HEC MONTRÉAL

Size Distortions in Robust Estimators: Implications for Asset Pricing

par

Nicolas Harvie

Vincent Grégoire
HEC Montréal
Codirecteur de recherche

Anthony Sanford
HEC Montréal
Codirecteur de recherche

Sciences de la gestion
(Spécialisation Finance)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences
(M. Sc.)

December 2023

Résumé

Les prédicteurs de rendements excédentaires présentent persistence et une variance changeante dans le temps, impliquant la nécessité d'erreurs consistantes à l'hétéroskédasticité et l'autocorrélation (HAC) dans les tests linéaires d'évaluation d'actif. En utilisant des simulations, nous démontrons que bien qu'ils conduisent à d'importantes améliorations, de telles corrections échouent à fournir des propriétés de taille adéquates sous l'hypothèse nulle d'absence de rendements anormaux. Même les estimateurs robustes optimalement spécifiés souffrent de distorsions de taille, impliquant que les meilleurs HAC restent imparfaits. Nous proposons une standardisation de l'estimateur robuste atténuant le problème, sans le résoudre entièrement. Nous trouvons que entre 2006 et 2021, plus de 20% d'une large variété de prédicteurs de rendement excédentaires diffèrent en statut de significativité au niveau standard de 5% en comparant cet estimateur à l'OLS, et plus de 30% à un niveau plus restrictif.

Mots-clés

Anomalies, évaluation d'actifs, autocorrélation, hétéroskédasticité, estimation robuste

Abstract

Predictors of excess returns exhibit persistence and time-varying variance, implying the need for heteroskedasticity and autocorrelation-consistent errors (HAC) in linear tests. Using simulations, we show that although they lead to important improvements, such corrections fail to provide adequate size properties under the null hypothesis of zero abnormal returns. Even optimally specified robust estimators suffer from size distortions, implying that the best HACs remain imperfect. We propose a standardization of the robust estimator that alleviates the problem, albeit not completely. We find that between 2006 and 2021, more than 20% of a wide panel of predictors differ in significance status at the standard 5% level in comparing this estimator to OLS, and more than 30% at a more restrictive level.

Keywords

Anomalies, asset pricing, autocorrelation, heteroskedasticity, robust estimation

Contents

List of Tables

List of Figures

List of Acronyms

AP-HAC	Asset Pricing Heteroskedasticity and Autocorrelation Consistent
AR	AutoRegressive
BLUE	Best Linear Unbiased Estimator
CL	Christoffersen and Langlois, 2013
CZ	Chen and Zimmermann, 2021
DGP	Data Generating Process
GARCH	Generalized AutoRegressive Conditional Heteroskedasticity
HAC	Heteroskedasticity and Autocorrelation Consistent
HXZ	Hou, Xue, and Zhang, 2020
MHT	Multiple Hypothesis Testing
NGARCH	Nonlinear Generalized AutoRegressive Conditional Heteroskedasticity
NW	Newey and West, 1986
OLS	Ordinary Least Squares
QS	Quadratic Spectral
SKEWT	Skewed Student t
SE	Standard Error

SW	Stock and Watson, 2003
TH	Tukey-Hanning
WN	White noise

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Vincent Grégoire and Professor Anthony Sanford, for their invaluable guidance and support throughout my research. Their insights and expertise have been instrumental in shaping this work. I am also deeply grateful to my wonderful girlfriend, Josianne, for her love, encouragement, and patience. Her support has been a constant source of motivation and inspiration. Additionally, I would like to thank my dad, Denis, for his unconditional belief in me. Without his guidance and encouragement, I would not be where I am today. Last but not least, I want to extend my appreciation to my dear friend Felix for his wisdom, loyalty and unwavering belief in the future.

Introduction

Improvements to the statistical methodology associated with linear tests of factor models in finance have often been directed towards correcting for various biases (B. T. Kelly, Pruitt, and Su, 2019; Feng, Giglio, and Xiu, 2020; Giglio and Xiu, 2021; Gu, B. Kelly, and Xiu, 2021; Giglio, B. Kelly, and Xiu, 2022). Recent evidence has shown that predictors of excess returns[1] show persistence and time-varying volatility (Christoffersen and Langlois, 2013; Gupta and B. Kelly, 2019; Arnott, Clements, Kalesnik, and J. Linnainmaa, 2021; Ehsani and J. T. Linnainmaa, 2022), implying the need for heteroskedasticity and autocorrelation-consistent (HAC) estimation in asset pricing tests. Although HAC corrections have been the subject of discussions in the past (Petersen, 2008; Gow, Ormazabal, and Taylor, 2010), empirical and theoretical advancements in the literature have not been substantial in the last couple of decades.

We first demonstrate that OLS estimation in the context of linear asset pricing tests falsely rejects the null hypothesis of no abnormal returns much more often than it should under classic confidence interval-based tests in the presence of autocorrelation and heteroskedasticity. We show that this effect leads to widespread incorrect inference in simulated settings due to the inflated rejection rate of standard asset pricing tests. Furthermore, we show that although robust estimators significantly improve the size properties of these tests in the same settings, they all fail even at optimal conditions to achieve the theoretically expected rejection rate. Repercussions are even more serious when considering the multiple hypothesis testing (MHT) problem that exists in asset pricing, where size distortions have an exponential effect. Our results also suggest potential implications for other subfields of asset pricing, empirical

[1] By predictors, we refer to variables constructed in the spirit of Fama and French, 1992, usually closely related to firms' fundamentals, price-derived elements and other characteristics that are often utilized in traditional factor models.

corporate finance and more generally any field of study that conducts inference in linear tests using HAC estimators.

We simulate predictor returns using multiple econometric models as data-generating processes (DGP). We calibrate those models using a panel of predictors obtained from Chen and Zimmermann, 2021 (CZ), scaled to $\alpha = 0$. These predictors are estimated using the models suggested in Christoffersen and Langlois, 2013. By scaling the alpha of predictors used in estimation to zero, we ensure that any deviations from the expected rejection rate in linear tests on simulated predictors are solely attributable to their time series characteristics. We then simulate predictors of excess returns based on those parameters. In this manner, we allow for robust inference as $n \to \infty$ (where n corresponds to the number of simulated predictors) for the linear asset pricing tests under simulations generated by realistic model assumptions. We then perform linear tests that account for autocorrelation and heteroskedasticity to determine an optimal yet realistic approach.

Traditional methods in finance to account for heteroskedasticity and autocorrelation in linear regression usually involve heteroskedasticity and autocorrelation-consistent errors (HAC) in the spirit of Newey and West (NW) (1986), with flexibility in their two main parameters: kernel choice and bandwidth (Lazarus, Lewis, Stock, and Watson, 2018). We employ this framework by conducting a horse race between estimators that incorporate a wide array of kernel choices and different discretely-valued bandwidths. With traditional OLS, we find significantly inflated empirical rejection rates for linear asset pricing tests neighboring 7% and 1.7% at the 5% and 1% levels respectively, for simulated predictors generated by models allowing for autocorrelation and heteroskedasticity. Although the gains in employing HAC estimators in such settings are found to be important, we find that no such estimator achieves the aforementioned expected rejection rates.

We propose the use of an empirically motivated table that recommends a simple modification to the threshold t-statistic under the NW estimator with different bandwidths. Although this is an imperfect solution, we believe that any improvement toward consistency and standardization in HAC practice is well-warranted. We also present results suggesting that some settings for HAC errors may worsen the problem, which adds to the value of standardization. Although our demonstration is empirical rather than formal, we consider it an important first step toward solving the problem of autocorrelation and heteroskedasticity in linear asset pricing tests.

We provide evidence that this minor methodological change has an important impact on asset pricing test results by applying it to predictors documented in the literature and collected by Chen and Zimmermann, 2021. To do so, we simply regress the actual predictor time series one by one on a constant using the optimal estimator found in the table and report the significance status obtained. To account for alpha-decay, we restrict our sample size to after 2005 as implied in Chen and Velikov, 2023. When comparing the empirically optimal estimator to OLS in this period, we find 10 (20.08%) anomalies to differ in significance status at the $p = 0.05$ level and 7 (31.81%) at the $p = 0.01$ level on 48 and 22 respectively. Additionally, we show that accounting for autocorrelation and heteroskedasticity does not only affect a few predictors but shifts the whole distribution of their test statistics towards zero.

Literature Review

Factor models in asset pricing have long been important for researchers to try and explain cross-sectional returns on assets. An essential contribution in such a literature, Sharpe (1965) and Lintner (1965) Capital Asset Pricing model, sought to explain returns as a linear function of market risk, scaled by a coefficient idiosyncratic to every asset they obtain by running an ordinary least squares regression. Strides were made when Fama and French (1992) suggested two additional factors to the model closely related to firms' fundamental characteristics of size and value.

In doing so, they not only expanded on the previous benchmark but instigated a gold standard for the discovery of new predictors potentially explanatory of excess returns on assets. Grouping the cross-section of firms by their ranking on some particular feature often linked to financial ratios or price-related elements, they found that the resulting premium obtained by subtracting the returns of the conditionally high firms from the conditionally low firms in terms of this particular variable could potentially be associated with exceeding returns, and tested for statistical significance.

With this linear methodology based on portfolio sorts, researchers simply had to determine a particular firm characteristic they thought could be associated with exceeding returns by setting the null hypothesis as $H_0 : \alpha = 0$ and the alternative as $H_a : \alpha \neq 0$, create a predictor by theoretically buying and selling the appropriate assets and test it using a simple linear regression.

As the discovery of components explanatory of cross-sectional returns could (and still can) be associated with the elaboration of successful trading strategies that integrate them, high incentives combined with a straightforward methodology led to the proliferation of such components. However initially productive, it became apparent at the start of the millennia that this quest for clarity in expressing cross-sectional excess returns as functions of firms' characteristics would rather lead the field astray. As early as 2011, Cochrane (2011) stated in his presidential address that the effort to understand cross-sectional excess returns by anomaly research would lead researchers adrift with the explosion of new variables.

