

HIGH-DIMENSIONAL PREDICTIVE REGRESSION WITH MIXED
INTEGRATION ORDERS AND ENDOGENEITY

by

Alien Shrestha

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Applied Mathematics (Statistics Concentration)

Charlotte

2026

Approved by:

Dr. Taufiquar Khan

Jiancheng Jiang

Shaoyu Li

Eliana Christou

Weidong Tian

ABSTRACT

ALIEN SHRESTHA. High-dimensional predictive regression with mixed integration orders and endogeneity. (Under the direction of DR. TAUFUQUAR KHAN)

Predictive regression models are a standard tool for forecasting asset returns using lagged state variables. Existing approaches face three limitations: (1) they typically examine predictors individually rather than jointly, (2) they depend upon the order of integration of the state variables, which is often unknown, and (3) they do not adequately address endogeneity in high-dimensional settings. We propose a unified error projection-based framework that handles multiple predictors with mixed integration orders without requiring prior knowledge of their stationarity or endogeneity status. Our methodology extends weighted estimation equations to high-dimensional settings by incorporating regularization to address the dimensionality inflation inherent in endogeneity correction. We show that, with probability approaching one, the penalized estimating equation has a root that identifies the nonzero coefficients, that these coefficients converge at rates depending on the order of integration of the corresponding predictor, and that they are asymptotically normal after normalization, with the number of predictors allowed to grow exponentially with the sample size. We also show that the threshold criterion used to classify predictors in the univariate setting degenerates with several predictors, and select the threshold by a largest-gap rule instead. We develop algorithms that integrate coordinate descent optimization with random weighted bootstrap inference to provide parameter estimates and confidence regions. Simulation studies show good performance in both low- and high-dimensional scenarios. An empirical application to 50 years of stock return data using eight financial and macroeconomic predictors across three economic periods illustrates the approach and shows that joint modeling yields different conclusions about return predictability than single-variable methods.

DEDICATION

To my parents Hari Kumar Shrestha, Radha Shrestha, and my brother Alish Shrestha for their unconditional love and support.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to Dr. Jiancheng Jiang for his continuous guidance, support, and mentorship throughout my doctoral journey. It is in large part because of his dedication as a mentor that I have reached this stage. I will always remember his readiness, availability, and patience in tackling even the smallest problems during the early years of my program. Even after his relocation to China, he worked closely with the department to ensure my transition proceeded smoothly and that I was shielded from any administrative burden the move might have caused. He remained highly active following his relocation, regularly checking in on my progress and offering constant feedback and advice on my research that was instrumental in bringing this dissertation to completion. Beyond the dissertation itself, he has also been a generous source of encouragement and advice as I pursue my postdoctoral career. I am deeply grateful for his mentorship and for the time and effort he has invested in my development throughout my doctoral studies.

I would also like to express my deepest gratitude to Dr. Taufiqar Khan for his gracious support and mentorship during my doctoral journey. Dr. Khan has been a tremendously positive influence ever since I began my doctoral journey at UNC Charlotte. Beyond his guidance and feedback on academic and research matters, his willingness to listen to and acknowledge my struggles outside of academics helped me build the mental and psychological fortitude that proved so crucial as I began my doctoral research and other endeavors. He has also been an instrumental figure in my postgraduate career, offering constant support, words of encouragement, and valuable feedback on my candidacy profile for future ventures.

Furthermore, I am deeply grateful to have had Dr. Shaoyu Li, Dr. Eliana Christou, and Dr. Weidong Tian serve as committee members for my dissertation. Their gracious support and feedback have allowed me to complete my research and this dissertation to the best of my ability. I am also grateful to the Department of Mathematics

and Statistics and its faculty members for providing the academic and financial support I needed throughout my doctoral journey, through their graduate assistantship program. In particular, I would like to express my sincere thanks to Dr. Shaozhong Deng and Dr. Duan Chen for their timely support and assistance with all the administrative needs and concerns a PhD student faces along the way. I am also deeply grateful to UNC Charlotte's EPIC (Energy Production and Infrastructure Center) for supporting me through a research assistantship during the first year and a half of my doctoral studies.

Finally, I would like to acknowledge my family and friends for their love, support, and patience during this journey. In particular, I want to thank my family for their unconditional love and unwavering patience, which became the bedrock of my willpower and determination to finish this journey. I would also like to thank my friends, especially Radhakrishna Adhikari, Susan Subedi, Sushant Baskota, and Sakul Mahat Chettri, for their unwavering support and constant encouragement, which propelled me to reach this stage as I close this chapter of my life and begin a new one. Last but not least, I would like to express my deepest thanks to everyone who has wished me well on this journey so far and in the future.

TABLE OF CONTENTS

LIST OF TABLES	xi
LIST OF ABBREVIATIONS	xii
CHAPTER 1: INTRODUCTION	1
1.1. Background	2
1.1.1. Time Series Regression and Predictive Models	2
1.1.2. Stationarity and the Order of Integration	3
1.1.3. Endogeneity in Predictive Regressions	4
1.1.4. High-Dimensional Regression and Regularization	5
1.2. Related Literature and the Gap Addressed by This Dissertation	7
1.3. The Approach of This Dissertation	9
1.4. Contributions of the Dissertation	10
1.5. Organization of the Dissertation	12
CHAPTER 2: MODEL FORMULATION AND REGULARIZATION	13
2.1. Model Formulation	13
2.1.1. Least Squares Estimation	15
2.1.2. High-Dimensional Weighted Estimation Equation (HD-WEE)	16
2.1.3. Selecting the Classification Threshold	18
2.2. Model Regularization	21
2.2.1. Adaptive LASSO Regularization	22
2.2.2. SCAD Regularization	24
2.2.3. Implementation	26

CHAPTER 3: ASYMPTOTIC THEORY	28
3.1. Setting and notation	29
3.2. Primitive conditions	30
3.3. The estimator as a sparse root of the penalized estimating equation	34
3.4. Sparsity and block-wise rates	35
3.5. Oracle asymptotic normality	37
CHAPTER 4: RANDOM WEIGHTED BOOTSTRAP AND SIMULATIONS	39
4.1. Random Weighted Bootstrap	39
4.1.1. Bootstrap Procedure	40
4.1.2. Confidence Interval Construction	43
4.2. Simulations	44
4.2.1. Simulation Design	45
4.2.2. Selecting the Threshold	47
4.2.3. Baseline Results	49
4.2.4. High-Dimensional Results	52
CHAPTER 5: EMPIRICAL APPLICATION: EXCESS STOCK RETURN PREDICTABILITY	56
5.1. Methodological Framework	56
5.2. Data	57
5.3. Empirical Implementation	58
5.4. Empirical Results	60
5.5. Comparison with Single-Variable Analysis	63

CHAPTER 6: CONCLUSION AND FUTURE WORK	66
6.1. Summary of Contributions	66
6.2. Limitations	68
6.3. Future Work	68
APPENDIX A: EXTERNAL RESULTS USED	76
APPENDIX B: VERIFICATION OF $(P4^{**})$: STATIONARY DESIGN	78
APPENDIX C: VERIFICATION OF $(P4^{**})$: CROSS-GRAM BOUNDS, MIXED DESIGN	80
APPENDIX D: VERIFICATION OF $(P4^{**})$: THE SCORE EVENT, MIXED DESIGN	85
APPENDIX E: PROOFS OF THE SPARSITY AND ORACLE NOR- MALITY THEOREMS	87

LIST OF TABLES

TABLE 4.1: Classification by rule for selecting the threshold c^*	47
TABLE 4.2: Baseline simulation: estimates of β and bootstrap coverage	49
TABLE 4.3: Baseline simulation: estimates of η and variable selection	51
TABLE 4.4: High-dimensional simulation: estimates of β and variable selection	52
TABLE 4.5: High-dimensional simulation: estimates of η	53
TABLE 4.6: High-dimensional simulation: coverage of five bootstrap intervals	54
TABLE 5.1: Classification of the predictors and estimated persistence, by period	59
TABLE 5.2: Return predictability across three economic periods	62
TABLE 5.3: Correlations and variance inflation factors of the predictors, by period	65

LIST OF ABBREVIATIONS

LASSO	Least Absolute Shrinkage and Selection Operator
OLS	Ordinary Least Squares
SCAD	Smoothly Clipped Absolute Deviation
WEE	Weighted Estimating Equation

CHAPTER 1: INTRODUCTION

Predictive regression is a tool often used for exploring the predictive ability of state variables to forecast asset returns, financial variables and other economic trends [Rapach and Zhou, 2013]. Consider a predictive regression model:

$$y_t = \mathbf{x}_t^\top \boldsymbol{\beta} + \epsilon_t$$

where $\mathbf{x}_t$ is a vector of state (i.e., predictor) variables, $\boldsymbol{\beta}$ is the parameter vector that projects the relationship of the state variables onto the response variable y_t , and ϵ_t represents the time series error. The objective of a predictive regression model is to estimate and provide inference on these parameters, which is then used to conduct forecasts and predictions on the response variable.

Section 1.1 builds up the statistical background needed to state that objective precisely and to see why it is harder than it looks: what distinguishes a time series regression from an ordinary one (Section 1.1.1), what it means for a predictor to be stationary or non-stationary and why the distinction matters (Section 1.1.2), what endogeneity is and why it routinely contaminates predictive regressions (Section 1.1.3), and why estimation with many predictors requires regularization (Section 1.1.4). Readers already fluent in these topics may wish to skip ahead to Section 1.2, which reviews the existing approaches and the gap they leave. Section 1.3 describes how this dissertation fills that gap, Section 1.4 lists its contributions, and Section 1.5 outlines the remaining chapters.

1.1 Background

1.1.1 Time Series Regression and Predictive Models

Regression analysis, in its most general form, relates a response variable to a set of predictors through their conditional expectation, and estimates the relationship by minimizing a loss function — typically squared error — between the fitted and observed values of the response. In a *cross-sectional* regression, observations are exchangeable: there is no natural ordering among firms, households, or individuals in a sample, and predictors are contemporaneous with the response they explain. A *time series* regression differs in two essential ways. First, observations are ordered in time and generally dependent: today’s value of a macroeconomic or financial series is correlated with yesterday’s, so the errors ϵ_t and often the predictors themselves can no longer be treated as independent draws. Second, and particularly in the *predictive* regressions that are the subject of this dissertation, the predictors $\mathbf{x}_t$ are typically lagged relative to the response: $\mathbf{x}_t$ (or, in some notations, $\mathbf{x}_{t-1}$) is observable at the time the forecast of y_t is made, so that β answers the question “how much does information available today move tomorrow’s outcome,” rather than “how do two variables co-move at the same point in time.” This forecasting orientation is what gives the model its name and its appeal in finance and macroeconomics, where the practical objective is rarely to explain y_t after the fact but to forecast it in advance [Rapach and Zhou, 2013].

Under the classical time series regression assumptions — predictors that are stationary and (strictly or weakly) exogenous, together with a well-behaved error process — ordinary least squares (OLS) applied to the model above remains consistent and asymptotically normal, and standard confidence intervals and hypothesis tests behave as expected. The remainder of this section shows why those classical assumptions are frequently violated in economic and financial data.

1.1.2 Stationarity and the Order of Integration

A time series is (weakly) *stationary* if its mean, variance, and autocovariance structure do not change over time: it fluctuates around a fixed level and its statistical behavior in one decade looks like its behavior in another. A convenient way to see how stationarity can fail is through a first-order autoregressive, or AR(1), model for a single predictor,

$$x_t = \rho x_{t-1} + u_t,$$

where u_t is a mean-zero shock. If $|\rho| < 1$, the effect of any past shock u_{t-k} on x_t decays geometrically at rate ρ^k : the series reverts toward its mean, has a constant, finite variance, and is said to be *stationary*, or *integrated of order zero*, written I(0). If instead $\rho = 1$, the model becomes a random walk, $x_t = x_{t-1} + u_t$, so that x_t is simply the running sum $\sum_{j \leq t} u_j$ of all past shocks: nothing decays, every shock has a permanent effect, and the variance of x_t grows linearly with t rather than settling down. Such a series is *non-stationary*; because it must be differenced once, $x_t - x_{t-1} = u_t$, to arrive at a stationary series, it is said to be *integrated of order one*, written I(1). Values of ρ close to but strictly less than one — *near-unit-root* or NI(1) processes — behave like a blend of the two, reverting so slowly that, in finite samples, they can be difficult to tell apart from a true unit root.

Traditionally, regression analysis begins with testing for the stationarity of the state variables, as many estimation and inference procedures rely on this property. ARIMA regression models, for example, require stationarity, and for non-stationary variables, procedures such as differencing are applied to achieve stationarity. To determine which variables are stationary, economists often rely on both economic theory and statistical tests [Dickey and Fuller, 1979]. Even with these tools, however, uncertainty remains regarding the true integration order of many economic time series. For example, variables such as the unemployment rate [Clark, 1989, Nelson and Plosser, 1982,

Song and Wu, 2007], exchange rates [Roll, 1979, Adler and Lehmann, 1983, Mark, 1990, Chowdhury and Sdogati, 1993], and stock prices [Phillips and Lee, 2013, 2016] have been argued to follow both stationary ($I(0)$) and non-stationary ($I(1)$) processes in the literature [Cavanagh et al., 1995].

This distinction is not a mere technicality. While the order of integration does not pose a significant problem for parameter estimation — OLS remains consistent for β regardless — inference relies heavily on the stationary nature of the predictors, because standard OLS assumptions break down when regressors contain unit root $I(1)$ processes. When predictors contain unit roots, OLS estimators no longer converge to Normal distributions but rather to functionals of Brownian motion, rendering standard confidence intervals and hypothesis tests invalid [Phillips, 1987, Park and Phillips, 1988]. Furthermore, standard bootstrap methods are also inconsistent in the presence of mixed integration orders: while they improve finite-sample performance for stationary predictors, they remain inconsistent at the unit root boundary [Park and Phillips, 1988, Basawa et al., 1991]. A predictive regression that mixes $I(0)$, $I(1)$, and $NI(1)$ predictors — the empirically realistic case, given the examples above — therefore cannot rely on a single, one-size-fits-all asymptotic theory.

1.1.3 Endogeneity in Predictive Regressions

In addition to the challenges posed by non-stationarity, endogeneity in the model further complicates estimation and inference. Endogeneity refers to the correlation between a predictor variable and the error term, and it can arise for several familiar reasons — an omitted variable that drives both the predictor and the response, simultaneity between the two, or measurement error in the predictor. In predictive regressions specifically, endogeneity often arises mechanically: a persistent predictor’s own innovations u_t can be correlated with the return innovation ϵ_t , a phenomenon commonly known as *Stambaugh bias* [Stambaugh, 1999]. Whatever its source, endogeneity leads to biased and inconsistent parameter estimates under standard es-

timization methods. This issue is common in predictive models, particularly in fields such as corporate finance and accounting [Roberts and Whited, 2012]. Approaches to correcting it are reviewed in Section 1.2.

1.1.4 High-Dimensional Regression and Regularization

The discussion so far has considered a single predictor. Empirically, however, researchers rarely believe that only one state variable matters, and a joint model of many predictors is the natural response. As the number of predictors p grows, ordinary least squares deteriorates: its variance grows with p , and once p is comparable to or larger than the sample size n , the design matrix $\mathbf{X}^\top \mathbf{X}$ becomes ill-conditioned or singular and OLS is no longer even well defined.

Regularization, or *penalized regression*, addresses this by adding a penalty on the size of $\boldsymbol{\beta}$ to the least-squares objective,

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta}} \sum_t (y_t - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \sum_j p_\lambda(|\beta_j|),$$

where $p_\lambda(\cdot)$ is a penalty function and λ a tuning parameter controlling its strength. Introducing a penalty trades a small amount of bias for a large reduction in variance, which generally improves out-of-sample prediction; when the penalty is chosen to be sparsity-inducing, it additionally performs automatic variable selection, shrinking the coefficients on unimportant predictors exactly to zero. Ridge regression [Hoerl and Kennard, 1970] uses the penalty $p_\lambda(|\beta_j|) = \lambda \beta_j^2$: this stabilizes estimation under collinearity and high dimensionality but shrinks every coefficient smoothly toward zero without ever setting one to exactly zero, so it does not perform variable selection. The Least Absolute Shrinkage and Selection Operator (LASSO) of Tibshirani [1996] instead uses the ℓ_1 penalty $p_\lambda(|\beta_j|) = \lambda |\beta_j|$, whose kink at zero does set coefficients exactly to zero, combining estimation and selection in a single convex optimization problem. The convenience of convexity comes at a cost, however: the same constant

penalty that correctly zeroes out noise coefficients also shrinks every true, nonzero coefficient toward zero by roughly λ , so LASSO estimates of the active coefficients are biased even asymptotically, and consistent selection under correlated designs requires restrictive conditions on the predictors. The *adaptive LASSO* of Zou [2006] addresses the bias problem by reweighting the penalty using an initial consistent estimate, $p_\lambda(|\beta_j|) = \lambda w_j |\beta_j|$ with weights w_j inversely related to that initial estimate, so that large coefficients are penalized less and small ones more; under suitable conditions this restores selection consistency and asymptotically unbiased estimation of the nonzero coefficients in fixed-dimensional settings.

The smoothly clipped absolute deviation (SCAD) penalty of Fan and Li [2001] pursues the same goal — an unbiased estimator of the large, truly active coefficients — through a different route: rather than reweighting a convex penalty, SCAD is itself non-convex. It behaves like the LASSO penalty (quadratic, in the sense of matching its derivative) near zero, transitions through a region penalizing moderately sized coefficients, and then flattens out entirely for coefficients beyond a threshold, so that once a coefficient is judged confidently nonzero, the penalty stops pulling it toward zero at all. An estimator is said to possess the *oracle property* if it (i) selects the correct set of nonzero coefficients with probability tending to one, and (ii) estimates those nonzero coefficients as efficiently, asymptotically, as if the true sparsity pattern were known in advance — that is, as if the researcher could have simply run OLS on the correct, much smaller subset of predictors from the start. Fan and Li [2001] show that SCAD achieves the oracle property in fixed-dimensional settings, and Fan and Lv [2011] extend the result to high-dimensional, NP-dimensional designs where the number of predictors d may grow (sub-)exponentially in the sample size n . Both results, however, are established for stationary designs whose estimating equation is the gradient of a scalar objective function.

1.2 Related Literature and the Gap Addressed by This Dissertation

The theoretical literature is aware of these shortcomings. Campbell and Yogo [2006] address endogeneity for a single predictor by constructing a Bonferroni confidence interval/hypothesis test method that is resistant to endogeneity-induced bias. However, it does not correct the estimation itself and is only valid for inference under a single predictor, and assumes that the innovations (ϵ_t, u_t) are jointly normally distributed. Gippel et al. [2014] also highlight the limitations of conventional approaches to fixing endogeneity. It surveys and unifies standard tools such as 2SLS, IV, and differenced/system GMM into a joint framework, discusses their limitations, and ultimately proposes natural experiments as a solution. Amihud and Hurvich [2004], Amihud et al. [2009], and Cai and Wang [2014] propose a linear projection and least squares estimation approach that removes endogeneity by projecting ϵ_t onto the innovations of the predictor process. The method projects the error as a linear combination of the innovation of the single predictor variable and a white noise term; least squares estimation is then applied to the original model, which is now valid since the error has been transformed from a correlated error into a white noise error. However, like the previous methods discussed, it fails to accommodate joint predictability from multiple state variables at once.

Despite the extensive body of work on endogeneity surveyed above, and the equally extensive literature on non-stationarity and unit roots discussed in Section 1.1.2, few studies address the estimation and inference problem when endogenous regressors include a mix of stationary ($I(0)$), near-integrated ($NI(1)$), and non-stationary ($I(1)$) processes. Hong et al. [2024] propose weighted estimation equation (WEE) and random weighted bootstrap methods to perform inference in a unified framework incorporating both $I(0)$ and $I(1)$ variables. To be more precise, the paper proposes a data-dependent weighting scheme applied to the estimating equation. Unlike a traditional weighted least squares approach, where the weights are fixed through-

out the sample, Hong et al. [2024]’s weighting scheme is dynamic: it is recalculated for each observation based on the stationarity classification of each variable. The resulting estimator is shown to converge to the true parameter vector regardless of each variable’s integration order; however, because its asymptotic variance takes an unknown, non-standard form, Hong et al. [2024] implement a random weighted bootstrap for inference rather than a closed-form variance estimator. The full framework thus provides a unifying estimation and inference procedure for a single state variable regardless of its stationary nature. Its limitation is the single variable: the approach cannot assess the joint predictability of multiple state variables and may suffer from modeling bias due to missing important regressors, and although Hong et al. [2024] show how to handle endogeneity in single-predictor models using linear projection, extending this to multiple predictors creates substantial computational and methodological challenges.

Recent work on LASSO for high-dimensional predictive regression with unit-root regressors establishes consistency: Mei and Shi [2024] considers a high-dimensional regression with a mix of stationary, non-stationary, and cointegrated predictors under the LASSO framework, deriving new deviation-bound and restricted-eigenvalue conditions tailored to the unit-root setting. However, Mei and Shi [2024] establishes only estimation consistency; it does not establish the asymptotic oracle property needed for inference. More recently, Gao et al. [2026] proposed the XDLASSO estimator, which provides valid inference through a desparsified LASSO that corrects for both shrinkage bias and the Stambaugh bias induced by nonstationary regressors. This framework combines Phillips’s IVX self-generated-instrument approach with desparsified-LASSO debiasing and, like our proposed model, does not require prior knowledge of each variable’s stationarity. Both papers, however, treat the stationary predictors as exogenous and focus on estimation and inference for a LASSO fit, and Gao et al. [2026] target asymptotic normality for a coefficient of interest

rather than oracle-property sparse recovery. Recent advances in post-selection inference [Lee et al., 2016, Tibshirani et al., 2016] provide valid inference after variable selection in fixed, stationary settings, but existing implementations do not accommodate integrated regressors or the mixed-order environments common in financial applications.

Taken together, each strand of this literature handles part of the problem. The projection methods correct for endogeneity but only for a single predictor; Hong et al. [2024] accommodate mixed orders of integration, again for a single predictor; and the high-dimensional LASSO methods allow many predictors with unit roots but treat the stationary predictors as exogenous. No existing method delivers estimation and inference for a predictive regression with many predictors of unknown and mixed integration orders whose innovations may be correlated with the regression error. That is the gap this dissertation addresses.

1.3 The Approach of This Dissertation

We extend the univariate framework of Hong et al. [2024] to a multivariate predictive regression whose predictors may be stationary, near-integrated or integrated, without requiring the order of integration of any predictor to be known. Following Amihud and Hurvich [2004], we project the regression error onto the innovations of every predictor. This adds the first differences $\Delta x_{i,t}$ as regressors and leaves an error uncorrelated with the predictor innovations. The purpose of the projection is not to determine which predictors are endogenous. It is to remove the bias that correlated innovations would otherwise induce, so that the predictive coefficients can be estimated, and inference carried out, without assuming which predictors are exogenous.

The price of the projection is dimensionality: it at least doubles the number of parameters, from $p + 1$ to $2p + 1$. We therefore estimate the projected model by weighted estimating equations penalized with SCAD [Fan and Li, 2001], solved by coordinate descent, and compare it with the Adaptive LASSO [Tibshirani, 1996, Zou,

2006] and with unweighted OLS. The penalty sets unneeded correction terms to zero and so keeps the number of estimated parameters manageable. Retaining a correction term that is not needed costs some precision but introduces no bias, so the penalty economizes on parameters rather than determining the validity of the estimates. We use SCAD rather than the LASSO because its flat region leaves sufficiently large coefficients unpenalized, which is what the oracle property requires.

