimators. And then set the sample average of $\psi\left(W_i, \lambda_k, \hat{\gamma}_\ell, \hat{\alpha}_\ell\right)$ to zero to solve for λ_k. Furthermore, the explicit form of $\hat{\lambda}_k$ is

$$\hat{\lambda}_k = \frac{1}{n}\sum_{l=1}^{L}\sum_{i\in I_l}\left\{m\left(W_i, \hat{\gamma}_l\right) + \hat{\alpha}_l\left(X_i\right)\left[\bar{r}_i - \hat{\gamma}_l\left(X_i\right)\right]\right\}. \tag{12}$$

For asymptotic inference, the variance estimator for each observation can be

$$\widehat{\psi}_{il} = m\left(W_i, \hat{\gamma}_l\right) - \hat{\lambda}_k + \hat{\alpha}_l\left(X_i\right)\left[\bar{r}_i - \hat{\gamma}_l\left(X_i\right)\right], i \in I_l, (l = 1, 2, \cdots, L), \tag{13}$$

where $\bar{r}_i$ is full sample time-series average of the ith test asset.

The asymptotic variance of $\sqrt{n}\left(\hat{\lambda}_k - \lambda_k\right)$ is given by

$$\widehat{V} = \frac{1}{n}\sum_{l=1}^{L}\sum_{i \in I_l}\widehat{\psi}_{il}^2. \tag{14}$$

As Theorem 3 shown in Chernozhukov et al. (2022), under proper assumptions,

$$\sqrt{n}\left(\hat{\lambda}_k - \lambda_k\right) \xrightarrow{d} N(0, V), \hat{V} \xrightarrow{p} V = E\left[\psi_0(W)^2\right]. \tag{15}$$

Asymptotic Gaussianity and consistency ensure we can identify factors given many existing benchmark factors. Note that to obtain a consistent estimator of λ_k, r_L needs to converge to zero more slowly than $\sqrt{ln(p)/n}$. We adopt the iterative tuning procedure in Chernozhukov et al. (2022) for choosing the regularization parameter r_L.

3 | SIMULATION

3.1 | Simulation process

We examine the performance of the ADML estimator in a simulation experiment. The simulation process parallels that in Feng et al. (2020) but allows for more general model settings. As in Feng et al. (2020), factors to be tested (denoted as g_t) include (1) useful factors (non-zero SDF loading), (2) redundant factors (zero SDF loading but correlated with useful factors), and (3) useless factors (zero SDF loading and not correlated with useful factors). Control factors h_t include four useful factors and $p-4$ redundant or useless factors. The test assets expected returns $E(r_t)$ and returns r_t are generated by these factors. Online Appendix A shows the detailed simulation process. The total number of Monte Carlo runs is 2000.

We use essential factors to calibrate the simulation parameters as in Feng et al. (2020). Specifically, the Fama–French five factors are used to calibrate the five useful factors. χ is the OLS estimator coefficient from the regression C_g on C_h. λ is the OLS estimator coefficient from the regression $\bar{r}$ on C_g and C_h. We apply the ADML to the simulated data and conduct inference with various parameters settings with $p = 75$, 100, and 125, the number of test assets $n = 150$, 300, and 450 and the length of time series $T = 240$, 360, and 480.

3.2 | Simulation results

Figure 1 reports histograms of the standardized ADML and plug-in estimates using estimated standard errors. The left column reports the histograms of the ADML estimates along with the standard normal density plotted in solid line. The ADML estimator uses LASSO to estimate Riesz representor α_0 and the neural network with three units in the hidden layer to estimate γ_0. The LASSO tuning parameters are chosen by the iterative tuning procedure of Chernozhukov et al. (2022) with a maximum number of iterations at 10. The right column reports the estimates without the influence function (plug-in). The top row reports the standardized estimates for $n = 150$, the middle row for $n = 300$, and the last row for $n = 450$. In the simulation, $T = 480$ and $p = 100$.

Figure 1 shows that inference without the influence function (plug-in estimator) can lead to large biases as documented by Chernozhukov et al. (2022). By contrast, with the influence function, the ADML estimator produces an unbiased and asymptotically normal distribution, which benefits from the double-robust property. In Figure B1 in the online Appendix, we use neural networks with one, two, four, and five units in the hidden layer to estimate γ_0 and find similar results. Besides, we test the asymptotic performance of the ADML estimator using various sets of parameters in Figure B2 in the online Appendix. The results show that these different parameter sets will have little impact on the consistency and effectiveness of statistical inference.

Furthermore, we test whether the ADML method can consistently uncover those useful factors via Monte Carlo simulation. Table B1 in the online Appendix shows the average selection ratios for the useful, redundant, and useless fac-

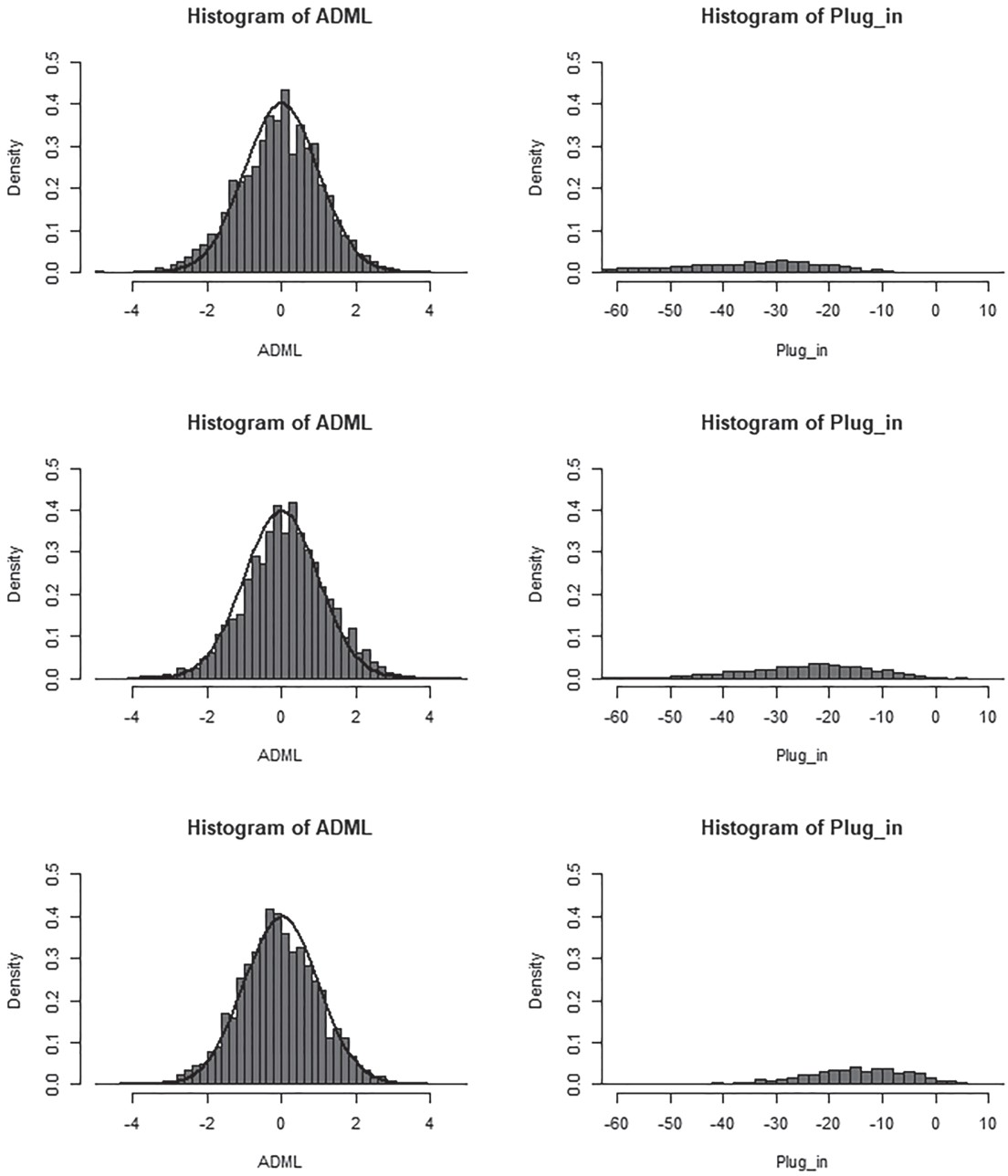

FIGURE 1 Histograms of the standardized estimates in simulations. *Note*: This figure reports histograms of the standardized automatic debiased machine learning (ADML) and plug-in estimates using estimated standard errors. The left column reports the ADML estimates compared with the standard normal density in solid lines. The ADML estimator uses LASSO to estimate Riesz representor α_0 and the neural network with three units in the hidden layer to estimate γ_0. The LASSO tuning parameters are chosen by the iterative tuning procedure of Chernozhukov et al. (2022) with a maximum number of iterations at 10. The top row reports the standardized estimates for $n = 150$, the middle row for $n = 300$, and the last row for $n = 450$. In the simulation, $T = 480$ and $p = 100$. The true average partial effect of C_{g1} is 37.34.

tors in 200 simulation runs[3] under a variety of parameter settings, that is, the number of factors $p = 75$, 100, and 125, the number of test assets $n = 150$, 300, and 450, and the length of time series $T = 240$, 360, and 480. Factors with t-statistic greater than 1.96 are counted as being selected. As shown in Table B1 in the online Appendix, the average selection probabilities for the useful factors are close to 1 under all parameter settings. By contrast, the average probabilities for the

[3]The simulation procedures under various parameter settings are computationally burdensome. To save commutating time, in this simulation study, we limited the number of simulations to be 200. We can see that even with 200 simulated samples, the selection probabilities for the useful factors are already approaching 1. Increasing the number of simulations will yield better results.

ADML method to select the redundant and useless factors are close to 0. These simulation results indicate that the ADML method is highly effective in identifying the useful factors.

In sum, the simulated distributions of estimates and factor selection results with various parameter sets confirm the strong performance of the ADML for identifying the significant incremental contribution of factors.

4 | EMPIRICAL ANALYSIS

4.1 | Data

The factor library and test portfolios come from Feng et al. (2020). The factor zoo contains 150 risk factors monthly from June 1976 to December 2017. The factors include several published factors, such as liquidity factor from Pástor and Stambaugh (2003), q-factors from Hou et al. (2015), Fama–French five factors from Fama and French (2015), betting against beta factor of Frazzini and Pedersen (2014), the intermediary asset pricing factors of He et al. (2017), and the quality minus junk factor of Asness et al. (2019). In addition, factor proxies of 135 long–short value-weighted portfolios constructed by firm characteristics surveyed in Hou et al. (2020) and Green et al. (2017) are also included. Table B2 in the online appendix lists the name and definition of all the factors. As in Hou et al. (2020), we divide these factors into six categories: momentum, value-versus-growth, investment, profitability, intangibles, and trading frictions.

The empirical analysis rests on a large set of characteristic-sorted portfolios in line with the suggestion of Lewellen et al. (2010). The baseline results use a set of 750 portfolios which include 3×2 portfolios sorted by size and book-to-market ratio, 3×2 portfolios sorted by size and operating profitability, 3×2 portfolios sorted by size and investment, 3×2 portfolios sorted by size and short-term reversal on prior (1–1) return, 3 × 2 portfolios sorted by size and momentum on prior (2–12) return, and 3×2 portfolios sorted by size and long-term reversal on prior (13–60) return and 714 additional 3×2 portfolios cover additional characteristics. To avoid an insufficient number of stocks in a given portfolio, we retain portfolios with more than 10 stocks. In addition, the 202 portfolios employed by Giglio and Xiu (2021) and 1825 5 × 5 sorted portfolios are utilized for the robustness checks.