Growing concerns about methodological flaws in asset pricing research would also lay some doubt as to the validity of these past discoveries. Researchers saw applicable to asset pricing research the multiple hypothesis testing (MHT) problem, by which joint tests of the same hypothesis would eventually lead to extremely high odds of falsely rejecting the null (Harvey, Liu, and Saretto, 2020; Chordia, Goyal, and Saretto, 2020; Harvey and Liu, 2020). While numerous efforts have been made to limit the effects of the MHT concern, such as a simple increase in the standard threshold t-statistic for significance (Harvey, 2017; Chordia, Goyal, and Saretto, 2020) or more complex methodological changes (Feng, Giglio, and Xiu, 2020; Giglio and Xiu, 2021), the problem remains at the forefront of asset pricing research.

Specifically, suppose we have m hypotheses, each associated with an asset pricing test in the spirit of what was discussed previously. Further suppose that each test has the same null hypothesis H_0 and alternative hypothesis H_a, as is the case in anomaly research. In this setting, the p-value for each hypothesis is the probability of observing a test statistic at least as extreme as the one obtained, assuming that the null hypothesis is true. Considering that the number of predictors currently associated with excess returns at the time of writing surpasses several hundred, one can only imagine how many such tests have been undertaken historically without being reported. Coupled with a threshold for significance traditionally set at 0.05, it is evident why such a problem is still widely recognized as one of the most important for asset pricing research.

Although it represents an important challenge for anomaly research, the MHT problem is certainly not the only methodological flaw in asset pricing research that arose recently in the literature. There also have been important efforts oriented towards reducing omitted variable bias that exists in linear asset pricing tests by introducing summarized versions of other factors as control sets in addition to the conventional Fama-French specification. Put simply, any of the aforementioned tests fails to introduce the appropriate controls in its setting without losing parsimony. As the set of controls should theoretically include all the significant predictors (and possibly others that are yet undiscovered) to accurately assess the marginal significance of the tested predictor, this configuration de facto introduces an omitted variable bias which affects conclusions drawn from the hypothesis test. Solutions most often favored to partially solve this concern usually consist in the application of a factorization method applied to knowingly significant predictors to generate the optimal control set as a benchmark without losing parsimony of the model (B. T. Kelly, Pruitt, and Su, 2019; Feng, Giglio, and Xiu, 2020; Gu, B. Kelly, and Xiu, 2021; Giglio and Xiu, 2021). Although extremely important, this line of research does not consider the possible flaws that could arise in inference due to the characteristics of the anomaly time series itself, which will be a significant driver of our contribution.

Other concerns include but are not limited to misleading conclusions due to malpractices from researchers in response to incentives (Harvey, 2021), idiosyncratic differences in methodology amongst the set of researchers (Menkveld, Dreber, Holzmeister, Huber, Johannesson, Kirchler, Razen, and Weitzel, 2022) and slight but significant dispersion in data employed (Akey, Robertson, and Simutin, 2022).

All these considerations, in line with the academia-wide concerns surrounding the replicability of scientific research (Ioannidis, 2005; Baker, 2016), led academics in finance to re-evaluate the validity of previously discovered anomalies and attempt to replicate the results obtained by past research. Hou, Xue and Zhang (HXZ) (2020) document the existence of 437 anomaly variables, ranging from accounting measures to price dynamics. Replicating these anomalies, they maintain that while some do seem to hold the test of significance, most of them fail to live up to the standard they have been raised at in the literature. Contrarily to those findings, more recent replicability research by Chen and Zimmerman (2021) finds that the vast majority of market anomalies issued from past research can be replicated, attributing the differing results of HXZ to important deviations in methodology, such as removing micro-cap stocks from their analysis and extending the sample of anomalies to unclear predictors. Although it is probably the case that practical applications of asset pricing research might omit micro-caps from their investment universe, Chen and Zimmerman's analysis emphasizes that academic replication should be done by conserving the precise methodology employed by past researchers. From a conceptual perspective, Chen (2021) also suggests that attributing all discovered significant anomalies to p-hacking and data mining might be misguided with a thought experiment involving simple and intuitive calculations. As dozens of published and subsequently replicated anomalies t-statistics exceed 6.0, basic statistical knowledge would indicate that the infinitesimal odds of joint false positives are extremely unlikely. Chen suggests that such an unlikely scenario would have required 10,000 researchers to generate 8 factors every day for hundreds of years, which is highly improbable.

Although these recent findings seem favorable toward past anomaly research, there is still an important degree of doubt in academia as to the future of this strand of inquiry. However, notwithstanding evidence concerning the validity of market anomalies usually integrated into factor models, their application in the industry has seen tremendous growth over the years. Style investing, which we can closely associate with asset allocation according to certain factors, has gained increasing popularity in the asset management realm, both amongst institutions (Froot and Teo, 2008), sophisticated investors (Baquero and Verbeek, 2008) and individual investors (Bender, Briand, Melas, Subramanian, et al., 2013). Questions such as how to build style portfolios (Israel, Jiang, and Ross, 2017) and how to design optimal strategies using common factors (Blitz and Vidojevic, 2019) have been prominent in the literature oriented towards industry applications. Intuitively, style investing requires investors to buy and sell stocks in the spirit of the Fama and French methodology outlined previously to capture the premium associated with the resulting long-short portfolio. We suggest that due to the increasing popularity of factor models, any improvement to anomaly-detection methodology that is implementable such as the one we suggest is highly warranted.

Although some doubt remains as to the real-world profitability of anomaly-inspired strategies due to trading costs, short sale costs and post-publication decay (Korajczyk and Sadka, 2004; McLean and Pontiff, 2016; Muravyev, Pearson, and Pollet, 2022), such an exercise lends itself to credible grounds in the context of pragmatic asset allocation. This shift from considering anomalies as purely statistical artifacts to components suggesting asset allocation styles for investors with different objectives brought researchers to take a closer look at the properties of the predictors themselves rather than in the stocks that constitute them. If anomalies that constitute portfolios held by asset managers were to hold econometric properties that materially affect how style investing performs, the impact would be significant considering the soar in its popularity. For example, Christoffersen and Langlois (2013) find important nonlinearities in the joint distribution of Fama-French Factors premia, with important implications as to the risk implied by holding a portfolio of these diversified factors. While asset managers may believe that their assets are protected by the apparent lack of correlation between the components that compose their portfolios, such a belief could be illusory as the diversification properties of these components fail to live up to expectations during market turmoil, where diversification benefits are most in demand.

Time series analysis of anomaly premia has also led to important discoveries, such as the existence of positive autocorrelation in factor returns themselves (Gupta and B. Kelly, 2019; Arnott, Clements, Kalesnik, and J. Linnainmaa, 2021), which led researchers to suggest that strategies that employ price momentum are in fact proxying for momentum in firm fundamentals (Novy-Marx, 2015; Ehsani and J. T. Linnainmaa, 2022). While the existence of persistence has been well known for individual stock returns (Jegadeesh and Titman, 1993) and mutual funds returns (Carhart, 1997), evidence of such econometric processes for anomaly portfolios is a lot more recent. An important implication of this concept is that predictor momentum, rather than the predictor itself, could be the root cause of statistical significance when tested in linear settings. In our context, falsely attributing significance to a predictor of exceeding returns due to autocorrelation and heteroskedasticity could lead practitioners and researchers to draw erroneous conclusions based on inflated t-statistics, with the most apparent outcome being inadequate investments without knowledge of such.

Chapter 1

Methodology

Traditional methodology for anomaly detection in asset pricing usually involves one of two procedures. One option is the Fama-Macbeth (1973) two-stages least square regression with the suggested anomaly portfolio time series as the independent variable and accompanying controls if desired. If the coefficient on the variable of interest is marginally significant given a confidence level, it suggests that it is priced in the excess returns of the market and hence can be considered an anomaly. The other option implies sorting the cross-section of stocks on the desired variable and creating a long-short portfolio that expresses the premia associated with this variable in the spirit of Fama and French, 1992. The time series is then tested for non-zero alpha using least square regression with accompanying optional controls at the researchers' discretion. If the intercept is significantly different from zero, it suggests that the variable displays non-zero excess returns above the benchmark and can be considered a predictor.

As outlined previously, past research has documented that actual predictors exhibit persistence and time-varying variance, which can be directly associated with autocorrelation and heteroskedasticity. We know that the traditional linear regression method employed in the portfolio-sorting methodology outlined previously must satisfy the Gauss-Markov assumptions. These assumptions are necessary for the OLS estimator to be the best linear unbiased estimator (BLUE) of the population parameters. In an anomaly research context, applying regular OLS in linear tests fails to account for autocorrelation and heteroskedasticity in the dependent variable, breaking the Gauss-Markov assumptions of orthogonality and homoskedasticity of errors. Violations of these assumptions can lead to biased and inefficient estimators. We show in future results that this leads to inflated t-statistics estimates.

1.1 Overview of HAC estimators

As predictors of excess returns incorporate autoregressive and heteroskedastic properties, inference from linear tests on their time series is traditionally conducted with robust estimators to avoid drawing erroneous conclusions, hence their importance. However, as we will show, correct specification is essential in such estimation and even in the optimal case, HACs remain estimations that fail to fully correct for size distortions. In finance, approaches reliant on HACs usually follow the principles established by Newey and West (1986). These methods offer flexibility through the selection of a suitable kernel and bandwidth. Specifically, NW first addresses heteroskedasticity by employing the methodology found in White (1980).

To correct for autocorrelation, NW makes use of the Bartlett kernel, a traditional methodology to estimate long-run covariance. This kernel remains a popular choice in the financial literature to correct autocorrelation, as it weighs the covariance of the residuals in a uniformly decreasing manner. This configuration generalizes the autoregressive component of a time series as progressively weaker as one goes further from the observation of interest.