The weighting of Hong et al. [2024] depends on a threshold c^* that classifies each predictor as stationary or persistent. The criterion Hong et al. [2024] use to choose it degenerates once there are several predictors, and block cross-validation, the natural alternative for time series, over-selects; we locate the threshold instead at the largest gap in the sorted classification statistics. For inference we extend the random weighted bootstrap of Hong et al. [2024] to the penalized, high-dimensional estimator and construct confidence intervals using the basic bootstrap interval [Efron and Tibshirani, 1993, DiCiccio and Efron, 1996], which makes no normality assumption and performs best of the five constructions we compare in simulations.

The oracle theory for SCAD in Fan and Li [2001] and Fan and Lv [2011] does not apply to this estimator. The regressors mix stationary and integrated components, and because the weights are coordinate-specific the weighted estimating equation is not the gradient of any objective function. We therefore characterize the estimator directly as a sparse root of the penalized estimating equation and establish its asymptotic properties following the D_n -normalization approach of Hong et al. [2021] and Hong et al. [2024]. Finally, we apply the approach to the joint predictability of U.S. stock returns by eight financial and macroeconomic predictors over three economic periods since 1969 [Lettau and Van Nieuwerburgh, 2008], where it leads to conclusions that differ from the single-variable analysis of Hong et al. [2024].

1.4 Contributions of the Dissertation

The contributions of this dissertation are:

1. **A multivariate framework for mixed integration orders and endogeneity.** We extend the univariate weighted estimating equation approach of Hong et al. [2024] to multiple predictors, with an error-projection correction for every predictor and SCAD regularization of the resulting $2p + 1$ parameters, so that neither the order of integration nor the endogeneity of any predictor needs to be specified in advance.
2. **A threshold rule for the multivariate weighting.** We show that the efficiency criterion Hong et al. [2024] use to choose the classification threshold c^* reduces, with several predictors, to the total precision of the estimator, so that it selects no downweighting in every replication of our simulations, and that block cross-validation over-selects. We replace both with a largest-gap rule motivated by the separation of the classification statistic established in Chapter 3, and document its finite-sample performance; a proof of its consistency is left open.
3. **Asymptotic theory for mixed integration orders and a growing number of predictors.** We prove sparsity and *block-wise* convergence rates ($\sqrt{n}$ for stationary active coefficients, n for integrated ones) together with oracle asymptotic normality, with $\log d = O(n^\nu)$. Verifying the underlying concentration condition requires a dual cross-Gram normalization not needed in the univariate setting of Hong et al. [2024].
4. **Bootstrap inference for the regularized, high-dimensional estimator.** We extend the random weighted bootstrap of Hong et al. [2024] to this setting, re-solving the penalized estimating equation in each replicate, and compare five interval constructions. The basic interval gives the most accurate coverage, while intervals centered on the bootstrap mean perform poorly for coefficients the penalty sets to zero. Validity of the bootstrap for the penalized estimator

is supported by simulation rather than proved.

5. **An empirical application to stock return predictability.** Applying the methodology to fifty years of S&P 500 return data across three economic periods, we show that joint and single-variable analyses lead to different conclusions about which predictors carry predictive content. Several joint coefficients have signs opposite to those expected from single-variable analyses, which is consistent with strong correlation among the predictors, so the joint results are best read as statements about sets of related predictors.

1.5 Organization of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 introduces the predictive regression model, the error-projection correction, the weighted estimating equation and its regularization. Chapter 3 develops the asymptotic theory for the resulting estimator, with the supporting concentration-bound verifications in Appendices A–D and the theorem proofs in Appendix E. Chapter 4 develops the random weighted bootstrap and reports simulation evidence under low- and high-dimensional designs, together with a comparison of rules for selecting the threshold c^* . Chapter 5 applies the methodology to fifty years of S&P 500 return data, and Chapter 6 concludes and discusses directions for future work.

CHAPTER 2: MODEL FORMULATION AND REGULARIZATION

This chapter develops the model formulation motivated in Chapter 1. Section 2.1 introduces the predictive regression model and extends the linear-projection-based endogeneity correction of Amihud and Hurvich [2004] to the weighted estimating equation (WEE) framework of Hong et al. [2024], allowing multiple predictors of unknown and possibly mixed integration order to be corrected for endogeneity jointly. Section 2.2 then addresses the resulting increase in dimensionality by incorporating SCAD and adaptive LASSO regularization via coordinate descent.

2.1 Model Formulation

Consider a p -variate predictive regression model with predictor variables comprising both stationary and non-stationary processes:

$$y_t = \beta_0 + \sum_{i=1}^p \beta_i x_{i,t-1} + \epsilon_t, \quad (2.1)$$

for $t = 1, 2, \dots, n$, where y_t is the response variable at time t , and $x_{i,t-1}$ denotes the lagged predictor variables. The predictor process is assumed to follow:

$$x_{i,t} = \rho_i x_{i,t-1} + u_{i,t}, \quad (2.2)$$

where $u_{i,t}$ represents the innovation term with $E[u_{i,t}] = 0$ and finite variance, and ρ_i determines the order of integration of $x_{i,t}$ (stationary if $|\rho_i| < 1$, near-integrated if $\rho_i = 1 + \frac{c_i}{n}$ for some constant c_i , and unit root if $|\rho_i| = 1$). The variable $x_{i,t}$ may represent an endogenous variable, meaning the error term ϵ_t can be correlated with $u_{i,t}$. For example, in asset return prediction, dividend yield may be endogenous if

both returns and dividends respond to common unobserved market sentiment. More formally, $\text{Cov}(\epsilon_t, u_{i,t}) \neq 0$ for some i , depending on the endogenous nature of the predictors.

To address this endogeneity, we follow the linear projection approach developed by Amihud and Hurvich [2004] and apply a linear projection of ϵ_t onto $u_{i,t}$ for $i = 1, \dots, p$:

$$\epsilon_t = \sum_{i=1}^p \eta_i u_{i,t} + v_t, \quad (2.3)$$

where v_t is a white noise term with $E[v_t] = 0$, finite variance, and uncorrelated with all variables and their lags. This projection effectively removes the endogenous components from the error term by decomposing ϵ_t into a part that is correlated with the predictor innovations (captured by the $\eta_i u_{i,t}$ terms) and an orthogonal error v_t .

To derive the estimable form of our model, we rewrite Eq. (2.2) as:

$$u_{i,t} = x_{i,t} - \rho_i x_{i,t-1} = \Delta x_{i,t} - (\rho_i - 1)x_{i,t-1}$$

and substituting this into Eq. (2.1) using Eq. (2.3) yields:

$$\begin{aligned} y_t &= \beta_0 + \sum_{i=1}^p \beta_i x_{i,t-1} + \sum_{i=1}^p \eta_i (\Delta x_{i,t} - (\rho_i - 1)x_{i,t-1}) + v_t \\ &= \beta_0 + \sum_{i=1}^p [\beta_i - \eta_i(\rho_i - 1)]x_{i,t-1} + \sum_{i=1}^p \eta_i \Delta x_{i,t} + v_t. \end{aligned} \quad (2.4)$$

Define $\alpha_i = \beta_i - \eta_i(\rho_i - 1)$ and $\alpha_0 = \beta_0$, where α_i represents the “corrected” coefficient (after adjusting for endogeneity) and η_i captures the strength of the endogenous relationship. The model can be rewritten as:

$$y_t = \alpha_0 + \sum_{i=1}^p \alpha_i x_{i,t-1} + \sum_{i=1}^p \eta_i \Delta x_{i,t} + v_t. \quad (2.5)$$

The transformed model in Eq. (2.5) simultaneously handles endogeneity and mixed

integration orders in a single framework. Importantly, this formulation does not assume all variables are endogenous. If a variable is exogenous, $\text{Cov}(\epsilon_t, u_{i,t}) = 0$, implying $\eta_i = 0$. Thus, if all predictors are exogenous, i.e., $\eta_i = 0$ for all i , Eq. (2.5) simplifies to the traditional predictive regression model:

$$y_t = \alpha_0 + \sum_{i=1}^p \alpha_i x_{i,t-1} + v_t. \quad (2.6)$$

Hence, the formulation in Eq. (2.5) can be viewed as a generalized form of predictive regression that accounts for potential endogeneity while having the traditional approach (where endogeneity is absent) as a special case.

2.1.1 Least Squares Estimation

Given that the transformed model in Eq. (2.5) explicitly accounts for endogeneity via the linear projection, we can estimate it using Ordinary Least Squares (OLS). However, the estimation requires a two-step procedure since the autoregressive parameters ρ_i must be estimated before the main regression.

To begin, let $\mathbf{z}_t$ denote the combined vector of regressors:

$$\mathbf{z}_t = (1, x_{1,t-1}, \dots, x_{p,t-1}, \Delta x_{1,t}, \dots, \Delta x_{p,t})^\top.$$

Further, define the parameter vector $\boldsymbol{\gamma} = (\alpha_0, \alpha_1, \dots, \alpha_p, \eta_1, \dots, \eta_p)^\top$. The OLS estimator $\hat{\boldsymbol{\gamma}}$ is obtained by solving the normal equations:

$$\sum_{t=1}^n \mathbf{z}_t (y_t - \mathbf{z}_t^\top \boldsymbol{\gamma}) = \mathbf{0}. \quad (2.7)$$

The closed-form solution is:

$$\hat{\boldsymbol{\gamma}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \quad (2.8)$$

where $\mathbf{Z}$ is the $n \times (2p+1)$ design matrix with rows $\mathbf{z}_t^\top$, and $\mathbf{y} = (y_1, \dots, y_n)^\top$ is the

response vector.

In practice, we must first estimate the autoregressive parameters ρ_i from Eq. (2.2) separately for each predictor $i = 1, \dots, p$ using OLS on the individual AR(1) regressions. The normal equation, which recognizes we do not have data at time period 0, is:

$$\sum_{t=2}^n x_{i,t-1}(x_{i,t} - \rho_i x_{i,t-1}) = 0.$$

Solving for ρ_i yields:

$$\hat{\rho}_i = \frac{\sum_{t=2}^n x_{i,t-1}x_{i,t}}{\sum_{t=2}^n x_{i,t-1}^2}.$$

After obtaining the autoregressive parameter estimates, we recover the original structural parameters from the relationship $\alpha_i = \beta_i - \eta_i(\rho_i - 1)$:

$$\begin{aligned}\hat{\beta}_0 &= \hat{\alpha}_0 \\ \hat{\beta}_i &= \hat{\alpha}_i + \hat{\eta}_i(\hat{\rho}_i - 1) \quad \text{for } i = 1, \dots, p\end{aligned}$$

The recovered parameters $\hat{\beta}_i$ represent the original predictive coefficients from Eq. (2.1). When a predictor is exogenous ($\hat{\eta}_i = 0$), we have $\hat{\beta}_i = \hat{\alpha}_i$, recovering the standard OLS coefficient. While OLS can handle endogeneity through the linear projection, it fails to deliver valid inference when predictors have mixed orders of integration. The non-standard limiting distributions that arise at the unit root boundary [Phillips, 1987, Park and Phillips, 1988] make standard OLS confidence intervals unreliable, motivating the weighted estimation approach developed by Hong et al. [2024] and extended in the following subsection.

2.1.2 High-Dimensional Weighted Estimation Equation (HD-WEE)

Unlike standard OLS, which treats all observations equally regardless of the persistence properties of the regressors, the HD-WEE method applies adaptive observation weights that downweight non-stationary predictors due to their different asymp-

otic behavior, restoring valid inference under mixed integration orders. Consider the weighted estimation equation:

$$\sum_{t=1}^n \mathbf{\Omega}_t \mathbf{z}_t (y_t - \mathbf{z}_t^\top \boldsymbol{\gamma}) = \mathbf{0}, \quad (2.9)$$

where $\mathbf{z}_t = (1, x_{1,t-1}, \dots, x_{p,t-1}, \Delta x_{1,t}, \dots, \Delta x_{p,t})^\top$ is the regressor vector from Eq. (2.5), and

$$\mathbf{\Omega}_t = \text{diag}(1, w_{1,t}, \dots, w_{2p,t})$$

is a diagonal weight matrix where each element $w_{i,t}$ corresponds to the i -th component of $\mathbf{z}_t$. The weighted estimator for $\boldsymbol{\gamma}$ is given by:

$$\hat{\boldsymbol{\gamma}}_w = \left(\sum_{t=1}^n \mathbf{\Omega}_t \mathbf{z}_t \mathbf{z}_t^\top \right)^{-1} \left(\sum_{t=1}^n \mathbf{\Omega}_t \mathbf{z}_t y_t \right). \quad (2.10)$$

The innovation lies in the weight specification. The weights are defined as:

$$w_{i,t} = \begin{cases} 1, & i \in I, \\ (1 + \|\mathbf{z}_{t,I^c}\|^2)^{-\frac{1}{2}}, & i \in I^c, \end{cases} \quad (2.11)$$

where stationary variables (indexed by set I) receive full weight, while near-integrated and integrated variables (indexed by I^c) receive downweighted observations. The index set I is determined by:

$$I = \left\{ i : n^{-\frac{1}{2}} \log(n) \max_{1 \leq t \leq n} |z_{i,t}| < c^* \right\},$$

which selects variables whose scaled maximum absolute value falls below threshold c^* . Intuitively, non-stationary variables grow without bound and will have large maximum values over the sample, while stationary variables remain bounded. For instance, in the empirical application of Chapter 5, every predictor except inflation

is placed in I^c in each of the three periods under the scale-free form of the statistic described in Section 2.1.3 (Table 5.1).

We estimate the autoregressive parameters with the same weights and the same classification of predictors into I and I^c as in the main HD-WEE estimation. After determining $\hat{\alpha}_i$ and $\hat{\eta}_i$ through $\hat{\gamma}_w$, we estimate ρ_i by solving the weighted least squares equation:

$$\sum_{t=2}^n w_{i,t-1} x_{i,t-1} (x_{i,t} - \rho_i x_{i,t-1}) = 0, \quad (2.12)$$

where the weights $w_{i,t-1}$ are determined via Eq. (2.11).

Finally, we recover the original structural parameters using the same relationship as in the OLS case:

$$\begin{aligned} \hat{\beta}_{0,w} &= \hat{\alpha}_{0,w} \\ \hat{\beta}_{i,w} &= \hat{\alpha}_{i,w} + \hat{\eta}_{i,w}(\hat{\rho}_{i,w} - 1) \quad \text{for } i = 1, \dots, p. \end{aligned} \quad (2.13)$$

The weighting allows the same estimator to be used whatever the order of integration of each $x_{i,t}$, without pre-testing for stationarity, and so permits joint inference across predictors with different degrees of persistence [Hong et al., 2024].

2.1.3 Selecting the Classification Threshold

The threshold parameter c^* governs the boundary between stationary and non-stationary classifications and must be carefully selected. Hong et al. [2024] propose selecting c^* by maximizing estimation efficiency, which is computationally tractable in their univariate setting. The variance matrix of the weighted estimates is

$$\Sigma_{\hat{\gamma}_w} = \mathbf{B}_n^{-1} \mathbf{A}_n \mathbf{B}_n^{-\top} \sigma_v^2, \quad \mathbf{A}_n = \sum_{t=1}^n (\Omega_t \mathbf{z}_t)^{\otimes 2}, \quad \mathbf{B}_n = \sum_{t=1}^n \Omega_t \mathbf{z}_t \mathbf{z}_t^\top,$$

where $a^{\otimes 2} = a a^\top$ for any matrix a . Since σ_v^2 does not depend on c^* , Hong et al. [2024] choose c^* to maximize $\|\mathbf{A}_n^{-1/2} \mathbf{B}_n\|_F$, where $\|\cdot\|_F$ is the Frobenius norm.

In our multivariate extension this criterion degenerates. Writing Σ for the variance matrix $\Sigma_{\hat{\gamma}_w}$ above,

$$\|\mathbf{A}_n^{-1/2}\mathbf{B}_n\|_F^2 = \text{tr}(\mathbf{B}_n^\top \mathbf{A}_n^{-1} \mathbf{B}_n) = \sigma_v^2 \text{tr}(\Sigma^{-1}), \quad (2.14)$$

so the criterion is the total precision of the estimator. With a single predictor this is a scalar and the criterion trades efficiency against robustness as intended. With several predictors of different orders of integration it is a sum over coordinates converging at different rates: by the block-wise rates of Lemma 3.4.2, $(\Sigma^{-1})_{jj}$ is of order n for $j \in I$ and of order n^2 for $j \in I^c$, so the trace is dominated by the integrated coordinates. Downweighting those coordinates is precisely what reduces their precision, and hence the criterion. The maximum is therefore attained at $I^c = \emptyset$, where HD-WEE reduces to OLS, and this is what the criterion selects in every replication of both simulation designs of Section 4.2.2. The obstacle is structural rather than computational.

Expanding window block cross-validation is the natural alternative for dependent data, and avoids the look-ahead bias of standard K -fold cross-validation [Racine, 2000, Bergmeir and Benítez, 2012]. It partitions the series into K contiguous blocks and expands the training set: blocks 1 through k train the model and block $k + 1$ is evaluated, for $k = 1, \dots, K - 1$, and c^* is chosen to minimize the mean squared prediction error averaged over the $K - 1$ evaluation blocks. It is also unsuitable here, for a related reason. Downweighting observations costs prediction accuracy while buying validity, so a criterion minimizing mean squared prediction error is biased against downweighting and selects an empty I^c in 40 per cent of baseline replications. Confining the search to non-degenerate classifications removes that outcome but not the mismatch: in the high-dimensional design the selected classifications are then far too large, averaging nine downweighted predictors out of twenty-five against a truth of two (Section 4.2.2).

We therefore choose c^* in a way that exploits this separation directly. The idea is simple: sort the classification statistics of the p predictors from smallest to largest, find the largest jump between consecutive values, and place the threshold there. Formally, let $T_i = n^{-1/2} \log(n) \max_{t \leq n} |x_{i,t-1}|$ for $i = 1, \dots, p$, with order statistics $T_{(1)} \leq \dots \leq T_{(p)}$, and set

$$k^* = \arg \max_{1 \leq k < p} \frac{T_{(k+1)}}{T_{(k)}}, \quad \widehat{c} = T_{(k^*+1)}, \quad (2.15)$$

so that $I^c = \{i : T_i \geq \widehat{c}\}$ contains the predictors above the largest gap. Because the gap is measured as a ratio rather than a difference, the rule is unaffected by the overall level of the statistics.

The rule is motivated by the behavior of the statistic established in Chapter 3 (Remark 3.2.3). For a stationary predictor, T_i vanishes at rate $n^{-1/2} \log(n) \log(nd)$; for an integrated predictor, it diverges at rate $\log n$. Within each group the statistics are of the same order. The sorted statistics therefore form two clusters, and the ratio between the clusters grows with n while the ratios within them do not, so the largest ratio eventually falls between the two. Unlike a fixed threshold, the rule does not require the location of that boundary to be known in advance. By construction it always places at least one predictor in each of I and I^c , so it cannot return the empty classification that the efficiency criterion selects in every replication of the simulations and that unrestricted cross-validation often selects. The corresponding cost is that it would downweight at least one predictor even if all were stationary.

We do not prove that the rule is consistent. The theory of Chapter 3, like that of Hong et al. [2024], is for a fixed c^* , and shows that any fixed $c^* > 0$ classifies the predictors correctly with probability tending to one. The evidence for the rule itself is therefore its finite-sample performance, reported in Section 4.2.2: it recovers the true classification in 85 and 87 per cent of replications of the two simulation designs,

and places the exact unit root in I^c in 99 per cent.

One practical caveat concerns units. In finite samples T_i depends on the scale of x_i : the same predictor has a maximum one hundred times smaller when measured in decimals than in percent, so when predictors are measured on different scales their ordering by T_i can reflect units as much as persistence. In that case we use the scale-free statistic $T_i/\text{SD}(\Delta x_i)$, which divides by the sample standard deviation of the first differences. This leaves the asymptotic separation intact, since $\text{SD}(\Delta x_i)$ converges to σ_{u_i} for a stationary predictor and is bounded and bounded away from zero in probability for an integrated one. The rescaling affects only the classification; the predictors enter the estimating equation in their original units. In the simulations of Chapter 4, where all innovations have unit variance, the two forms behave similarly and both are reported. In the empirical application of Chapter 5, where log valuation ratios sit alongside interest rates measured in decimals, we use the scale-free form.

2.2 Model Regularization

The linear projection corrects for the correlation between the predictor innovations and the regression error, but it at least doubles the number of parameters, from $p + 1$ in the original model to $2p + 1$ in the transformed model of Eq. (2.5). With the $p = 8$ predictors of our empirical application the transformed model has 17 parameters, and in the high-dimensional simulation design with $p = 25$ it has 51. This inflation causes three problems in finite samples. Many predictors may not need a correction term ($\eta_i = 0$), so including a difference term for every predictor invites overfitting. Estimation becomes numerically unstable, or infeasible, as $2p + 1$ approaches n . And irrelevant parameters reduce estimation efficiency and complicate inference.

We therefore regularize both the OLS and the HD-WEE estimators, shrinking unneeded correction terms η_i and irrelevant level coefficients α_i to zero. The aim is to keep the number of estimated parameters manageable, not to classify the predictors as endogenous or exogenous. Retaining a correction term whose true value is zero

costs some precision but does not bias the estimates of the remaining coefficients, whereas setting a needed correction term to zero would. The regularization therefore matters for efficiency, and the simulations of Section 4.2 check that it does not discard needed terms.

We consider two penalties, the Adaptive LASSO and SCAD, and apply each within both estimation frameworks, giving four combinations: OLS-LASSO, OLS-SCAD, WEE-LASSO and WEE-SCAD. Comparing them separates the contribution of the weighting from that of the penalty. For each penalty we present the HD-WEE formulation; the OLS version follows by setting $\mathbf{\Omega}_t = \mathbf{I}$.

Both penalized estimators are stated as estimating equations rather than as penalized criteria. With coordinate-specific weights in $\mathbf{\Omega}_t$, the weighted estimating equation is not the first-order condition of any objective function, although setting $\mathbf{\Omega}_t = \mathbf{I}$ recovers the usual penalized least squares criteria. The implications for the asymptotic theory are taken up in Chapter 3.

2.2.1 Adaptive LASSO Regularization

The Adaptive LASSO [Zou, 2006] extends the LASSO [Tibshirani, 1996] by weighting the penalty on each coefficient inversely to an initial estimate, so that large coefficients are shrunk less than small ones. Under regularity conditions it has the oracle property in stationary settings: it selects the correct model with probability tending to one, and its estimates of the nonzero coefficients are asymptotically as efficient as if the true model were known [Zou, 2006].