4.2 | Identify factors sequentially

The first problem is whether the ADML estimator can identify the marginal contribution of a factor g_t after controlling for some other factors h_t. Assuming that all the factors discovered up to now contain pricing information, we can test each factor's incremental pricing effect by including all the other factors as control variables. The useful factors must significantly explain the cross-sectional stock return. In this sense, we test the marginal contribution of the 150 factors sequentially relative to all the other factors.[4]

Table 1 reports the marginal contribution of 150 factors from July 1976 to December 2017, sequentially relative to the other 149 factors. The test assets include a set of 750 3 × 2 bivariate-sorted portfolios. λ_k represents the average partial effect of each factor using the ADML estimator. ADML estimator uses the LASSO to estimate Riesz representor α_0 and neural networks with one to five units in the hidden layer to estimate γ_0. The LASSO tuning parameters are chosen by the iterative tuning procedure of Chernozhukov et al. (2022) with the maximum number of iterations at 10. Bold font indicates the Bonferroni corrected significant factor with the absolute value of t-statistic larger than 3.59.[5]

We take the results of the neural network with three units in the hidden layer (NN3) as the benchmark and uncover some interesting findings. First, in each category, more than one factor has a significant marginal contribution relative to all factors of the same category. Three of the 12 momentum factors, seven of the 24 value-versus-growth factors, 11 of the 39 investment factors, four of the 16 profitability factors, 10 of the 38 intangible related factors, and four of the 21 trading friction factors have the absolute value of t-statistics exceeding 3.59, which confirm the incremental effect of these factors. Second, many factors have no significant incremental contribution relative to other factors. These factors cannot be the true factors contained in SDF because every useful factor must have a marginal contribution no matter what controls we

[4]Our approach differs from that of Feng et al. (2020), who focus on testing the pricing power of new factors compared with previously proposed ones. Instead, our emphasis lies in identifying all factors that can explain cross-sectional returns, and therefore, we utilize all the other factors as controls.

[5]As in Harvey et al. (2016), we employ the Bonferroni correction and set the significance cut-off at α/n (n is the number of tested factors) to avoid the multiple testing problem.

TABLE 1 Identify factors sequentially.

	NN1		NN2		NN3		NN4		NN5	
Variable	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**
sue	2.95	0.78	2.39	0.66	2.78	0.78	1.24	0.39	2.65	0.75
STR	−0.01	−0.70	**0.37**	**22.24**	−0.05	−2.56	0.00	0.13	**0.17**	**9.33**
mom6m	2.36	2.19	2.50	2.33	2.17	2.10	2.67	2.85	2.68	2.78
mom36m	**−0.94**	**−50.38**	**−0.91**	**−52.40**	**−0.70**	**−46.19**	**−0.59**	**−38.69**	**−0.52**	**−25.13**
UMD	**6.10**	**4.49**	**6.20**	**4.87**	**5.73**	**4.63**	**5.66**	**4.62**	**5.74**	**4.50**
nincr	0.53	0.44	0.62	0.51	0.68	0.57	−0.69	−0.59	0.29	0.25
indmom	1.04	0.63	0.58	0.38	0.79	0.52	1.00	0.65	0.91	0.60
chmom	2.19	1.76	2.67	2.49	2.94	2.94	2.87	2.43	3.12	2.72
rs	2.33	0.79	−0.28	−0.10	0.92	0.33	0.06	0.02	0.10	0.04
aeavol	**−5.80**	**−7.94**	**−5.37**	**−7.68**	**−5.75**	**−8.21**	**−5.26**	**−7.06**	**−5.12**	**−6.82**
ear	−3.15	−2.97	−3.07	−3.04	−3.04	−3.00	−2.54	−2.52	−2.95	−3.01
rsup	7.05	2.57	6.84	2.51	7.20	2.54	6.81	2.52	7.74	2.87
MktRf	**0.50**	**4.53**	**0.60**	**5.54**	**0.59**	**5.70**	**0.60**	**5.76**	**0.46**	**4.23**
ep	**1.52**	**7.41**	**1.37**	**6.70**	**1.37**	**7.09**	**1.34**	**6.77**	**1.32**	**6.87**
dy	−0.15	−1.67	−0.19	−1.98	−0.19	−1.96	−0.22	−2.44	−0.22	−2.43
LTR	3.17	2.37	3.75	2.80	3.41	2.20	3.77	2.29	3.72	2.10
lev	**2.41**	**4.70**	**2.11**	**4.38**	**1.92**	**3.90**	**1.82**	**3.79**	**1.90**	**3.99**
depr	0.14	0.62	0.15	0.65	0.19	0.86	0.19	0.87	0.20	0.99
pchdepr	−2.50	−1.07	−2.59	−1.12	−2.29	−0.95	−2.23	−0.93	−2.29	−0.96
HML	1.85	2.43	1.61	2.20	1.21	1.59	1.49	2.11	1.43	1.94
cp	**1.42**	**5.57**	**1.42**	**5.51**	**1.26**	**4.92**	**1.27**	**4.72**	**1.36**	**5.23**
divi	0.16	0.32	0.17	0.35	−0.05	−0.11	−0.20	−0.42	−0.08	−0.16
divo	−0.06	−0.14	−0.13	−0.34	−0.44	−1.16	−0.41	−1.12	−0.35	−0.97
sp	**9.07**	**5.16**	**9.56**	**5.53**	**8.42**	**5.20**	**9.55**	**5.83**	**8.92**	**5.53**
bm_ia	3.13	3.51	**3.67**	**4.12**	**3.39**	**4.10**	**3.78**	**4.64**	**3.62**	**4.16**
cfp_ia	**9.58**	**5.18**	**8.56**	**5.04**	**8.65**	**5.42**	**8.65**	**5.94**	**9.24**	**5.78**
cfp	0.49	1.09	0.34	0.78	0.45	1.15	0.42	1.01	0.35	0.87
tb	1.63	2.25	1.18	1.68	1.05	1.38	1.03	1.42	0.79	1.13
op	0.14	0.79	0.10	0.63	0.11	0.65	0.14	0.86	0.18	1.14
nop	0.25	1.61	0.23	1.50	0.23	1.46	0.30	2.04	0.26	1.70
ndp	−1.16	−2.89	−1.02	−2.88	−1.04	−2.71	−0.95	−2.76	−0.91	−2.68
ebp	1.60	3.43	1.23	2.70	1.25	2.85	0.94	2.24	0.94	2.15
cashpr	**2.63**	**3.82**	2.12	3.25	1.85	2.62	1.72	2.55	1.76	2.54
em	4.16	2.16	3.13	1.70	4.71	2.49	5.16	2.70	4.53	2.60
HML_Devil	2.26	1.74	1.86	1.48	1.91	1.51	2.03	1.63	1.88	1.49
QMJ	**1.12**	**3.70**	0.93	3.18	0.85	2.81	0.62	2.11	0.71	2.31
acc	5.52	1.99	4.94	1.82	4.96	1.72	6.38	2.40	6.51	2.43
chinv	0.35	0.18	1.11	0.57	2.04	1.09	0.68	0.35	0.72	0.38
chtx	−4.66	−2.91	**−6.12**	**−3.84**	**−5.66**	**−3.84**	**−5.61**	**−3.83**	**−5.62**	**−3.90**
grltnoa	**5.32**	**6.04**	**3.75**	**4.30**	**3.39**	**3.95**	3.16	3.50	**3.14**	**3.61**
grltnoa_hxz	8.47	3.58	7.96	3.41	7.79	3.44	7.35	3.23	6.89	3.06
rd	−0.23	−0.24	0.72	0.75	0.89	0.95	0.69	0.75	0.62	0.68
cinvest	2.40	1.09	1.87	0.88	2.36	1.21	2.98	1.54	3.13	1.57
cinvest_a	4.33	1.64	4.91	1.79	3.47	1.32	3.42	1.33	3.25	1.23
noa	**13.78**	**5.12**	**13.20**	**5.02**	**13.04**	**5.33**	**13.57**	**5.47**	**13.69**	**5.48**
dnoa	**5.85**	**4.12**	**7.64**	**5.42**	**7.90**	**6.00**	**8.08**	**5.93**	**8.61**	**6.43**
egr	−0.02	−0.02	1.02	0.97	0.57	0.55	0.54	0.52	1.02	0.98
lgr	0.97	1.04	1.13	1.28	1.10	1.21	0.72	0.83	0.52	0.59
dcoa	0.10	0.13	1.10	1.42	0.73	1.02	0.92	1.29	1.07	1.55
dcol	−0.26	−0.40	0.44	0.72	0.34	0.55	0.09	0.15	0.18	0.30
dwc	**−8.53**	**−4.47**	**−7.22**	**−3.92**	**−9.06**	**−5.30**	**−8.79**	**−4.75**	**−8.12**	**−4.49**
dnca	2.77	1.15	2.24	0.94	2.37	0.97	2.03	0.85	2.61	1.08
dncl	1.20	1.07	1.28	1.21	0.99	0.93	1.07	0.97	1.13	1.02
dnco	−0.54	−1.08	−0.41	−0.84	−0.50	−0.99	−0.22	−0.41	−0.01	−0.02

TABLE 1 Continued.

	NN1		NN2		NN3		NN4		NN5	
Variable	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**
dfin	2.73	0.74	0.47	0.14	1.38	0.41	0.39	0.12	0.38	0.11
ta	−2.45	−0.82	−2.52	−0.94	−3.78	−1.38	−3.17	−1.20	−4.25	−1.57
dsti	−12.42	−2.29	−8.84	−1.76	−9.77	−1.85	−7.64	−1.52	−9.52	−1.81
dfnl	**4.73**	**3.84**	3.06	2.59	3.17	2.60	2.72	2.16	2.60	2.06
egr_hxz	1.38	1.62	2.04	2.49	2.37	2.98	2.28	2.71	2.53	3.15
grcapx	−2.42	−1.23	−0.46	−0.25	−1.44	−0.77	−0.10	−0.06	−0.49	−0.27
pchcapx3	1.34	1.60	1.34	1.64	1.50	1.90	1.51	1.85	1.34	1.68
cei	**−0.14**	**−9.75**	**−0.27**	**−12.05**	**−0.41**	**−14.55**	**−0.47**	**−25.19**	**−0.36**	**−19.94**
nef	−0.46	−2.41	−0.53	−2.72	−0.49	−2.55	−0.52	−2.81	−0.49	−2.70
ndf	**3.66**	**83.39**	**3.65**	**86.24**	**3.28**	**69.03**	**3.36**	**73.37**	**3.37**	**92.01**
nxf	**1.45**	**4.79**	0.73	2.44	0.72	2.34	0.77	2.60	0.72	2.50
roic	2.31	3.55	1.54	2.57	1.30	2.16	1.28	2.12	1.21	2.04
chcsho	**1.75**	**5.49**	**1.61**	**4.82**	**1.56**	**4.64**	**1.66**	**5.21**	**1.74**	**5.62**
dpia	8.46	3.21	7.27	2.94	7.35	3.04	7.22	2.93	7.16	2.83
pchcapx	−1.95	−1.47	−0.41	−0.32	0.29	0.23	0.10	0.08	−0.14	−0.11
cdi	0.53	0.42	1.29	1.07	1.79	1.48	2.33	1.88	2.30	1.87
ivg	**9.08**	**5.33**	**8.22**	**4.86**	**8.73**	**5.22**	**8.36**	**5.24**	**8.22**	**5.05**
poa	4.36	2.33	5.35	2.95	4.77	2.74	4.30	2.51	3.82	2.28
hire	−0.66	−1.12	−0.76	−1.31	−0.40	−0.66	−0.30	−0.56	−0.41	−0.72
CMA	**2.78**	**4.45**	**2.71**	**4.38**	**2.91**	**4.63**	**3.03**	**4.81**	**3.06**	**4.92**
HXZ_IA	**3.86**	**6.12**	**2.96**	**4.89**	**2.98**	**4.80**	**3.32**	**5.40**	**3.37**	**5.45**
sgr	−0.01	−0.02	0.42	0.82	0.28	0.54	0.09	0.17	0.33	0.67
cto	−4.86	−3.09	−4.45	−2.90	−4.57	−3.21	−3.97	−2.90	−4.20	−3.06
os	−1.69	−2.58	−1.16	−1.68	−0.50	−0.77	−1.09	−1.63	−0.92	−1.33
zs	**2.56**	**5.01**	**2.31**	**4.78**	**1.91**	**4.01**	**2.13**	**4.36**	**2.01**	**4.24**
ps	−0.46	−0.62	−1.21	−1.61	−1.41	−1.97	−1.97	−2.78	−1.76	−2.32
chempia	1.47	0.73	0.53	0.30	1.92	1.09	1.50	0.85	1.03	0.58
ms	1.97	2.05	1.74	1.88	1.33	1.45	1.37	1.44	1.48	1.61
rna	−0.28	−0.65	−0.37	−0.98	−0.36	−0.99	−0.53	−1.48	−0.35	−0.95
pm	−0.67	−1.71	−0.91	−2.40	−0.93	−2.40	−0.84	−2.13	−0.81	−2.12
ato	−0.34	−0.51	−0.79	−1.51	−0.77	−1.31	−0.73	−1.38	−0.69	−1.28
chatoia	**8.53**	**7.46**	**8.16**	**7.41**	**8.67**	**8.38**	**8.11**	**7.93**	**8.22**	**8.01**
chpmia	0.03	0.03	−1.00	−1.05	−0.70	−0.70	−0.77	−0.82	−0.97	−1.00
roaq	−2.51	−2.95	−1.88	−2.13	−1.38	−1.72	−1.29	−1.58	−1.38	−1.68
gma	2.33	1.68	2.03	1.56	1.50	1.11	1.62	1.23	1.26	0.98
RMW	**3.58**	**5.42**	**3.21**	**5.04**	**2.90**	**4.46**	**2.84**	**4.36**	**3.11**	**4.98**
HXZ_ROE	**3.13**	**3.83**	**3.29**	**4.22**	**3.28**	**4.06**	**3.18**	**4.12**	**2.90**	**3.83**
cashdebt	**−5.53**	**−4.86**	**−4.51**	**−4.01**	**−4.60**	**−4.15**	**−4.07**	**−3.81**	**−3.94**	**−3.73**
currat	−0.14	−0.50	0.01	0.04	0.02	0.08	0.03	0.13	0.07	0.27
pchcurrat	**17.62**	**3.72**	16.83	3.48	16.50	3.46	15.82	3.48	16.19	3.53
pchquick	−3.94	−0.94	−6.71	−1.60	−7.47	−1.84	−9.67	−2.41	−8.02	−1.98
pchsaleinv	**5.94**	**9.81**	**7.31**	**11.82**	**7.67**	**11.71**	**8.09**	**12.75**	**7.85**	**12.80**
quick	−0.76	−3.01	**−0.93**	**−3.90**	**−0.96**	**−3.77**	**−1.01**	**−4.27**	−0.86	−3.44
salecash	−0.54	−1.96	−0.51	−1.85	−0.49	−1.97	−0.42	−1.60	−0.35	−1.38
saleinv	1.43	2.63	1.61	3.01	1.68	3.13	1.54	2.81	1.44	2.67
salerec	0.56	1.39	0.09	0.23	0.12	0.31	0.11	0.27	0.13	0.34
pchgm_pchsale	**−6.69**	**−4.81**	**−5.29**	**−3.78**	**−6.29**	**−4.41**	−5.43	−3.59	**−5.35**	**−4.02**
pchsale_pchinvt	**8.71**	**5.34**	**9.85**	**6.33**	**10.91**	**7.09**	**11.15**	**7.11**	**11.64**	**7.72**
pchsale_pchrect	1.68	0.82	1.26	0.62	1.63	0.78	1.57	0.80	1.53	0.73
pchsale_pchxsga	−1.54	−0.94	−0.85	−0.53	−1.43	−0.95	−0.24	−0.15	−0.07	−0.05
etr	−1.03	−0.45	−0.84	−0.41	−0.92	−0.40	−1.60	−0.69	−1.19	−0.48
lfe	**−13.04**	**−5.37**	**−13.13**	**−5.37**	**−11.92**	**−4.78**	**−11.62**	**−4.89**	**−11.46**	**−4.68**
pchcapx_ia	2.88	1.45	2.89	1.57	2.95	1.63	2.84	1.46	2.95	1.48