Combining the White matrix and a particular HAC kernel (of which the Bartlett outlined previously is an example) lends itself to the generalized form of the HAC estimator. The form can be specified by replacing the weighing function $K(x)$ with other kernels, such as the ones presented in Table 1 of the Appendix.

$$\begin{aligned}\hat{\Omega}_{\text{HAC}} = &\left[X' \cdot \operatorname{diag}\left(\widehat{\epsilon}_t^2\right) \cdot X\right] \\ &+ \sum_{j=1}^{l} \left(K(x)\right) \cdot \sum_{t=j+1}^{T} \left(X_t \widehat{\epsilon}_t \widehat{\epsilon}_{t-j} X_{t-j}' + X_{t-j} \widehat{\epsilon}_{t-j} \widehat{\epsilon}_t X_t'\right) \qquad (1.1)\end{aligned}$$

Where X_t is a $N \times 1$ vector, so $X_t \widehat{\epsilon}_t \widehat{\epsilon}_{t-j} X_{t-j}' = (\widehat{\epsilon}_t \widehat{\epsilon}_{t-j}) \cdot X_t X_{t-j}'$ is $N \times N$ and l is the bandwidth, also referred to as truncation lag. Although the Bartlett kernel is known for its simplicity and is widely utilized in many financial applications, there is currently no empirical research that ascertains its optimality in an asset pricing context. Literature on kernel density estimation of long-run covariance has suggested many different weighting schemes, that are often ignored (Parzen, 1962; Tukey, 1967; Andrews, 1991). Additionally, there is no definite standard to select for the bandwidth parameter required in the estimator. Practitioners and researchers are hence left with employing either an intuitive approximation (i.e. based on the periodicity of the data, their "domain knowledge" or anything else), a "rule-of-thumb" based approach taken from a general reference or the default value of their statistical package.

As we show in the next sections, although they have widely been used without precise justification in financial research, both kernel and bandwidth choice affect inference results significantly in the context of linear asset pricing tests when data exhibits autocorrelation and heteroskedasticity. Additionally, while HACs are often considered adequate solutions, we demonstrate empirically that they remain estimations that fail to fully correct those biases. Although our analysis is specific to an asset pricing application, it also generalizes to many areas in financial research where inference might be affected by autocorrelation and heteroskedasticity, such as mutual funds evaluation, empirical corporate finance and others. We hope that our research will contribute to raising concerns about the adequate usage of HAC estimation in financial research and that it will prompt econometricians to seek a better solution.

To compare and contrast various estimator specifications, we conduct a horse race applying the possible combinations of kernels and bandwidths, of which the specifics can be found in Appendix 2.2, on simulated predictors. In comparing with the white noise model, we are then in a good position to assess the size distortions of our tests. Given the potential difficulties or impracticality associated with altering kernels and bandwidth parameters, we also offer a temporary solution to alleviate the problem and standardize HAC estimation. To do so, we recommend the use of an empirically informed table that suggests a straightforward adjustment to the threshold t-statistics within the Newey-West framework for varying bandwidths.

We then show the impact of applying this framework to the knowingly clear predictors gathered from the CZ database. To do so, we simply regress the predictor time series, one by one, on a constant using the optimal estimator and report the t-statistics obtained, which we will show differ largely from the ones obtained originally by CZ.

1.2 Data

Our dataset is composed of the time series for the clear predictors replicated by Chen & Zimmerman (2021) until the end of 2022. Clear predictors are market anomalies that have been documented in asset pricing literature and are considered as having surpassed the threshold for statistical significance in-sample both at the time of their finding and after replication.

First, the authors use the portfolio sorting method outlined previously to construct a long-short portfolio for each anomaly. Then, they use the OLS estimator to determine the magnitude and significance of the anomaly's $\hat{\alpha}$ by setting its time series as the dependent variable and a constant as the independent variable. This constant $\hat{\alpha}$ hence refers to the estimated mean excess returns of the predictor over the risk-free rate and is conventionally accompanied by its p-value and t-statistic. Note that while CZ's replication specifically uses the in-sample data for each anomaly to compute significance to obtain replicable results of the original study, we opt for using the available data in its entirety. In other words, while CZ selects each timeframe according to the original studies, we use each predictor as early as it is available up to the most recent iteration of the dataset. This is principally done to obtain the most representative picture of the predictors' processes. For this same reason, we omit in our estimation the predictors that exhibit a t-statistic below 1.96 in their full sample, which shrinks the panel from 207 variables to 173.

Predictor returns are computed monthly as floating values with 6 decimals. We conserve the monthly frequency as most asset pricing tests are done with this periodicity in the literature. Although the dataset spans from January 1926 to December 2021 in the current iteration, most of the anomalies start exhibiting returns between 1940 and 1980 due to the prior data unavailability in the signal construction stage. As our analysis will require the estimation of econometric processes on the factors, we must ensure the sample size is sufficiently large. Out of our 173 factors, all have more than 197 subsequent data points available, with 152 having more than 500. The mean amount of observations per predictor is 803. Specifics concerning the predictors and extensive details as to how they are constructed can be found at Open Asset Pricing (Chen and Zimmermann, 2021).

As mentioned in the introduction, we are specifically interested in assessing the size of linear-based asset pricing tests using simulations. This requires in the calibration step for our clear predictors to exhibit an α of 0, which, by definition, is not the case. We address this problem by employing a similar methodology to what is found in Fama and French's landmark study on mutual funds evaluation, where they scale fund returns by subtracting the time series by their alpha component directly (Fama and French, 2010). That is, they compute $\hat{\alpha}$ for every mutual fund via least squares estimation on a constant, and then subtract the same time series respectively by their estimated $\hat{\alpha}$. We do a similar procedure on predictor returns used to estimate models that will generate the simulations, fixing their α to zero. Doing so ensures that any departure from the conventional rejection rate of our asset pricing tests is solely an effect of the time series properties of anomaly returns.

	Original	Scaled
N	787	803
Mean	0.0051	9.78903e-20
Standard Deviation	0.03754	0.03683
Minimum	-0.21529	-0.21700
25%	-0.01281	-0.01781
Median%	0.00458	-0.00056
75%	0.02229	0.01707
Maximum	0.26288	0.25948

Table 1.1:

Summary Statistics for CZ anomalies. The first column represents the initial set gathered by CZ, while the second represents the sample of scaled anomalies that are kept to generate the simulated anomalies.

Table 1.1 presents summary statistics for the original and scaled anomalies from CZ. Average statistics are computed by taking the metrics for every anomaly in each panel (original and scaled) and averaging those findings. Notably, the rightward column exhibits a significantly smaller mean, due to the scaling process discussed previously. The average number of observations per predictor time series is represented by N.

1.3 Simulation settings

We use six univariate models as DGPs to simulate predictors' returns. Our main model draws from the class of univariate processes Christoffersen and Langlois (2013) (CL) use to model factor returns. Opting for an iterative approach, they seek to fit the time series of the three Fama-French and Momentum factors most accurately with popular econometric models allowing autocorrelation and time-varying variance. By iterative approach, it is meant that they start with the simplest tests for autocorrelation and heteroskedasticity, quickly excluding the possibility of the White Noise (WN), GARCH and AR models to accurately fit the factors in isolation. They then explore the autoregressive conditionally heteroskedastic model (AR-GARCH), further including nonlinear dynamics (AR-NGARCH) and skewed-t innovations (AR-NGARCH-SKEWT). Studying conditional mean, residuals and autocorrelation functions, they find that the best model to represent the factors' time series is a combination of a mean AR model of order 3, the nonlinear NGARCH of Engle and Ng (1993) and Hansen's (1994) skewed-t innovations. Details concerning the skewed-t innovations can be found in Section 2.2 of the Appendix. This model allows for autoregressive and conditionally heteroskedastic processes with leverage and skewness, features commonly held to constitute financial time series and specifically asset returns. Although their analysis is centered on the FF3 and momentum factors, we suggest that it can be generalized to other predictors.

Mapping our approach to CL, we opt for a step-by-step approach, building the AR-NGARCH-SKEWT piecewise and providing regression results for simulations generated by the simplest model to the most complex. Doing so allows us to show the rejection rate of the OLS estimator for a wide range of processes, including very simple ones. The six (6) models acting as DGPs are presented in Table 1.2. While the first five are implicit in CL's methodology, the last (and most accurate) model can be found directly in Section 2 of their work.

Model	Mean Process	Volatility Process
WN	$y_t = \mu + \sigma_t e_t$	Constant
GARCH	$y_t = \mu + \sigma_t e_t$	$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2$
AR	$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \sigma_t e_t$	Constant
AR-GARCH	$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \sigma_t e_t$	$\sigma_t^2 = \omega + \alpha \epsilon_{t-1}^2 + \beta \sigma_{t-1}^2$
AR-NGARCH	$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \sigma_t e_t$	$\sigma_t^2 = \omega + \alpha \sigma_{t-1}^2 \left(\epsilon_{jt-1} - \theta\right)^2 + \beta \sigma_{t-1}^2$
AR-NGARCH-SKEWT	$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \sigma_t e_t$	$\sigma_t^2 = \omega + \alpha \sigma_{t-1}^2 \left(\epsilon_{jt-1} - \theta\right)^2 + \beta \sigma_{t-1}^2$

Table 1.2:

Specifics of models calibrated on the sample of predictors and then used for simulations. Note that shocks are normally distributed in the simulations with a mean of 0 and variance of 1 $e_t \sim \mathrm{N}(0,1)$, unless they are SKEWT, where they are $e_t \sim \mathrm{SKEWT}(0,1,\lambda,\nu)$. Also note that μ is fixed at 0.

For each model and each predictor, we estimate the parameters on the predictors' time series, yielding an array of parameter values for each relevant parameter. Then, we take the median value of these arrays of parameters and simulate representative anomalies with the resulting values. For example, our first set of simulations is issued from a WN process with constant variance $\epsilon_t \sim \text{WN}(0, \sigma^2)$ which constitutes our benchmark when assessing the size properties of linear tests. As such a model breaks no Gauss-Markov assumptions, we would expect that it presents an empirical rejection rate with the OLS estimator that is adequate as the set of tests grows large. Table 1.3 outlines the values for the parameters employed to simulate predictors by model.

	WN	GARCH	AR	AR-GARCH	AR-NGARCH	AR-NGARCH-SKEWT
μ	-5.42101e-20	-0.000507979	3.96889e-06	-0.000488651	1.68979e-17	1.68979e-17
σ	0.00118301		0.00116287			
ϕ_1			0.0745727	0.0730808	0.078694	0.078694
ϕ_2			0.01855	0.0104689	0.0156102	0.0156102
ϕ_3			-0.012924	-0.0124273	-0.00667979	-0.00667979
ω		3.89026e-05		3.7322e-05	3.27716e-05	3.27716e-05
α		0.136723		0.120465	0.136734	0.136734
β		0.8		0.819737	0.806392	0.806392
θ					-0.138116	-0.138116
λ						1.02049
ν						6.12561

Table 1.3:

Parameters for simulations of predictors by model. Note that the μ parameters are specifically fixed at 0 in the simulations.