To state the estimator, we write the transformed model of Eq. (2.5) compactly as

$$y_t = \sum_{i=0}^{2p} \gamma_i z_{i,t} + v_t, \quad (2.16)$$

where $z_{0,t} = 1$, $z_{i,t} = x_{i,t-1}$ and $z_{p+i,t} = \Delta x_{i,t}$ for $i = 1, \dots, p$, and γ stacks the level coefficients $\alpha_0, \alpha_1, \dots, \alpha_p$ and the correction coefficients $\eta_1, \dots, \eta_p$. The Adaptive

LASSO estimator $\tilde{\gamma}_\lambda$ solves the penalized estimating equation

$$\sum_{t=1}^n \mathbf{\Omega}_t \mathbf{z}_t \{y_t - \mathbf{z}_t^\top \boldsymbol{\gamma}\} - n\lambda \boldsymbol{\varpi} \circ \text{sgn}(\boldsymbol{\gamma}) = \mathbf{0}, \quad \varpi_i = \frac{1}{|\hat{\gamma}_{i,\text{init}}|}, \quad (2.17)$$

where $\hat{\gamma}_{\text{init}}$ is the initial estimate that defines the adaptive weights $\boldsymbol{\varpi}$, $\circ$ denotes the Hadamard product, and λ is the tuning parameter. The intercept is not penalized. In the HD-WEE case, $\mathbf{\Omega}_t$ is the weight matrix of Eq. (2.11) and $\hat{\gamma}_{\text{init}}$ is the unpenalized HD-WEE estimate of Eq. (2.10). In the OLS case, $\mathbf{\Omega}_t = \mathbf{I}$, the initial estimate is the OLS estimate, and Eq. (2.17) is the first-order condition of the standard Adaptive LASSO objective. The equation is solved by coordinate descent [Friedman et al., 2010], with λ chosen by the Schwarz information criterion of Section 2.2.3:

Algorithm 1 Adaptive LASSO Algorithm

- 1: Compute initial estimates: $\hat{\gamma}_{OLS}$ via OLS on Eq. (2.16) for the OLS case, or $\hat{\gamma}_{HD-WEE}$ via Eq. (2.10) for the HD-WEE case.
 - 2: **for** $\lambda \in \Lambda$ **do**
 - 3: Solve the penalized estimating equation in Eq. (2.17) to obtain $\tilde{\gamma}_\lambda$.
 - 4: Evaluate $SIC(\lambda)$ using Eq. (2.22).
 - 5: **end for**
 - 6: Select $\lambda_{\text{opt}} = \arg \min_{\lambda \in \Lambda} SIC(\lambda)$.
 - 7: Obtain final estimate $\tilde{\gamma}_{\lambda_{\text{opt}}}$.
 - 8: Recover original parameters: $\hat{\beta}_0 = \hat{\alpha}_0$ and $\hat{\beta}_i = \hat{\alpha}_i + \hat{\eta}_i(\hat{\rho}_i - 1)$ for $i = 1, \dots, p$.
-

In finite samples the adaptive weights are finite, so some shrinkage remains on every nonzero coefficient. The same is true of the LASSO-based estimators of Mei and Shi [2024] and Gao et al. [2026] discussed in Section 1.2. That shrinkage carries through the recovery formula $\hat{\beta}_i = \hat{\alpha}_i + \hat{\eta}_i(\hat{\rho}_i - 1)$ into the structural coefficients, and it is visible in the simulations of Section 4.2, where the Adaptive LASSO estimates of the smallest nonzero coefficient are biased downward while the SCAD estimates are

not. This motivates the SCAD penalty.

2.2.2 SCAD Regularization

The smoothly clipped absolute deviation (SCAD) penalty [Fan and Li, 2001] removes the shrinkage of large coefficients altogether. Its derivative has three regions: for $|\gamma_i| \leq \lambda$ it applies LASSO shrinkage; for $\lambda < |\gamma_i| \leq a\lambda$ the shrinkage decreases as the coefficient grows; and for $|\gamma_i| > a\lambda$ there is no shrinkage at all [Fan and Li, 2001, Fan and Lv, 2010]. Coefficients that are large relative to λ are therefore left unpenalized, which is what gives SCAD the oracle property: asymptotically it selects the correct model and estimates the nonzero coefficients as if the true model were known. The coordinate descent update also accommodates the observation-specific weights of HD-WEE directly, since each coordinate step is a weighted least squares problem [Breheny and Huang, 2011, Friedman et al., 2010].

The SCAD-penalized HD-WEE estimator $\hat{\gamma}_\lambda$ solves

$$\sum_{t=1}^n \mathbf{\Omega}_t \mathbf{z}_t (y_t - \mathbf{z}_t^\top \boldsymbol{\gamma}) - nP'_\lambda(|\boldsymbol{\gamma}|) \circ \text{sgn}(\boldsymbol{\gamma}) = \mathbf{0}, \quad (2.18)$$

where P'_λ is the derivative of the SCAD penalty function P_λ . As with the Adaptive LASSO, the intercept is not penalized, and setting $\mathbf{\Omega}_t = \mathbf{I}$ gives the OLS-SCAD estimator. We solve Eq. (2.18) by coordinate descent [Friedman et al., 2010, Breheny and Huang, 2011]. At each coordinate step, the unpenalized weighted least squares solution is

$$a_i = \frac{\sum_{t=1}^n w_{i,t} z_{i,t} \left(y_t - \sum_{j \neq i} \hat{\gamma}_j z_{j,t} \right)}{\sum_{t=1}^n w_{i,t} z_{i,t}^2}, \quad (2.19)$$

where $w_{i,t}$ are the HD-WEE weights of Eq. (2.11), equal to one throughout in the

OLS case. With the soft-thresholding function

$$S(a, b) = \begin{cases} a - b & \text{if } a > b, \\ 0 & \text{if } |a| \leq b, \\ a + b & \text{if } a < -b, \end{cases}$$

the SCAD update is

$$\hat{\gamma}_i = \begin{cases} S(a_i, \lambda) & \text{if } |a_i| \leq 2\lambda, \\ \frac{S\left(a_i, \frac{a\lambda}{a-1}\right)}{1 - \frac{1}{a-1}} & \text{if } 2\lambda < |a_i| \leq a\lambda, \\ a_i & \text{if } |a_i| > a\lambda, \end{cases} \quad (2.20)$$

with $a = 3.7$ as recommended by Fan and Li [2001]. The three cases correspond to the three regions of the penalty. The algorithm is as follows:

Algorithm 2 SCAD Regularization Algorithm (HD-WEE and OLS)

- 1: Compute initial estimates: $\hat{\gamma}_{HD-WEE}$ via Eq. (2.10) for the HD-WEE case, or $\hat{\gamma}_{OLS}$ via Eq. (2.8) for the OLS case.
 - 2: **for** $\lambda \in \Lambda$ **do**
 - 3: Initialize $\hat{\gamma} \leftarrow \hat{\gamma}_{\text{init}}$.
 - 4: **repeat**
 - 5: **for** $i \in \{1, \dots, 2p\}$ **do**
 - 6: Compute the coordinate solution a_i using Eq. (2.19).
 - 7: Apply the SCAD update $\hat{\gamma}_i$ using Eq. (2.20).
 - 8: **end for**
 - 9: **until** convergence of $\hat{\gamma}$
 - 10: Evaluate $SIC(\lambda)$ using Eq. (2.22).
 - 11: **end for**
 - 12: Select $\lambda_{\text{opt}} = \arg \min_{\lambda \in \Lambda} SIC(\lambda)$.
 - 13: Obtain final estimate $\hat{\gamma}_{\lambda_{\text{opt}}}$.
 - 14: Recover original parameters: $\hat{\beta}_0 = \hat{\alpha}_0$ and $\hat{\beta}_i = \hat{\alpha}_i + \hat{\eta}_i(\hat{\rho}_i - 1)$ for $i = 1, \dots, p$.
-

WEE-SCAD combines the three elements of the framework: the weighting for inference under mixed orders of integration, the penalty for a manageable number of parameters, and the flat region of SCAD for nearly unbiased estimates of the nonzero coefficients. OLS-SCAD shares the last two but not the first, so it remains exposed to the non-standard limit at the unit root; in the simulations of Section 4.2 this appears as undercoverage of the unit-root coefficient in the baseline design. Chapter 3 establishes the asymptotic properties of WEE-SCAD.

2.2.3 Implementation

Three details of the implementation are common to both penalties.

Standardization. Both penalized estimating equations are solved on standardized regressors. For each coordinate i , let

$$s_i = \left(n^{-1} \sum_{t=1}^n w_{i,t} z_{i,t}^2 \right)^{1/2}, \quad \tilde{z}_{i,t} = z_{i,t} / s_i, \quad (2.21)$$

so that $n^{-1} \sum_t w_{i,t} \tilde{z}_{i,t}^2 = 1$. This is the normalization under which the SCAD update of Eq. (2.20) solves the i -th row of Eq. (2.18) exactly. The adaptive weights of Eq. (2.17) are computed on the same scale, as $\varpi_i = |s_i \hat{\gamma}_{i,\text{init}}|^{-1}$. The scales s_i are computed once, on the original sample, and are held fixed in the bootstrap of Chapter 4.

The tuning grid. The penalty enters both estimating equations as $n\lambda$, and the grid Λ is built on that scale. Its largest value is

$$\lambda_{\max} = \max_j n^{-1} \left| \sum_{t=1}^n w_{j,t} \tilde{z}_{j,t} R_{0,t} \right|,$$

where $R_{0,t}$ is the residual from the intercept-only fit. This is the smallest value of λ at which every penalized coefficient is zero.

Selecting λ . For either penalty, λ is chosen from Λ by the Schwarz information criterion (SIC) [Chen and Chen, 2008, Wang et al., 2007],

$$SIC(\lambda) = n \log \left(n^{-1} \sum_{t=1}^n \{y_t - \mathbf{z}_t^\top \hat{\gamma}_\lambda\}^2 \right) + C_n \|\hat{\gamma}_\lambda\|_0, \quad (2.22)$$

where $C_n = \max\{\log n, \log(\log n) \cdot \log(2p)\}$, $\hat{\gamma}_\lambda$ is the penalized estimate at λ , and $\|\hat{\gamma}_\lambda\|_0$ is its number of nonzero slope coefficients, excluding the intercept. The selected value is $\lambda_{\text{opt}} = \arg \min_{\lambda \in \Lambda} SIC(\lambda)$. We use the log form of Wang et al. [2007], in which the penalty C_n is on the same scale as the fit term. The residual sum of squares is unweighted, in the original sample and in the bootstrap replicates alike, since a version weighted by $N(1, 1)$ bootstrap weights can be negative.

CHAPTER 3: ASYMPTOTIC THEORY

The regularization schemes of Section 2.2 were motivated by the oracle property of the SCAD penalty, which Fan and Li [2001] and Fan and Lv [2011] establish for stationary designs. That theory does not cover the present setting, in which the regressor vector $\mathbf{z}_t$ mixes stationary and integrated components and the estimating equation carries observation- and coordinate-specific weights. This chapter supplies the missing theory. Theorem 3.4.1 establishes sparsity and block-wise convergence rates, and Theorem 3.5.1 establishes oracle asymptotic normality. Each theorem is stated with an outline of its proof. The full proofs are in Appendix E; they rely on a concentration condition, (P4**), which is verified in Appendices B–D using the external results collected in Appendix A.

Correspondence with the notation of Sections 2.1 and 2.2. The results below are stated for a generic predictive regression $y_t = \mathbf{x}_{t-1}^\top \beta_0 + v_t$ with d candidate predictors. In the notation of Section 2.1, the regressor vector is $\mathbf{x}_{t-1} \equiv \mathbf{z}_t$ and the parameter vector is $\beta_0 \equiv \boldsymbol{\gamma} = (\alpha_0, \alpha_1, \dots, \alpha_p, \eta_1, \dots, \eta_p)^\top$, so that $d = 2p + 1$ and the coefficients on both the levels $x_{i,t-1}$ and the endogeneity-correction terms $\Delta x_{i,t}$ are treated jointly. The error v_t is the orthogonalized error of Eq. (2.3), which the linear projection renders uncorrelated with the innovations by construction.

The persistence partition corresponds to the index sets of Eq. (2.11): the stationary block $\mathcal{I}_2$ is the set I of that equation, and the nonstationary block $\mathcal{I}_1$ is its complement I^c . Note that the first differences $\Delta x_{i,t}$ are stationary even when the level $x_{i,t-1}$ is integrated, so those coordinates belong to $\mathcal{I}_2$; the block $\mathcal{I}_1$ contains only levels. Finally, the weights $\omega_{t,j}$ below are the $w_{i,t}$ of Eq. (2.11), written with the time index first for consistency with the matrix $\boldsymbol{\Omega}_t$.

3.1 Setting and notation

The model is predictive,

$$y_t = x_{t-1}^\top \beta_0 + v_t, \quad t = 1, \dots, n,$$

with d candidate predictors, $\log d = O(n^\nu)$ for some $\nu \in (0, 1)$, and sparsity $s = \|\beta_0\|_0$ fixed. The index set $\{1, \dots, d\}$ is partitioned in two independent ways.

By persistence. The nonstationary block $\mathcal{I}_1$ collects the levels of the I(1)/NI(1) predictors, and the stationary block $\mathcal{I}_2$ the remainder. Following Hong et al. [2021] and Hong et al. [2024], the weighted estimating equation is

$$\sum_{t=1}^n \Omega_t x_{t-1} (y_t - x_{t-1}^\top \beta) = 0, \quad \Omega_t = \text{diag}(1, \omega_{t,1}, \dots, \omega_{t,d}),$$

where $\omega_{t,j} = 1$ for $j \in \mathcal{I}_2$ and $\omega_{t,j} = (1 + \|X_{t-1, \mathcal{I}_1}\|^2)^{-1/2}$ for $j \in \mathcal{I}_1$. The weights are written here in terms of the true blocks $\mathcal{I}_1$ and $\mathcal{I}_2$. In practice they are computed from the data-driven classification of Eq. (2.11), which coincides with the true blocks on an event whose probability tends to one. Hong et al. [2021], (A.13)–(A.14), and Hong et al. [2024] establish this for fixed d , and Remark 3.2.3 sketches why it continues to hold when d grows with n . Restricting attention to this event contributes the $o(1)$ terms in the probability statements below. We write Ω_t for the diagonal *matrix* of weights and $\omega_{t,j}$ for its j -th diagonal *entry*, and set

$$\tilde{x}_{t-1,j} = \omega_{t,j} x_{t-1,j},$$

so that $|\tilde{x}_{t-1,j}| < 1$ and $\sum_{t \leq n} \tilde{x}_{t-1,j}^2 < n$ hold *deterministically* for $j \in \mathcal{I}_1$, while $\tilde{x}_{t-1,j} = x_{t-1,j}$ for $j \in \mathcal{I}_2$.

By sparsity. The active set is $\mathcal{M}_0 = \text{supp}(\beta_0)$, with $|\mathcal{M}_0| = s$ fixed, and $\mathcal{M}_0^c$ is its complement. Every quantity below is classified by both partitions. In particular,

$|\mathcal{M}_0^c \cap \mathcal{I}_2| \asymp d$ grows with n , while $|\mathcal{M}_0^c \cap \mathcal{I}_1| \leq d_1 = |\mathcal{I}_1|$ and $|\mathcal{M}_0| = s$ are fixed.

The score at the truth has j -th component $\xi_j = \sum_{t=1}^n \tilde{x}_{t-1,j} v_t$, and we write ξ_A for the subvector indexed by A . The block normalizer is

$$D_n = \text{diag}(\sqrt{n} \text{ on } \mathcal{I}_2, n \text{ on } \mathcal{I}_1),$$

with $D_{n,11}$ its restriction to $\mathcal{M}_0$. We set $d_n = \frac{1}{2} \min_{j \in \mathcal{M}_0} |\beta_{0,j}|$, and the SCAD penalty p_{λ_n} has parameter $a > 2$. As in Hong et al. [2024], the threshold c^* is fixed and λ_n is a deterministic sequence; the data-driven choices of Sections 2.1.3 and 2.2.3 are not covered by the theory.

3.2 Primitive conditions

Assumption 1 (Stationary block: (S1)–(S4)). *(S1)* $(X_{t,\mathcal{I}_2}, v_t)$ is strictly stationary and geometrically α -mixing: $\alpha(h) \leq c_\varrho e^{-c_\beta h}$.

(S2) $\mathbb{E}[x_{t,j}] = 0$ and $\sup_{j \in \mathcal{I}_2} \mathbb{E}[\exp(|x_{t,j}|/K_x)] < \infty$. Also, $\mathbb{E}[v_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\exp(|v_t|/K_v) \mid \mathcal{F}_{t-1}] \leq C_0$ a.s.

(S3) With γ the combined exponent of Merlevède et al. [2011], $\frac{1-\gamma}{2-\gamma} < \varrho < \frac{1}{2}$ and $0 < \nu < 1 - 2\varrho$. Under (S1)–(S2) one has $\gamma = 1/3$ for products of two sub-exponential factors, hence $\varrho \in (2/5, 1/2)$ and $\nu < 1/5$.

(S4) $\sup_{j,\ell} |\mathbb{E}[x_{t,j} x_{t,\ell}]| \leq \bar{C} < \infty$.

Assumption 2 (Nonstationary block: (N1)–(N4)). *(N1)* *Cross-block independence.*

$\{X_{t,\mathcal{I}_2}\}$ and $\{v_t\}$ are independent of the innovations $\{U_t\}$ of the nonstationary block. Only the following consequence is used: conditionally on $\sigma(\{U_t\})$, the stationary block retains its geometric α -mixing and its uniform sub-exponential tail bound.

(N2) *Fixed nonstationary dimension.* $|\mathcal{I}_1| = d_1$ is fixed, independent of n and d .

(N3) *I(1)/NI(1) structure.* For $i \in \mathcal{I}_1$, $x_{t,i} = \rho_i x_{t-1,i} + u_{t,i}$ with $\rho_i = 1 + c_i/n$, and

$\Sigma_0 = \text{Var}(U_t)$ positive definite (Hong et al., 2024, (A0)), excluding cointegration within $\mathcal{I}_1$ and ensuring $\|B(r)\| > 0$, so that $B^* = B/\|B\|$ is well defined.

(N4) *Strong approximation.* The $u_{t,i}$ satisfy the hypotheses of Lemma A1 in the online appendix of Hong et al. [2021] (see also Cai et al., 2015): stationary α -mixing, $\mathbb{E}|u_{t,i}|^r < \infty$ for some $r > 2 + \delta^*$ with $0 < \delta^* \leq 2$, and summable mixing coefficients. Consequently, with $U_{n,i}(r) = n^{-1/2}x_{[nr],i}$ and $R_{n,i} := U_{n,i} - U_{c_i}$,

$$\sup_{0 \leq r \leq 1} |R_{n,i}(r)| = O(n^{-\theta^*}(\log n)^{\lambda^*}) =: \rho_n \text{ a.s.}, \quad \theta^* = \frac{1}{2} - \frac{1}{2+\delta^*} > 0.$$

Assumption 3 ((P4**): concentration). Let $\kappa_* = 1/(2C_2V_*^2)$, with C_2 from Theorem 1 of Merlevède et al. [2011] and V_*^2 the bound of their Remark 3, and $u_n = \kappa_*^{-1/2}n^{1/2-\varrho}(\log n)^{1/2}$. Define

$$\begin{aligned} \mathcal{E}_2 &= \left\{ \|\xi_{\mathcal{M}_0^c}\|_\infty \leq u_n \sqrt{n} \right\}, \\ \mathcal{E}_3' &= \left\{ \max_{j \in \mathcal{M}_0^c, \ell \in \mathcal{M}_0 \cap \mathcal{I}_2} \left| \frac{1}{n} \sum_{t=1}^n \tilde{x}_{t-1,j} x_{t-1,\ell} \right| \leq C_X \right\}, \\ \mathcal{E}_3'' &= \left\{ \max_{j \in \mathcal{M}_0^c, \ell \in \mathcal{M}_0 \cap \mathcal{I}_1} \left| n^{-3/2} \sum_{t=1}^n \tilde{x}_{t-1,j} x_{t-1,\ell} \right| \leq C_X \right\}. \end{aligned}$$

For every $\varepsilon > 0$ there exists $C_X = C_X(\varepsilon)$ such that

$$\liminf_{n \rightarrow \infty} \mathbb{P}(\mathcal{E}_2 \cap \mathcal{E}_3' \cap \mathcal{E}_3'') \geq 1 - \varepsilon.$$

Although it is stated as an assumption, (P4**) is not maintained: it is *verified* from (S1)–(S4) and (N1)–(N4), in Appendix B for the stationary design and in Appendices C and D for the mixed design.

Remark 3.2.1 (Why the two cross-Gram events carry different normalizations). For $\ell \in \mathcal{I}_1$ the column factor $x_{t-1,\ell}$ is a raw $\text{I}(1)$ variable, of order $\sqrt{t}$, while the row factor is bounded by the weight. Hence $\sum_{t \leq n} \tilde{x}_{t-1,j} x_{t-1,\ell} = O(\sum_{t \leq n} \sqrt{t}) = O(n^{3/2})$,

and the entry normalized by n^{-1} *diverges* like $\sqrt{n}$: a fixed-constant bound at that normalization fails with probability tending to one whenever $\mathcal{M}_0^c \cap \mathcal{I}_1$ and $\mathcal{M}_0 \cap \mathcal{I}_1$ are both non-empty. The power $n^{-3/2}$ is the unique one leaving the entry $O_P(1)$ and nondegenerate, and it agrees with the D_n convention of Hong et al. [2024]. The weight bounds only the row factor, not the column factor.

The ε -form of (P4**) is needed because several of the normalized cross-Gram entries converge to nondegenerate *random* limits, so the constant C_X that bounds them with probability $1 - \varepsilon$ grows as $\varepsilon \rightarrow 0$ and no single constant serves every ε . The event $\mathcal{E}_2$, whose threshold u_n diverges, does hold with probability tending to one.

Assumption 4 ((R): rate conditions). $\lambda_n \rightarrow 0$; $\lambda_n / (n^{-\varrho}(\log n)^{1/2}) \rightarrow \infty$; $d_n/\lambda_n \rightarrow \infty$; and $\log d = O(n^\nu)$ with $\nu < 1 - 2\varrho$.

The lower bound $\lambda_n \gg n^{-\varrho}(\log n)^{1/2}$ is exactly the rate at which $\|\xi_{\mathcal{M}_0^c}\|_\infty$ concentrates on $\mathcal{E}_2$, while $d_n/\lambda_n \rightarrow \infty$ places the active coordinates in the flat region of the SCAD penalty; λ_n thus occupies the corridor between the noise level of the inactive scores and the weakest true signal. The implementation solves the penalized equation on the standardized regressors of Eq. (2.21), which is equivalent to replacing λ_n by $\lambda_n s_j$ in the conditions for coordinate j . The theorems continue to hold under the additional condition $\inf_j \mathbb{E} x_{t,j}^2 > 0$, which keeps the scales s_j of the stationary coordinates bounded away from zero; for the nonstationary coordinates s_j grows with n , which only relaxes the conditions.

Assumption 5 ((H'): coercivity at the D_n normalization). With

$$B_{n,11} = n^{-1/2} \sum_{t=1}^n \tilde{x}_{t-1, \mathcal{M}_0} x_{t-1, \mathcal{M}_0}^\top D_{n,11}^{-1},$$

we have $B_{n,11} = O_P(1)$ and $B_{n,11}^{-1} = O_P(1)$. Equivalently, for every $\varepsilon > 0$ there is $c_H = c_H(\varepsilon) > 0$ with $\liminf_n \mathbb{P}(\lambda_{\min}(B_{n,11}^{\text{sym}}) \geq c_H) \geq 1 - \varepsilon$, where $B_{n,11}^{\text{sym}} = \frac{1}{2}(B_{n,11} + B_{n,11}^\top)$.