TABLE 1 Continued.

	NN1		NN2		NN3		NN4		NN5	
Variable	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**
adm	−3.62	−2.97	−3.88	−3.11	−3.30	−2.29	−3.17	−2.43	−3.24	−2.54
rdm	0.99	1.79	1.05	1.94	1.04	1.98	0.75	1.44	1.14	2.30
rds	0.17	0.35	0.02	0.05	−0.06	−0.12	−0.12	−0.25	−0.25	−0.56
kz	**3.48**	**5.14**	**2.98**	**4.56**	**2.85**	**4.51**	**2.35**	**3.59**	**2.61**	**4.01**
ob_a	0.31	0.32	−0.01	−0.01	0.07	0.09	0.24	0.27	−0.19	−0.26
roavol	**−0.45**	**−3.61**	**−0.49**	**−3.94**	**−0.43**	**−3.69**	−0.42	−3.42	−0.43	−3.53
pricedelay	0.33	0.39	1.42	1.75	1.12	1.48	1.10	1.37	1.13	1.42
age	−0.67	−2.78	−0.75	−3.19	−0.79	−3.51	−0.76	−3.31	−0.67	−2.97
herf	−1.78	−1.01	−1.79	−1.07	−1.71	−0.98	−1.30	−0.76	−0.78	−0.45
ww	**−1.71**	**−4.27**	−1.12	−3.37	−1.18	−3.19	−1.13	−2.97	−1.20	−3.26
tang	−0.61	−1.38	−1.20	−2.91	−1.19	−2.89	−1.08	−2.60	−0.97	−2.25
moms12m	0.84	0.67	1.16	1.00	0.90	0.78	0.90	0.80	1.11	0.98
absacc	**−2.22**	**−4.44**	**−2.18**	**−4.51**	**−2.24**	**−4.58**	**−2.03**	**−4.15**	**−1.88**	**−3.85**
invest	7.48	2.91	7.35	3.02	6.55	2.65	7.09	3.05	7.16	2.90
realestate_hxz	1.94	1.36	1.85	1.42	1.42	1.03	1.43	1.07	1.44	1.06
pctacc	0.85	0.34	−1.03	−0.44	−0.75	−0.32	−1.96	−0.79	−2.29	−0.97
ol	**3.07**	**98.26**	**2.29**	**69.16**	**2.53**	**68.84**	**2.56**	**72.51**	**2.53**	**64.64**
cash	−0.46	−1.65	−0.45	−1.57	−0.49	−1.89	−0.46	−1.82	−0.39	−1.57
orgcap	0.55	0.70	0.54	0.77	0.77	1.01	0.85	1.16	0.77	1.08
gad	−0.04	−0.02	−0.47	−0.19	0.52	0.21	0.09	0.03	0.11	0.05
ala	0.56	1.25	−0.09	−0.21	−0.09	−0.21	−0.16	−0.38	−0.13	−0.31
convind	**−2.45**	**−3.86**	−1.90	−3.05	−1.94	−3.11	−1.66	−2.76	−1.75	−2.91
beta	0.01	0.11	−0.05	−0.60	−0.04	−0.50	−0.06	−0.64	−0.08	−0.90
pps	**1.30**	**3.60**	0.90	2.26	1.18	2.83	1.16	2.77	1.11	2.85
baspread	−0.15	−1.33	−0.19	−1.68	−0.13	−1.11	−0.18	−1.65	−0.19	−1.67
SMB	**1.80**	**5.18**	**1.49**	**4.34**	**1.63**	**4.30**	**1.48**	**3.92**	**1.35**	**4.16**
IPO	0.11	0.36	0.18	0.61	0.47	1.55	0.27	0.87	0.37	1.19
turn	−0.15	−1.10	−0.11	−0.78	−0.07	−0.49	−0.05	−0.36	−0.12	−0.83
mve_ia	**3.28**	**4.24**	**2.82**	**3.79**	**2.98**	**3.93**	**3.22**	**4.13**	**2.68**	**3.66**
dolvol	−0.07	−0.26	−0.20	−0.83	−0.18	−0.71	−0.18	−0.72	−0.16	−0.65
std_dolvol	5.12	2.43	5.59	2.62	5.30	2.94	5.38	2.49	4.75	2.12
std_turn	0.17	1.10	0.27	1.61	0.25	1.59	0.22	1.36	0.12	0.77
ill	0.98	1.23	−0.22	−0.29	−0.22	−0.27	−0.61	−0.76	−0.63	−0.78
LIQ_PS	**2.90**	**4.13**	2.20	3.13	**2.65**	**3.91**	**3.12**	**4.63**	**2.96**	**4.33**
idiovol	0.27	2.58	0.32	2.90	0.29	2.82	0.25	2.42	0.23	2.32
retvol	−0.08	−0.61	−0.09	−0.71	−0.10	−0.80	−0.05	−0.44	−0.04	−0.37
zerotrade	−0.11	−1.04	−0.03	−0.28	−0.03	−0.24	−0.06	−0.52	−0.08	−0.79
sin	1.40	1.59	1.15	1.40	1.45	1.76	1.43	1.74	1.20	1.45
stdcf	1.19	3.42	0.87	2.50	0.93	2.55	0.99	2.80	1.03	3.02
stdacc	1.29	3.44	1.13	2.92	**1.33**	**3.72**	1.18	3.28	1.08	2.94
maxret	0.01	0.08	0.05	0.30	0.07	0.43	0.05	0.29	0.04	0.25
BAB	1.87	2.31	1.78	2.26	1.91	2.46	1.89	2.47	2.28	3.03
Intermediary	−0.36	−3.57	0.00	0.00	−0.10	−1.00	−0.05	−0.56	−0.09	−0.90

Note: This table reports the marginal contribution of 150 factors from July 1976 to December 2017, sequentially relative to the other 149 factors. The test assets include a set of 750 3 × 2 bivariate-sorted portfolios. λ_k represents the average partial effect of each factor using the ADML estimator. ADML estimator uses the LASSO to estimate Riesz representor α_0 and neural networks with one to five units in the hidden layer to estimate γ_0. The LASSO tuning parameters are chosen by the iterative tuning procedure of Chernozhukov et al. (2022) with a maximum number of iterations at 10.

use. Third, there are more than 30 significant factors even if we choose all other factors as controls, consistent with the view that a few factors cannot summarize the cross-sectional expected stock returns (Kozak et al., 2020). The estimation results of neural networks using other units in the hidden layer are consistent with those of NN3, indicating the robustness of the estimation.

Unlike the classical machine learning method as in Gu et al. (2020), the ADML estimator can obtain unbiased estimation, construct confidence intervals and make statistical inferences, which can help compare the factors' significance and

TABLE 2 Spanning test.

	$\lvert\bar{\alpha}\rvert$	$p < 0.01$	$p < 0.05$	No. of factors	$\lvert\bar{\alpha}\rvert$	$p < 0.01$	$p < 0.05$	No. of factors	$\lvert\bar{\alpha}\rvert$	$p < 0.01$	$p < 0.05$	No. of factors
	Panel A: All				Panel B: Momentum				Panel C: Value-versus-growth			
ADML3	0.14	16	40	147	0.11	0	1	11	0.23	4	8	24
ADML5	0.14	16	42	145	0.10	0	1	11	0.23	4	12	24
ADML6	0.13	17	39	144	0.10	0	1	11	0.22	4	9	24
FF3	0.18	41	71	147	0.14	1	3	12	0.18	4	9	22
FF5	0.15	33	52	145	0.12	0	2	12	0.16	5	8	22
FF6	0.13	26	50	144	0.06	1	1	11	0.16	5	9	22
OLS6	0.16	31	44	144	0.14	2	3	11	0.31	13	15	24
DSL6	0.13	25	43	144	0.17	1	4	12	0.14	3	4	22
	Panel D: Investment				Panel E: Profitability				Panel F: Intangibles			
ADML3	0.13	8	19	38	0.12	2	2	16	0.09	0	6	37
ADML5	0.12	8	19	37	0.11	2	2	16	0.08	0	4	36
ADML6	0.12	9	18	37	0.11	2	2	15	0.08	0	5	36
FF3	0.15	13	23	39	0.20	7	9	16	0.11	6	10	38
FF5	0.13	11	16	38	0.13	2	6	15	0.16	10	13	38
FF6	0.11	9	14	38	0.11	1	6	15	0.15	9	14	38
OLS6	0.10	6	9	36	0.12	3	4	15	0.09	3	7	38
DSL6	0.10	11	15	37	0.22	7	10	16	0.08	1	6	38
	Panel G: Trading frictions											
ADML3	0.19	2	4	21								
ADML5	0.20	2	4	21								
ADML6	0.19	2	4	21								
FF3	0.35	10	17	20								
FF5	0.20	5	7	20								
FF6	0.17	1	6	20								
OLS6	0.25	4	6	20								
DSL6	0.17	2	4	19								

Note: This table reports the spanning test results of various factor models. Column "$\lvert\bar{\alpha}\rvert$" reports the mean absolute alpha of factors being tested. Column "$p < 0.01$" ("$p < 0.05$") reports the number of alphas with p-value less than 0.01 (0.05). Column "No. of factors" reports the number of factors being tested. We report the results for the Fama–French three (FF3), five (FF5), and six factors (FF6); the top three, five, and six most significant factors selected by the ADML with NN3 (denoted as ADML3, ADML5, and ADML6, respectively); and the most significant six factors selected by the OLS (OLS6) and by the double-selection LASSO (DSL6). The t-statistics are adjusted for heteroscedasticity and autocorrelations.

importance. Some factors such as net debt finance, operating leverage, 36-month momentum, composite equity issuance, % change sales-to-inventory, and industry-adjusted change in asset turnover appear statistically significant. They are useful in explaining the cross-sectional expected return. While some factors, such as current ratio, order backlog, dividend initiation, R&D-to-sales, growth in advertising expense, and book-asset liquidity, appear insignificant in explaining the cross-section when controlling other factors, which deem these factors redundant or useless.