Although all our DGPs are flexible, allowing both choice of sample size and number of simulations, we use a fixed number of 100,000 simulations for each process with finite sample sizes of 800 and 200. The important number of simulated predictors is based on the necessity to allow for generalizable conclusions. The first sample size is set to best mimic our predictors' time series, of which the N was found to be 803 on average (see Table 1.1), and the second is to showcase the impact of our analysis on a small sample representing the smallest N, 197. Table 1.4 outlines summary statistics for simulated predictors returns generated by different models.

	WN	GARCH	AR	AR-GARCH	AR-NGARCH	AR-NGARCH-SKEWT
Mean	-4.58685e-07	2.30484e-05	-1.94265e-05	8.46778e-06	-5.98989e-06	1.13915e-05
Std	0.034415	0.0247927	0.0342025	0.0250707	0.0246105	0.0238848
Minimum	-0.132351	-0.152359	-0.131475	-0.142549	-0.160672	-0.223566
25%	-0.0232067	-0.0154762	-0.023094	-0.0158524	-0.0152088	-0.0130129
50%	7.34251e-06	2.09892e-05	-1.82895e-05	1.7104e-05	-3.49784e-05	-0.000214145
75%	0.0232079	0.0155182	0.0230669	0.0158686	0.0151503	0.0128123
Maximum	0.133422	0.1508	0.132016	0.144102	0.164263	0.230665

Table 1.4: Summary statistics for simulated predictors' returns generated by different models.

Chapter 2

Discussion

Here we report and discuss the findings associated with the implications of autocorrelation and heteroskedasticity on linear asset pricing tests while providing empirical grounds for a related correction. While our analysis starts with simulated processes to reach generalizable conclusions on simulated predictors, we also show in an empirical exercise the repercussions of said correction on the actual corpus of significant anomalies gathered by CZ.

2.1 Simulation Results

Figure 2.1 and Figure 2.2 show the effect of the kernel and bandwidth interactions on the rejection rates of linear asset pricing tests. Throughout the discussion, we invite the reader to consult Appendix 2.2 for details on the different kernels presented, where we include graphs and equations on each of them.[1] Conventional econometric wisdom suggests that an increasing bandwidth parameter increases bias and diminishes variance until over-rejections worsen. However, what occurs is that $\hat{\Omega}$ eventually places the full weight on all the sample autocovariances, which reduce to zero as $n \to \infty$. This leads the estimators to produce increasing rejections as bandwidth increases after attaining a minimum.

[1] For a theoretical discussion on the effect of the bandwidth parameter on the HAC estimator's variance/bias tradeoff and its impact on inference, which complements nicely the interpretation of the figures, we direct the reader to Kiefer and Vogelsang, 2005 at page 1139 to 1141 and S. Ng and Perron, 1996.

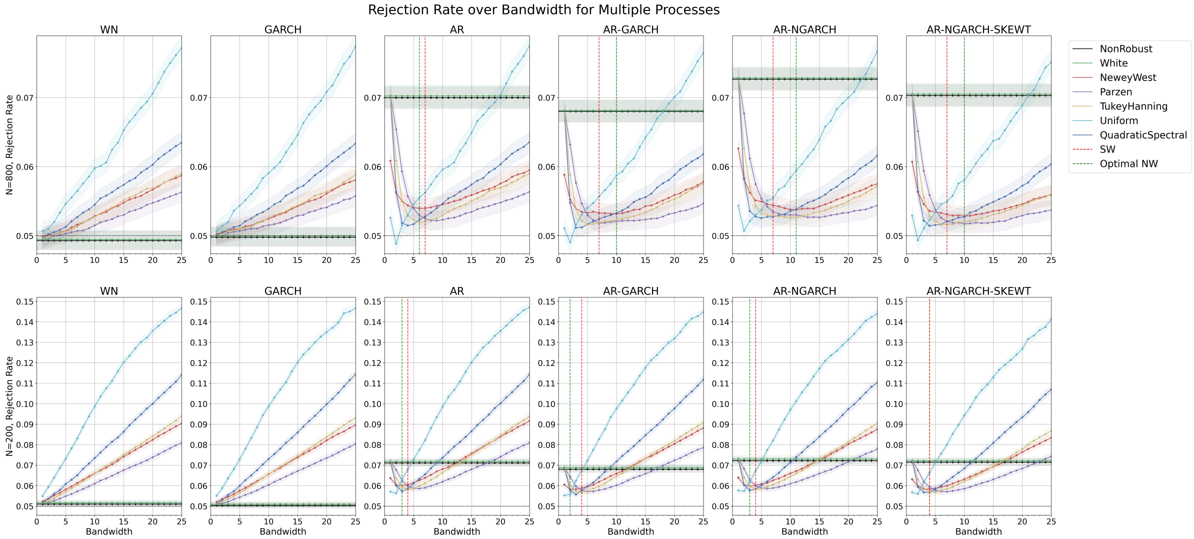

Figure 2.1:

Empirical rejection rate of linear-based asset pricing tests for $N = 800$ and $N = 200$, at $p = 0.05$, given the interaction between kernel and bandwidth parameters. The assumed model for the simulated predictors can be found above each of the columns. The x axis represents the bandwidth of the estimator while the y axis represents the associated rejection rate. Kernel specification can be found in the legend, given by different colors, where the specifics surrounding them can be found in Appendix 2.2. Highlighted zones around the lines are the confidence intervals of our results based on simulations. Dotted vertical lines of different colors represent the Stock-Watson bandwidth and empirically determined optimal bandwidths.

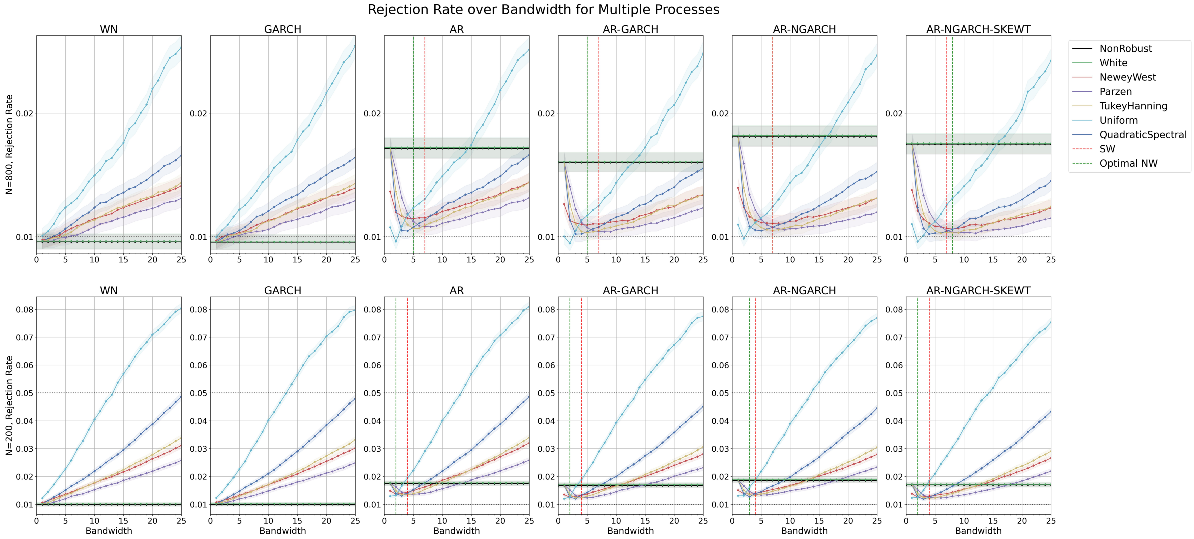

Figure 2.2:

Empirical rejection rate of linear-based asset pricing tests for $N = 800$ and $N = 200$, at $p = 0.01$, given the interaction between kernel and bandwidth parameters. The assumed model for the simulated predictors can be found above each of the columns. The x axis represents the bandwidth of the estimator while the y axis represents the associated rejection rate. Kernel specification can be found in the legend, given by different colors, where the specifics surrounding them can be found in Appendix 2.2. Highlighted zones around the lines are the confidence intervals of our results based on simulations. Dotted vertical lines of different colors represent the Stock-Watson bandwidth and empirically determined optimal bandwidths.

Each row represents a particular sample size, denoted by N, and each column corresponds to a model assumption used as DGP, from simplest to most intricate as described in Table 1.2. Within each subplot, each line corresponds to the empirical rejection rate obtained by using a specific estimator outlined in the legend, both at the $p = 0.05$ and $p = 0.01$ levels. Respectively, Default is the least square estimator, White is the heteroskedasticity-robust estimator, with HAC estimators being Uniform, Newey-West, Quadratic-Spectral (QS), Parzen and Tukey-Hanning (TH). Bandwidth is set from 1 to 25, which allows for showing the rejection rate of the estimators as this important parameter increases. The highlighted zone around each line corresponds to the confidence interval of the rejection rates obtained via simulations for every estimator. We add dotted vertical lines of different colors to represent the optimal bandwidth parameters obtained by simulations at both levels and the Stock-Watson suggested bandwidth.

In line with our expectations, a WN process of the predictors being tested results in an approximately adequate type 1 error rate for OLS and White. For robust estimators however, the rejection rate worsens significantly with an increasing bandwidth parameter for both sample sizes, which suggests that HAC estimators do not reduce to the OLS estimator when autocorrelation and heteroskedasticity are not intended features of the time series being studied. The size distortions are even more pronounced for the smaller sample size $N = 200$, with rejection rates reaching nearly 15% at a bandwidth of 25, while reaching approximately 7.8% for the larger sample size $N = 800$. A great deal of caution must therefore be exercised before applying such corrections if the econometric properties of the data are unknown. The picture is very similar in the context of processes following GARCH volatility and constant mean. Both the OLS and White estimators are well specified, with other estimators worsening with increasing bandwidth parameters. The surprising finding is that heteroskedasticity in isolation doesn't seem to create important size distortions in the context of linear asset pricing tests. As a result, White's estimator, whose objective is to limit the effects of heteroskedasticity in such tests, has indistinguishable impact[2] on the rejection rate when compared with the least squares estimator.