Remark 3.2.2 (Status and form of (H')). (H') is a maintained assumption, the analogue at the D_n normalization of the coercivity conditions of Hong et al. [2021] (Lemma A.3) and Hong et al. [2024]. The normalization by $D_{n,11}$ is necessary: at the uniform n^{-1} scaling the I(1) block of the active Gram diverges, by Remark 3.2.1. Because the limit of the nonstationary block is a random matrix, the correct form is tightness of the inverse, $B_{n,11}^{-1} = O_P(1)$, rather than a bound $\lambda_{\min} \geq c_H$ at a fixed c_H .

Remark 3.2.3 (Classification of the predictors when d grows: a sketch). The blocks $\mathcal{I}_1$ and $\mathcal{I}_2$ are the true classification. With c^* fixed and d fixed, Hong et al. [2021] and Hong et al. [2024] show that the data-driven set I of Eq. (2.11) coincides with it with probability tending to one. The following argument indicates why this continues to hold when d grows; it is offered as a supporting argument rather than a formal result. The two blocks require separate treatment.

For the stationary block what is needed is that the statistic

$$T_j = n^{-1/2} \log(n) \max_{t \leq n} |x_{t,j}|$$

vanish uniformly in j . Assumption (S2) bounds the sub-exponential norm uniformly over $j \in \mathcal{I}_2$, so $\mathbb{P}(|x_{t,j}| > x) \leq C e^{-x/K_x}$ with constants free of j , and a union bound over the nd pairs (j, t) , which requires no independence, gives $\max_{j \in \mathcal{I}_2} \max_{t \leq n} |x_{t,j}| = O_P(\log(nd))$. Under (R), $\log(nd) = \log n + O(n^\nu) = O(n^\nu)$, so

$$\max_{j \in \mathcal{I}_2} T_j = O_P(n^{\nu-1/2} \log n) \longrightarrow 0$$

for any $\nu < 1/2$, and in particular under $\nu < 1/5$. Every stationary coordinate therefore falls below any fixed $c^* > 0$ with probability tending to one.

For the nonstationary block the requirement is the opposite, that the statistic diverge, and for this it is not enough that $n^{-1/2} \max_t |x_{t,i}| = O_P(1)$, since an $O_P(1)$ sequence may approach zero. A lower bound is needed. By (N3)–(N4), $n^{-1/2} x_{[nr],i}$

is uniformly close to the diffusion process $U_{c_i}(r)$ of Hong et al. [2024], Lemma 1, so by the continuous mapping theorem $n^{-1/2} \max_{t \leq n} |x_{t,i}| \Rightarrow \sup_{r \in [0,1]} |U_{c_i}(r)|$, whose distribution places no mass at zero because Σ_0 is positive definite. Hence for every $\varepsilon > 0$ there is a $\kappa > 0$ such that $\mathbb{P}(n^{-1/2} \max_t |x_{t,i}| > \kappa) \geq 1 - \varepsilon$ for all large n , and on that event $T_i \geq \kappa \log n \rightarrow \infty$. Since d_1 is fixed by (N2), the union over $i \in \mathcal{I}_1$ is finite and the growth of d plays no role in this half of the argument; ν enters the classification only through the stationary union bound.

The two halves give the separation that motivates the largest-gap rule of Eq. (2.15): the sorted statistics fall into two groups whose ratio grows with n . The separation is also invariant to rescaling each predictor by a fixed constant, since a constant factor is absorbed into the O_P bounds and affects neither limit. In finite samples it is not, which is why the empirical application uses the scale-free statistic $T_i/\text{SD}(\Delta x_i)$ of Chapter 2. That division also changes neither limit: $\text{SD}(\Delta x_i)$ converges to σ_{u_i} for a stationary coordinate and is $O_P(1)$ and bounded away from zero for an integrated one, by the same continuous-mapping argument.

3.3 The estimator as a sparse root of the penalized estimating equation

The weights in Ω_t are coordinate-specific: the i -th row of the estimating equation reads $\sum_t \omega_{t,i} x_{t-1,i} (y_t - x_{t-1}^\top \beta) = 0$. Consequently the matrix multiplying β has (i, m) entry $\sum_t \omega_{t,i} x_{t-1,i} x_{t-1,m}$ and (m, i) entry $\sum_t \omega_{t,m} x_{t-1,m} x_{t-1,i}$, and these differ whenever $\omega_{t,i} \neq \omega_{t,m}$, which is the situation here: the stationary block carries weight 1 and the nonstationary block carries ω_t .

A vector field is the gradient of a scalar function only if its Jacobian is symmetric. The Jacobian in β of the estimating equation is $-\sum_t \Omega_t x_{t-1} x_{t-1}^\top$, which is not symmetric; hence *no scalar objective Q_n exists whose gradient is the estimating equation*, and the estimator cannot be characterized as a local maximizer. This does not matter without a penalty, since the unpenalized system is linear with a unique solution. It matters once SCAD is added, since the penalized system is nonlinear and the question

“which root?” is ordinarily settled by reference to an objective.

We therefore characterize $\hat{\beta}$ directly as a *sparse root of the penalized estimating equation*: existence on the active subspace by a Brouwer inward-pointing argument, sparsity by verifying the Karush–Kuhn–Tucker conditions of the penalized system. As a result, the argument does not rely on the characterization of strict local maximizers in Theorem 1 of Fan and Lv [2011], which is stated for penalized likelihoods.

The penalized system. Write

$$\Psi_n(\beta_1) = \sum_{t=1}^n \tilde{x}_{t-1, \mathcal{M}_0} (y_t - x_{t-1, \mathcal{M}_0}^\top \beta_1) - n \bar{p}_{\lambda_n}(\beta_1), \quad \bar{p}_{\lambda_n}(\beta_1) = (p'_{\lambda_n}(|\beta_j|) \text{sgn}(\beta_j))_{j \in \mathcal{M}_0},$$

for the active-block restriction. Substituting $y_t = x_{t-1}^\top \beta_0 + v_t$ and setting $\delta = \beta_1 - \beta_{10}$ gives the *exact* identity

$$\Psi_n(\beta_{10} + \delta) = \xi_{\mathcal{M}_0} - n H_{n,11} \delta - n \bar{p}_{\lambda_n}(\beta_{10} + \delta), \quad H_{n,11} = n^{-1} \sum_{t=1}^n \tilde{x}_{t-1, \mathcal{M}_0} x_{t-1, \mathcal{M}_0}^\top. \quad (3.1)$$

Because the model is linear in β , the residual is linear in δ , so (3.1) holds exactly and no Taylor expansion is needed.

3.4 Sparsity and block-wise rates

Theorem 3.4.1 (Sparsity and parametric rate). *Under (S1)–(S4), (N1)–(N4), (P4**), (R) and (H'), for every $\varepsilon > 0$ there is $\tau < \infty$ such that, with probability at least $1 - \varepsilon - o(1)$, the penalized estimating equation admits a root $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2) \in \mathbb{R}^d$ with*

(a) (Sparsity) $\hat{\beta}_2 = 0$;

(b) ($\sqrt{n}$ -consistency) $\|\hat{\beta}_1 - \beta_{10}\|_2 \leq \tau/\sqrt{n} = O_P(n^{-1/2})$.

Proof strategy. The proof has three steps. First, a root $\hat{\beta}_1$ exists on the active subspace with $\|\hat{\beta}_1 - \beta_{10}\|_2 = O_P(n^{-1/2})$. On a sphere $\|\delta\|_2 = \tau/\sqrt{n}$ around β_{10} , the quadratic term of $\delta^\top \Psi_n(\beta_{10} + \delta)$ dominates the linear score term for τ large enough,

by (H') and the active-block central limit theorem of Hong et al. [2021]. The vector field therefore points inward on the sphere, and Brouwer's theorem gives a root inside it. Second, the penalty term vanishes on the active coordinates, because $d_n/\lambda_n \rightarrow \infty$ places each of them in the flat region of the SCAD penalty. Third, sparsity, $\hat{\beta}_2 = 0$, follows by verifying the Karush–Kuhn–Tucker conditions of the penalized system directly, in place of conditions (7)–(9) of Fan and Lv [2011], which presuppose an objective function. The substantive step bounds the subgradient of each inactive coordinate by a score term, controlled by $\mathcal{E}_2$ of Assumption 3, plus a cross term. The cross term is bounded entrywise using the block-wise rates of Lemma 3.4.2 and the cross-Gram events $\mathcal{E}'_3$ and $\mathcal{E}''_3$; the $n^{3/2}$ normalization of $\mathcal{E}''_3$ is offset by the faster $O_P(n^{-1})$ rate of the integrated active coordinates. Full details are given in Appendix E. $\square$

As in Fan and Li [2001], Theorem 3.4.1 does not say that the root found by coordinate descent is this root; it says that a root with these properties exists. The SCAD algorithm of Section 2.2.2 starts from the unpenalized HD-WEE estimate, which is consistent, so the root it finds is the natural candidate for the one in the theorem, but we do not prove that it is. The coefficients set to zero include the correction coefficients η_i whose true value is zero, since these belong to $\mathcal{M}_0^c$.

Lemma 3.4.2 (Block-wise rates). *On the sparsity event and under (H'), $D_{n,11}(\hat{\beta}_1 - \beta_{10}) = O_P(1)$; that is,*

$$\hat{\beta}_\ell - \beta_{0\ell} = O_P(n^{-1/2}) \quad (\ell \in \mathcal{M}_0 \cap \mathcal{I}_2), \quad \hat{\beta}_\ell - \beta_{0\ell} = O_P(n^{-1}) \quad (\ell \in \mathcal{M}_0 \cap \mathcal{I}_1).$$

Proof. The active-block equation, with the penalty eliminated as in the proof above, reads $\sum_t \tilde{x}_{t-1, \mathcal{M}_0} x_{t-1, \mathcal{M}_0}^\top (\hat{\beta}_1 - \beta_{10}) = \xi_{\mathcal{M}_0}$; inserting $D_{n,11}^{-1} D_{n,11}$ and dividing by $\sqrt{n}$ gives the sandwich identity $B_{n,11} D_{n,11} (\hat{\beta}_1 - \beta_{10}) = n^{-1/2} \xi_{\mathcal{M}_0}$, and (H') together with the active-block CLT yields $D_{n,11} (\hat{\beta}_1 - \beta_{10}) = O_P(1)$. See Appendix E for the full argument. $\square$

The aggregate bound of Theorem 3.4.1(b) does not imply Lemma 3.4.2: an ℓ_2 bound is dominated by the slowest coordinates and yields only $O_P(n^{-1/2})$ for every coordinate, discarding the faster rate of the nonstationary ones. The distinction matters: the KKT verification in Appendix E relies on the $O_P(n^{-1})$ rate of the I(1) active coordinates for the cross-term bound in Step 3.

Because s is fixed and the expansion (3.1) is exact, the second-order remainder and the Lyapunov sum that restrict Theorem 4 of Fan and Lv [2011] to $s = o(n^{1/3})$ do not arise here.

3.5 Oracle asymptotic normality

Let

$$A_{n,11} = n^{-1} \sum_{t=1}^n (\tilde{x}_{t-1, \mathcal{M}_0})^{\otimes 2}, \quad B_{n,11} = n^{-1/2} \sum_{t=1}^n \tilde{x}_{t-1, \mathcal{M}_0} x_{t-1, \mathcal{M}_0}^\top D_{n,11}^{-1}.$$

Theorem 3.5.1 (Oracle asymptotic normality). *Under the assumptions of Theorem 3.4.1,*

$$A_{n,11}^{-1/2} B_{n,11} D_{n,11} (\hat{\beta}_1 - \beta_{10}) \xrightarrow{d} N(0, \sigma_v^2 I_s).$$

Proof strategy. On the sparsity event $\{\hat{\beta}_2 = 0\}$, whose probability is at least $1 - \varepsilon - o(1)$ by Theorem 3.4.1, the exact identity (3.1) collapses: the cross-block contribution vanishes because $\hat{\beta}_2 - \beta_{20} = 0$, and the penalty term vanishes because the active coordinates eventually sit in the flat region of the SCAD penalty, as in the proof of Theorem 3.4.1. What remains is the linear sandwich identity $B_{n,11} D_{n,11} (\hat{\beta}_1 - \beta_{10}) = n^{-1/2} \xi_{\mathcal{M}_0}$ of (E.1). Pre-multiplying by $A_{n,11}^{-1/2}$ and applying the martingale central limit theorem to $n^{-1/2} \xi_{\mathcal{M}_0}$ gives the stated normal limit, as in the proof of Theorem 2 of Hong et al. [2024] and the s -dimensional restriction of Theorem 3.1 of Hong et al. [2021]. The self-normalization by $A_{n,11}^{-1/2}$ is what makes the limit free of the random quantities associated with the nonstationary block. The full argument is given in Appendix E. $\square$

The theorem delivers the oracle property in the form appropriate to a mixed design: stationary coefficients are estimated at rate $\sqrt{n}$ and nonstationary ones at rate n , and after block normalization the joint limit is standard normal. Under homoskedasticity $\hat{\sigma}_v^2 = n^{-1} \sum_t \hat{v}_t^2$ is consistent once $\hat{\beta}_1 \rightarrow_p \beta_{10}$ and $\hat{\beta}_2 = 0$ are available; under heteroskedasticity the sandwich form of Hong et al. [2024], Theorem 3 applies instead.

Remark 3.5.2 (Scope: the projected coefficients versus the recovered structural coefficients). Theorems 3.4.1 and 3.5.1 are stated for the generic β_0 , which under the correspondence set out at the start of this chapter is the projected coefficient vector $\gamma = (\alpha_0, \dots, \alpha_p, \eta_1, \dots, \eta_p)^\top$, jointly estimated by the penalized WEE. The structural coefficients reported in Tables 4.2–5.2 are recovered afterward via $\hat{\beta}_i = \hat{\alpha}_i + \hat{\eta}_i(\hat{\rho}_i - 1)$, using an autoregressive estimate $\hat{\rho}_i$ obtained separately (Eq. (2.12)) and, for $i \in \mathcal{I}_1$, itself subject to non-standard unit-root asymptotics. We do not extend Theorems 3.4.1 and 3.5.1 through this recovery map. Confidence intervals for β do not require such an extension: the random weighted bootstrap of Section 4.1 re-estimates $\hat{\rho}_i^*$ within each replicate and recovers $\hat{\beta}^*$ by the identical two-step map used for the point estimate (Algorithm 3), so the reported intervals inherit their validity from the bootstrap’s replication of the full estimation procedure rather than from a closed-form asymptotic distribution for $\hat{\beta}$ itself.

Together, Theorems 3.4.1 and 3.5.1 establish that the WEE-SCAD estimator recovers the correct sparsity pattern and achieves oracle asymptotic normality under mixed orders of integration, with the number of predictors growing as $\log d = O(n^\nu)$. Both results rest on the concentration condition (P4**), verified in Appendices B–D. The theory is for fixed tuning parameters and for the projected coefficients γ rather than the recovered structural ones. Chapter 4 develops the random weighted bootstrap used to construct confidence intervals for β in practice, and reports simulation evidence on variable selection and bootstrap coverage at realistic sample sizes.

CHAPTER 4: RANDOM WEIGHTED BOOTSTRAP AND SIMULATIONS

This chapter develops the inference procedure for the estimator of Chapter 2 and evaluates it by simulation. Section 4.1 introduces a random weighted bootstrap that approximates the sampling distribution of the structural coefficients and compares five ways of turning the bootstrap replicates into confidence intervals. Section 4.2 compares rules for selecting the threshold c^* and then evaluates WEE-SCAD against the other estimators under a baseline and a high-dimensional design.

4.1 Random Weighted Bootstrap

Theorem 3.5.1 gives a normal limit for the nonzero coefficients of $\boldsymbol{\gamma}$ after normalization, but it does not by itself deliver confidence intervals for $\boldsymbol{\beta}$. The limiting variance depends on the unknown σ_v^2 and on the matrices $A_{n,11}$ and $B_{n,11}$, which are random in the limit when predictors are I(1) or NI(1). The structural coefficients are recovered through $\hat{\rho}_i$, whose limiting distribution is non-standard for nonstationary predictors [Phillips, 1987, Park and Phillips, 1988]. And the theorem describes the oracle root of the estimating equation, whereas the estimator computed in a finite sample places positive probability on zero for coefficients near the threshold. We therefore use a random weighted bootstrap to approximate the sampling distribution of $\hat{\boldsymbol{\beta}}$ and construct confidence intervals. Rather than resampling observations, which would break the temporal dependence of the series, the random weighted bootstrap reweights the estimating equation itself, so the data and their dependence structure are left intact [Hong et al., 2024].

4.1.1 Bootstrap Procedure

We bootstrap directly from the estimating equation that defines the estimator. For the unpenalized HD-WEE estimator this is

$$\sum_{t=1}^n \zeta_t \mathbf{\Omega}_t \mathbf{z}_t (y_t - \mathbf{z}_t^\top \boldsymbol{\gamma}) = \mathbf{0}, \quad (4.1)$$

where ζ_t are i.i.d. random weights with mean and variance one, discussed below, with the closed-form solution

$$\hat{\boldsymbol{\gamma}}_w^* = \left(\sum_{t=1}^n \zeta_t \mathbf{\Omega}_t \mathbf{z}_t \mathbf{z}_t^\top \right)^{-1} \left(\sum_{t=1}^n \zeta_t \mathbf{\Omega}_t \mathbf{z}_t y_t \right). \quad (4.2)$$

For the penalized estimators the replicate re-solves the penalized estimating equation with the same random weights,

$$\sum_{t=1}^n \zeta_t \mathbf{\Omega}_t \mathbf{z}_t (y_t - \mathbf{z}_t^\top \boldsymbol{\gamma}) - n P'_{\lambda^*}(|\boldsymbol{\gamma}|) \circ \text{sgn}(\boldsymbol{\gamma}) = \mathbf{0}, \quad (4.3)$$

where P'_{λ^*} is the SCAD penalty derivative of Eq. (2.18), replaced by $\lambda^* \boldsymbol{\varpi}$ for the Adaptive LASSO of Eq. (2.17), and λ^* is re-selected by the SIC of Eq. (2.22) over the grid Λ within each replicate, with the unweighted residual sum of squares described in Section 2.2.3. The replicate is solved by the same coordinate descent used in the original sample. There is no closed form in this case and Eq. (4.2) does not apply. For the Adaptive LASSO, the initial estimate that defines the adaptive weights $\boldsymbol{\varpi}$ is recomputed from the ζ -weighted equation in each replicate. Three quantities are held fixed at their original-sample values across all replicates: the classification I^c and the threshold c^* that produced it, the standardization scales s_i of Eq. (2.21), and the grid Λ . The only source of variation across replicates is therefore $\{\zeta_t\}$. Because the bootstrap weights change the weighted column norms of the standardized regressors, the thresholds of the SCAD update in Eq. (2.20) are multiplied by $h_i = n \left(\sum_t \zeta_t w_{i,t} \tilde{z}_{i,t}^2 \right)^{-1}$,

which equals one in the original sample. Setting $\mathbf{\Omega}_t = \mathbf{I}$ throughout yields the OLS bootstrap as a special case.

Several distributions have been used for the bootstrap weights, including the Dirichlet, exponential and gamma [Efron and Tibshirani, 1993]. Hong et al. [2024] show that, for weights with mean and variance one, the conditional distribution of the bootstrap estimator matches the limiting distribution of the estimator under mixed orders of integration, and report that their results are unchanged when exponential weights replace normal ones. We use $N(1, 1)$ weights in the simulations of Section 4.2 and $\text{Exp}(1)$ weights in the empirical application of Chapter 5. Exponential weights are always positive, whereas about one $N(1, 1)$ weight in six is negative; at the smaller sample sizes of the empirical application this can make the weighted system singular (Section 5.3), while at $n = 450$ no $N(1, 1)$ replicate failed in the simulations.

The result of Hong et al. [2024] establishes the validity of the random weighting for the *unpenalized* estimator. We are not aware of a corresponding result for the penalized estimator under mixed orders of integration, and this dissertation does not prove one. The evidence for the intervals reported here is therefore their finite-sample coverage in Section 4.2 rather than a validity theorem, a limitation we return to in Chapter 6.

For the autoregressive parameters, we apply the same random weighting scheme:

$$\sum_{t=2}^n \zeta_t w_{i,t-1} x_{i,t-1} (x_{i,t} - \rho_i x_{i,t-1}) = 0, \quad (4.4)$$

where $w_{i,t-1}$ are the HD-WEE weights from Eq. (2.11), with $w_{i,t-1} = 1$ for all t in the OLS case. We recover the bootstrap versions of the original structural parameters

using the same transformation as in the main estimation:

$$\begin{aligned}\widehat{\beta}_{0,w}^* &= \widehat{\alpha}_{0,w}^* \\ \widehat{\beta}_{i,w}^* &= \widehat{\alpha}_{i,w}^* + \widehat{\eta}_{i,w}^* (\widehat{\rho}_{i,w}^* - 1) \quad \text{for } i = 1, \dots, p.\end{aligned}\tag{4.5}$$

The bootstrap replicates $\{\widehat{\beta}_w^{*(b)}\}_{b=1}^B$ approximate the sampling distribution of $\widehat{\beta}_w$ and form the basis for confidence interval construction described in Section 4.1.2.

The full bootstrap procedure is summarized as follows:

Algorithm 3 Random Weighted Bootstrap Algorithm

- 1: Compute the point estimate $\widehat{\beta}$ using the relevant method: $\widehat{\beta}_{HD-WEE}$ via Eq. (2.13) for the HD-WEE case, or $\widehat{\beta}_{OLS}$ via the OLS recovery formula for the OLS case. For the OLS case, set $\mathbf{\Omega}_t = \mathbf{I}$ throughout.
 - 2: **for** $b = 1, \dots, B$ **do**
 - 3: Generate i.i.d. bootstrap weights $\{\zeta_t\}_{t=1}^n$ with mean and variance one: $N(1, 1)$ in the simulations, $\text{Exp}(1)$ in the empirical application.
 - 4: Re-solve the estimating equation of the relevant method with weights $\{\zeta_t\}$: Eq. (4.1) for the unpenalized estimators, or Eq. (4.3) for the penalized estimators, re-selecting $\lambda^{*(b)}$ by the SIC of Eq. (2.22) over the grid Λ . Hold I^c , c^* and the scales s_i at their original-sample values. Denote the result $\widehat{\gamma}_w^{*(b)}$.
 - 5: Recover $\widehat{\beta}_w^{*(b)}$ from $\widehat{\gamma}_w^{*(b)}$ and $\widehat{\rho}_w^{*(b)}$ via Eq. (4.5).
 - 6: **end for**
 - 7: Discard any replicate whose weighted system is singular or whose coordinate descent fails to converge, and report the number discarded.
 - 8: Collect bootstrap replicates $\{\widehat{\beta}_w^{*(b)} - \widehat{\beta}_w\}_{b=1}^B$ and construct confidence intervals using the methods described in Section 4.1.2.
-

4.1.2 Confidence Interval Construction

Given the bootstrap replicates $\{\hat{\beta}_w^{*(b)}\}_{b=1}^B$, we construct $(1 - \alpha) \times 100\%$ confidence intervals for each component β_i using five methods [Efron and Tibshirani, 1993, DiCiccio and Efron, 1996]. Let $\hat{\beta}$ denote the point estimate, $\bar{\beta}^* = B^{-1} \sum_{b=1}^B \hat{\beta}^{*(b)}$ the bootstrap mean, and q_τ^* the τ -quantile of the bootstrap distribution $\{\hat{\beta}^{*(b)}\}_{b=1}^B$.