Overall, the procedure above provides a valuable framework to apply the ADML method to identify factors sequentially and avoid the regularization and overfitting bias in a general framework.

4.3 | Spanning test

In the previous subsection, we find useful factors that can significantly explain the cross-sectional stock returns. In this subsection, we explore whether the most significant factors identifies by the ADML perform well in explaining all the remaining factors in a factor spanning test. To keep similar parsimony, we assemble the most significant three to six factors to form factor models and compare their performance with the well-known Fama–French sparse factor models.

Table 2 reports the spanning test results of various factor models. Column "$|\bar{\alpha}|$" reports the mean absolute alpha of factors being tested. Column "$p < 0.01$" ("$p < 0.05$") reports the number of alphas with p-value less than 0.01 (0.05). Column "No. of factors" reports the number of factors being tested. We first compare model performances of the Fama–French three (FF3), five (FF5), and six factors (FF6) with the top three, five, and six most significant factors selected by the ADML with NN3[6] (denoted as ADML3, ADML5, ADML6, respectively).

Panel A of Table 2 shows the model performance tested using all the factors. The results show that the ADML3, ADML5, and ADML6 outperform the FF3, FF5, and FF6 models, respectively. The ADML3, ADML5, and ADML6 models leave 16, 16, and 17 factors unexplained with p-value less than 0.01, while the numbers of unexplained factors for FF3, FF5, and FF6 models are 41, 33, and 26, respectively. The average absolute factor alphas are 0.18%, 0.15%, and 0.13% per month for the FF3, FF5, and FF6 models. By contrast, the absolute alphas are 0.14%, 0.14%, and 0.13% per month for the ADML3, ADML5, and ADML6 models.

Panels B to G of Table 2 give model performance in explaining factors within each of the six Hou et al. (2020) categories. The ADML factor models outperform the FF models in all categories, except for the value-versus-growth category. Especially for the intangible category, the ADML factor models can explain all the other intangible factors at the 1% level, and the absolute alphas are only 0.09%, 0.08%, and 0.08% per month. By contrast, the FF3, FF5, and FF6 models do not perform well, leaving six, 10, and nine significant intangible factors unexplained.

Besides, we compare the pricing errors of factors selected by the OLS and the double-selection LASSO (Feng et al., 2020) under the linear SDF assumption with factors selected by the ADML under the nonlinear SDF assumption. Table 2 shows that the six-factor model selected by the ADML method (ADML6) outperforms the six-factor models selected by the OLS (OLS6) and by the double-selection LASSO (DSL6). The average magnitudes of the absolute factor alphas are 0.13%, 0.16%, and 0.13% for ADML6, OLS6, and DSL6, respectively. The ADML6 cannot fully explain 17 (39) factors, with significant alphas at the 1% (5%) level. By contrast, the number of unexplained factors for the OLS6 is 31 (44) at the 1% (5%) significance level, and the number for the DSL6 is 25 (43). Overall, it indicates that the factors deemed useful by the ADML under the nonlinear SDF assumption can better explain the other factors even in the linear spanning test.[7]

4.4 | Recursive estimates

Furthermore, we examine how each factor's incremental contribution relative to the other factors and the number of significant factors in each category vary over time. We identify each factor by the ADML using recursive (expanding-window) subsamples. The first subsample starts at July 1976 and ends at December 1998. The following subsample extends the sample period by 1 year, and we repeat the analysis to obtain 20 subsample estimates for each factor.

Figure 2 reports the factor significance rate of recursive ADML estimates. In the figure, factors are identified by their ID, and each column represents the proportion of the absolute value of t-statistics exceeding 3.59 in 20 estimates. Figure 2

[6]The six most significant factors identified by the ADML with NN3 are net debt finance, operating leverage, 36-month momentum, composite equity issuance, % change sales-to-inventory, and industry-adjusted change in asset turnover.

[7]Freyberger et al. (2020), who emphasize stock return prediction under both nonlinear and linear model structures, also find that the predictors identified by the nonlinear model slightly outperform the predictors selected by the linear model in the linear predictive model setting.

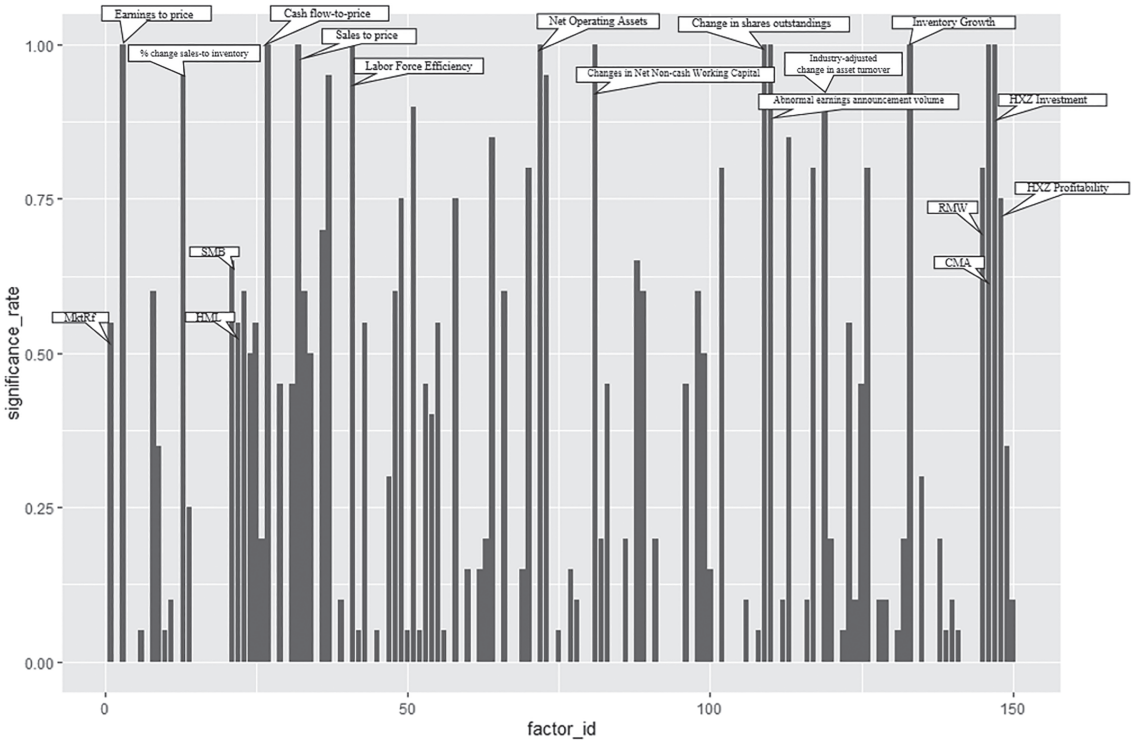

FIGURE 2 Expanding-window estimates: Factor significance rate. *Note*: This figure shows the factor significance rate of expanding-window ADML estimates. In particular, we identify each factor by ADML estimator in different recursive subsamples. Each subsample starts at July 1976, firstly ends at December 1998, and is increased by 1 year, obtaining 20 estimates. The significance rate of each factor among these 20 estimates are reported. The ADML estimator uses LASSO to estimate Riesz representor α_0 and the neural network with three units in the hidden layer to estimate γ_0. The LASSO tuning parameters are chosen by the iterative tuning procedure of Chernozhukov et al. (2022) with a maximum number of iterations at 10.

shows that many factors are not significant in all periods (some breakpoints on the horizontal axis mean that the factor significance ratio of these IDs is 0), which means that these factors have little incremental contribution relative to other factors. By contrast, some factors can explain cross-sectional stock returns at all times relative to all other factors, such as earnings to price, cash flow-to-price, sales to price, labor force efficiency, net operating assets, changes in net non-cash working capital to price, change in shares outstanding, abnormal earnings announcement volume, and inventory growth. Pricing factors in prominent asset pricing models, such as SMB, HML, RMW, and CMA in Fama and French (2015) five-factor model and IA and ROE factors in Hou et al. (2015) four-factor model, show strong robustness in explaining the cross-section.

Figure 3 illustrates the time-series performance of expanding-window ADML estimates in each category to illustrate the time-series pattern. On the whole, the number of significant factors is decreasing over time. In particular, the significance declines over time, especially after 2003, which is consistent with the findings of Green et al. (2017) and may reflect significant changes that occurred in the information and trading environment, such as the passing of the Sarbanes–Oxley Act, acceleration of 10-Q, and 10-K filing requirements by the SEC and the introduction of auto quoting by the NYSE. However, in contrast to the linear estimators of OLS and double-selection LASSO of Feng et al. (2020), the difference is that we find more significant factors (around 30 to 50),[8] which may arise due to the average partial effect we estimate that captures the interaction and nonlinear relationship of factors. Besides, for each category, the significance of investment and profitability factors decreased significantly in recent 20 years, while the factors of other categories only decreased slightly and steadily.[9]

[8]Green et al. (2017) find that 12 characteristics are independent determinants in non-microcap stocks from 1980 to 2014. The number of independent determinants sharply fell to 2 after 2003.

[9]In Figure B3 in the online Appendix, we employ neural networks with one, two, four, and five units in the hidden layer and obtain the expanding-window estimates. The results are robust to different neural network models, and the stable factors are highly similar.

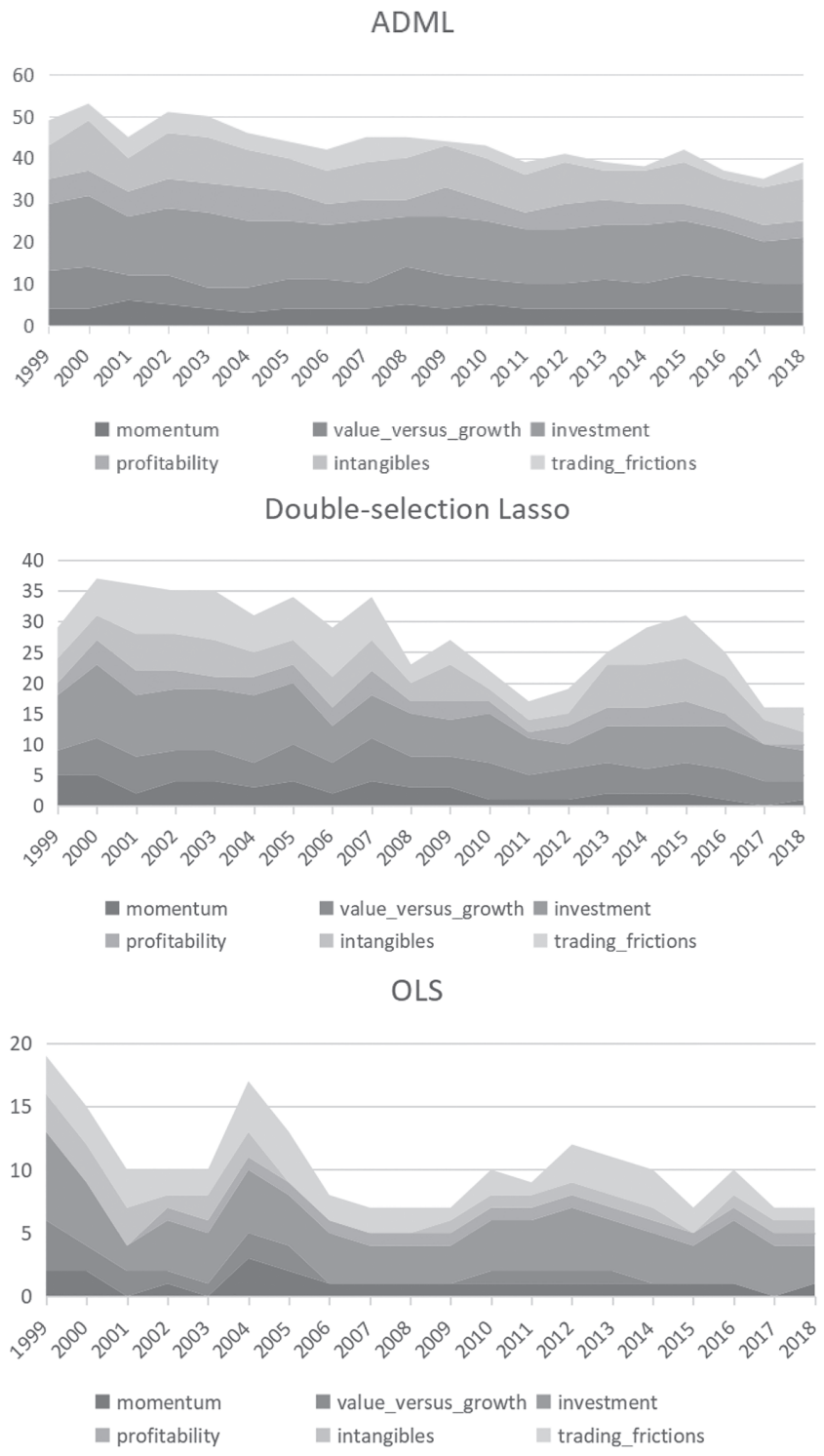

FIGURE 3 Expanding-window estimates: Time-series performance. *Note*: This figure reports the time-series performance in each category. We identify each factor by ADML, double-selection LASSO of Feng et al. (2020), and OLS estimator in different recursive subsamples. Each subsample starts at July 1976, firstly ends at December 1998, and is increased by 1 year, obtaining 20 estimates. The significant factor numbers are reported by category.