[2] Although the difference is not nil, as shown in our numerical results which are summarized here visually for readability.

The other four model assumptions present strikingly different results from the first two, which is of great interest as they can be most associated with anomaly processes. First and foremost, there are only small differences in the empirical type 1 error rate curves between the four model assumptions, which suggests that size distortions occur mostly due to the autoregressive property of the time series under study. Although the interaction between mean, volatility and distributional assumptions plays a key role in shaping the moments of returns, it is clear that the autoregressive component has the most impact on size distortions of linear asset pricing tests. While this may prove unanticipated, it follows closely with the previously reported minute impact of White's covariance matrix on results when the assumed process is constant GARCH.

For each of the four processes at sample size $N = 800$ tested at $p = 0.05$ using OLS, the false positive rate hovers around 7%, which is 40% above what classic confidence interval-based testing would suggest for a 95% test. In other words, this finding suggests that for a hypothetical set of 100 predictors with true α of 0 being tested via least squares, 7 would be significant on average at the 5% level rather than 5, uniquely due to their unaccounted-for econometric properties. Tested at $p = 0.01$, the rate of false positives remains consistently close to 1.7%, representing an approximate 70% overestimation compared to what conventional confidence interval-based testing would anticipate for a 99% level test. Strikingly, no HAC succeeds in fully correcting for autocorrelation and heteroskedasticity. That is, the estimators and their confidence intervals which provide the least size distortions do not touch the expected rejection rates, even less yield conservative estimates. The picture is even exacerbated at $N = 200$, where OLS presents similar false positive rates at both $p = 0.01$ and $p = 0.05$, but where all HAC estimators and their confidence intervals fail to make contact with the aforementioned levels in an even greater fashion. Supposing that we assume the benchmark AR-NGARCH-SKEWT model, even employing the empirically optimal Newey-West estimator would yield a size distortion of approximately 20%, a problem for which we currently have no solution.

Note that such important distortions in rejection rates result from a relatively weak autocorrelation as shown in Table 1.3, which renders these findings even more concerning. Although our approach is general in scope, some particular financial time series exhibit much higher levels of autocorrelation, which may affect inference even more pronouncedly. Our results suggest that much more concern should be directed towards the handling of such properties using HAC and that doing so should not be considered a fail-safe solution.

Following on the multiple hypothesis problem in anomaly research outlined previously, such distortions also have implications in the context of joint tests. In Figure 2.3 can be found the incidence of MHT given a test with an adequate versus distorted rejection rate. Relatively to the settings with accurate rejection rates, the settings with the inadequate rejection rate show a probability of type 1 error not only much more important from the start but increasing in a faster fashion with the number of tests. At $p = 0.05$, given 10 tests undertaken, the probability of type 1 error for the former setting is 40.12%, while it is 51.60% for the latter. Given 50 tests, both rise to 92.30% and 97.34% respectively. At $p = 0.01$, given 10 tests undertaken, the probability of type 1 error for the adequate setting is 10.46%, while it is 15.31% for the distorted setting. Given 50 tests, both rise to 40.10% and 53.73% respectively.

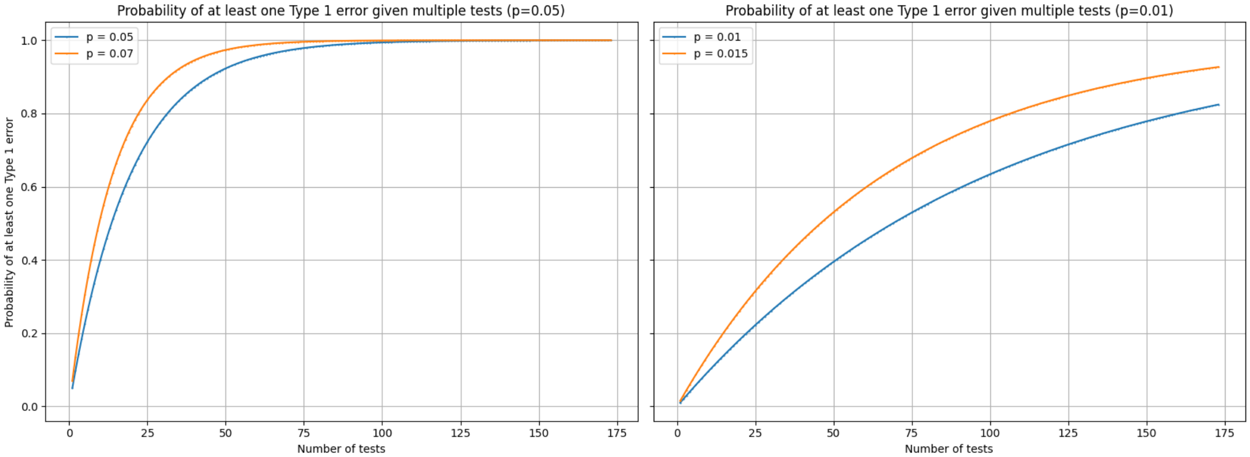

Figure 2.3:

Incidence of the MHT problem given an adequate rejection rate ($p = 0.05$ and $p = 0.01$) versus distorted ($p = 0.07$ and $p = 0.015$). The x axis represents the number of tests undertaken and the y axis is the probability of making at least one type 1 error.

As discussed previously, the MHT problem has been the subject of careful study in asset pricing due to its important implications on inference. Noticeably, Chordia, Goyal, and Saretto, 2020 recently found that accounting for this flaw in anomaly-detection methodology requires an important increase in threshold t-statistics: from 1.96 to 3.8 and 3.4 for time series and cross-sectional regressions respectively. They further estimate the expected proportion of type 1 error that researchers would produce if they failed to account for MHT to be around 45%. However, as their methodology assumes an adequate rejection rate in the case of independent tests, which we have shown is not the case due to autocorrelation and heteroskedasticity, we suggest that their results could be underestimations. Taking into account the initially inflated rejection rate that results from these properties, one could expect even stricter thresholds for significance to be required to correct for the multiple hypothesis testing problem.

However, Figure 2.1 and Figure 2.2 also suggest that a partial correction based on autocorrelation-robust covariance matrices is possible. Although such a correction does not aim at solving the multiple hypothesis testing problem, it can help to approach the initial expected rejection rate. We present the evolution in rejection rate over a series of bandwidth parameters by model assumption to conclude the validity of HAC estimators paving the way to a solution. Our first contender is the uniform kernel, which is the only one to fully correct for autocorrelation and heteroskedasticity in linear asset pricing tests, with only very few lags. This is due to the uniform kernel fully incorporating the autocorrelation structure of the residuals within its estimation, with weights initialized at $w = 1$ until the bandwidth lag is reached. Although efficient at $l = 1$, this structure comes at a significant cost, with a rejection rate increasingly distant from the objective as bandwidth grows due to covariance matrix misspecification. Choosing such a setting would require an unambiguous a priori knowledge of the true econometric structure of the time series being tested to avoid worsening the problem, which is often not the case in practice. For the researcher and practitioner, we suggest that a kernel offering more flexibility would be better suited. Such a candidate could be the very popular NW estimator, incorporating a Bartlett kernel with linearly decreasing weights. In all models, the NW significantly improves the type 1 error rate of the tests given a suitable bandwidth measure. Note that in contrast to its success in popularity over other non-uniform estimators presented, the NW is often the worst at optimal bandwidth in terms of approaching the intended 5% and 1% type 1 error rate. All the Quadratic Spectral (QS), Parzen and Tukey-Hanning (TH) perform better at optimal lags, which indicates that their weight structures better fit the autoregressive component of the simulated predictors' processes. The best estimators under ideal conditions are empirically determined to be the Parzen and Tukey-Hanning, although we emphasize that the difference between non-uniform estimators is relatively small in consideration of their significant gains over OLS.

From a practical perspective, we recognize that it is highly unlikely that practitioners and researchers will have access to a full apparatus to precisely determine the optimal kernel and bandwidth to use when undertaking linear tests. That is, they will not go through the ordeal of simulating thousands of time series with multiple models parameterized according to their series of interests and undergo linear tests with multiple HAC estimators to find the absolute best interaction of bandwidth and kernel given their context. Also, we acknowledge that they might not have full versatility in terms of estimators within their technical framework. Quite noticeably, the leading Python statistical library Statsmodels only includes the uniform and Newey-West HAC estimators (Seabold and Perktold, 2010). As mentioned previously, although the NW is not the absolute best at optimal conditions when compared to non-uniform kernels, we believe that its accessibility and interpretability more than compensate for its shortfalls. Hence, we suggest as an alternative to a more precise correction, a pragmatic and empirically informed correction based on this popular estimator.

	$N = 800$		$N = 200$	
Bandwidth	$p = 0.05$	$p = 0.01$	$p = 0.05$	$p = 0.01$
1	2.04734	2.69596	2.06223	2.70031
2	2.01118	2.64627	2.03578	2.66557
3	1.99736	2.62012	**2.03047**	**2.65861**
4	1.99066	2.60906	2.03204	2.66687
5	1.98871	2.60349	2.03768	2.67868
6	1.98705	2.59952	2.03874	2.68929
7	1.98691	2.59735	2.04475	2.7125
8	1.9834	**2.59473**	2.05171	2.72629
9	**1.98091**	2.59745	2.06071	2.73878
10	1.98264	2.59863	2.07016	2.75414

Table 2.1:

Adequate t-statistic threshold for a $p = 0.05$ or $p = 0.01$ test with NW estimator given bandwidth and sample size N. The assumed model assumption is AR-NGARCH-SKEWT. The most accurate given the specified threshold is shown in bold.

Table 2.1 reads as follows: given an assumed AR-GARCH-SKEWT model, our benchmark model, each cell represents the adequate threshold t-statistic required by bandwidth value to obtain an empirically correct rejection rate of $p = 0.01$ or $p = 0.05$ given a set of 100,000 tests on simulated series following this given model assumption. For example, testing an anomaly for significance in a large sample while having established time series properties following an AR-NGARCH-SKEWT at $p = 0.05$, the optimal bandwidth would be 9 with a threshold t-statistic of approximately $t = 1.98$. A more developed example of employing this table will be featured in the coming Section 2.2, where it will be used to determine the impact of the methodological change on a panel of predictors. Corroborating results found in Figure 2.1 and Figure 2.2, the t-statistics initially decrease and then increase after attaining a minima. Note that none of these bandwidth parameters obtain a conservative nor valid threshold of $t = 1.96$ nor $t = 2.574$ for these models, closely mapping our previous suggestion that such a solution remains imperfect.