The **percentile** method uses the bootstrap quantiles directly:

$$CI_{\text{perc}} = [q_{\alpha/2}^*, q_{1-\alpha/2}^*]. \quad (4.6)$$

The **mean-centered** method reflects the bootstrap distribution around its mean and recenters the result at the point estimate:

$$CI_{\text{mc}} = [\hat{\beta} - (q_{1-\alpha/2}^* - \bar{\beta}^*), \hat{\beta} - (q_{\alpha/2}^* - \bar{\beta}^*)]. \quad (4.7)$$

The **bias-corrected** (BC) method adjusts the quantile levels using a bias-correction factor $\hat{z}_0$:

$$CI_{\text{BC}} = [q_{\alpha_1}^*, q_{\alpha_2}^*], \quad \alpha_1 = \Phi(2\hat{z}_0 + z_{\alpha/2}), \quad \alpha_2 = \Phi(2\hat{z}_0 + z_{1-\alpha/2}), \quad (4.8)$$

with Φ denoting the standard normal CDF and $z_\tau = \Phi^{-1}(\tau)$. This is the bias-corrected and accelerated (BCa) interval of DiCiccio and Efron [1996] with the acceleration factor $\hat{a}$ set to zero; the acceleration is not estimated.

The **normal approximation** method uses the bootstrap standard deviation to form a normal approximation interval:

$$CI_{\text{normal}} = [\hat{\beta} - z_{1-\alpha/2} \cdot \hat{\sigma}^*, \hat{\beta} + z_{1-\alpha/2} \cdot \hat{\sigma}^*], \quad (4.9)$$

where $\hat{\sigma}^* = \text{SD}(\hat{\beta}^{*(b)})$ is the bootstrap standard deviation.

The **basic** (reverse percentile) method reflects the bootstrap quantiles about the point estimate:

$$CI_{\text{basic}} = \left[2\hat{\beta} - q_{1-\alpha/2}^*, 2\hat{\beta} - q_{\alpha/2}^* \right]. \quad (4.10)$$

We adopt the basic interval of Eq. (4.10) as the primary method of inference. The bias-corrected and normal approximation intervals rely on normal approximations, for $\hat{z}_0$ and for the multiplier $z_{1-\alpha/2}$ respectively, which are doubtful under mixed orders of integration: the limit of $\hat{\rho}_i$ is non-standard, and the penalized estimator places positive probability on zero. The percentile, basic and mean-centered intervals make no such assumption.

Of these three, the basic interval is the direct inversion of the bootstrap principle. If $\hat{\beta} - \beta$ has approximately the distribution of $\hat{\beta}^* - \hat{\beta}$, inverting gives $[\hat{\beta} - (q_{1-\alpha/2}^* - \hat{\beta}), \hat{\beta} - (q_{\alpha/2}^* - \hat{\beta})]$, which is Eq. (4.10); this subtracts the bootstrap bias $\bar{\beta}^* - \hat{\beta}$. The mean-centered interval of Eq. (4.7) measures the same deviations from $\bar{\beta}^*$ rather than from $\hat{\beta}$, and so does not make that correction. The two are asymptotically equivalent, but they differ when the penalty places mass at zero: $\bar{\beta}^*$ is then pulled toward zero, and the mean-centered interval carries that pull into the interval. Section 4.2.3 shows that the difference matters in practice, particularly for coefficients whose true value is zero. We therefore report basic intervals throughout and give the other four in the simulation tables for comparison.

4.2 Simulations

We evaluate the finite-sample performance of the methodology by Monte Carlo simulation. Section 4.2.1 describes the two designs and the performance measures. Section 4.2.2 compares the rules for selecting the threshold c^* discussed in Section 2.1.3, and Sections 4.2.3 and 4.2.4 report estimation, variable selection and bootstrap coverage in the baseline and high-dimensional designs.

4.2.1 Simulation Design

We compare six estimators: OLS and unpenalized WEE on the transformed model of Eq. (2.5), and the four penalized combinations of Section 2.2, OLS-LASSO, OLS-SCAD, WEE-LASSO and WEE-SCAD.

Baseline design. The data generating process follows a predictive regression model with mixed integration orders and selective endogeneity:

$$y_t = \beta_0 + \beta_1 x_{1,t-1} + \beta_2 x_{2,t-1} + \beta_3 x_{3,t-1} + \beta_4 x_{4,t-1} + \epsilon_t, \quad (4.11)$$

where the true parameter values are $\beta_0 = 2$, $\beta_1 = 0.5$, $\beta_2 = 1$, $\beta_3 = 1.5$, and $\beta_4 = -1$. The predictor variables are designed to span the full spectrum of integration orders relevant to financial time series:

$$x_{1,t} = 0.4x_{1,t-1} + u_{1,t} \quad (\text{stationary, } \rho_1 = 0.4)$$

$$x_{2,t} = 0.8x_{2,t-1} + u_{2,t} \quad (\text{stationary, } \rho_2 = 0.8)$$

$$x_{3,t} = x_{3,t-1} + u_{3,t} \quad (\text{unit root, } \rho_3 = 1.0)$$

$$x_{4,t} = 0.99x_{4,t-1} + u_{4,t} \quad (\text{near-integrated, } \rho_4 = 0.99)$$

where $u_{i,t} \stackrel{i.i.d.}{\sim} N(0, 1)$ for $i = 1, 2, 3, 4$. This design captures the mixed integration environment that motivates the HD-WEE framework, with predictors ranging from moderate stationarity through near-unit-root behavior to exact integration.

Two of the predictors have innovations that are correlated with the regression error and two do not. Through the projection

$$\epsilon_t = 0.5u_{2,t} + u_{3,t} + v_t, \quad v_t \stackrel{i.i.d.}{\sim} N(0, 1), \quad (4.12)$$

$\eta_1 = 0$, $\eta_2 = 0.5$, $\eta_3 = 1.0$ and $\eta_4 = 0$, so two correction terms are needed and two

are not. The design therefore also checks whether the penalized methods set the two unneeded terms to zero while keeping the two that are needed.

High-dimensional design. The second design adds 21 noise predictors, $x_{5,t}, \dots, x_{25,t}$, with $\beta_i = 0$ and $\eta_i = 0$, so the transformed model has $2(25) + 1 = 51$ parameters. Each follows a stationary AR(1) process $x_{i,t} = \rho_i x_{i,t-1} + u_{i,t}$ with $u_{i,t} \stackrel{i.i.d.}{\sim} N(0, 1)$ and $\rho_i \sim \text{Uniform}(0, 0.7)$, redrawn in every replication. The first four predictors and the error are as in the baseline design.

Estimation and performance measures. In both designs we generate 500 observations in each replication and discard the first 50 to remove the effect of the initial conditions, leaving $n = 450$ observations for estimation; all reported results refer to this sample size. Each design has 1,000 Monte Carlo replications. The tuning parameter λ is selected by the SIC of Eq. (2.22) and the threshold c^* by the largest-gap rule of Eq. (2.15). For each of the four penalized methods we construct bootstrap confidence intervals with $B = 1,000$ replicates and $N(1, 1)$ weights, using all five constructions of Section 4.1.2. The Monte Carlo standard error of a coverage probability near 0.95 is about ± 0.007 .

For each coefficient, the mean squared error (MSE) is the average over replications of the squared difference between the estimate and the true value. False positives (FP) and false negatives (FN) are counted as coefficient–replication pairs, over every truly zero (respectively truly nonzero) coefficient in the design. In the baseline design the truly zero coefficients are η_1 and η_4 , so the FP denominator is $2 \times 1,000 = 2,000$; in the high-dimensional design each noise predictor adds a zero level coefficient and a zero correction coefficient. Counting pairs rather than replications keeps the two designs comparable, since a count of replications with at least one false positive saturates as the number of zero coefficients grows.

4.2.2 Selecting the Threshold

Table 4.1 shows how each rule for selecting c^* classifies the predictors over the same 1,000 replications of each design. No penalized fits or bootstrap replicates are involved, so the comparison isolates the classification step.

Table 4.1: How each threshold rule classifies, over 1,000 replications of each design with the truth $I^c = \{x_3, x_4\}$. “Empty” and “Full” are the fractions of replications selecting an empty I^c , which makes HD-WEE identical to OLS, or the full set of predictors. “Exact” recovers $\{x_3, x_4\}$; x_3 has $\rho = 1$ and x_4 $\rho = 0.99$, while x_1 ($\rho = 0.4$) and x_2 ($\rho = 0.8$) are stationary. Scaled rules divide the classification statistic by $\text{SD}(\Delta x_i)$. Cross-validation is restricted to non-degenerate classifications except where marked unrestricted.

Rule	Empty	Full	Exact	x_3	x_4	x_1	x_2	mean $ I^c $
<i>Baseline, $p = 4$</i>								
Cross-validation	0.000	0.000	0.286	0.904	0.688	0.002	0.304	1.90
scaled	0.000	0.000	0.283	0.907	0.688	0.000	0.312	1.91
unrestricted	0.401	0.112	0.161	0.546	0.446	0.113	0.231	1.34
Largest gap	0.000	0.000	0.845	0.986	0.889	0.000	0.030	1.91
scaled	0.000	0.000	0.834	0.993	0.911	0.000	0.070	1.97
Efficiency	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00
<i>High-dimensional, $p = 25$</i>								
Cross-validation	0.000	0.000	0.109	0.958	0.861	0.283	0.679	9.10
scaled	0.000	0.000	0.107	0.958	0.863	0.308	0.701	9.15
unrestricted	0.192	0.038	0.087	0.775	0.708	0.231	0.564	7.44
Largest gap	0.000	0.000	0.873	0.994	0.880	0.000	0.001	1.88
scaled	0.000	0.000	0.891	0.994	0.900	0.000	0.003	1.90
Efficiency	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.00

The efficiency criterion of Hong et al. [2024] selects an empty I^c in every replication of both designs, for the reason given in Section 2.1.3: it rewards the precision that downweighting gives up. Unrestricted cross-validation selects an empty or a full I^c in 51 per cent of baseline replications (40 per cent empty, 11 per cent full) and 23 per cent of high-dimensional ones. Restricting it to non-degenerate classifications does not solve the problem. In the baseline design it recovers the true classification in only 29 per cent of replications, placing the stationary x_2 ($\rho = 0.8$) in I^c in 30 per cent and leaving the near-integrated x_4 out in 31 per cent. In the high-dimensional design

it places x_2 in I^c in 68 per cent of replications and x_1 ($\rho = 0.4$) in 28 per cent, and selects 9.1 predictors on average against a truth of two. Downweighting a stationary predictor costs precision on its coefficient without any gain in validity.

The largest-gap rule recovers the true classification in 85 and 87 per cent of replications of the two designs. It never places x_1 in I^c , places x_2 there in at most 3 per cent, and places the exact unit root x_3 there in 99 per cent. When it misses, it usually leaves the near-integrated x_4 in I , which is the case where, as Section 4.2.3 shows, downweighting matters least. The scale-free variant behaves similarly, as expected when all innovations have unit variance. The largest-gap rule is therefore used in the remaining simulations and, in its scale-free form, in the empirical application of Chapter 5, where the predictors are measured in different units.

4.2.3 Baseline Results

Table 4.2: Baseline simulation results for $n = 450$ with $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top = (2.0, 0.5, 1.0, 1.5, -1.0)^\top$, 1,000 Monte Carlo replications. Monte Carlo standard deviations in parentheses, MSE in italics. The lower panel reports coverage of the basic bootstrap interval of Eq. (4.10) at the nominal 95% level, $B = 1,000$; the Monte Carlo standard error of a coverage estimate is ± 0.007 .

Method	β_0	β_1	β_2	β_3	β_4
OLS	2.005	0.499	0.997	1.496	-1.000
	(0.121)	(0.044)	(0.033)	(0.011)	(0.011)
	<i>0.0145</i>	<i>0.0019</i>	<i>0.0011</i>	<i>0.00013</i>	<i>0.00012</i>
WEE	2.005	0.499	0.997	1.497	-1.000
	(0.129)	(0.044)	(0.033)	(0.012)	(0.012)
	<i>0.0166</i>	<i>0.0019</i>	<i>0.0011</i>	<i>0.00014</i>	<i>0.00014</i>
OLS-LASSO	2.006	0.492	0.996	1.496	-1.000
	(0.120)	(0.045)	(0.033)	(0.011)	(0.011)
	<i>0.0145</i>	<i>0.0021</i>	<i>0.0011</i>	<i>0.00013</i>	<i>0.00012</i>
OLS-SCAD	2.005	0.499	0.997	1.496	-1.000
	(0.120)	(0.044)	(0.033)	(0.011)	(0.011)
	<i>0.0144</i>	<i>0.0019</i>	<i>0.0011</i>	<i>0.00013</i>	<i>0.00012</i>
WEE-LASSO	2.006	0.492	0.995	1.497	-0.999
	(0.129)	(0.045)	(0.033)	(0.012)	(0.012)
	<i>0.0166</i>	<i>0.0021</i>	<i>0.0011</i>	<i>0.00015</i>	<i>0.00014</i>
WEE-SCAD	2.005	0.499	0.997	1.497	-1.000
	(0.128)	(0.044)	(0.033)	(0.012)	(0.012)
	<i>0.0165</i>	<i>0.0019</i>	<i>0.0011</i>	<i>0.00014</i>	<i>0.00013</i>
<i>Coverage, basic interval</i>					
OLS-LASSO	0.940	0.928	0.946	0.931	0.950
OLS-SCAD	0.939	0.953	0.948	0.927	0.953
WEE-LASSO	0.935	0.938	0.948	0.944	0.941
WEE-SCAD	0.945	0.952	0.951	0.951	0.954

Estimation. All six methods recover the structural parameters accurately (Table 4.2).

Weighting costs some efficiency: the Monte Carlo standard deviation of $\hat{\beta}_0$ is 0.129 under WEE against 0.121 under OLS. The one systematic departure is the Adaptive LASSO's shrinkage of the smallest coefficient: both LASSO methods estimate

β_1 at 0.492 against a true value of 0.5, while both SCAD methods give 0.499. The same pattern appears for η_2 in Table 4.3. This is the bias that the SCAD penalty is designed to avoid [Fan and Li, 2001, Zou, 2006].

Bootstrap coverage. The lower panel of Table 4.2 compares the four penalized methods under the basic interval. The clearest difference is at the exact unit root, where WEE-SCAD covers β_3 at 0.951 against 0.927 for OLS-SCAD, a gain well outside Monte Carlo error. On the two stationary predictors, which the classification rule almost never downweights, the two methods agree to within 0.003. At the near-integrated x_4 the gain is negligible, 0.954 against 0.953, because OLS-SCAD already covers well there. The weighting therefore helps where persistence is strongest and changes little elsewhere.

The choice of interval also matters. For WEE-SCAD, averaged over the five β coefficients, the basic interval deviates least from nominal coverage (0.0026) and the percentile interval most (0.0080). The difference is largest for the two coefficients whose true value is zero, η_1 and η_4 , which these averages exclude: the basic interval covers them at 0.995 and 0.983, the mean-centered interval at only 0.382 and 0.300. This is the effect of the point mass at zero described in Section 4.1.2.

Variable selection. The unpenalized methods cannot set a coefficient to zero, so they retain nonzero estimates of η_1 and η_4 in every replication. All four penalized methods set these to zero in most replications, and none excludes a nonzero coefficient in any replication. False positives are rare for all four and slightly rarer for SCAD: 23 of 2,000 coefficient–replication pairs for both SCAD methods, against 31 and 34 for the two Adaptive LASSO methods (Table 4.3). The coordinate descent converged in every replication, and no bootstrap replicate failed.

Table 4.3: Baseline endogeneity parameter estimates for $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \eta_4)^\top = (0, 0.5, 1.0, 0)^\top$, $n = 450$, 1,000 replications. Monte Carlo standard deviations in parentheses, MSE in italics. False positives (FP) are counted as coefficient–replication pairs over the two truly zero coefficients, out of $2 \times 1,000 = 2,000$; false negatives (FN) over the two truly nonzero coefficients, out of 2,000.

Method	η_1	η_2	η_3	η_4
OLS	−0.000	0.500	1.000	−0.001
	(0.047)	(0.048)	(0.046)	(0.048)
	<i>0.0022</i>	<i>0.0023</i>	<i>0.0021</i>	<i>0.0023</i>
WEE	−0.000	0.500	1.000	−0.001
	(0.047)	(0.048)	(0.046)	(0.049)
	<i>0.0022</i>	<i>0.0023</i>	<i>0.0021</i>	<i>0.0024</i>
OLS-LASSO	0.001	0.492	0.996	−0.000
	(0.012)	(0.049)	(0.047)	(0.016)
	<i>0.0001</i>	<i>0.0024</i>	<i>0.0022</i>	<i>0.0003</i>
	FP: 31 / 2,000 (1.6%)			FN: 0
OLS-SCAD	0.000	0.500	1.000	−0.000
	(0.010)	(0.047)	(0.046)	(0.016)
	<i>0.0001</i>	<i>0.0022</i>	<i>0.0021</i>	<i>0.0002</i>
	FP: 23 / 2,000 (1.2%)			FN: 0
WEE-LASSO	0.000	0.491	0.996	−0.000
	(0.013)	(0.049)	(0.047)	(0.016)
	<i>0.0002</i>	<i>0.0025</i>	<i>0.0022</i>	<i>0.0003</i>
	FP: 34 / 2,000 (1.7%)			FN: 0
WEE-SCAD	0.000	0.500	1.000	−0.000
	(0.010)	(0.047)	(0.046)	(0.016)
	<i>0.0001</i>	<i>0.0022</i>	<i>0.0021</i>	<i>0.0003</i>
	FP: 23 / 2,000 (1.2%)			FN: 0

4.2.4 High-Dimensional Results

Table 4.4: High-dimensional simulation results for $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^\top = (2.0, 0.5, 1.0, 1.5, -1.0)^\top$ with $n = 450$, $p = 25$ predictors ($2p + 1 = 51$ parameters after the endogeneity correction), 1,000 Monte Carlo replications, $B = 1,000$ bootstrap samples. Monte Carlo standard deviations in parentheses, MSE in italics. False positives are counted as coefficient–replication pairs: out of $21 \times 1,000 = 21,000$ for β and $23 \times 1,000 = 23,000$ for η .

Method	β_0	β_1	β_2	β_3	β_4
OLS	2.001	0.501	0.998	1.496	−1.000
	(0.123)	(0.045)	(0.038)	(0.011)	(0.012)
	<i>0.0151</i>	<i>0.0020</i>	<i>0.0014</i>	<i>0.00014</i>	<i>0.00014</i>
	β FP: all 21,000		η FP: all 23,000		FN: 0
WEE	2.001	0.501	0.998	1.498	−1.000
	(0.132)	(0.045)	(0.038)	(0.012)	(0.013)
	<i>0.0174</i>	<i>0.0020</i>	<i>0.0014</i>	<i>0.00014</i>	<i>0.00017</i>
	β FP: all 21,000		η FP: all 23,000		FN: 0
OLS-LASSO	2.001	0.472	0.992	1.496	−0.999
	(0.110)	(0.046)	(0.036)	(0.010)	(0.011)
	<i>0.0120</i>	<i>0.0029</i>	<i>0.0013</i>	<i>0.00013</i>	<i>0.00012</i>
	β FP: 124 (0.59%)		η FP: 64 (0.28%)		FN: 0
OLS-SCAD	2.001	0.498	0.998	1.496	−1.000
	(0.109)	(0.046)	(0.035)	(0.010)	(0.011)
	<i>0.0119</i>	<i>0.0021</i>	<i>0.0012</i>	<i>0.00012</i>	<i>0.00012</i>
	β FP: 44 (0.21%)		η FP: 22 (0.10%)		FN: 0
WEE-LASSO	2.003	0.473	0.992	1.495	−0.994
	(0.130)	(0.046)	(0.036)	(0.012)	(0.013)
	<i>0.0169</i>	<i>0.0028</i>	<i>0.0013</i>	<i>0.00016</i>	<i>0.00020</i>
	β FP: 163 (0.78%)		η FP: 75 (0.33%)		FN: 0
WEE-SCAD	2.002	0.498	0.998	1.498	−1.000
	(0.117)	(0.046)	(0.035)	(0.011)	(0.012)
	<i>0.0137</i>	<i>0.0021</i>	<i>0.0012</i>	<i>0.00012</i>	<i>0.00014</i>
	β FP: 46 (0.22%)		η FP: 23 (0.10%)		FN: 0

Table 4.5: High-dimensional endogeneity parameter estimates for $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \eta_4)^\top = (0, 0.5, 1.0, 0)^\top$, $n = 450$, $p = 25$, 1,000 replications. Monte Carlo standard deviations in parentheses, MSE in italics. Entries of exactly zero indicate that the coefficient was set to zero in every replication.

Method	η_1	η_2	η_3	η_4
OLS	0.002	0.499	1.001	−0.000
	(0.050)	(0.049)	(0.050)	(0.048)
	<i>0.0025</i>	<i>0.0024</i>	<i>0.0025</i>	<i>0.0023</i>
WEE	0.002	0.499	1.001	0.000
	(0.050)	(0.049)	(0.051)	(0.048)
	<i>0.0025</i>	<i>0.0024</i>	<i>0.0026</i>	<i>0.0023</i>
OLS-LASSO	−0.000	0.462	0.984	0.000
	(0.005)	(0.052)	(0.049)	(0.000)
	<i>0.00003</i>	<i>0.0041</i>	<i>0.0026</i>	<i>0.0000</i>
OLS-SCAD	0.000	0.494	1.001	0.000
	(0.000)	(0.053)	(0.048)	(0.000)
	<i>0.0000</i>	<i>0.0029</i>	<i>0.0023</i>	<i>0.0000</i>
WEE-LASSO	−0.000	0.463	0.983	0.000
	(0.006)	(0.051)	(0.049)	(0.000)
	<i>0.00003</i>	<i>0.0040</i>	<i>0.0027</i>	<i>0.0000</i>
WEE-SCAD	0.000	0.494	1.001	0.000
	(0.000)	(0.053)	(0.048)	(0.000)
	<i>0.0000</i>	<i>0.0028</i>	<i>0.0023</i>	<i>0.0000</i>

Estimation. All six methods recover the four nonzero coefficients despite the 21 irrelevant predictors, with Monte Carlo standard deviations close to those of the baseline design (Table 4.4). The Adaptive LASSO’s shrinkage of β_1 is larger here, 0.472–0.473 against 0.498 for SCAD. Among the penalized methods, weighting again costs some precision on the intercept, with standard deviations of 0.117–0.130 against 0.109–0.110.

Variable selection. No method excludes a nonzero coefficient. Among the penalized methods, false positives on $\boldsymbol{\beta}$ occur in 0.21 and 0.22 per cent of coefficient–replication pairs for OLS-SCAD and WEE-SCAD, and in 0.59 and 0.78 per cent for OLS-LASSO and WEE-LASSO, with the same ordering on $\boldsymbol{\eta}$. Both SCAD methods set η_1 and η_4

to zero in every replication (Table 4.5), consistent with Theorem 3.4.1.

Table 4.6: Coverage probabilities for the high-dimensional setting ($p = 25$, $2p + 1 = 51$ parameters) across all five bootstrap CI methods. MC SE: ± 0.007 .