4.5 | Comparison to the Chinese stock market

This subsection investigates the difference of significant factors to explain the cross-sectional expected returns between the Chinese and US stock markets. As one of the largest emerging markets in the world, China's stock market has a large proportion of small investors and a high limit to arbitrage, making it systematically different from the US stock market. Following Hou et al. (2020) and Chen et al. (2022), we construct 132 firm characteristics in the Chinese stock market using the trading and financial data from the China Stock Market & Accounting Research Database (CSMAR). The detailed definitions of these characteristics are reported in Table B3 in the online Appendix. We construct 132 factors based on

TABLE 3 Identify factors sequentially in Chinese stock market.

	NN1		NN2		NN3		NN4		NN5	
Variable	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**
abr	1.48	1.96	−0.21	−0.25	−0.22	−0.27	−0.23	−0.28	−0.17	−0.22
abtur	**−5.27**	**−3.85**	**−5.58**	**−3.83**	**−5.89**	**−3.83**	**−5.95**	**−3.94**	**−6.02**	**−4.01**
age	0.46	0.70	0.35	0.52	−0.19	−0.29	−0.12	−0.18	−0.26	−0.39
ala	**11.22**	**6.60**	**7.45**	**4.63**	**7.71**	**4.86**	**7.54**	**4.74**	**7.32**	**4.58**
alaq	0.17	0.14	0.64	0.51	0.67	0.55	0.49	0.40	0.68	0.55
alm	0.14	0.11	1.07	0.82	1.06	0.83	0.83	0.64	1.16	0.91
almq	0.47	0.42	0.65	0.55	0.62	0.52	0.65	0.56	0.67	0.56
am	1.39	1.04	1.74	1.23	1.90	1.38	1.91	1.35	1.87	1.33
ami	**2.11**	**4.80**	0.90	1.99	1.00	2.20	0.91	2.05	0.90	1.99
amq	−0.70	−0.79	−0.93	−1.00	−0.63	−0.67	−0.67	−0.74	−0.61	−0.67
ato	3.23	1.13	2.47	0.90	2.63	0.96	3.19	1.18	3.17	1.17
atoq	−0.66	−0.40	0.44	0.27	0.04	0.02	0.32	0.19	0.51	0.32
beta	**2.00**	**6.29**	**1.95**	**5.65**	**1.72**	**5.07**	**1.63**	**4.85**	**1.52**	**4.49**
beta_	**0.98**	**3.91**	0.55	2.08	0.48	1.90	0.57	2.22	0.55	2.11
betad	**2.94**	**5.63**	1.79	3.20	1.85	3.33	1.88	3.37	1.88	3.31
betafp	**−1.22**	**−4.29**	−0.82	−2.82	−0.89	−3.15	−0.85	−2.99	−0.87	−3.04
bl	**−2.47**	**−3.69**	−0.65	−0.90	−0.51	−0.73	−0.30	−0.42	−0.12	−0.17
blq	1.11	1.51	1.53	1.99	1.77	2.34	1.95	2.54	1.99	2.63
bm	**−0.10**	**−20.46**	−0.01	−1.93	**0.03**	**5.19**	0.01	1.58	0.02	3.51
bmj	**2.80**	**3.59**	1.92	2.29	1.72	2.06	1.85	2.24	1.86	2.24
bmq	−0.23	−0.18	−0.18	−0.14	−0.03	−0.02	0.19	0.15	0.18	0.14
cdi	−2.21	−0.91	−3.60	−1.61	−3.52	−1.58	−3.08	−1.40	−3.78	−1.67
cei	7.18	2.99	8.28	3.47	**9.32**	**3.95**	**8.52**	**3.58**	**9.06**	**3.77**
cla	−1.06	−2.19	−0.84	−1.64	−0.85	−1.70	−0.69	−1.33	−0.70	−1.37
claq	1.52	2.64	−0.06	−0.10	−0.30	−0.51	−0.21	−0.34	−0.08	−0.13
cop	−0.97	−2.02	−0.69	−1.37	−0.76	−1.54	−0.67	−1.34	−0.62	−1.25
cp	−0.08	−0.05	2.78	1.80	2.97	1.93	2.77	1.81	2.48	1.62
cpq	2.26	2.39	2.85	2.96	2.80	3.01	2.71	2.93	2.63	2.80
cs	4.21	2.11	4.30	2.10	4.24	1.96	4.48	2.21	4.40	2.26
cta	−2.43	−2.11	−1.38	−1.12	−1.64	−1.35	−1.41	−1.13	−1.47	−1.18
cto	3.37	2.87	2.44	1.91	2.36	1.82	2.18	1.71	2.14	1.67
ctoq	4.00	2.93	2.76	2.03	2.40	1.79	2.46	1.83	2.19	1.65
cvd	−1.42	−1.80	−1.09	−1.45	−1.29	−1.70	−0.98	−1.25	−1.30	−1.63
cvt	**−1.62**	**−69.75**	**−1.76**	**−75.81**	**−1.81**	**−72.85**	**−1.82**	**−71.54**	**−1.87**	**−72.17**
db	−0.12	−0.17	−0.12	−0.16	−0.20	−0.28	−0.13	−0.18	−0.18	−0.25
dcoa	**−18.12**	**−4.19**	−13.33	−3.43	**−13.95**	**−3.65**	−13.14	−3.42	**−13.71**	**−3.60**
dcol	**22.45**	**5.36**	**19.58**	**5.14**	**17.09**	**4.68**	**16.38**	**4.53**	**16.30**	**4.54**
dfin	2.96	1.65	1.06	0.58	0.73	0.41	0.48	0.26	0.55	0.30
dfnl	**7.05**	**3.56**	**6.91**	**3.56**	**6.92**	**3.59**	**7.01**	**3.62**	**6.87**	**3.57**
dgs	8.53	3.20	**9.99**	**3.97**	8.62	3.44	8.33	3.38	8.45	3.43
dlti	−2.28	−1.74	−0.93	−0.73	−0.70	−0.55	−0.89	−0.69	−0.62	−0.49
dm	**−5.36**	**−4.38**	−3.69	−3.27	−3.03	−2.70	−3.05	−2.70	−2.99	−2.67
dmq	**−0.62**	**−88.47**	**−0.41**	**−51.57**	**−0.30**	**−43.04**	**−0.28**	**−44.35**	**−0.25**	**−42.36**
dnca	**−4.08**	**−11.05**	**−3.00**	**−8.45**	**−2.91**	**−8.14**	**−2.73**	**−7.98**	**−2.66**	**−7.83**
dnco	−1.85	−1.45	−1.52	−1.10	−1.30	−0.97	−1.03	−0.76	−0.81	−0.60
dnoa	**14.16**	**5.01**	9.56	3.42	**9.62**	**3.58**	8.69	3.23	8.04	3.03
dpia	**−3.29**	**−115.63**	**−2.91**	**−74.18**	**−3.05**	**−72.95**	**−2.93**	**−75.36**	**−2.89**	**−72.78**
droa	3.64	1.96	5.21	2.65	4.85	2.57	5.27	2.74	5.23	2.77
droe	4.70	2.00	2.88	1.29	2.47	1.14	2.33	1.05	2.42	1.10
dsa	8.11	2.46	3.60	1.11	2.88	0.91	2.56	0.80	2.68	0.84
dsi	**5.96**	**5.59**	**11.09**	**11.85**	**10.56**	**11.10**	**10.08**	**10.93**	**10.06**	**10.96**
dss	2.23	1.16	1.84	0.97	2.08	1.10	1.86	1.01	1.58	0.86
dsti	**−10.31**	**−6.49**	**−10.24**	**−5.86**	**−9.17**	**−5.34**	**−9.11**	**−5.25**	**−9.40**	**−5.48**
dtv	**−3.16**	**−5.78**	**−2.08**	**−3.79**	−1.78	−3.30	−1.76	−3.25	−1.78	−3.32

TABLE 3 Continued.