However, we see this framework as a significant improvement in contrast to what is suggested in the literature and programming documentation, while remaining realistically implementable. Indeed, if researchers or practitioners are not inclined to use a non-standard table as suggested by Kiefer (2005), they are prompted to either guess or apply a rule of thumb derived mathematically. Most-oft suggested rules of thumb corrections respectively employ the NW estimator with a bandwidth parameter set at $0.75 \cdot (n^{1/3})$ (Stock and Watson, 2003) (SW), or $4 \cdot \left(\frac{n}{100}\right)^{\frac{2}{9}}$ (Wooldridge, 1996), with n representing the number of observations in the time series.

In our case, such rules of thumb would imply bandwidth parameters of 7 and 6 when $N = 800$ and 4 and 5 when $N = 200$ respectively. While these parameters seem rather accurate given the conclusions drawn previously, doubts can be cast as to their optimality in all conditions. There are multiple instances of settings, as shown in Figure 2.1 and Figure 2.2, where these rules fail to corroborate with the empirically found optimal bandwidth. Figure 2.4 presents the required p-value and their corresponding t-statistics thresholds to obtain significance based on different model assumptions. As in the previous figure, each row represents a particular sample size, with each subplot presenting inference results obtained by the estimators for a particular process versus the expected theoretical rejection rate of linear tests. Note that the BLUE line refers to the expected rejection rate at a given p, hence its slope of 1. We add to the OLS estimator the Stock and Watson "rule of thumb" estimator discussed previously to assess its adequacy. We also add an extreme example, the Newey-West estimator with 60 lags (5 years of monthly data), which has been used for example in an asset pricing context by Frazzini and Pedersen, 2014. By doing so, we seek to demonstrate the impact of a misspecified robust estimator on rejection rates.

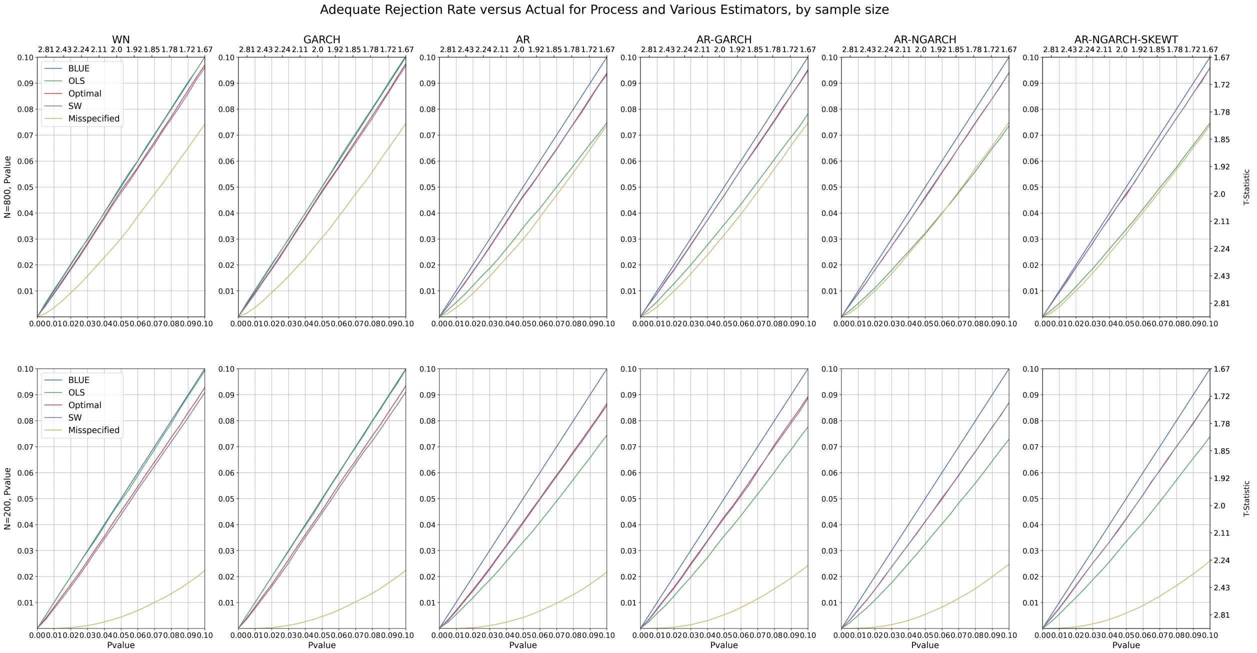

Figure 2.4:

Empirical rejection rate versus expected theoretical rejection rate of certain popular estimators given different model assumptions. T-statistics are provided on the upper and right axes, while p-values are provided on the lower and left axes.

Noticeably, adequate threshold statistics to obtain consistent estimates are higher than the expected $t = 1.96$ for all estimators. For example, the last subplots of both rows show that a linear test employing OLS would require a t-statistic of 2.11 ($p = 0.035$) when $N = 800$ and 2.16 ($p = 0.03$) when $N = 200$ to establish consistent significance at the 5% level for our assumed benchmark model. This finding suggests that any variable estimated with OLS in the asset pricing literature with a t-statistic below 2.16 (and not 1.96) should be considered non-significant at the 5% level. Even the empirically optimal and Stock-Watson estimators require modification in threshold over the range of p for all four relevant processes at both sample sizes. These results corroborate the previously demonstrated result that even well-specified HACs are insufficient to provide adequate size for linear tests. All subplots demonstrate that employing a misspecified HAC estimator, Newey-West with a bandwidth of 60 in our example, significantly exacerbates the aforementioned effects. Consistent t-statistics for a 5%

test approach 2.20 in larger samples (even worse than OLS) and go as far as 2.76 in smaller samples for an assumed AR-NGARCH-SKEWT model. This can be explained by the same process outlined in the early part of the methodology: parametrically imposing a weighting structure on the variance-covariance matrix of the residuals in the estimation stage has the potential to misspecify the HAC matrix if the bandwidth choice is incorrect and an even greater degree in a smaller sample. In other words, employing a robust estimator represents a double-edged sword, as an inconsiderate choice for the bandwidth parameter can also result in incorrect inference.

As such estimators are often accepted unquestionably and/or without explicit mention in the literature, one can only imagine the ramifications of such important distortions in the ratio of false positives. One might infer that this is not only valid for linear asset pricing tests but also for other applications in empirically oriented research involving financial time series. Note that this is all in assuming that HACs are perfect tools, which we have shown is not even the case. Only very few articles in finance exhaustively specify and justify the nature of the HAC estimators they employ in linear settings they conduct to draw inferences. That is if they use these robust estimators at all. We now know that such practice is misguided and has an important impact on conclusions drawn from scientific research. Let this be a testament to the importance of caution in the choice of bandwidth parameter.

2.2 Empirical exercise

In this section, we seek to determine the impact of our suggested HAC correction found in Table 2.1 for autocorrelation and heteroskedasticity, on the corpus of significant predictors gathered by Chen and Zimmermann, 2021. This exercise aims to show the deviations incurred from using an empirically optimal robust linear regression methodology for inference paired with an increase in t-statistics thresholds, specifically in an asset pricing context. As we have previously shown that such an approach remains imperfect, we offer it as a temporary solution, primarily aimed at standardization of HAC practice. This perspective could be extended, and such an exercise is also within the scope of other subfields of financial research that draw inferences from linear tests on time series suspected to have autocorrelation and time-varying volatility.

First and foremost, using the full sample provided by CZ until the most recent update of the database, we note that 173 are significant at $t = 1.96$ rather than the 207 they advertise as clear predictors. As mentioned previously in Section 1.2, this is due to some predictors losing significance while considered in their full sample rather than strictly from the sample used in the original study. Taking the perspective of a financial researcher interested in anomaly detection, we employ the bandwidth parameters and thresholds in Table 2.1. Note that the terms *differ* and *divergent* will be used for anomalies that change from significant to non-significant or the reverse given the specified threshold.

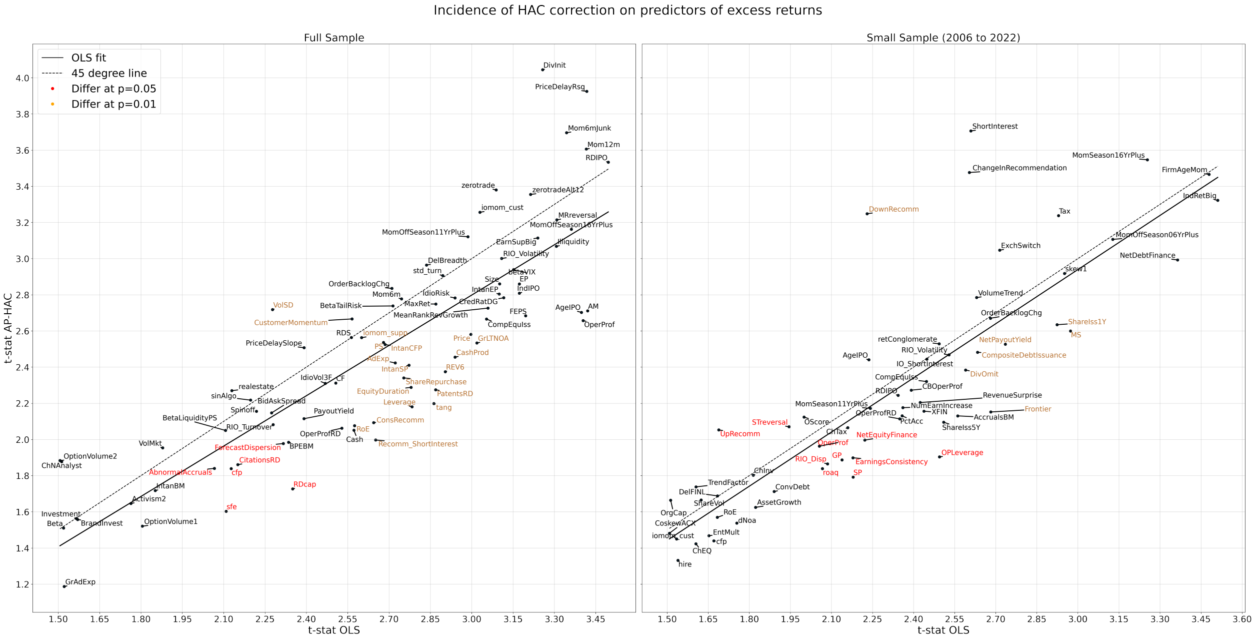

Figure 2.5:

Incidence of HAC correction on predictors of excess returns. The x axes represent the t-statistic obtained by the OLS estimator, while the y axes represent the t-statistic obtained by the empirically optimal robust estimator. The dotted 45-degree lines represent the equivalency of both estimators, and the plain lines represent the least square fit between estimators. Anomalies that differ at the $p = 0.05$ level are shown in red, and at $p = 0.01$, in orange. Both axes have been restricted for readability, and as anomalies presenting very high t-statistics are not at risk of differing.