Method	Parameter	Percentile	Basic	Mean-centered	BC	Normal
OLS-LASSO	β_0 (intercept)	0.959	0.966	0.963	0.967	0.967
	β_1 ($\rho = 0.4$)	0.812	0.969	0.958	0.949	0.948
	β_2 ($\rho = 0.8$)	0.916	0.933	0.929	0.938	0.929
	β_3 ($\rho = 1.0$)	0.931	0.943	0.942	0.942	0.942
	β_4 ($\rho = 0.99$)	0.929	0.940	0.938	0.942	0.939
OLS-SCAD	β_0 (intercept)	0.965	0.962	0.964	0.965	0.969
	β_1 ($\rho = 0.4$)	0.930	0.963	0.968	0.958	0.964
	β_2 ($\rho = 0.8$)	0.938	0.939	0.938	0.934	0.938
	β_3 ($\rho = 1.0$)	0.939	0.943	0.946	0.942	0.945
	β_4 ($\rho = 0.99$)	0.936	0.941	0.937	0.938	0.941
WEE-LASSO	β_0 (intercept)	0.906	0.953	0.945	0.949	0.940
	β_1 ($\rho = 0.4$)	0.837	0.969	0.957	0.953	0.953
	β_2 ($\rho = 0.8$)	0.912	0.938	0.931	0.934	0.930
	β_3 ($\rho = 1.0$)	0.898	0.949	0.938	0.944	0.938
	β_4 ($\rho = 0.99$)	0.827	0.936	0.907	0.923	0.907
WEE-SCAD	β_0 (intercept)	0.964	0.963	0.964	0.965	0.965
	β_1 ($\rho = 0.4$)	0.926	0.966	0.969	0.963	0.969
	β_2 ($\rho = 0.8$)	0.935	0.939	0.937	0.937	0.939
	β_3 ($\rho = 1.0$)	0.946	0.952	0.951	0.946	0.954
	β_4 ($\rho = 0.99$)	0.948	0.946	0.945	0.945	0.947

Bootstrap coverage. Table 4.6 reports coverage under all five intervals. At the exact unit root WEE-SCAD again covers best, but its gain over OLS-SCAD under the basic interval is 0.009, within Monte Carlo error, so in this design the effect of the weighting is modest. The percentile interval, which makes no bias correction, undercovers the Adaptive LASSO's shrunken β_1 , at 0.812 under OLS-LASSO and 0.837 under WEE-LASSO, while the basic interval covers it at 0.969 under both.

Across both designs, WEE-SCAD has the smallest average deviation from nominal coverage under the basic interval, 0.0026 in the baseline and 0.0092 here, against 0.0084 to 0.0138 for the other three penalized methods, and its coverage changes least

across the five interval constructions. The margins between the methods are small. Together with its nearly unbiased estimates and low false-positive rate, however, they make WEE-SCAD the best-performing of the four penalized methods in these designs. Chapter 5 applies it to stock return predictability.

CHAPTER 5: EMPIRICAL APPLICATION: EXCESS STOCK RETURN PREDICTABILITY

Prediction of equity returns is one of the most studied problems in financial economics. Welch and Goyal [2008] showed that many seemingly predictive relationships disappear when proper econometric methods are applied. Three features of the problem make it a natural application of the methods developed in this dissertation: the order of integration of the standard predictors is uncertain, their innovations are correlated with return innovations, and they are better examined jointly than one at a time. We analyze U.S. equity return predictability using the dataset of Welch and Goyal [2008], updated to 2021.

5.1 Methodological Framework

The model is the predictive regression of Section 2.1, with the monthly excess stock return r_t as the response:

$$r_t = \beta_0 + \sum_{i=1}^p \beta_i x_{i,t-1} + \epsilon_t, \quad t = 1, 2, \dots, n,$$

where $x_{i,t-1}$ is the i th lagged predictor. Each predictor follows an autoregressive process,

$$x_{i,t} = \rho_i x_{i,t-1} + u_{i,t},$$

where ρ_i determines its order of integration. Because the innovations of predictors such as the valuation ratios are correlated with return innovations [Campbell and

Yogo, 2006, Stambaugh, 1999], we use the linear projection of Section 2.1,

$$\epsilon_t = \sum_{i=1}^p \eta_i u_{i,t} + v_t.$$

which gives the transformed model

$$r_t = \alpha_0 + \sum_{i=1}^p \alpha_i x_{i,t-1} + \sum_{i=1}^p \eta_i \Delta x_{i,t} + v_t, \quad (5.1)$$

where $\alpha_i = \beta_i - \eta_i(\rho_i - 1)$ and the correction coefficient η_i captures the correlation between the return innovation and the innovation of the i th predictor.

5.2 Data

The data are monthly and span 50 years. The response r_t is the log return on the CRSP value-weighted S&P 500 index, including dividends, minus the log one-month Treasury bill rate, as in Hong et al. [2024].

We use $p = 8$ financial and macroeconomic predictors. Two series in the original dataset are excluded because they are nearly or exactly linear combinations of included ones: the dividend–price ratio (dp) differs from the dividend yield (dy) only in the timing of the price in the denominator, and the dividend payout ratio (de) is exactly dp minus the earnings–price ratio (ep). Including them would make the joint design nearly singular.

The eight predictors are: the dividend yield (dy), the difference between the log of dividends and the log of lagged prices; the earnings–price ratio (ep), the difference between the log of earnings and the log of prices; the book-to-market ratio (bm), the ratio of book value to market value for the Dow Jones Industrial Average; net equity expansion (ntis), the ratio of the 12-month moving sum of net issues by NYSE-listed stocks to the total end-of-year market capitalization of NYSE stocks; the term spread (tms), the difference between the long-term yield on government bonds and

the Treasury bill rate; the default yield spread (dfy), the difference between BAA- and AAA-rated corporate bond yields; inflation (infl), the monthly inflation rate computed from the Consumer Price Index for All Urban Consumers published by the Bureau of Labor Statistics; and the Treasury bill rate (tbl), the 3-month Treasury bill secondary market rate from FRED. All series are from the dataset of Welch and Goyal [2008], as in Hong et al. [2024]. The long-term yield (lty) is also excluded, because $tms = lty - tbl$ holds exactly in these data, in levels and in differences, so the three coefficients are not separately identified.

5.3 Empirical Implementation

We analyze three economic periods that reflect different policy regimes and market conditions. Period I (December 1969 – March 1987) is the pre-crash era of high inflation, the oil crises, and the monetary policy changes under Federal Reserve chairs Arthur Burns and Paul Volcker. Period II (April 1987 – December 2000) covers the post-crash recovery, including the technology boom, financial deregulation and the so-called “Great Moderation”. Period III (January 2001 – December 2021) covers the post-dot-com era, including the financial crisis, unconventional monetary policy and the COVID-19 pandemic.

In each period we estimate the model by SCAD-penalized HD-WEE, with λ selected by the SIC and c^* by the largest-gap rule of Eq. (2.15) applied to the scale-free statistic of Section 2.1.3. Table 5.1 reports the resulting classification of each predictor, under both the unscaled and the scale-free statistic, together with its estimated autoregressive coefficient $\hat{\rho}_i$. Inflation is the only predictor whose persistence is far from unity, with $\hat{\rho}_i$ between 0.31 and 0.59, and it is the only predictor the scale-free rule leaves in I ; every other predictor has $\hat{\rho}_i \geq 0.91$ in every period.

Table 5.1: Classification of predictors into I^c by period, under the classification statistic of Section 2.1.3 and under the scale-free variant that divides it by the sample standard deviation of Δx_i . $\hat{\rho}_i$ is the estimated autoregressive coefficient from Eq. (2.12). A bullet marks placement in I^c . The long-term yield lty is excluded throughout because $\text{tms} = \text{lty} - \text{tbl}$ holds exactly, in levels and in differences, so only two of the three are identified.

Predictor	$\hat{\rho}_i$			I^c , unscaled			I^c , scale-free		
	I	II	III	I	II	III	I	II	III
dy	0.978	1.002	0.965	•	•	•	•	•	•
ep	0.994	0.989	0.978	•	•	•	•	•	•
bm	0.988	0.993	0.944	•	•		•	•	•
ntis	0.976	0.982	0.980				•	•	•
tbl	0.972	0.990	0.983				•	•	•
tms	0.910	0.981	0.976				•	•	•
infl	0.590	0.307	0.520						
dfy	0.942	0.952	0.980				•	•	•
n	206	163	250						

The table also shows why the scale-free form is needed. The unscaled statistic grows with $\max_t |x_{i,t}|$, and the valuation ratios are log ratios of order one while tbl, tms, infl and dfy are rates in decimals, so under the unscaled statistic the ordering reflects units as much as persistence. In Period II, for example, dy has an unscaled statistic of 0.316 and tbl one of 0.014, although their $\hat{\rho}_i$ are 1.002 and 0.990. The unscaled rule disagrees with $\hat{\rho}_i$ on four or five predictors per period, treating tbl, tms, ntis and dfy as stationary, whereas the scale-free rule agrees with $\hat{\rho}_i$ for all eight predictors in all three periods.

The separation on which the rule relies is weaker in these samples than in the simulations. The ratio at the selected gap is 2.0 in Period I and 2.8 in Period II, but only 1.4 in Period III, where the scale-free statistics of the persistent predictors range from 2.7 to 7.4 and inflation, at 2.0, is the only clear outlier. With 163 to 250 observations, a stationary but highly persistent series also has a statistic well above zero, so the classification in these samples rests mainly on the contrast between inflation and the other predictors.

Inference uses the random weighted bootstrap of Section 4.1 with $B = 1,000$ repli-

cates and exponential weights, $\zeta_t \sim \text{Exp}(1)$. As discussed in Section 4.1.1, $N(1, 1)$ weights are equally valid but can be negative, and with 17 parameters and 163 to 250 observations this is enough to make the weighted system singular in some replicates. Under $N(1, 1)$ weights, 2, 26 and 81 of the 1,000 replicates failed in Periods I, II and III, and in Period III the bootstrap standard deviations of the remaining replicates were about an order of magnitude larger than under exponential weights, with identical point estimates. With exponential weights no replicate failed in any period.

5.4 Empirical Results

Table 5.2 reports the estimates for the three periods, with 95% basic bootstrap intervals, consistent with the simulation evidence of Section 4.2. A predictor is considered significant if its basic interval excludes zero. The table also reports the bootstrap mean and standard deviation of each coefficient and its zero-fraction, the proportion of bootstrap replicates in which it is set to zero, as stability diagnostics. The percentile interval cannot be used to judge significance for a coefficient whose zero-fraction exceeds 2.5%, since one of its endpoints is then exactly zero by construction; this is the point mass at zero that motivates the basic interval (Section 4.1.2).

Period I (1969–1987). Five predictors are significant along with the intercept: *dy*, *ep*, *ntis*, *tms* and *dfy*. They fall into two groups. For *ep*, *ntis* and *dfy* the zero-fractions are below 5% and the bootstrap means lie within about half a bootstrap standard deviation of the estimates. For *dy* and *tms* the zero-fractions are about 32%, and the bootstrap mean of *tms* (0.092) lies more than one bootstrap standard deviation below its estimate (0.208), so these two should be read with caution. Book-to-market is not significant, and the bill rate and inflation are set to zero in most replicates, with zero-fractions of 82% and 78%. The correction terms for *dy*, *ep* and *bm* are significant.

Period II (1987–2000). Three predictors are significant along with the intercept: dy , ep and tms . The earnings–price ratio and the term spread are stable, with zero-fractions below 1% and bootstrap means within 0.03 bootstrap standard deviations of the estimates. The dividend yield is less so, with a zero-fraction of 9% and a bootstrap mean 0.44 standard deviations from its estimate. Inflation and the default spread are set to zero, and book-to-market, net issuance and the bill rate enter only through their correction terms. The correction terms for ep and bm are significant.

Period III (2001–2021). Three predictors are significant along with the intercept, dy , bm and $ntis$, and ep is significant at the margin, its interval having a lower endpoint of 0.0004. All four are stable, with zero-fractions of 5% or less and bootstrap means within about 0.2 bootstrap standard deviations of the estimates. The bill rate, the term spread, inflation and the default spread are not significant. The correction terms for ep , bm , tbl , tms and dfy are significant.

Table 5.2: Return predictability across three economic periods using SCAD-penalized HD-WEE, with the scale-free classification rule of Table 5.1. Estimates in bold have a 95% basic bootstrap interval excluding zero.

Variable	Estimate	Boot. mean	Boot. SD	95% basic CI	Zero %
Period I ($n = 206$)					
Intercept	0.0034	0.0035	0.0006	(0.0022 , 0.0044)	0.0
dy	-0.0365 [†]	-0.0226	0.0187	(-0.0731 , -0.0167)	31.9
ep	0.0408 [†]	0.0357	0.0154	(0.0190 , 0.0819)	0.0
bm	-0.0227 [†]	-0.0286	0.0192	(-0.0517, 0.0105)	0.0
ntis	0.2246 [†]	0.2019	0.0714	(0.1432 , 0.4492)	4.5
tbl	0.0804 [†]	0.0237	0.0545	(-0.0183, 0.1608)	81.6
tms	0.2077 [†]	0.0924	0.1029	(0.0815 , 0.4153)	31.8
infl	0	-0.0161	0.0438	—	77.5
dfy	-1.0365 [†]	-0.8872	0.2850	(-1.7525 , -0.6450)	1.7
<i>Significant $\hat{\eta}_i$: dy, ep, bm</i>					
Period II ($n = 163$)					
Intercept	0.0072	0.0071	0.0009	(0.0055 , 0.0090)	0.0
dy	-0.0494 [†]	-0.0414	0.0183	(-0.0988 , -0.0336)	9.2
ep	0.0633	0.0630	0.0146	(0.0337 , 0.0929)	0.0
bm	0.0022 [†]	-0.0327	0.0553	(-0.0061, 0.1651)	13.3
ntis	0.0083 [†]	0.0831	0.1853	(-0.6068, 0.0225)	21.7
tbl	-0.0032 [†]	0.0792	0.2198	(-0.7664, 0.0191)	51.7
tms	1.2017	1.1958	0.2251	(0.8088 , 1.7358)	0.4
infl	0	0.0117	0.1262	—	88.7
dfy	0	0.9091	1.4232	—	64.0
<i>Significant $\hat{\eta}_i$: ep, bm</i>					
Period III ($n = 250$)					
Intercept	0.0061	0.0060	0.0017	(0.0027 , 0.0097)	0.0
dy	0.1507	0.1409	0.0502	(0.0713 , 0.3013)	4.7
ep	0.0288	0.0269	0.0185	(0.0004 , 0.0595)	3.0
bm	-0.2635	-0.2528	0.1275	(-0.5657 , -0.0406)	0.0
ntis	0.8924	0.8782	0.2924	(0.3887 , 1.7847)	4.0
tbl	-0.2028	-0.2542	0.1783	(-0.4056, 0.2308)	4.0
tms	-0.0872	-0.1290	0.1528	(-0.1789, 0.4439)	4.7
infl	0.1287 [†]	0.3954	0.4785	(-1.3377, 0.2575)	39.0
dfy	-1.9837 [†]	-1.4960	1.4097	(-4.2331 , 0.3121)	15.3
<i>Significant $\hat{\eta}_i$: ep, bm, tbl, tms, dfy</i>					

Notes: $B = 1,000$, $\zeta_t \sim \text{Exp}(1)$; no replicate was discarded in any period. An estimate of 0 means the coefficient was set to zero by SCAD; its bootstrap mean and zero-fraction are still reported because it is nonzero in some replicates. For bm, ntis and tbl in Period II and for tms and infl in Period III the level coefficient is zero ($\hat{\alpha}_i = 0$), and the reported $\hat{\beta}_i$ comes entirely from the correction term $\hat{\eta}_i(\hat{\rho}_i - 1)$. The long-term yield lty is excluded because $\text{tms} = \text{lty} - \text{tbl}$ holds exactly in these data. [†] Bootstrap bias above 0.3 bootstrap standard deviations or zero-fraction above 15%.

5.5 Comparison with Single-Variable Analysis

We now compare these results with the single-variable WEE analysis of Hong et al. [2024], which examined the same predictors one at a time over the same three periods, without regularization.

In Period I, Hong et al. [2024] find dfy , $ntis$ and tms significant, and the dividend payout ratio significant at the 10% level; that variable is not among the eight we use. All three of their significant predictors are also significant in the joint model, which in addition finds ep significant, and dy with the caution noted above.

In Period II, Hong et al. [2024] find no significant predictors, whereas the joint model finds dy , ep and tms significant. In this period, predictability is detected only when the predictors are modeled jointly.

In Period III, Hong et al. [2024] find lty and tbl significant. The long yield is not separately identified in our specification, since we drop it to break the exact relation $tms = lty - tbl$, so its information enters only through tms and tbl . Neither is significant once the valuation ratios and net issuance are included; the joint model instead finds dy , bm and $ntis$ significant, with ep at the margin.

The two analyses therefore lead to different conclusions about which predictors carry independent predictive content. Some predictors that are significant on their own lose significance once the others are accounted for, and some that are insignificant on their own become significant jointly. The two analyses are best read together rather than as substitutes. The earnings–price ratio is significant in all three periods, though only marginally after 2001; Gao et al. [2026], using a different estimator and a large set of macroeconomic controls, find a positive coefficient on the log earnings–price ratio for S&P 500 returns that is marginally significant before 1994 and insignificant afterward.

Signs of the joint coefficients. Several joint estimates have signs opposite to their univariate counterparts and to what the predictability literature leads one to expect.

The dividend yield is negative in Periods I and II, book-to-market is negative in Periods I and III, and net equity issuance is positive and the default spread negative in both Periods I and III. For issuance, the market-timing argument of Baker and Wurgler [2000], carried over to net equity expansion by Welch and Goyal [2008], implies a negative coefficient. Two explanations can be examined directly from the estimates.

The first is the recovery formula $\hat{\beta}_i = \hat{\alpha}_i + \hat{\eta}_i(\hat{\rho}_i - 1)$, and it is not responsible: the correction term never reverses the sign of a nonzero level coefficient in any period. The correction is not always small, however. It accounts for about half of the bill-rate coefficient in Period III, and for the whole of each coefficient whose level coefficient is zero (listed in the notes to Table 5.2).

The second is that the joint estimates are partial rather than marginal coefficients. Each measures the predictive content of one variable with the others held fixed, and when the predictors are strongly correlated a partial coefficient can have the opposite sign to the marginal one. The correlations are strong in the first two periods (Table 5.3): the variance inflation factors of the valuation ratios are 9.2 to 13.9 in Period I, those of *dy*, *bm* and *tbl* are 17.4, 18.4 and 16.5 in Period II, and the correlation between net issuance and the bill rate reaches -0.95 in Period II. In Period III every variance inflation factor is 5.1 or less, so correlation among the predictors explains less of the pattern there.

The penalty selects among coefficients but does not resolve which of two correlated predictors carries the signal. The joint results are therefore best read as statements about the predictive content of the set of valuation variables rather than of each one separately. Chapter 6 draws these findings together with the theoretical and simulation results of the preceding chapters.

Table 5.3: Correlation matrices of the predictor levels and variance inflation factors, by period. The long-term yield lty is excluded because $tms = lty - tbl$ holds exactly in these data.

<i>Period I (1969–1987), n = 206</i>								
	dy	ep	bm	ntis	tbl	tms	infl	dfy
dy	1.00	0.90	0.84	−0.27	0.63	−0.11	0.28	0.47
ep	0.90	1.00	0.91	−0.14	0.65	−0.27	0.49	0.25
bm	0.84	0.91	1.00	0.07	0.55	−0.32	0.48	0.28
ntis	−0.27	−0.14	0.07	1.00	−0.17	−0.30	0.18	−0.29
tbl	0.63	0.65	0.55	−0.17	1.00	−0.56	0.38	0.41
tms	−0.11	−0.27	−0.32	−0.30	−0.56	1.00	−0.55	0.16
infl	0.28	0.49	0.48	0.18	0.38	−0.55	1.00	−0.17
dfy	0.47	0.25	0.28	−0.29	0.41	0.16	−0.17	1.00
VIF	9.2	13.9	10.7	1.8	4.6	3.3	2.0	2.2
<i>Period II (1987–2000), n = 163</i>								
	dy	ep	bm	ntis	tbl	tms	infl	dfy
dy	1.00	0.78	0.92	−0.21	0.26	0.46	0.24	0.49
ep	0.78	1.00	0.71	−0.58	0.64	−0.01	0.29	0.36
bm	0.92	0.71	1.00	−0.34	0.38	0.39	0.31	0.71
ntis	−0.21	−0.58	−0.34	1.00	−0.95	0.64	−0.27	−0.44
tbl	0.26	0.64	0.38	−0.95	1.00	−0.63	0.33	0.42
tms	0.46	−0.01	0.39	0.64	−0.63	1.00	−0.03	0.15
infl	0.24	0.29	0.31	−0.27	0.33	−0.03	1.00	0.23
dfy	0.49	0.36	0.71	−0.44	0.42	0.15	0.23	1.00
VIF	17.4	7.2	18.4	11.4	16.5	8.1	1.2	3.6
<i>Period III (2001–2021), n = 250</i>								
	dy	ep	bm	ntis	tbl	tms	infl	dfy
dy	1.00	−0.18	0.59	−0.53	−0.36	0.14	−0.23	0.65
ep	−0.18	1.00	0.43	0.11	0.13	−0.23	0.01	−0.54
bm	0.59	0.43	1.00	−0.01	−0.32	0.21	−0.12	0.22
ntis	−0.53	0.11	−0.01	1.00	−0.10	0.36	0.09	−0.40
tbl	−0.36	0.13	−0.32	−0.10	1.00	−0.54	0.09	−0.19
tms	0.14	−0.23	0.21	0.36	−0.54	1.00	−0.05	0.27
infl	−0.23	0.01	−0.12	0.09	0.09	−0.05	1.00	−0.26
dfy	0.65	−0.54	0.22	−0.40	−0.19	0.27	−0.26	1.00
VIF	5.1	3.7	4.6	2.7	1.7	2.0	1.1	3.3

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Summary of Contributions

This dissertation develops a unified framework for predictive regression that addresses three econometric challenges together: (1) uncertain orders of integration of the predictors, (2) potential endogeneity of multiple state variables, and (3) high-dimensional parameter spaces. The methodology extends the WEE approach of Hong et al. [2024] from single-predictor models to high-dimensional settings in which multiple predictors with mixed orders of integration may be endogenous.

The key step is to combine the linear projection approach to endogeneity with regularization within the WEE framework. The projection removes the bias from correlated innovations without requiring the researcher to specify which predictors are endogenous, at the cost of at least doubling the parameter space; the regularization keeps that enlarged parameter space manageable by setting unneeded correction terms to zero. No *a priori* assumptions about the stationarity or endogeneity of the predictors are needed. Chapter 3 shows that, with probability at least $1 - \varepsilon - o(1)$, the penalized estimating equation has a root that sets the zero coefficients to zero and estimates the nonzero coefficients as if the true model were known, at rate $\sqrt{n}$ for stationary predictors and n for integrated predictors (Theorem 3.4.1 and Lemma 3.4.2), and that these estimates are asymptotically normal after normalization (Theorem 3.5.1). The number of predictors may grow exponentially with the sample size.