	NN1		NN2		NN3		NN4		NN5	
Variable	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**
dwc	**7.16**	**91.21**	**5.24**	**59.77**	**4.93**	**60.22**	**4.64**	**58.96**	**4.27**	**52.52**
ebp	2.57	2.52	1.19	1.13	1.49	1.49	1.68	1.64	1.89	1.87
ebpq	−3.73	−2.54	−3.15	−2.33	−2.99	−2.26	−2.73	−2.00	−2.74	−2.05
em	**−2.67**	**−5.21**	**−2.06**	**−3.81**	**−1.92**	**−3.62**	−1.71	−3.24	−1.61	−3.01
emq	**−3.13**	**−4.84**	**−2.79**	**−4.40**	**−2.65**	**−4.27**	**−2.35**	**−3.84**	**−2.24**	**−3.60**
ep	**2.11**	**3.93**	1.53	2.75	1.39	2.58	1.24	2.26	1.16	2.15
epq	2.07	2.90	2.18	3.21	1.99	3.04	1.84	2.85	1.74	2.60
etr	−9.67	−3.28	**−10.71**	**−3.76**	**−10.36**	**−3.70**	**−10.35**	**−3.67**	−9.92	−3.53
gla	1.17	2.16	0.46	0.77	0.40	0.70	0.40	0.69	0.32	0.56
glaq	1.04	2.34	0.62	1.30	0.63	1.37	0.59	1.26	0.57	1.21
gpa	0.57	1.19	0.11	0.20	0.21	0.41	0.17	0.33	0.17	0.33
ia	−4.24	−2.90	−3.55	−2.45	−3.48	−2.45	−3.58	−2.54	−3.73	−2.69
iaq	**−4.12**	**−3.88**	−3.29	−3.35	**−3.62**	**−3.56**	−3.45	−3.36	−3.58	−3.50
ir	**−3.65**	**−4.89**	−2.36	−2.94	−2.38	−3.02	−2.50	−3.15	−2.25	−2.81
isc	−1.67	−1.36	−3.61	−3.19	−3.42	−3.02	−3.45	−3.10	−3.51	−3.17
isff	−7.06	−2.76	−5.68	−2.28	−5.47	−2.29	−5.64	−2.30	−5.70	−2.36
isq	**11.87**	**6.01**	**13.73**	**7.74**	**13.08**	**7.65**	**12.48**	**7.13**	**11.86**	**6.81**
iv	−0.51	−0.82	−1.07	−1.79	−0.74	−1.26	−0.61	−1.05	−0.58	−0.97
ivc	**−3.50**	**−6.02**	**−3.15**	**−5.32**	**−2.81**	**−4.84**	**−2.75**	**−4.63**	**−2.72**	**−4.68**
ivff	**−5.24**	**−10.62**	**−3.88**	**−7.54**	**−3.51**	**−7.05**	**−3.41**	**−6.70**	**−3.32**	**−6.52**
ivg	−1.81	−1.24	−3.00	−2.08	−2.92	−2.07	−2.87	−2.00	−2.92	−2.04
ivq	**−5.60**	**−10.42**	**−4.11**	**−7.54**	**−3.59**	**−6.54**	**−3.55**	**−6.47**	**−3.49**	**−6.41**
kz	−1.97	−1.51	−0.75	−0.56	−0.95	−0.72	−0.62	−0.47	−0.83	−0.63
kzq	**6.03**	**4.31**	2.91	2.18	2.85	2.15	2.57	1.96	2.73	2.10
mdr	−0.91	−1.75	−1.70	−3.40	−1.66	−3.35	−1.51	−3.06	−1.49	−3.00
me	**−3.35**	**−10.18**	**−1.50**	**−4.46**	**−1.58**	**−4.70**	**−1.62**	**−4.84**	**−1.57**	**−4.62**
ndp	**6.61**	**4.61**	**6.35**	**4.16**	**6.73**	**4.33**	**6.61**	**4.36**	**6.40**	**4.20**
ndpq	−0.12	−0.08	1.97	1.52	2.23	1.75	2.21	1.69	2.27	1.74
noa	**5.93**	**4.02**	3.40	2.23	3.49	2.37	3.61	2.46	3.18	2.17
nop	**−0.38**	**−23.07**	**−1.12**	**−42.23**	**−1.17**	**−45.03**	**−1.24**	**−45.46**	**−1.19**	**−46.44**
nopq	**3.24**	**121.80**	**2.03**	**96.74**	**2.02**	**91.01**	**1.97**	**84.33**	**2.02**	**92.92**
nsi	**10.47**	**3.75**	7.07	2.55	6.73	2.49	6.80	2.55	6.45	2.41
ocp	**4.19**	**5.18**	2.39	2.85	2.23	2.69	2.03	2.43	2.04	2.45
ocpq	**−5.11**	**−134.74**	**−3.27**	**−77.53**	**−3.44**	**−75.99**	**−3.18**	**−74.75**	**−3.11**	**−76.72**
ola	0.73	1.83	0.78	1.85	0.80	1.90	0.67	1.57	0.68	1.62
olaq	0.99	2.47	0.88	2.11	0.89	2.16	0.80	1.95	0.81	1.99
ole	0.31	0.80	0.56	1.36	0.55	1.33	0.51	1.25	0.51	1.25
oleq	0.91	2.25	0.83	1.96	0.79	1.89	0.80	1.90	0.75	1.79
oll	**14.77**	**6.18**	**10.58**	**4.49**	**9.10**	**4.01**	**8.62**	**3.76**	**8.44**	**3.68**
olq	8.17	3.42	7.94	3.23	7.71	3.12	7.61	3.10	7.42	2.99
op	0.64	0.55	−0.38	−0.33	−0.55	−0.47	−0.52	−0.45	−0.61	−0.52
opa	0.34	1.02	0.44	1.24	0.41	1.15	0.41	1.14	0.43	1.22
ope	0.12	0.35	0.24	0.68	0.22	0.62	0.30	0.83	0.21	0.59
opq	**5.46**	**4.02**	**5.44**	**4.19**	**5.64**	**4.53**	**5.49**	**4.37**	**5.53**	**4.48**
pm	0.34	0.74	0.85	1.74	0.69	1.45	0.70	1.46	0.70	1.45
pmq	−0.56	−1.09	−0.04	−0.07	−0.09	−0.16	−0.17	−0.32	−0.09	−0.17
poa	5.45	1.26	3.30	0.85	3.27	0.85	2.95	0.78	2.87	0.74
pps	**−3.88**	**−4.75**	−2.60	−3.44	**−3.08**	**−4.05**	−2.64	−3.55	**−2.82**	**−3.80**
r6	0.10	0.15	−0.31	−0.48	−0.44	−0.69	−0.39	−0.60	−0.54	−0.85
r11	−1.00	−1.61	−0.58	−0.93	−0.45	−0.73	−0.55	−0.88	−0.47	−0.75
r1a	2.68	1.52	1.73	1.29	1.40	1.01	1.53	1.08	1.60	1.10
r1n	−0.54	−0.81	−0.78	−1.14	−0.76	−1.14	−0.76	−1.13	−0.88	−1.30
r_2_5_a	−1.89	−0.84	−1.67	−0.80	−1.41	−0.73	−1.41	−0.74	−1.71	−0.95
r_2_5_n	0.36	0.40	−0.49	−0.53	−1.21	−1.28	−1.16	−1.24	−1.03	−1.14

TABLE 3 Continued.

	NN1		NN2		NN3		NN4		NN5	
Variable	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**	λ_k	***t*-stat**
rev	0.72	0.69	0.57	0.54	0.28	0.28	0.02	0.02	0.24	0.22
rna	0.66	1.51	1.08	2.36	0.95	2.09	0.92	2.02	1.02	2.24
rnaq	0.81	1.61	0.89	1.74	0.77	1.51	0.65	1.28	0.77	1.48
roa	0.72	1.39	0.55	1.03	0.56	1.09	0.45	0.87	0.47	0.91
roe	0.58	1.06	0.53	0.94	0.52	0.94	0.52	0.95	0.52	0.95
rs	**9.93**	**3.88**	**9.93**	**4.70**	**9.96**	**4.83**	**10.12**	**4.90**	**9.82**	**4.74**
sg	−0.52	−0.58	−1.70	−1.84	−1.82	−2.01	−1.94	−2.11	−1.82	−1.96
sgq	2.63	1.37	1.36	0.77	1.20	0.69	0.53	0.29	0.91	0.53
sp	**−0.99**	**−94.15**	**−0.37**	**−35.88**	**−0.43**	**−50.49**	**−0.41**	**−38.16**	**−0.43**	**−47.01**
spq	**1.65**	**133.15**	**1.23**	**85.19**	**1.22**	**84.52**	**1.18**	**79.89**	**1.13**	**77.43**
sr	−0.12	−0.10	−1.08	−0.88	−1.76	−1.46	−1.79	−1.48	−1.67	−1.38
srev	0.93	0.72	0.24	0.21	0.28	0.25	0.04	0.03	0.04	0.03
sue	10.24	3.33	**11.10**	**3.83**	**12.20**	**4.01**	**13.01**	**4.34**	**12.58**	**4.34**
tail	**9.43**	**10.90**	**10.34**	**13.63**	**10.30**	**14.58**	**10.26**	**14.45**	**10.13**	**14.22**
tan	**2.70**	**123.86**	**2.16**	**100.90**	**1.95**	**95.47**	**1.62**	**87.70**	**1.63**	**84.58**
tanq	4.15	1.94	2.24	1.03	2.10	0.99	2.06	0.96	1.96	0.92
tbi	**−3.42**	**−4.35**	**−3.36**	**−4.02**	**−3.40**	**−4.22**	**−3.35**	**−4.13**	**−3.07**	**−3.83**
tbiq	**3.97**	**4.94**	2.06	2.45	1.89	2.30	2.00	2.43	1.98	2.38
tes	**11.11**	**4.72**	5.82	2.58	5.26	2.40	5.21	2.38	5.27	2.42
ts	**−11.91**	**−5.11**	**−10.82**	**−5.13**	**−10.03**	**−4.81**	**−10.39**	**−4.92**	**−10.00**	**−4.74**
tur	−0.62	−1.41	−0.33	−0.71	−0.21	−0.47	−0.11	−0.26	−0.36	−0.83
tv	**−1.95**	**−5.65**	**−1.56**	**−4.31**	**−1.55**	**−4.45**	**−1.37**	**−3.92**	**−1.40**	**−3.99**
w52	**−1.89**	**−4.10**	−1.38	−2.86	−1.44	−3.04	−1.41	−3.02	−1.41	−3.04
wwq	**1.18**	**76.62**	**0.47**	**31.20**	**0.62**	**41.34**	**0.65**	**42.80**	**0.62**	**39.70**

Note: This table reports the marginal contribution of 132 factors in the Chinese stock market relative to the other factors. The test assets include a set of 762 3 × 2 bivariate-sorted portfolios. λ_k represents the average partial effect of each factor using the ADML estimator. ADML estimator uses the LASSO to estimate Riesz representor α_0 and neural networks with one to five units in the hidden layer to estimate γ_0. The LASSO tuning parameters are chosen by the iterative tuning procedure of Chernozhukov et al. (2022) with a maximum number of iterations at 10. Bold font indicates the Bonferroni corrected significant factor with the absolute value of *t*-statistic larger than 3.55.

these firm characteristics and a total of 762 3 × 2 bivariate-sorted portfolios as test assets, following the construction methods of Feng et al. (2020).[10] The sample period is from January 2000 to December 2021.

Table 3 reports the marginal contribution of 132 factors in the Chinese stock market relative to the other factors estimated via the ADML. We uncover some interesting similarities and differences between the Chinese and US stock market. Various factors related to valuation ratios, profitability, and investment are significant in both stock markets. However, momentum and reversal factors, such as momentum factor (UMD) and 36-month momentum (mom36m), are significant in explaining the cross-sectional stock returns in the US stock market. By contrast, momentum-related factors such as prior 6-month returns (r6), prior 11-month returns (r11), and short-term reversal (srev) have insignificant incremental pricing power relative to other factors. These findings are consistent with Li et al. (2010) and Cheung et al. (2015), confirming the weak momentum effect in China's stock market.

In addition, the lottery-related factors, such as idiosyncratic volatility per the Fama and French (1993) three-factor model (ivff) and idiosyncratic skewness per the *q*-factor model (isq), are significant in explaining the cross-sectional expected returns in China's stock market. By contrast, the marginal pricing power of lottery-related factors, such as maximum daily return, is weak in the US stock market. China's stock market is dominated by retail investors rather than institutions (Liu et al., 2019), and the short selling limit exacerbates the limit to arbitrage (Chang et al., 2014), making behavioral mispricing persist. Our findings support the well-documented gambling preference in China's stock market in recent studies (Yao et al., 2019; Zhu et al., 2021).

Sentiment-based factors, such as 1-month abnormal turnover (abtur), have strong marginal pricing power relative to other factors in China's stock market. By contrast, turnover-related factors are insignificant in explaining the cross-section

[10]Specifically, we sort stocks based on these 132 characteristics and then build long–short value-weighted portfolios (top 30%–bottom 30%) as factors. For test assets, we use 3 × 2 portfolios sorted by size and these characteristics, eliminate the portfolio with missing values, and finally obtain a total of 762 portfolios as the test assets.