In Figure 2.5 can be found the incidence of our correction on the predictors of excess returns when taken in their full sample. The x axes represent the t-statistic obtained by the OLS estimator, while the y axes represent the t-statistic obtained by the empirically optimal robust estimator. The dotted 45-degree lines would represent the equivalency of both estimators, and the plain lines represent the least square fit between estimators. We show the anomalies that differ at the $p = 0.05$ level in red, and at $p = 0.01$ in orange. Both axes have been restricted for readability, and as anomalies presenting very high t-statistics are not at risk of differing.

6 anomalies on 173 (3.46%) differ at the 1.96 level when applying the optimal NW correction with a bandwidth of 9 and a threshold of $t = 1.98$. 19 anomalies on 151 (12.58%) differ at the 2.574 level when applying the correction with a bandwidth of 8 and a threshold of $t = 2.594$. Based on the OLS fit, there is an empirical ground on which to suggest that rather than affecting only a few anomalies, HAC corrections shift the whole distribution of t-statistics downwards for the sample of clear predictors. As this shift is not of great importance in terms of absolute t-statistics, prior results where divergences do not seem numerous can be explained by the fact that most significant predictors present t-statistics that are far superior to the suggested thresholds when analyzed in their full sample. In other words, when using all available data, predictors when tested for α present t-statistics that are much above the barriers for significance. Therefore, they are not at risk of differing in significance status due to this methodological change.

However, we know from recent asset pricing literature that most anomalies are subject to alpha decay due to publication and increasing possibility of arbitrage (Schwert, 2003; Marquering, Nisser, and Valla, 2006; Jones and Pomorski, 2017; Guerard and Markowitz, 2018; Jacobs and Müller, 2020; Pénasse, 2022; Chen and Velikov, 2023). There is consensus in the literature that most predictors of exceeding returns performed extremely well in the period leading to the golden age of anomaly research and became progressively weaker in the coming of the second millennia due to crowding arising out of the soar of computationally and quantitatively oriented investing.

As anomalies weaken with time, we suggest that HAC corrections would become more influential due to the average t-statistic of the still significant predictors being closer to the benchmark statistical thresholds for the reasons outlined previously. To test this hypothesis, we follow Chen and Velikov, 2023 and restrict our time frame between 2006-01-01 and 2022-01-01, which roughly approximates the size of our small sample $N = 200$. This reduces our significant anomalies to 48 at the 1.96 level and 22 at the 2.574 level. Applying the well-specified Newey-West correction with 3 lags, we respectively find 10 (20.08%) and 7 anomalies (31.81%) to differ at those levels, using modified thresholds of $t = 2.03$ and $t = 2.658$. These findings demonstrate that as the distribution of t-statistics associated with clear anomalies shifts towards the thresholds naturally with time, correction for autocorrelation and heteroskedasticity will have increasing importance as small deviations will become more meaningful.

Conclusion

Given the recent evidence on predictors' persistence and time-varying volatility, we demonstrate empirically that linear asset pricing tests employing least squares estimation are misspecified as they do not account for autocorrelation and heteroskedasticity. Using simulations based on the underlying econometric processes of clear predictors of excess returns, we demonstrate that asset pricing tests based on portfolio sorts and OLS display an inflated type 1 error rate. Furthermore, we reach the important conclusion that HAC estimators widely held as standards to correct for such biases fail to provide estimates without size distortions, which compels for a better approach. To alleviate the problem, we guide the path towards a temporary standardization of HAC estimation employed in asset pricing. Applying said correction to clear predictors gathered by Chen & Zimmerman (2021), we demonstrate that the impact of said correction is significant, particularly considering alpha decay. Our work not only aligns with the extensive body of literature dedicated to enhancing statistical methodology in asset pricing but also paves the way for further exploration of scenarios in financial research where autocorrelation and heteroskedasticity may pose significant statistical challenges.

Bibliography

Akey, Pat, Adriana Robertson, and Mikhail Simutin (2022). “Noisy factors”. In: Rotman School of Management Working Paper Forthcoming.

Andrews, Donald WK (1991). “Heteroskedasticity and autocorrelation consistent covariance matrix estimation”. In: *Econometrica: Journal of the Econometric Society*, pp. 817–858.

Arnott, Robert D., Mark Clements, Vitali Kalesnik, and Juhani Linnainmaa (2021). “Factor momentum”. In: Available at SSRN 3116974.

Baker, Monya (2016). “1,500 scientists lift the lid on reproducibility”. In: *Nature* 533.7604, p. 7604.

Baquero, Guillermo and Marno Verbeek (2008). “Style Investing: Evidence from Hedge Fund Investors”. In: Available at SSRN 1102712.

Bender, Jennifer, Remy Briand, Dimitris Melas, Raman Aylur Subramanian, et al. (2013). “Foundations of factor investing”. In: Available at SSRN 2543990.

Blitz, David and Milan Vidojevic (2019). “The characteristics of factor investing”. In: *The Journal of Portfolio Management* 45.3, pp. 69–86.

Carhart, Mark M. (1997). “On persistence in mutual fund performance”. In: *The Journal of Finance* 52.1, pp. 57–82.

Chen, Andrew Y. (2021). “The Limits of p-Hacking: Some Thought Experiments”. In: *The Journal of Finance* 76.5, pp. 2447–2480.

Chen, Andrew Y. and Mihail Velikov (2023). “Zeroing in on the expected returns of anomalies”. In: *Journal of Financial and Quantitative Analysis* 58.3, pp. 968–1004.

Chen, Andrew Y. and Tom Zimmermann (2021). “Open source cross-sectional asset pricing”. In: *Critical Finance Review, Forthcoming.*

Chordia, Tarun, Amit Goyal, and Alessio Saretto (2020). “Anomalies and false rejections”. In: *The Review of Financial Studies* 33.5, pp. 2134–2179.

Christoffersen, Peter and Hugues Langlois (2013). “The joint dynamics of equity market factors”. In: *Journal of Financial and Quantitative Analysis* 48.5, pp. 1371–1404.

Cochrane, John H. (2011). “Presidential address: Discount rates”. In: *The Journal of Finance* 66.4, pp. 1047–1108.

Ehsani, Sina and Juhani T. Linnainmaa (2022). “Factor momentum and the momentum factor”. In: *The Journal of Finance* 77.3, pp. 1877–1919.

Engle, Robert F. and Victor K. Ng (1993). “Measuring and testing the impact of news on volatility”. In: *The journal of finance* 48.5, pp. 1749–1778.

Fama, Eugene F. and Kenneth R. French (1992). “The cross-section of expected stock returns”. In: *The Journal of Finance* 47.2, pp. 427–465.

— (2010). “Luck versus skill in the cross-section of mutual fund returns”. In: *The journal of finance* 65.5, pp. 1915–1947.

Fama, Eugene F. and James D. MacBeth (1973). “Risk, return, and equilibrium: Empirical tests”. In: *Journal of Political Economy* 81.3, pp. 607–636.

Feng, Guanhao, Stefano Giglio, and Dacheng Xiu (2020). “Taming the factor zoo: A test of new factors”. In: *The Journal of Finance* 75.3, pp. 1327–1370.

Frazzini, Andrea and Lasse Heje Pedersen (2014). “Betting against beta”. In: *Journal of Financial Economics* 111.1, pp. 1–25.

Froot, Kenneth and Melvyn Teo (2008). “Style investing and institutional investors”. In: *Journal of Financial and Quantitative Analysis* 43.4, pp. 883–906.

Giglio, Stefano, Bryan Kelly, and Dacheng Xiu (2022). “Factor Models, Machine Learning, and Asset Pricing”. In: *Annual Review of Financial Economics* 14, pp. 337–368.

Giglio, Stefano and Dacheng Xiu (2021). “Asset pricing with omitted factors”. In: *Journal of Political Economy* 129.7, pp. 1947–1990.

Gow, Ian D., Gaizka Ormazabal, and Daniel J. Taylor (2010). “Correcting for cross-sectional and time-series dependence in accounting research”. In: *The Accounting Review* 85.2, pp. 483–512.

Gu, Shihao, Bryan Kelly, and Dacheng Xiu (2021). “Autoencoder asset pricing models”. In: *Journal of Econometrics* 222.1, pp. 429–450.

Guerard, John and Harry Markowitz (2018). “The existence and persistence of financial anomalies: What have you done for me lately?” In: *Financial Planning Review* 1.3-4, e1022.

Gupta, Tarun and Bryan Kelly (2019). “Factor momentum everywhere”. In: *The Journal of Portfolio Management* 45.3, pp. 13–36.

Hansen, Bruce E. (1994). “Autoregressive conditional density estimation”. In: *International Economic Review*, pp. 705–730.

Harvey, Campbell R. (2017). “Presidential address: The scientific outlook in financial economics”. In: *The Journal of Finance* 72.4, pp. 1399–1440.

— (2021). “Be skeptical of asset management research”. In: Available at SSRN 3906277.

Harvey, Campbell R. and Yan Liu (2020). “False (and missed) discoveries in financial economics”. In: *The Journal of Finance* 75.5, pp. 2503–2553.

Harvey, Campbell R., Yan Liu, and Alessio Saretto (2020). “An evaluation of alternative multiple testing methods for finance applications”. In: *The Review of Asset Pricing Studies* 10.2, pp. 199–248.