The simulations show good finite-sample performance in both the baseline and the high-dimensional designs. No method excludes a genuinely nonzero coefficient in either design. The penalties differ in selection and bias: with $p = 25$ predictors and $2p + 1 = 51$ parameters, the SCAD methods record false positives in 0.21 to 0.22 per

cent of coefficient–replication pairs on β , against 0.59 to 0.78 per cent for the Adaptive LASSO, and the Adaptive LASSO shrinks the smallest nonzero coefficient, giving 0.472 to 0.473 for $\beta_1 = 0.5$ against 0.498 under SCAD. Under the basic bootstrap interval, WEE-SCAD has the smallest average deviation from nominal coverage of the four penalized methods in both designs, although the margins are small. The weighting improves coverage at the exact unit root by 0.024 in the baseline design; in the high-dimensional design the corresponding gain is 0.009, within Monte Carlo error. The choice of interval also matters: the percentile interval undercovers the Adaptive LASSO’s small coefficient substantially, while the basic interval does not. For the threshold c^* , the efficiency criterion of Hong et al. [2024] degenerates in the multivariate setting and block cross-validation over-selects, whereas the largest-gap rule recovers the true classification in 85 and 87 per cent of replications.

The empirical application to stock return predictability, using eight financial and macroeconomic predictors over three sub-periods, illustrates the approach. In Period I the dividend yield, earnings–price ratio, net equity issuance, term spread and default spread are significant, the dividend yield and term spread less firmly than the others; in Period II the dividend yield, earnings–price ratio and term spread are significant; and in Period III the dividend yield, book-to-market and net equity issuance are significant, with the earnings–price ratio at the margin. The correction terms of the earnings–price and book-to-market ratios are significant in every period, and those of the bill rate, term spread and default spread also in Period III. Inference uses exponential rather than $N(1, 1)$ bootstrap weights, since negative weights made the weighted system singular in some replicates. Joint and single-variable analyses differ: all three predictors that Hong et al. [2024] find significant in Period I remain significant in the joint model, and in Period II, where they find none, predictability appears only jointly. Several joint coefficients have signs opposite to those expected from single-variable analyses, which is consistent with strong correlation among the

predictors, particularly in the first two periods. The joint results are therefore best read as statements about sets of related predictors, and the two analyses are best read together.

6.2 Limitations

The asymptotic theory is for the coefficients of the transformed model with fixed tuning parameters. Extending it to the recovered structural coefficients and to data-driven choices of the threshold and penalty parameters is left for future work; in particular, the consistency of the largest-gap rule is not established. Nor is there a proof that the random weighted bootstrap is valid for the penalized estimator: the result of Hong et al. [2024] covers the unpenalized case, and the evidence for our intervals is the simulation coverage reported in Chapter 4. Finally, the separation on which the largest-gap rule relies is weak at the sample sizes of the empirical application. In the most recent period the selected gap is a ratio of only 1.4, and the classification rests mainly on the contrast between inflation and the other predictors.

6.3 Future Work

A first direction concerns the threshold c^* . By construction, the largest-gap rule places at least one predictor in each of I and I^c , so it would downweight some predictor even if all were stationary (Section 2.1.3). A data-dependent minimum value of c^* , below which no predictor is downweighted, would remove this restriction. How to construct such a bound, and whether the resulting rule classifies the predictors consistently, remain open questions.

A second direction concerns the constancy of the coefficients. Throughout this dissertation, each coefficient β_i is treated as constant within each sample period. The empirical application of Chapter 5 suggests that this is a simplification, since the significance of the predictors shifts across the three sub-periods: the default spread is significant only in Period I and book-to-market only in Period III, and the dividend

yield changes sign, from negative in Periods I and II to positive in Period III. Rather than pre-selecting a small number of discrete regimes, a natural extension is to let each coefficient vary continuously with time, $\beta_i(t)$, and let the data reveal the shape of that variation.

We have developed a preliminary time-varying-coefficient (TVC) extension of the WEE-SCAD framework along these lines. The model retains the endogeneity correction used throughout this dissertation, projecting the model error onto each predictor's innovations plus an independent white noise term, but allows the projected coefficients $\alpha_i(t)$ to be smooth, unspecified functions of time rather than constants, which turns the estimation problem into a semiparametric varying-coefficient model. We estimate $\alpha_i(t)$ by *profile least squares*: a local linear kernel-weighted regression estimates each coefficient function at a given point in time with the projection parameters η_i held fixed, an ordinary least squares step then re-estimates η_i from the resulting residuals, and the two steps are iterated to convergence. Inference again uses a random weighted bootstrap, applied at each stage of the profiling procedure, in the same spirit as the bootstrap of Chapter 4. In preliminary simulations with one stationary ($\rho = 0.6$) and one near-unit-root ($\rho = 0.99$) predictor and true coefficient functions of varying shape, the estimated coefficient functions track their true counterparts closely, and the bootstrap confidence bands have coverage close to the nominal 95% level.

This extension is preliminary and is not developed into a full chapter of this dissertation. Several steps remain before it would meet the standard of the constant-coefficient framework: extending the current two-predictor formulation to the high-dimensional setting; incorporating SCAD or Adaptive LASSO regularization into the profile least squares step, so that unneeded correction terms can again be set to zero; and, most importantly, establishing the asymptotic theory, namely consistency and a limiting distribution for the profile estimator and validity of the corresponding

bootstrap, analogous to Theorems 3.4.1 and 3.5.1.

REFERENCES

- M. Adler and B. Lehmann. Deviations from purchasing power parity in the long run. *Journal of Finance*, 38:1471–1487, 1983.
- Y. Amihud and C. M. Hurvich. Predictive regressions: A reduced-bias estimation method. *Journal of Financial and Quantitative Analysis*, 39:813–841, 2004.
- Y. Amihud, C. M. Hurvich, and Y. Wang. Multiple-predictor regressions: Hypothesis testing. *Review of Financial Studies*, 22:413–434, 2009.
- M. Baker and J. Wurgler. The equity share in new issues and aggregate stock returns. *Journal of Finance*, 55(5):2219–2257, 2000.
- I. V. Basawa, A. K. Mallik, W. P. McCormick, J. H. Reeves, and R. L. Taylor. Bootstrap for unstable first-order autoregressive processes. *Annals of Statistics*, 19(2):1098–1101, 1991.
- C. Bergmeir and J. M. Benítez. On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191:192–213, 2012.
- P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253, 2011.
- Z. Cai and Y. Wang. Testing predictive regression models with nonstationary regressors. *Journal of Econometrics*, 178:4–14, 2014.
- Z. Cai, Y. Wang, and Y. Wang. Testing instability in a predictive regression model with nonstationary regressors. *Econometric Theory*, 31:953–980, 2015.
- J. Y. Campbell and M. Yogo. Efficient tests of stock return predictability. *Journal of Financial Economics*, 81:27–60, 2006.

- Christopher L Cavanagh, Graham Elliott, and James H Stock. Inference in models with nearly integrated regressors. *Econometric Theory*, 11(5):1131–1147, 1995.
- J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.
- A. R. Chowdhury and F. Sdogati. Purchasing power parity in the major EMS countries: The role of price and exchange rate adjustment. *Journal of Macroeconomics*, 15:25–45, 1993.
- P. K. Clark. Trend reversion in real output and unemployment. *Journal of Econometrics*, 40:15–32, 1989.
- T. J. DiCiccio and B. Efron. Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228, 1996.
- D. A. Dickey and W. A. Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 74:427–431, 1979.
- B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York, 1993.
- J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.
- J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.
- J. Fan and J. Lv. Non-concave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57:5467–5484, 2011.

- J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.
- Z. Gao, J. H. Lee, Z. Mei, and Z. Shi. LASSO inference for high dimensional predictive regressions. *Journal of Econometrics*, 255:106240, 2026.
- J. K. Gippel, T. M. Smith, and Y. Zhu. Endogeneity in accounting and finance research: Natural experiments as a state-of-the-art solution, 2014. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.1748604>.
- A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- S. Hong, J. Jiang, X. Jiang, and Z. Xiao. Unifying inference for semiparametric regression. *The Econometrics Journal*, 24:482–501, 2021.
- S. Hong, D. J. Henderson, J. Jiang, and Q. Ni. Unifying estimation and inference for linear regression with stationary and integrated or near-integrated variables. *Journal of Financial Econometrics*, 22(5):1397–1420, 2024.
- J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the LASSO. *Annals of Statistics*, 44(3):907–927, 2016.
- Martin Lettau and Stijn Van Nieuwerburgh. Reconciling the return predictability evidence. *Review of Financial Studies*, 21(4):1607–1652, 2008.
- N. C. Mark. Real and nominal exchange rates in the long run: An empirical investigation. *Journal of International Economics*, 28:115–136, 1990.
- Z. Mei and Z. Shi. On LASSO for high dimensional predictive regression. *Journal of Econometrics*, 242(2):105809, 2024.

- F. Merlevède, M. Peligrad, and E. Rio. A Bernstein type inequality and moderate deviations for weakly dependent sequences. *Probability Theory and Related Fields*, 151:435–474, 2011.
- C. R. Nelson and C. R. Plosser. Trends and random walks in macroeconomic time series: Some evidence and implications. *Journal of Monetary Economics*, 10:139–162, 1982.
- J. Y. Park and P. C. B. Phillips. Statistical inference in regressions with integrated processes: Part 1. *Econometric Theory*, 4(3):468–497, 1988.
- P. C. B. Phillips. Time series regression with a unit root. *Econometrica*, 55(2):277–301, 1987.
- P. C. B. Phillips and J. H. Lee. Predictive regression under various degrees of persistence and robust long-run horizon regression. *Journal of Econometrics*, 177:250–264, 2013.
- P. C. B. Phillips and J. H. Lee. Robust econometric inference with mixed integrated and mildly explosive regressors. *Journal of Econometrics*, 192:433–450, 2016.
- J. Racine. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics*, 99(1):39–61, 2000.
- David Rapach and Guofu Zhou. Forecasting stock returns. *Handbook of Economic Forecasting*, 2:328–383, 2013.
- M. R. Roberts and T. M. Whited. Endogeneity in empirical corporate finance, 2012. Simon School Working Paper No. FR 11-29. Available at SSRN: <http://dx.doi.org/10.2139/ssrn.1748604>.
- R. Roll. Violations of purchasing power parity and their implications for efficient

- international commodity markets. In M. Sarnat and G. P. Szego, editors, *International Finance and Trade*, volume 1, chapter 6. Cambridge, MA, 1979.
- F. M. Song and Y. Wu. Hysteresis in unemployment: Evidence from 48 U.S. states. *Economic Inquiry*, 35:235–243, 2007.
- R. F. Stambaugh. Predictive regressions. *Journal of Financial Economics*, 54:375–421, 1999.
- R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.
- R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.
- H. Wang, R. Li, and C. L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.
- I. Welch and A. Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21:1455–1508, 2008.
- H. Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

APPENDIX A: EXTERNAL RESULTS USED

Merlevède et al. [2011], Theorem 1. For a centred sequence (X_j) satisfying (2.5) $\tau(x) \leq \exp(-cx^{\gamma_1})$, (2.6) $\sup_k \mathbb{P}(|X_k| > t) \leq \exp(1 - t^{\gamma_2})$ and (2.7) $1/\gamma = 1/\gamma_1 + 1/\gamma_2$ with $\gamma < 1$, there exist $C_1, \dots, C_4 > 0$ depending only on c, γ, γ_1 such that for $n \geq 4$ and $x > 0$,

$$\begin{aligned} \mathbb{P}\left(\sup_{j \leq n} \left| \sum_{t=1}^j X_t \right| \geq x\right) &\leq \underbrace{n \exp\left(-\frac{x^\gamma}{C_1}\right)}_{\text{I}} + \underbrace{\exp\left(-\frac{x^2}{C_2(1+nV)}\right)}_{\text{II}} \\ &\quad + \underbrace{\exp\left(-\frac{x^2}{C_3 n} \exp\left\{\frac{x^{\gamma(1-\gamma)}}{C_4(\log x)^\gamma}\right\}\right)}_{\text{III}}, \end{aligned} \quad (\text{A.1})$$

with V defined by truncation in their (1.10). Their Remark 1 shows that (2.6) together with strong mixing $\varrho(n) \leq \exp(-cn^{\gamma_1})$ implies (2.5); their Remark 2 permits rescaling of the tail constants, affecting only C_1 – C_4 ; their Remark 3 bounds V by a quantile integral, finite here and denoted V_* .

Term II is the Gaussian regime (many small contributions), Term I the large-deviation regime (a single large summand, the prefactor n being a union bound over which one), and Term III an intermediate correction. Which term binds depends on the threshold: at $x \asymp \sqrt{n \log n}$ and at $x = u_n \sqrt{n}$ Term II binds, while at $x \asymp n$ Term I binds.

Term III is negligible at every threshold used below. All thresholds satisfy $x \gtrsim \sqrt{n \log n}$; the outer factor of Term III's exponent is $x^2/(C_3 n) \gtrsim \log n$ and the inner exponential has argument $x^{\gamma(1-\gamma)}/\{C_4(\log x)^\gamma\} \rightarrow \infty$ (with $\gamma = 1/3$, $\gamma(1-\gamma) = 2/9$, so the argument grows like $n^{1/9}/(\log n)^{1/3}$). Term III is therefore smaller than any power of n and than any exponential bound obtained below, and is omitted from the comparisons.

Lemma A1 of the 2021 Online Appendix (Cai et al., 2015). Under (N4), $\sup_{0 \leq r \leq 1} |U_{n,i}(r) - U_{c_i}(r)| = O(n^{-\theta^*} (\log n)^{\lambda^*})$ almost surely. Three features are used: the bound has an explicit *rate*; it holds *almost surely*, hence consumes no probability budget and enters no union bound; and it is *uniform in r* , so the bound factors out of sums over t . By contrast the corresponding statement (A.8) of that appendix is $o_P(1)$ with no rate, because it carries kernel-estimation error from the partialling step; WEE-SCAD has no nonparametric component, so Lemma A1 may be applied directly.

APPENDIX B: VERIFICATION OF (P4**): STATIONARY DESIGN

In the stationary design all predictors are $I(0)$, all weights equal one, and $\xi_j = \sum_t x_{t-1,j} v_t$. The following is the specialization of Appendices C and D to that case, and is included because it is self-contained and because the mixed-case arguments reduce to it on the stationary block.

Summands. For the score, $\zeta_t = x_{t-1,j} v_t$ is centred by the tower property, $\mathbb{E}[\zeta_t] = \mathbb{E}[x_{t-1,j} \mathbb{E}(v_t \mid \mathcal{F}_{t-1})] = 0$ — note this uses $\mathbb{E}[v_t \mid \mathcal{F}_{t-1}] = 0$, the marginal condition $\mathbb{E}[x_{t,j}] = 0$ alone being insufficient, since a product of centred variables need not be centred. It inherits geometric α -mixing from (S1), being a measurable function of finitely many coordinates of a mixing process. Its tails follow from (S2) by the exponential-transform argument: since $u \mapsto e^{u/K_x}$ is increasing, $\{|x_{t-1,j}| > u\} = \{e^{|x_{t-1,j}|/K_x} > e^{u/K_x}\}$ and Markov's inequality gives $\mathbb{P}(|x_{t-1,j}| > u) \leq C_x e^{-u/K_x}$; likewise for v_t . If both factors are at most $\sqrt{u}$ then the product is at most u , so

$$\mathbb{P}(|\zeta_t| > u) \leq \mathbb{P}(|x_{t-1,j}| > \sqrt{u}) + \mathbb{P}(|v_t| > \sqrt{u}) \leq 2C e^{-\sqrt{u}/K},$$

by the union bound, which requires no independence. Hence ζ_t is sub-Weibull(1/2), and after rescaling (MPR Remark 2) satisfies (2.6) with $\gamma_2 = 1/2$; the supremum over k in (2.6) is immediate by stationarity, and the constants are uniform in j by the supremum in (S2). With $\gamma_1 = 1$ from (S1) and Remark 1, $1/\gamma = 1 + 2 = 3$, so $\gamma = 1/3 < 1$.

Inactive score. At $x = u_n \sqrt{n} = \kappa_*^{-1/2} n^{1-\varrho} (\log n)^{1/2}$, Term II of (A.1) has exponent $\kappa_*^{-1} \log n \cdot n^{1-2\varrho} / (C_2 V_*^2) = 2n^{1-2\varrho} \log n$, while Term I has exponent of order $n^{\gamma(1-\varrho)}$. By (S3), and since $2 - \gamma > 0$,

$$\varrho > \frac{1-\gamma}{2-\gamma} \iff 2\varrho - \varrho\gamma > 1-\gamma \iff \gamma(1-\varrho) > 1-2\varrho,$$

so Term I has the strictly larger power of n , decays faster, and is dominated; Term II binds. Absorbing Terms I and III into a factor of two, $\mathbb{P}(|\xi_j| > u_n \sqrt{n}) \leq 2 \exp(-n^{1-2\varrho} \log n)$, and Bonferroni over the $d - s$ inactive coordinates gives

$$\mathbb{P}(\mathcal{E}_2^c) \leq 2(d - s)e^{-n^{1-2\varrho} \log n} \leq 2 \exp\{Mn^\nu - n^{1-2\varrho} \log n\} \longrightarrow 0,$$

the bracket tending to $-\infty$ because $\nu < 1 - 2\varrho$ by (S3): writing it as $-n^{1-2\varrho} \log n \{1 - Mn^{-(1-2\varrho)-a} / \log n\}$, the exponent $(1-2\varrho)-a$ is strictly positive, so the braced factor tends to one.

Cross-Gram. The product $x_{t-1,j}x_{t-1,\ell}$ is not centred, so put $\psi_t = x_{t-1,j}x_{t-1,\ell} - \mathbb{E}[x_{t-1,j}x_{t-1,\ell}]$; it is centred by construction, mixing as before, and sub-Weibull(1/2) by the same product argument, subtraction of a constant bounded by $\bar{C}$ under (S4) leaving the tail class unchanged. At $x = C'_X n$, Term I has exponent of order $n^{1/3}$ and Term II of order n (one power of n cancelling against the n in $C_2(1 + nV_*^2)$); since $n \gg n^{1/3}$, Term I binds, and absorbing the prefactor,

$$\mathbb{P}\left(\left|\sum_t \psi_t\right| \geq C'_X n\right) \leq 2 \exp(\log n - cn^{1/3}) \leq e^{-c_3 n^{1/3}}.$$

By the triangle inequality and (S4), pointwise $|\frac{1}{n} \sum_t x_{t-1,j}x_{t-1,\ell}| \leq |\frac{1}{n} \sum_t \psi_t| + \bar{C}$, so $\{|\frac{1}{n} \sum_t x_{t-1,j}x_{t-1,\ell}| > C_X\} \subseteq \{|\frac{1}{n} \sum_t \psi_t| > C'_X\}$ with $C_X = C'_X + \bar{C}$, and the bound transfers by monotonicity. Bonferroni over the $(d - s)s$ pairs survives because $\nu < \gamma$, which follows from (S3) via the chain

$$\nu < 1 - 2\varrho < \frac{\gamma}{2 - \gamma} < \gamma, \tag{B.1}$$

the middle inequality being the rearrangement above and the last holding since $\gamma < 1$ implies $2 - \gamma > 1$.

APPENDIX C: VERIFICATION OF (P4**): CROSS-GRAM BOUNDS, MIXED
DESIGN

Lemma C.0.1. *Under (S1)–(S4) and (N1)–(N4), for every $\varepsilon > 0$ there is $C_X(\varepsilon)$ with $\liminf_n \mathbb{P}(\mathcal{E}'_3 \cap \mathcal{E}''_3) \geq 1 - \varepsilon$.*

The two events are separated by the *column* block, which fixes the normalization (Remark 3.2.1); within each, the proof splits by the *row* block, which fixes both the size of the union bound and the technique available. This gives four cases:

case	row	column	count
(a)	$\mathcal{M}_0^c \cap \mathcal{I}_2$	$\mathcal{M}_0 \cap \mathcal{I}_2$	$\asymp sd$, growing
(b)	$\mathcal{M}_0^c \cap \mathcal{I}_1$	$\mathcal{M}_0 \cap \mathcal{I}_1$	$\leq d_1 s$, fixed
(c)	$\mathcal{M}_0^c \cap \mathcal{I}_1$	$\mathcal{M}_0 \cap \mathcal{I}_2$	$\leq d_1 s$, fixed
(d)	$\mathcal{M}_0^c \cap \mathcal{I}_2$	$\mathcal{M}_0 \cap \mathcal{I}_1$	$\asymp sd$, growing

Cases (a) and (c) together give $\mathcal{E}'_3$; cases (b) and (d) give $\mathcal{E}''_3$.

C.1 Case (a)

Here $\tilde{x}_{t-1,j} = x_{t-1,j}$, and the argument is that of Appendix B: centre the product, apply (A.1) at $x = C'_X n$ where Term I binds, obtain $e^{-c_3 n^{1/3}}$ per pair, transfer to the uncentred entry with $C_X = C'_X + \bar{C}$, and take the union over $\asymp sd$ pairs, which survives by (B.1). This case contributes probability tending to one at a *fixed* constant and spends no part of the ε budget: the limit of the entry is the constant $\mathbb{E}[x_j x_\ell]$, not a random variable.

C.2 Case (b)

Write $W_n = n^{-3/2} \sum_t \tilde{x}_{t-1,j} x_{t-1,\ell} = \frac{1}{n} \sum_t \tilde{x}_{t-1,j} (n^{-1/2} x_{t-1,\ell})$.

Approximation and remainder. By (N4), $n^{-1/2} x_{t-1,\ell} = U_{c_\ell}(\frac{t-1}{n}) + R_{n,\ell}(\frac{t-1}{n})$ with $\sup_r |R_{n,\ell}| \leq \rho_n$ a.s., so

$$\left| W_n - \frac{1}{n} \sum_t \tilde{x}_{t-1,j} U_{c_\ell}\left(\frac{t-1}{n}\right) \right| \leq \frac{1}{n} \sum_t |\tilde{x}_{t-1,j}| \cdot \sup_r |R_{n,\ell}(r)| \leq \rho_n \rightarrow 0 \text{ a.s.},$$

using $|\tilde{x}_{t-1,j}| \leq 1$. No condition on θ^* beyond positivity is needed, and being almost sure the bound costs no probability.

Riemann convergence. For fixed ω , $U_{c_\ell}(\cdot, \omega)$ is an ordinary continuous function on the compact $[0, 1]$, hence uniformly continuous, and the displayed sum is a Riemann sum with mesh $1/n$; therefore it converges to $Z_{j\ell} = \int_0^1 B_j^*(r) U_{c_\ell}(r) dr$. Since this holds for every ω in a set of probability one, $W_n \rightarrow Z_{j\ell}$ a.s.

Second moment. $\mathbb{E}[Z_{j\ell}^2]$ has no closed form, $B_j^* = B_j/\|B\|$ being a ratio of Gaussian objects and hence not Gaussian; an upper bound suffices. Deterministically $|B_j^*(r)| \leq \|B^*(r)\| = 1$, so $|Z_{j\ell}| \leq \int_0^1 |U_{c_\ell}(r)| dr$ and, by Cauchy–Schwarz (the interval having length one) and Tonelli,

$$\mathbb{E}[Z_{j\ell}^2] \leq \int_0^1 \mathbb{E}[U_{c_\ell}(r)^2] dr = \int_0^1 r dr = \frac{1}{2} < \infty$$

for $c_\ell = 0$; for $c_\ell \neq 0$, $\mathbb{E}[U_{c_\ell}(r)^2] = (e^{2rc_\ell} - 1)/(2c_\ell)$ and the integral is again finite and explicit.