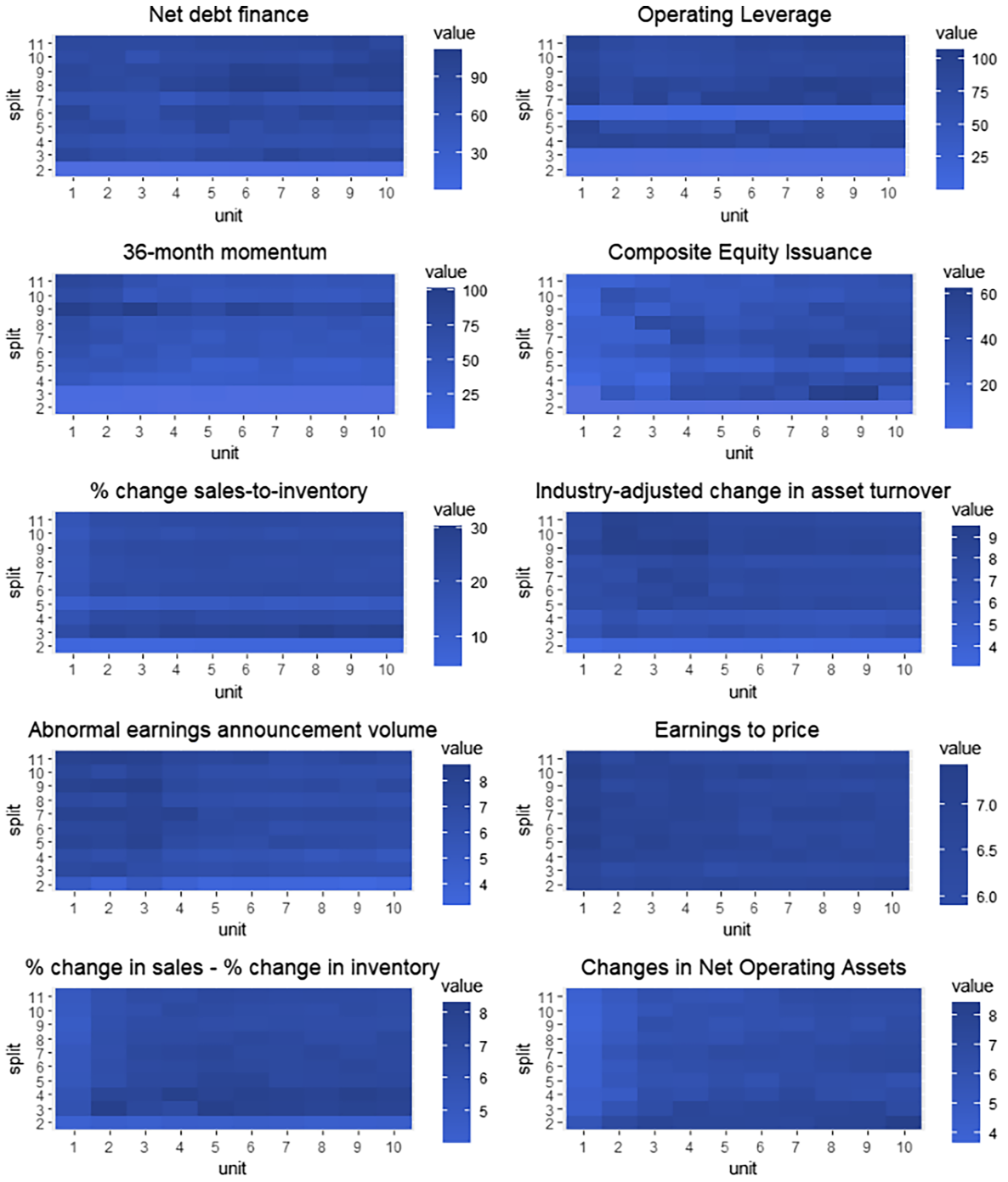

FIGURE 4 Robustness to different parameters: Significant factors. *Note*: This figure reports heat maps for ADML absolute t-statistics of the top 10 significant factors with different tuning parameters and numbers of sample-splitting subsets. The test assets include a set of 750 3×2 bivariate-sorted portfolios. The ADML estimator uses the LASSO estimator with tuning parameters chosen by the iterative tuning procedure of Chernozhukov et al. (2022) to estimate Riesz representor α_0. This figure shows the absolute t-statistics for different units in the hidden layer (complexity) on the X-axis and varying numbers of subsets on the Y-axis. The absolute value of the t-statistic greater than 3 is displayed in blue, and the value close to 0 is displayed in yellow.

of the US stock market. The high turnover comes from optimism toward the stock by sentiment-driven investors (Baker & Stein, 2004). Large sentiment-driven traders and short-sale impediments strengthen the pricing power of turnover-related factors in China's stock market, making the price of the high turnover stock higher than the fundamentals due to high sentiment and lowering its expected return. Our ADML estimator can well identify these factors and support Liu et al. (2019) that turnover is an essential factor in the Chinese stock market.

4.6 | Robustness

4.6.1 | Robustness to the hyperparameter settings

This subsection explores how the ADML estimates vary with the tuning parameters and the number of sample-splitting subsets. Neural networks, which are often used in estimating the risk premia, induce regularization and overfitting bias. Neural networks with more units in the hidden layer signify more complex models and are more likely to face overfitting issues. The ADML estimator applies the sample-splitting method to alleviate this concern, and the number of splitting subsets is a key hyperparameter. Therefore, we explore whether our findings are robust to varying the complexity of the neural networks and the number of splitting subsets.

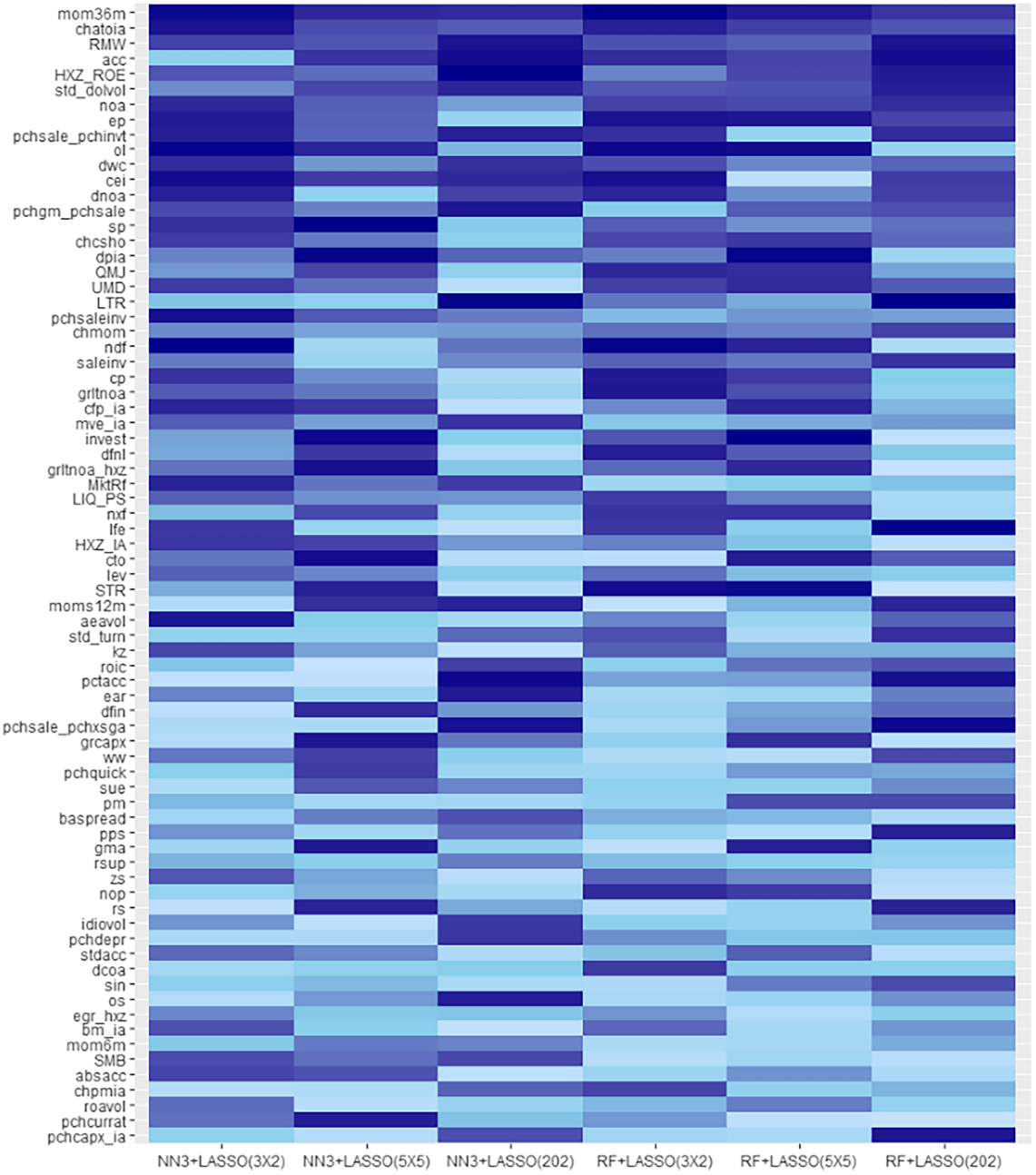

FIGURE 5

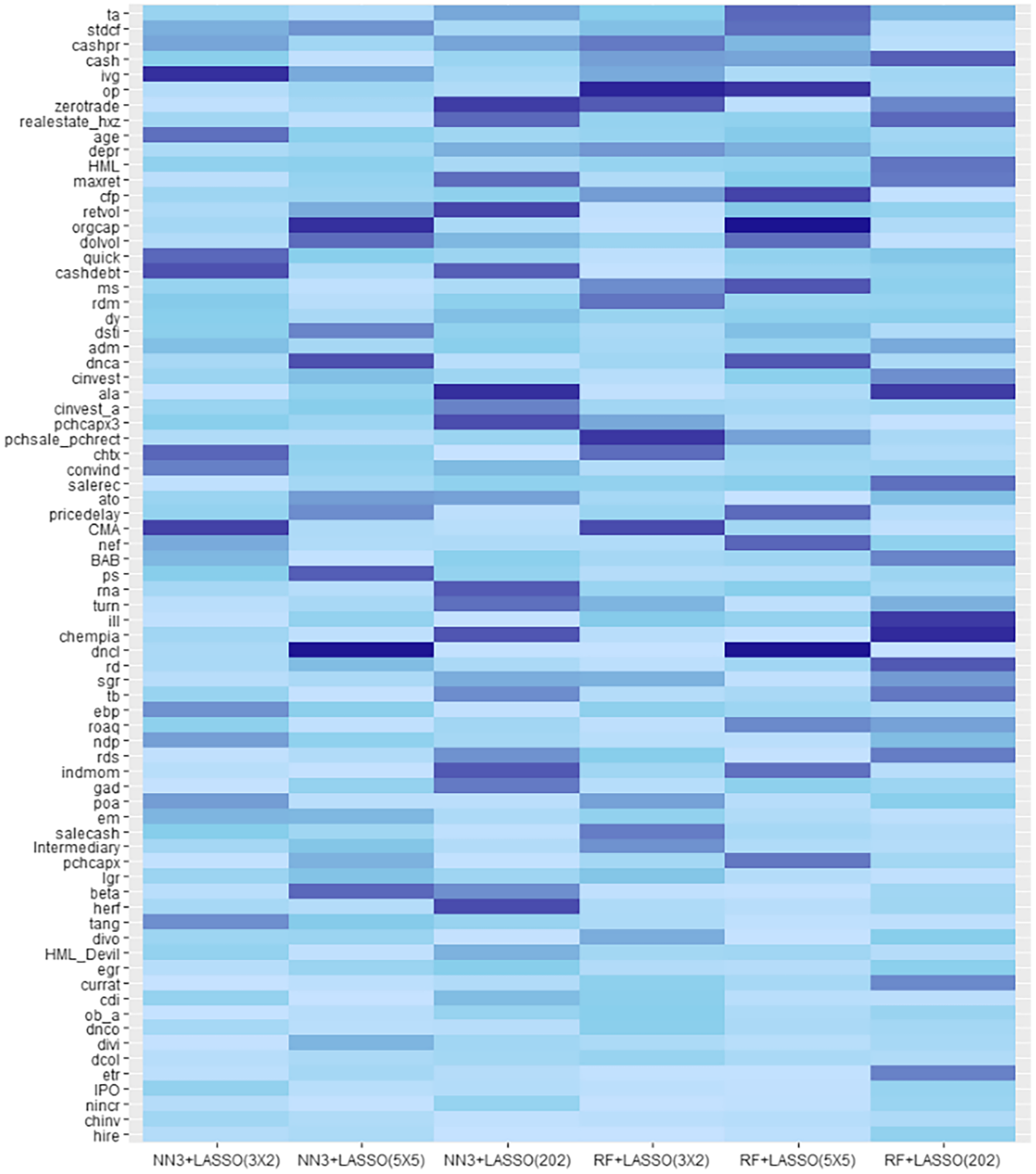

FIGURE 5 Factor importance. *Note*: This figure reports the factor importance with different choices of test assets and machine learning methods. Column (1) reports the results of the neural network with three units in the hidden layer to estimate γ_0 and 3×2 bivariate-sorted portfolios as test assets. In Column (2), we see the results from test assets of 5×5 bivariate-sorted portfolios. And then, Column (3) presents the results of 202 portfolios from Giglio and Xiu (2021) as test assets. Column (4) presents the results with random forests to estimate γ_0 and 3×2 bivariate-sorted portfolios as test assets. Finally, Columns (5) and (6) report the corresponding results for 5×5 bivariate-sorted portfolios and 202 portfolios as test assets. We rank the factor significance of each factor for each column and then sum the ranks. Factors sorted with the highest rankings are at the top, and the lowest rankings are at the bottom. With each column, the lightest to darkest colors show the relative significance from lowest to highest.