Hou, Kewei, Chen Xue, and Lu Zhang (2020). “Replicating anomalies”. In: *The Review of financial studies* 33.5, pp. 2019–2133.

Ioannidis, John PA (2005). “Why most published research findings are false”. In: *PLoS medicine* 2.8, e124.

Israel, Ronen, Sarah Jiang, and Adrienne Ross (2017). “Craftsmanship alpha: An application to style investing”. In: *The Journal of Portfolio Management* 44.2, pp. 23–39.

Jacobs, Heiko and Sebastian Müller (2020). “Anomalies across the globe: Once public, no longer existent?” In: *Journal of Financial Economics* 135.1, pp. 213–230.

Jegadeesh, Narasimhan and Sheridan Titman (1993). “Returns to buying winners and selling losers: Implications for stock market efficiency”. In: *The Journal of Finance* 48.1, pp. 65–91.

Jones, Christopher S. and Lukasz Pomorski (2017). “Investing in disappearing anomalies”. In: *Review of Finance* 21.1, pp. 237–267.

Kelly, Bryan T., Seth Pruitt, and Yinan Su (2019). "Characteristics are covariances: A unified model of risk and return". In: *Journal of Financial Economics* 134.3, pp. 501–524.

Kiefer, Nicholas M. and Timothy J. Vogelsang (2005). "A new asymptotic theory for heteroskedasticity-autocorrelation robust tests". In: *Econometric Theory* 21.6, pp. 1130–1164.

Korajczyk, Robert A. and Ronnie Sadka (2004). "Are momentum profits robust to trading costs?" In: *The Journal of Finance* 59.3, pp. 1039–1082.

Lazarus, Eben, Daniel J. Lewis, James H. Stock, and Mark W. Watson (2018). "HAR inference: Recommendations for practice". In: *Journal of Business & Economic Statistics* 36.4, pp. 541–559.

Lintner, John (1965). "Security prices, risk, and maximal gains from diversification". In: *The journal of finance* 20.4, pp. 587–615.

Marquering, Wessel, Johan Nisser, and Toni Valla (2006). "Disappearing anomalies: a dynamic analysis of the persistence of anomalies". In: *Applied financial economics* 16.4, pp. 291–302.

McLean, R. David and Jeffrey Pontiff (2016). "Does academic research destroy stock return predictability?" In: *The Journal of Finance* 71.1, pp. 5–32.

Menkveld, Albert J., Anna Dreber, Felix Holzmeister, Juergen Huber, Magnus Johannesson, Michael Kirchler, Michael Razen, and Utz Weitzel (2022). "Non-standard errors". In.

Muravyev, Dmitriy, Neil D. Pearson, and Joshua Matthew Pollet (2022). "Anomalies and their Short-Sale Costs". In: Available at SSRN 4266059. URL: https://ssrn.com/abstract=4266059.

Newey, Whitney K. and Kenneth D. West (1986). "A simple, positive semi-definite, heteroskedasticity and autocorrelation-consistent covariance matrix estimator and a direct test for heteroskedasticity". In: *Econometrica: journal of the Econometric Society*, pp. 817–838.

Ng, Serena and Pierre Perron (1996). "The exact error in estimating the spectral density at the origin". In: *Journal of Time Series Analysis* 17.4, pp. 379–408.

Novy-Marx, Robert (2015). "Fundamentally, momentum is fundamental momentum". In: National Bureau of Economic Research, No. w20984.

Parzen, Emanuel (1962). “On estimation of a probability density function and mode”. In: *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076.

Pénasse, Julien (2022). “Understanding alpha decay”. In: *Management Science* 68.5, pp. 3966–3973.

Petersen, Mitchell A (2008). “Estimating standard errors in finance panel data sets: Comparing approaches”. In: *The Review of financial studies* 22.1, pp. 435–480.

Schwert, G. William (2003). “Anomalies and market efficiency”. In: *Handbook of the Economics of Finance*. Vol. 1, pp. 939–974.

Seabold, Skipper and Josef Perktold (2010). “statsmodels: Econometric and statistical modeling with python”. In: *9th Python in Science Conference*.

Sharpe, William F. (1965). “Risk-aversion in the stock market: Some empirical evidence”. In: *The Journal of Finance* 20.3, pp. 416–422.

Sheppard, Kevin (Mar. 2021). *bashtage/arch: Release 4.18*. Version v4.18. DOI: 10.5281/zenodo.593254.

Stock, James H. and Mark W. Watson (2003). *Introduction to Econometrics*. Vol. 104. Boston: Addison Wesley.

Tukey, J.W. (1967). “An introduction to the calculations of numerical spectrum analysis”. In: *Spectral Analysis of Time Series*, pp. 25–46.

White, Halbert (1980). “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity”. In: *Econometrica: journal of the Econometric Society*, pp. 817–838.

Wooldridge, Jeffrey M. (1996). *Introductory Econometrics: A Modern Approach*. 3rd.

Appendix A – Hansen's skewed-t errors

The density of Hansen's skewed-t distribution (Hansen, 1994) is given by

$$g(z \mid \eta, \lambda) = \begin{cases} bc\left(1+\frac{1}{\eta-2}\left(\frac{bz+a}{1-\lambda}\right)^2\right)^{-(\eta+1)/2} & z < -a/b \\ bc\left(1+\frac{1}{\eta-2}\left(\frac{bz+a}{1+\lambda}\right)^2\right)^{-(\eta+1)/2} & z \geq -a/b \end{cases} \tag{1}$$

where $2 < \eta < \infty$, and $-1 < \lambda < 1$. The constants a, b, and c are given by

$$a = 4\lambda c\left(\frac{\eta-2}{\eta-1}\right), \tag{2}$$

$$b^2 = 1+3\lambda^2 - a^2 \tag{3}$$

$$c = \frac{\Gamma\left(\frac{\eta+1}{2}\right)}{\sqrt{\pi(\eta-2)\Gamma\left(\frac{\eta}{2}\right)}} \tag{4}$$

Which composes a proper density function with mean zero and unit variance.

Appendix B – Kernels

Appendix B.1. Kernel Functions

Kernel	Function
Uniform	$K(x) = \begin{cases} 1.0, & \text{if } j \leq l \\ 0, & \text{otherwise} \end{cases}$
Bartlett (Newey and West, 1986)	$K(x) = \begin{cases} 1-\|z\| & z \leq 1 \\ 0 & z > 1 \end{cases}$
Quadratic Spectral (Andrews, 1991)	$K(x) = \begin{cases} 1 & z = 0 \\ \frac{3}{x^2}\left(\frac{\sin x}{x} - \cos x\right), x = \frac{6\pi z}{5} & z > 0 \end{cases}$
Parzen (1962)	$K(x) = \begin{cases} 1-6z^2(1-z) & z \leq \frac{1}{2} \\ 2(1-z)^3 & \frac{1}{2} < z \leq 1 \\ 0 & z > 1 \end{cases}$
Tukey-Hanning (Tukey, 1967)	$K(x) = \begin{cases} \frac{1}{2} + \frac{1}{2}\cos \pi z & z \leq 1 \\ 0 & z > 1 \end{cases}$

Table 1: Kernel Functions for Density Estimation. Note that $z = |\frac{j}{l}|, j = 0, 1, \ldots, l$ where l is the bandwidth parameter (Sheppard, 2021).

Appendix B.2. Kernel Visualization

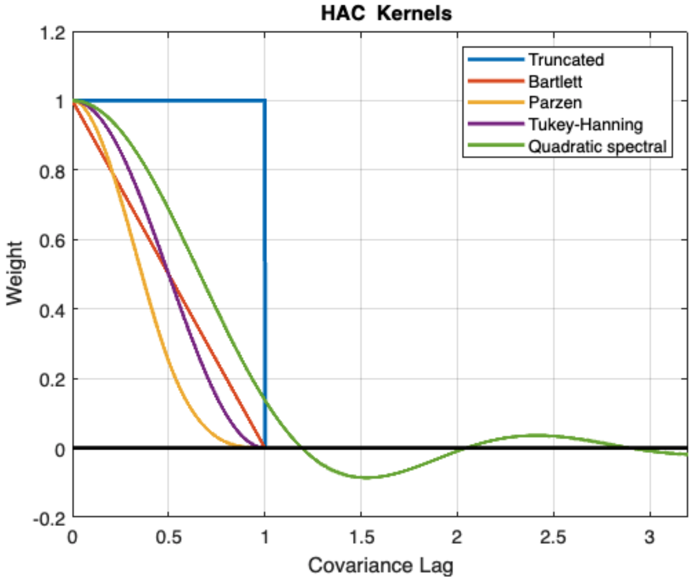

Figure 1: Kernel visualization by window type. Reproduced via Matlab Code found in MathWorks HAC Documentation

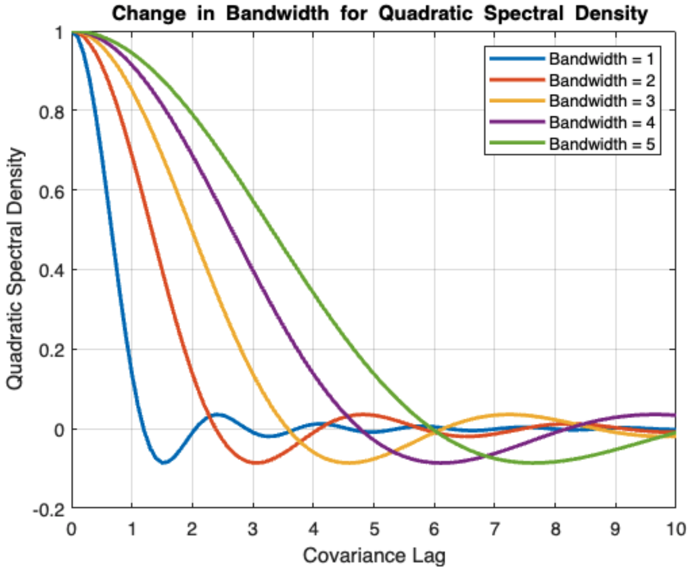

Figure 2: Kernel visualization by bandwidth. Reproduced via Matlab Code found in MathWorks HAC Documentation