From the limit to finite n . Since $W_n \rightarrow Z_{j\ell}$ a.s., for each $\delta > 0$ there is almost surely an N with $|W_n| \leq |Z_{j\ell}| + \delta$ for $n \geq N$, so $\{|W_n| > C_X + \delta\} \subseteq \{|Z_{j\ell}| > C_X\}$ eventually and $\limsup_n \mathbb{P}(|W_n| > C_X + \delta) \leq \mathbb{P}(|Z_{j\ell}| > C_X)$; the slack δ is absorbed into C_X . (This is why the conclusion is stated in $\liminf$ form: an almost-sure argument controls the limit superior of the failure probabilities, not each fixed n .)

Chebyshev and the union. Applying Markov’s inequality to $Z_{j\ell}^2$ and choosing $C_X = (3d_1s \mathbb{E}[Z^2]/\varepsilon)^{1/2}$ allocates $\varepsilon/(3d_1s)$ per pair, so the union over the at most d_1s pairs contributes at most $\varepsilon/3$.

C.3 Case (c)

This quantity is $\xi_{n,12}$ of Hong et al. [2021], Lemma A3, Step 2 (their (A.20)–(A.23)), in which the term $\varrho(z_{t-1}) = \mathbb{E}(X_{t-1,2} \mid z_{t-1})$ arises as the correction for their nonparametric component; WEE–SCAD has no such component, so that term is

absent and the blocking argument is otherwise unchanged, the simplification removing a term rather than altering the structure. Their result gives $\frac{1}{n} \sum_t \tilde{x}_{t-1,j} x_{t-1,\ell} = o_P(1)$ entrywise. That argument proceeds through the modulus of continuity of B^* and carries *no rate*, so it could not survive a growing union; under (N2) the family is finite, a finite union of $o_P(1)$ quantities is $o_P(1)$, and hence for any fixed $C_X > 0$ the corresponding failure probability tends to zero and is in particular at most $\varepsilon/3$ eventually.

C.4 Case (d)

The count grows, so Chebyshev is unavailable; the column is nonstationary, so (A.1) does not apply directly, I(1) processes not being α -mixing. The resolution removes the randomness from the column.

Step 1: approximation. By (N4),

$$n^{-3/2} \sum_t x_{t-1,j} x_{t-1,\ell} = \frac{1}{n} \sum_t x_{t-1,j} c_t + \frac{1}{n} \sum_t x_{t-1,j} R_{n,\ell} \left(\frac{t-1}{n} \right), \quad c_t = U_{c_\ell} \left(\frac{t-1}{n} \right).$$

The remainder is at most $\rho_n \cdot \frac{1}{n} \sum_t |x_{t-1,j}|$. Here the row factor is unweighted and unbounded, so the bound of case (b) is unavailable; instead, by stationarity, mixing and the finite first moment implied by (S2), the law of large numbers gives $\frac{1}{n} \sum_t |x_{t-1,j}| = O_P(1)$, and uniformly over the $\asymp d$ coordinates the same sub-exponential tails and a union bound give $\max_j \frac{1}{n} \sum_t |x_{t-1,j}| = O_P(\kappa_n)$ with κ_n growing at most like a small power of n ; since ρ_n decays polynomially at rate $\theta^* > 0$, $\rho_n \kappa_n \rightarrow 0$ and the slack is absorbed into C_X .

Step 2: the path event. Let $\Lambda_M = \{\sup_{0 \leq r \leq 1} |U_{c_\ell}(r)| \leq M\}$. Almost-sure continuity on a compact interval makes $\sup_r |U_{c_\ell}(r)|$ an a.s. finite random variable, so $\mathbb{P}(\Lambda_M^c) \rightarrow 0$ as $M \rightarrow \infty$; fix $M = M(\varepsilon/3)$, independent of n , with $\mathbb{P}(\Lambda_M^c) \leq \varepsilon/3$. This is where the entire ε budget of case (d) is spent.

Step 3: conditioning. Condition on $\sigma(\{U_t\})$. By (N1) the stationary block retains

its mixing and tails; the c_t are now *deterministic* and, on Λ_M , satisfy $|c_t| \leq M$. Thus (A.1) is never applied to a nonstationary object, but to a stationary mixing sequence multiplied by bounded deterministic constants.

Step 4: MPR's conditions. Multiplication by deterministic constants does not enlarge the generated σ -fields, $\sigma(c_t x_{t-1,j}) \subseteq \sigma(x_{t-1,j})$, so mixing is inherited unchanged and (2.5) follows via Remark 1. For the tails, on Λ_M ,

$$\mathbb{P}(|c_t x_{t-1,j}| > u) \leq \mathbb{P}(|x_{t-1,j}| > u/M) \leq C_x e^{-u/(MK_x)},$$

whose right-hand side is free of t . The summands are *not* identically distributed, the c_t differing across t ; but (2.6) requires only a uniform tail bound, $\sup_k \mathbb{P}(|X_k| > u) \leq \dots$, which this provides. Only one factor is random, so $\gamma_2 = 1$, and with $\gamma_1 = 1$ this gives $\gamma = 1/2$ — a better exponent than elsewhere.

Step 5: conditional bound and its removal. Applying (A.1) at $x = C'_X n$, Term I binds and $\mathbb{P}(A_{j\ell} \mid \sigma(\{U_t\})) \leq e^{-cn^{1/2}}$ on Λ_M , where $A_{j\ell} = \{|\sum_t c_t x_{t-1,j}| > C'_X n\}$. Since $\Lambda_M \in \sigma(\{U_t\})$, the tower property gives

$$\mathbb{P}(A_{j\ell} \cap \Lambda_M) = \mathbb{E}[\mathbf{1}_{\Lambda_M} \mathbb{P}(A_{j\ell} \mid \sigma(\{U_t\}))] \leq \mathbb{E}[\mathbf{1}_{\Lambda_M} e^{-cn^{1/2}}] \leq e^{-cn^{1/2}},$$

the bound being pullable from the expectation because $e^{-cn^{1/2}}$ is the same number for every realized path: c depends only on M , K_x , C_x and MPR's constants, none of which depend on ω .

Step 6: union and assembly. By Bonferroni over the $\asymp sd$ pairs and splitting on Λ_M — which is shared across pairs and therefore paid for once —

$$\mathbb{P}\left(\bigcup_{j,\ell} A_{j\ell}\right) \leq sd e^{-cn^{1/2}} + \frac{\varepsilon}{3}, \quad sd e^{-cn^{1/2}} \leq \exp\{Mn^\nu - cn^{1/2}\} \rightarrow 0$$

since $\nu < 1/5 < 1/2 = \gamma$.

C.5 Combination

Since $\mathcal{M}_0^c = (\mathcal{M}_0^c \cap \mathcal{I}_2) \cup (\mathcal{M}_0^c \cap \mathcal{I}_1)$, the maximum over $j \in \mathcal{M}_0^c$ in each event is the larger of the two maxima treated above; taking C_X to be the maximum of the four constants gives a single threshold. Writing $F_a, \dots, F_d$ for the four failure events at that threshold, $(\mathcal{E}'_3 \cap \mathcal{E}''_3)^c \subseteq F_a \cup F_b \cup F_c \cup F_d$ and

$$\limsup_n \mathbb{P}((\mathcal{E}'_3 \cap \mathcal{E}''_3)^c) \leq 0 + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} = \varepsilon,$$

case (a) contributing nothing to the budget. This proves Lemma C.0.1. ■

APPENDIX D: VERIFICATION OF (P4^{**}): THE SCORE EVENT, MIXED
DESIGN

Lemma D.0.1. *Under (S1)–(S4) and (N1)–(N4), $\liminf_n \mathbb{P}(\mathcal{E}_2) = 1$.*

Part 1: $j \in \mathcal{M}_0^c \cap \mathcal{I}_2$ (*count $\asymp d$, growing*). This is the argument of Appendix B, unchanged, giving $\mathbb{P}(\max_{j \in \mathcal{M}_0^c \cap \mathcal{I}_2} |\xi_j| > u_n \sqrt{n}) \rightarrow 0$.

Part 2: $j \in \mathcal{M}_0^c \cap \mathcal{I}_1$ (*count $\leq d_1$, fixed*). Here $\xi_j = \sum_t \tilde{x}_{t-1,j} v_t$ is a martingale, $\tilde{x}_{t-1,j}$ being $\mathcal{F}_{t-1}$ -measurable and $\mathbb{E}[v_t | \mathcal{F}_{t-1}] = 0$.

Convergence to a stochastic integral. Equations (A.41)–(A.42) of Hong et al. (2021) establish $n^{-1} \sum_t \tilde{X}_{t-1,1} \tilde{v}_t = \int_0^1 U_\gamma(r) dW_v(r) + o_P(1)$, their normalization being n^{-1} because their regressor is unweighted and their limit a Brownian bridge because of the kernel partialling. Two differences apply here: our regressor carries the weight, so $|\tilde{x}_{t-1,j}| < 1$ and the normalization is $n^{-1/2}$; and there is no partialling, so the limiting error process is a genuine Brownian motion and the integrand is B_j^* . The same argument gives

$$n^{-1/2} \xi_j \implies Z_j := \int_0^1 B_j^*(r) dW_v(r).$$

This convergence is *in distribution*: unlike the strong approximation of (N4) it does not hold pathwise, $\int f dW$ not being constructible as a pathwise Riemann–Stieltjes integral, Brownian paths having unbounded variation.

Second moment. By the Itô isometry and $|B_j^*(r)| \leq 1$,

$$\mathbb{E}[Z_j^2] = \int_0^1 \mathbb{E}[B_j^*(r)^2] dr \leq \int_0^1 1 dr = 1.$$

No Cauchy–Schwarz step is needed: the isometry moves the square inside directly, and the boundedness then applies to what remains. The ratio $B_j^* = B_j/\|B\|$ need never be evaluated, only bounded.

The growing threshold. Write $S_{j,n} = n^{-1/2}\xi_j$, so the event is $\{|S_{j,n}| \leq u_n\}$. Fix $K > 0$. Since $u_n \rightarrow \infty$ there is N with $u_n \geq K$ for $n \geq N$, whence $\{|S_{j,n}| > u_n\} \subseteq \{|S_{j,n}| \geq K\}$ for $n \geq N$. The set $\{x : |x| \geq K\}$ is closed, so the portmanteau theorem applies to it — a *fixed* set is required, which is why K is introduced, the set $\{|x| > u_n\}$ moving with n . Hence

$$\limsup_n \mathbb{P}(|S_{j,n}| > u_n) \leq \limsup_n \mathbb{P}(|S_{j,n}| \geq K) \leq \mathbb{P}(|Z_j| \geq K) \leq \frac{\mathbb{E}[Z_j^2]}{K^2} \leq \frac{1}{K^2}.$$

The left-hand side does not depend on K , so the inequality holds for every $K > 0$; letting $K \rightarrow \infty$ gives $\mathbb{P}(|\xi_j| > u_n\sqrt{n}) \rightarrow 0$. The growing threshold is what makes this work: an $O_P(1)$ limit is compared against a divergent bound, and no ε budget is required.

Union. The bound is the same for every j , since $\mathbb{E}[Z_j^2] \leq 1$ uniformly, so Bonferroni over the at most d_1 coordinates — a *fixed* number by (N2), so that a finite sum of vanishing terms vanishes and no rate comparison is needed — gives the result. Combining Parts 1 and 2 proves Lemma D.0.1. ■

APPENDIX E: PROOFS OF THE SPARSITY AND ORACLE NORMALITY THEOREMS

This appendix proves Theorem 3.4.1 (sparsity and block-wise consistency) and Theorem 3.5.1 (oracle asymptotic normality) from Chapter 3.

E.1 Proof of Theorem 3.4.1

Step 1: existence of a root on the active subspace

We use the standard consequence of Brouwer's fixed-point theorem: if $\Psi : \mathbb{R}^s \rightarrow \mathbb{R}^s$ is continuous on $\{\|\delta\|_2 \leq \rho\}$ and $\delta^\top \Psi(\beta_{10} + \delta) < 0$ whenever $\|\delta\|_2 = \rho$, then $\Psi(\beta_{10} + \cdot)$ has a zero in the interior. Take $\rho = \tau/\sqrt{n}$ and let $\|\delta\|_2 = \tau/\sqrt{n}$. By (3.1),

$$\delta^\top \Psi_n(\beta_{10} + \delta) = \underbrace{\delta^\top \xi_{\mathcal{M}_0}}_{(1)} - \underbrace{n \delta^\top H_{n,11} \delta}_{(2)} - \underbrace{n \delta^\top \bar{p}_{\lambda_n}(\beta_{10} + \delta)}_{(3)}.$$

Term (1), linear in τ . By Cauchy–Schwarz and the active-block central limit theorem (Hong et al. 2021, Theorem 3.1, applied to the s -dimensional restriction to $\mathcal{M}_0$, giving $n^{-1/2} \xi_{\mathcal{M}_0} = O_P(1)$ and, s being fixed, $\|\xi_{\mathcal{M}_0}\|_2 = O_P(\sqrt{n})$),

$$\delta^\top \xi_{\mathcal{M}_0} \leq \|\delta\|_2 \|\xi_{\mathcal{M}_0}\|_2 = \frac{\tau}{\sqrt{n}} \cdot O_P(\sqrt{n}) = \tau \cdot O_P(1),$$

the factors of $\sqrt{n}$ cancelling.

Term (2), quadratic in τ . A quadratic form depends only on the symmetric part of its matrix, which is why (H') is stated for $B_{n,11}^{\text{sym}}$; this is also where the non-symmetry of the weighted Gram is harmless. Writing $H_{n,11} = n^{-1/2} B_{n,11} D_{n,11}$ and using (H'), on an event of probability at least $1 - \varepsilon - o(1)$,

$$n \delta^\top H_{n,11} \delta \geq c_H n \|\delta\|_2^2 = c_H \tau^2.$$

Term (3), identically zero. On the sphere each coordinate satisfies $|\delta_j| \leq \tau/\sqrt{n}$,

while $\min_{j \in \mathcal{M}_0} |\beta_{0,j}| \geq 2d_n$, so $|\beta_{0,j} + \delta_j| \geq 2d_n - \tau/\sqrt{n}$. By (R), $\lambda_n \gg n^{-\varrho}(\log n)^{1/2}$ with $\varrho < 1/2$ and $d_n/\lambda_n \rightarrow \infty$, whence $d_n \gg n^{-1/2}$ and $\tau/\sqrt{n} = o(d_n)$ for fixed τ . Thus $|\beta_{0,j} + \delta_j| \geq d_n \gg a\lambda_n$ eventually, placing every active coordinate in the flat region of the SCAD penalty where $p'_{\lambda_n} = 0$; hence $\bar{p}_{\lambda_n}(\beta_{10} + \delta) = 0$. (This is the work done by the condition $d_n/\lambda_n \rightarrow \infty$ in (R): the signal dwarfs both the penalty and the perturbation.)

Choice of τ . Combining, on $\|\delta\|_2 = \tau/\sqrt{n}$,

$$\delta^\top \Psi_n(\beta_{10} + \delta) \leq \tau \cdot O_P(1) - c_H \tau^2.$$

Since $O_P(1)$ is not a fixed constant, τ is chosen on a high-probability event: given $\varepsilon > 0$, take K_ε with $\mathbb{P}(\|\xi_{\mathcal{M}_0}\|_2/\sqrt{n} > K_\varepsilon) \leq \varepsilon$ for large n and fix $\tau > K_\varepsilon/c_H$. On the intersection of that event with the event of (H'), $\delta^\top \Psi_n(\beta_{10} + \delta) \leq \tau K_\varepsilon - c_H \tau^2 < 0$ on the whole sphere, and Brouwer's theorem yields a root $\hat{\beta}_1$ in the interior, so $\|\hat{\beta}_1 - \beta_{10}\|_2 \leq \tau/\sqrt{n}$. This is part (b); no objective function has been used.

Step 2: block-wise rates

Proof of Lemma 3.4.2. The active-block equation, with the penalty eliminated by the flat-region argument of Step 1 (which uses only the aggregate rate, so that no circularity arises), reads $\sum_t \tilde{x}_{t-1, \mathcal{M}_0} x_{t-1, \mathcal{M}_0}^\top (\hat{\beta}_1 - \beta_{10}) = \xi_{\mathcal{M}_0}$. Inserting $D_{n,11}^{-1} D_{n,11}$ and dividing by $\sqrt{n}$,

$$B_{n,11} D_{n,11} (\hat{\beta}_1 - \beta_{10}) = n^{-1/2} \xi_{\mathcal{M}_0}. \quad (\text{E.1})$$

By (H'), $B_{n,11}^{-1} = O_P(1)$, and $n^{-1/2} \xi_{\mathcal{M}_0} = O_P(1)$ by the active-block CLT; hence $D_{n,11} (\hat{\beta}_1 - \beta_{10}) = O_P(1)$. Reading coordinatewise, the diagonal entries $\sqrt{n}$ and n give the two stated rates. $\square$

Step 3: sparsity via the KKT conditions

Extend $\hat{\beta}_1$ to $\hat{\beta} = (\hat{\beta}_1, 0)$. For $\hat{\beta}$ to be a sparse root of the penalized system it must satisfy

(K1) (*Active equation*) $\Psi_n(\hat{\beta}_1) = 0$ — by construction in Step 1.

(K2) (*Inactive subgradient*) $\|z\|_\infty < \rho'(0^+) = 1$, where $z = (n\lambda_n)^{-1}(\xi_{\mathcal{M}_0^c} - T_2)$: each inactive coordinate's estimating-equation value is dominated by the penalty's subgradient at zero, so that setting that coordinate to zero is consistent with the system.

(K3) (*Local behaviour of the penalty*) $\lambda_n \kappa(\rho; \hat{\beta}_1) \leq \lambda_{\min}(B_{n,11}^{\text{sym}})$, ensuring the root is isolated.

These play the role filled by conditions (7)–(9) of Fan and Lv [2011] in an objective-based treatment. The substantive verification is (K2).

At $(\hat{\beta}_1, 0)$ the inactive-coordinate value of the estimating equation is

$$\sum_{t=1}^n \tilde{x}_{t-1, \mathcal{M}_0^c} \left(v_t - x_{t-1, \mathcal{M}_0}^\top (\hat{\beta}_1 - \beta_{10}) \right) = \xi_{\mathcal{M}_0^c} - T_2, \quad T_2 = \left(\sum_{t=1}^n \tilde{x}_{t-1, \mathcal{M}_0^c} x_{t-1, \mathcal{M}_0}^\top \right) (\hat{\beta}_1 - \beta_{10}),$$

so $\|z\|_\infty \leq (n\lambda_n)^{-1} (\|\xi_{\mathcal{M}_0^c}\|_\infty + \|T_2\|_\infty)$.

The score term. On $\mathcal{E}_2$, $\|\xi_{\mathcal{M}_0^c}\|_\infty \leq u_n \sqrt{n} = \kappa_*^{-1/2} n^{1-\varrho} (\log n)^{1/2}$, so

$$(n\lambda_n)^{-1} \|\xi_{\mathcal{M}_0^c}\|_\infty \leq \frac{\kappa_*^{-1/2} n^{-\varrho} (\log n)^{1/2}}{\lambda_n} = o_P(1)$$

by the rate condition in (R).

The cross term. Bounded *entrywise* rather than by Cauchy–Schwarz, using Lemma 3.4.2 and the two cross-Gram events: uniformly in $j \in \mathcal{M}_0^c$, on $\mathcal{E}_3' \cap \mathcal{E}_3''$,

$$|T_{2,j}| \leq \sum_{\ell \in \mathcal{M}_0 \cap \mathcal{I}_2} C_X n \cdot O_P(n^{-1/2}) + \sum_{\ell \in \mathcal{M}_0 \cap \mathcal{I}_1} C_X n^{3/2} \cdot O_P(n^{-1}) = O_P(\sqrt{n}),$$

the $n^{3/2}$ column scale cancelling against the n^{-1} rate of the $I(1)$ active coordinates. Hence $(n\lambda_n)^{-1}\|T_2\|_\infty = O_P((\sqrt{n}\lambda_n)^{-1}) = o_P(1)$, since $\sqrt{n}\lambda_n \gg n^{1/2-\varrho}(\log n)^{1/2} \rightarrow \infty$ for $\varrho < 1/2$.

Combining, $\|z\|_\infty = o_P(1) < 1$ with probability tending to one, verifying (K2); no inactive coordinate can be moved from zero consistently with the system, so $\hat{\beta}_2 = 0$, which is part (a).

Condition (K3). For SCAD with $a > 2$ the local concavity $\kappa(\rho; \beta_1)$ vanishes once $|\beta_j| > a\lambda_n$ throughout the support. By part (b) and (R), $\|\hat{\beta}_1 - \beta_{10}\|_\infty = O_P(n^{-1/2}) = o_P(d_n)$ and $\min_{j \in \mathcal{M}_0} |\hat{\beta}_j| \geq d_n \gg a\lambda_n$ eventually, so $\kappa(\rho; \hat{\beta}_1) = 0$ and (K3) reduces to $\lambda_{\min}(B_{n,11}^{\text{sym}}) > 0$, which holds by (H'). $\blacksquare$

E.2 Proof of Theorem 3.5.1

Proof. Step 1: reduction on the sparsity event. On $\{\hat{\beta}_2 = 0\}$, whose probability is at least $1 - \varepsilon - o(1)$ by Theorem 3.4.1, the active-block equation is $\Psi_n(\hat{\beta}_1) = 0$. By the exact identity (3.1), the cross-block contribution vanishes identically, since $\hat{\beta}_2 - \beta_{20} = 0$; no probabilistic argument is needed for this cancellation.

Step 2: the penalty vanishes. As in Step 3 of the proof of Theorem 3.4.1, $\min_{j \in \mathcal{M}_0} |\hat{\beta}_j| \geq d_n \gg a\lambda_n$ eventually, so $\bar{p}_{\lambda_n}(\hat{\beta}_1) = 0$ with probability tending to one and (3.1) reduces to $\sum_t \tilde{x}_{t-1, \mathcal{M}_0} x_{t-1, \mathcal{M}_0}^\top (\hat{\beta}_1 - \beta_{10}) = \xi_{\mathcal{M}_0}$.

Step 3: normalization and the self-normalized CLT. By (E.1), $B_{n,11} D_{n,11} (\hat{\beta}_1 - \beta_{10}) = n^{-1/2} \xi_{\mathcal{M}_0}$; pre-multiplying by $A_{n,11}^{-1/2}$,

$$A_{n,11}^{-1/2} B_{n,11} D_{n,11} (\hat{\beta}_1 - \beta_{10}) = A_{n,11}^{-1/2} (n^{-1/2} \xi_{\mathcal{M}_0}).$$

Now $n^{-1/2} \xi_{\mathcal{M}_0} = n^{-1/2} \sum_t \tilde{x}_{t-1, \mathcal{M}_0} v_t$ is a martingale array with conditional covariance $A_{n,11} \sigma_v^2$, so the martingale central limit theorem — as applied in the proof of Theorem 2 of Hong et al. [2024], and in the s -dimensional restriction of Theorem 3.1 of Hong et al. [2021] — gives $A_{n,11}^{-1/2} (n^{-1/2} \xi_{\mathcal{M}_0}) \rightarrow_d N(0, \sigma_v^2 I_s)$, the self-normalization being

what renders the limit free of the random quantities associated with the nonstationary block. $\square$