Figure 4 reports heat maps for ADML absolute *t*-statistics of the top 10 significant factors with different tuning parameters and a number of subsets.[11] The test assets include a set of 750 3×2 bivariate-sorted portfolios. In order to satisfy the theoretical assumptions, the LASSO tuning parameter is fixed by the iterative tuning procedure of Chernozhukov

[11] The absolute value of *t*-statistic exceeding 3 displays in blue, and the value close to 0 displays in yellow.

et al. (2022) to estimate Riesz representor α_0. This figure shows the absolute t-statistics for different units in the hidden layer (complexity) on the X-axis and a different number of subsets on the Y-axis. The results show that for these significant factors, the statistical inferences are highly robust, which are greater than 3.0 for most pairs of tuning parameters and split subsets.

Overall, our results are highly robust to the complexity of the neural network and the number of sample-splitting subsets, showing the reliability of the ADML method.

4.6.2 | Robustness to choices of test assets and machine learning methods

In this subsection, we investigate the relative importance of factors with different choices of test assets and machine learning methods. Figure 5 reports the factor importance with different specifications. Column (1) reports the results of the neural network with three units in the hidden layer to estimate γ_0 and 3×2 bivariate-sorted portfolios as test assets. Column (2) reports the results from test assets of 5×5 bivariate-sorted portfolios. Column (3) reports the results of 202 portfolios from Giglio and Xiu (2021) as test assets. Column (4) reports the results with random forest to estimate γ_0 and 3×2 bivariate-sorted portfolios as test assets. Columns (5) and (6) report the corresponding results for 5×5 bivariate-sorted portfolios and 202 portfolios as test assets. We rank the factor significance of each factor for each column and then sum the ranks. Factors sorted with the highest rankings are at the top, and the lowest rankings are at the bottom. Each column's lightest to darkest colors denote the relative significance from lowest to highest.

Figure 5 demonstrates that specifications with different test assets and machine learning methods are generally in close agreement regarding the factor importance. The most significant factor can be divided into three groups. The first is the momentum and reversal, such as 36-month momentum (mom36m) and momentum factor (UMD) of Carhart (1997). The second is the valuation ratios or fundamental signals, such as percentage change in sales minus the percentage change in inventory (pchsale_pchinvt), working capital accruals (acc), earnings to price (ep), and so on. The third is the liquidity variables, such as composite equity issuance (cei) and liquidity volatility (std_dolvol).

Overall, although the significance of some factors changes slightly under some specifications, the most significant factors sleeted by the ADML remain robust.

5 | CONCLUSION

Compared with the traditional machine learning methods, ADML resolves machine learners' regularization and overfitting bias and supports standard statistical inference in high-dimensional settings. Applying the methodology to the simulated data, we find that the ADML estimator is unbiased and follows an asymptotic normal distribution. The traditional plug-in estimator, however, suffers from significant bias and distortion from normality.

We identify significant factors one by one using all the other factors as controls. The spanning test shows that the most significant factors selected by ADML outperform the Fama–French sparse factor models at a comparable level of parsimony, verifying that these factors are useful. However, contrary to prior studies, for example, Green et al. (2017), which document few factors with significant pricing power for the cross-sectional stock returns, we find that around 30 to 50 factors have significant pricing power across different periods. The differences mainly come from the limitation of the traditional linear framework, which cannot satisfactorily deal with the high-dimensional data and ignores nonlinearity in asset payoffs. The findings are robust to tuning parameter settings, different test assets, and choice of nonlinear machine learning methods.

Our discovery of a large number of significant factors in the nonlinear SDF setting is also consistent with sparse factors summarizing return predictability. Correlated factors may contain similar information for return prediction, but have distinct effects on the marginal utility of the investors. The sparsity is more prominent during the market booming periods (such as the post-2008 period), when few factors may provide sound return prediction. Different fund philosophies and styles may choose different combinations of significant factors in order to maintain competitive portfolio performance. Further investigation involving factor timing is worth potential future research.

ACKNOWLEDGMENTS

We thank two anonymous referees, Eric Ghysels (the editor), Lauren Cohen, Lin William Cong, Dashan Huang, Kunpeng Li, Yuxin Zhang, conference participants at the 4th International FinTech Research Forum at Renmin University of China and 2023 Guanghua International Symposium on Finance at Peking University, and seminar participants at the

University of Nottingham Ningbo China (UNNC). We acknowledge the financial support from the National Natural Science Foundation of China (Grants 72173127, 72192804, and 72303162) and Beijing Municipal Social Science Foundation (Grant 23JJC029). This research is also supported by the Public Computing Cloud at Renmin University of China and the FinTech Lab at Capital University of Economics and Business. The authors claim equal contribution to this article.

OPEN RESEARCH BADGES

This article has been awarded Open Data Badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. Data is available at http://dx.doi.org/10.15456/jae.2023335.2116565304.

DATA AVAILABILITY STATEMENT

The data and codes that support the findings of this study are publicly available in the Journal of Applied Econometrics Data Archive at https://doi.org/10.15456/jae.

REFERENCES

Adrian, T., Crump, R. K., & Vogt, E. (2019). Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds. *The Journal of Finance*, *74*(4), 1931–1973. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12776

Agarwal, V., & Naik, N. Y. (2004). Risks and portfolio decisions involving hedge funds. *Review of Financial Studies*, *17*(1), 63–98. https://doi.org/10.1093/rfs/hhg044

Asness, C. S., Frazzini, A., & Pedersen, L. H. (2019). Quality minus junk. *Review of Accounting Studies*, *24*(1), 34–112. https://doi.org/10.1007/s11142-018-9470-2

Baker, M., & Stein, J. C. (2004). Market liquidity as a sentiment indicator. *Journal of Financial Markets*, *7*(3), 271–299. https://www.sciencedirect.com/science/article/pii/S1386418103000466

Bansal, R., & Viswanathan, S. (1993). No arbitrage and arbitrage pricing: A new approach. *The Journal of Finance*, *48*(4), 1231–1262. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1993.tb04753.x

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of Finance*, *52*(1), 57–82. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1997.tb03808.x

Chang, E. C., Luo, Y., & Ren, J. (2014). Short-selling, margin-trading, and price efficiency: Evidence from the Chinese market. *Journal of Banking & Finance*, *48*, 411–424. https://www.sciencedirect.com/science/article/pii/ S0378426613003920

Chapman, D. A. (1997). Approximating the asset pricing kernel. *The Journal of Finance*, *52*(4), 1383–1410. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.1997.tb01114.x

Chen, D., Wu, K., & Zhu, Y. (2022). Stock return asymmetry in China. *Pacific-Basin Finance Journal*, *73*, 101757. https://www.sciencedirect.com/science/article/pii/S0927538X2200052X

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2017). Double/debiased/Neyman machine learning of treatment effects. *American Economic Review*, *107*(5), 261–65. https://www.aeaweb.org/articles?id=10.1257/aer.p20171038

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, *90*(4), 1501–1535. https://onlinelibrary.wiley.com/doi/abs/10.3982/ ECTA16294

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, *90*(3), 967–1027. https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA18515

Cheung, C., Hoguet, G. R., & Ng, S. (2015). Value, size, momentum, dividend yield, and volatility in China's A-share market. *The Journal of Portfolio Management*, *41*, 57–70.

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. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2011.01671.x

Daniel, K., & Moskowitz, T. J. (2016). Momentum crashes. *Journal of Financial Economics*, *122*(2), 221–247. https://www.sciencedirect.com/science/article/pii/S0304405X16301490

Dittmar, R. F. (2002). Nonlinear pricing kernels, kurtosis preference, and evidence from the cross section of equity returns. *The Journal of Finance*, *57*(1), 369–403. https://onlinelibrary.wiley.com/doi/abs/10.1111/1540-6261. 00425

Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, *33*(1), 3–56. https://www.sciencedirect.com/science/article/pii/0304405X93900235

Fama, E. F., & French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, *116*(1), 1–22. https://www.sciencedirect.com/science/article/pii/S0304405X14002323

Fama, E. F., & MacBeth, J. D. (1973). Risk, return, and equilibrium: Empirical tests. *Journal of Political Economy*, *81*(3), 607–636. https://doi.org/10.1086/260061

Feng, G., Giglio, S., & Xiu, D. (2020). Taming the factor zoo: A test of new factors. *The Journal of Finance*, *75*(3), 1327–1370. https://onlinelibrary.wiley.com/doi/abs/10.1111/jofi.12883

Frazzini, A., & Pedersen, L. H. (2014). Betting against beta. *Journal of Financial Economics*, *111*(1), 1–25. https://www.sciencedirect.com/science/article/pii/S0304405X13002675

Freyberger, J., Neuhierl, A., & Weber, M. (2020). Dissecting characteristics nonparametrically. *Review of Financial Studies*, *33*(5), 2326–2377. https://doi.org/10.1093/rfs/hhz123

Fung, W., & Hsieh, D. A. (2001). The risk in hedge fund strategies: Theory and evidence from trend followers. *Review of Financial Studies*, *14*(2), 313–341. https://doi.org/10.1093/rfs/14.2.313

Giglio, S., & Xiu, D. (2021). Asset pricing with omitted factors. *Journal of Political Economy*, *129*(7), 1947–1990. https://doi.org/10.1086/714090

Green, J., Hand, J. R. M., & Zhang, X. F. (2017). The characteristics that provide independent information about average U.S. monthly stock returns. *Review of Financial Studies*, *30*(12), 4389–4436.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, *33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Harvey, C. R., Liu, Y., & Zhu, H. (2016). ... and the cross-section of expected returns. *Review of Financial Studies*, *29*(1), 5–68. https://doi.org/10.1093/rfs/hhv059

Harvey, C. R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, *55*(3), 1263–1295. https://onlinelibrary.wiley.com/doi/abs/10.1111/0022-1082.00247

He, Z., Kelly, B., & Manela, A. (2017). Intermediary asset pricing: New evidence from many asset classes. *Journal of Financial Economics*, *126*(1), 1–35. https://www.sciencedirect.com/science/article/pii/S0304405X1730212X

Hou, K., Mo, H., Xue, C., & Zhang, L. (2018). Which factors? *Review of Finance*, *23*(1), 1–35. https://doi.org/10.1093/rof/rfy032

Hou, K., Xue, C., & Zhang, L. (2015). Digesting anomalies: An investment approach. *Review of Financial Studies*, *28*(3), 650–705. https://doi.org/10.1093/rfs/hhu068

Hou, K., Xue, C., & Zhang, L. (2020). Replicating anomalies. *Review of Financial Studies*, *33*(5), 2019–2133. https://doi.org/10.1093/rfs/hhy131

Jagannathan, R., & Korajczyk, R. A. (1986). Assessing the market timing performance of managed portfolios. *The Journal of Business*, *59*, 217–235.

Kozak, S., Nagel, S., & Santosh, S. (2020). Shrinking the cross-section. *Journal of Financial Economics*, *135*(2), 271–292. https://www.sciencedirect.com/science/article/pii/S0304405X19301655

Lewellen, J. (2015). The cross-section of expected stock returns. *Critical Finance Review*, *4*(1), 1–44. https://ideas.repec.org/a/now/jnlcfr/104.00000024.html

Lewellen, J., Nagel, S., & Shanken, J. (2010). A skeptical appraisal of asset pricing tests. *Journal of Financial Economics*, *96*(2), 175–194. https://www.sciencedirect.com/science/article/pii/S0304405X09001950

Li, B., Qiu, J., & Wu, S. (2010). Momentum and seasonality in Chinese stock markets. *Journal of Money, Investment and Banking*, *17*, 24–36. https://eprints.qut.edu.au/38422/

Liu, J., Stambaugh, R. F., & Yuan, Y. (2019). Size and value in China. *Journal of Financial Economics*, *134*(1), 48–69.

Pástor, L., & Stambaugh, R. F. (2003). Liquidity risk and expected stock returns. *Journal of Political Economy*, *111*(3), 642–685. https://doi.org/10.1086/374184

Yao, S., Wang, C., Cui, X., & Fang, Z. (2019). Idiosyncratic skewness, gambling preference, and cross-section of stock returns: Evidence from China. *Pacific-Basin Finance Journal*, *53*, 464–483. https://www.sciencedirect.com/science/article/pii/S0927538X1830180X

Zhu, H. B., Zhang, B., & Yang, L. H. (2021). The gambling preference and stock price: Evidence from China's stock market. *Emerging Markets Review*, *49*, 100803. https://www.sciencedirect.com/science/article/pii/S156601412100011X

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of the article.

How to cite this article: Maasoumi E., Wang J., Wang Z., & Wu K. (2024). Identifying factors via automatic debiased machine learning. *Journal of Applied Econometrics*, *39*(3), 438–461. https://doi.org/10.1002/jae.3031