The Statistics of Return Forecasting: Heavy Tails, Estimation, and the Limits of Predictive Inference*

Agustín Shehadi Candela
QUAFI Research
research@quafi.io

May 2026

Abstract

Forecasting asset returns is hard for reasons that are, at bottom, statistical. This paper gives a self-contained, mathematically rigorous account of why, and of the estimation theory that the difficulty demands. We formalize the forecasting problem for a stationary, ergodic return process, decompose predictive risk, and show that the predictable component is dwarfed by irreducible noise, so that estimation error governs out-of-sample performance and the finite-sample bias of predictive regression (Stambaugh, 1999) is first-order. We then develop the two technical obstacles that distinguish financial data from textbook settings. The first is *serial dependence*: valid inference about forecasts requires limit theory for dependent sequences and heteroskedasticity- and autocorrelation-consistent (HAC) variance estimation (Newey and West, 1987; Andrews, 1991), and the modern apparatus of predictive-ability testing (Diebold and Mariano, 1995; West, 1996; Giacomini and White, 2006). The second, and our central theme, is *heavy tails*: returns are regularly varying with a finite tail index, so high moments may not exist, classical central-limit arguments and sample-moment estimators become fragile (Mandelbrot, 1963; Cont, 2001; Davis and Mikosch, 1998), and one must turn to extreme-value theory — the Fisher–Tippett–Gnedenko and Pickands–Balkema–de Haan theorems (Gnedenko, 1943; Balkema and de Haan, 1974; Pickands, 1975) — and to tail-index estimation (Hill, 1975; Hall, 1982; McNeil and Frey, 2000). We give particular attention to *empirical confidence and prediction intervals*: when asymptotic normality fails, validity must be recovered by resampling that respects both dependence and heavy tails (block and stationary bootstraps, subsampling; Künsch, 1989; Politis and Romano, 1994; Politis et al.,

*QUAFI Working Paper Series, No. 2026-03. A self-contained, mathematically rigorous working paper. It is preliminary, circulated to inform and to invite comment; the views are the author's own. *Nothing herein is investment advice.* All cited results are attributed to their original published sources; classical theorems are stated without proof and referenced accordingly, while elementary results are proved for completeness. Any errors are mine.

1999) and by distribution-free conformal prediction with finite-sample coverage (Vovk et al., 2005; Lei et al., 2018), extended to the non-exchangeable, non-stationary regime that finance inhabits (Gibbs and Candès, 2021; Barber et al., 2023). We close with a rigorous workflow and a set of open problems. The unifying message is that honest predictive inference on returns is a problem of estimation under dependence and infinite-variance-adjacent tails, and that its solution is a disciplined combination of robust estimation, resampling, and distribution-free coverage.

Keywords: return forecasting; heavy tails; regular variation; extreme value theory; tail-index estimation; forecasting consistency; HAC inference; bootstrap; subsampling; conformal prediction; predictive regression.
JEL: C13, C14, C22, C53, C58, G17.
MSC (2020): 62M20, 62G32, 62E20, 91G70.

Contents

1 Introduction

The difficulty of forecasting asset returns is often described in economic terms — markets are efficient (Fama, 1970), edges are competed away — but its sharpest expression is statistical. The signal an analyst seeks is a small, slowly-varying conditional mean buried in noise whose distribution is heavy-tailed, serially dependent in its second moment, and non-stationary over long horizons. Each of these features breaks an assumption on which classical estimation and inference rest. The purpose of this paper is to set out, rigorously and from first principles, *which* assumptions break, *what* goes wrong when they do, and *which* estimation procedures restore valid inference — with sustained attention to the construction of trustworthy *empirical* confidence and prediction intervals.

The paper is organized around three obstacles and their remedies. Section 2 fixes the probabilistic setting — stationarity, ergodicity, mixing, and the martingale-difference structure of forecast errors — and Section 3 formalizes the *signal-to-noise* obstacle: a decomposition of predictive risk showing that estimation error, not the optimal forecast, dominates, and a precise statement of the predictive-regression bias of Stambaugh (1999). Section 4 treats *forecasting consistency*: the limit theory for dependent data that makes estimators and predictive-ability tests consistent, and the HAC variance estimation (Newey and West, 1987; Andrews, 1991) that makes their confidence statements valid. Section 5 is the technical core — *heavy tails*: regular variation, the equivalence between the tail index and the existence of moments, the extreme-value limit theorems, and the consequences for estimation when fourth or even second moments are in doubt. Section 6 surveys the state of the art in estimation under these conditions (Hill and POT tail estimators, conditional EVT, robust and quantile methods). Section 7 is devoted to *empirical confidence and prediction intervals*: why asymptotic intervals fail, how resampling under dependence and heavy tails restores them, and how distribution-free conformal prediction delivers finite-sample coverage even when distributional assumptions are abandoned. Section 8 assembles a defensible workflow; Section 9 states open problems; Section 10 concludes.

Our standard of rigor is the following: every nonstandard object is defined, classical theorems are stated precisely with assumptions and attributed to their sources (we do not reproduce their proofs), and elementary claims are proved.

2 Probabilistic setup

Fix a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Let $\{r_t\}_{t\in\mathbb{Z}}$ be a real-valued stochastic process of one-period (log) returns on an asset, and let $\{\mathcal{F}_t\}$ be a filtration with $\mathcal{F}_t = \sigma(r_s, \boldsymbol{x}_s : s \le t)$, where $\boldsymbol{x}_t \in \mathbb{R}^d$ collects any conditioning predictors observable at t.

Definition 1 (Stationarity and ergodicity)**.** The process $\{r_t\}$ is *strictly stationary* if, for every k and every h, $(r_{t_1}, \dots, r_{t_k}) \overset{d}{=} (r_{t_1+h}, \dots, r_{t_k+h})$. It is *covariance (weakly) stationary* if $\mathbb{E}[r_t^2] < \infty$,

$\mathbb{E}[r_t] = \mu$ does not depend on t, and $\mathrm{Cov}(r_t, r_{t+h}) = \gamma(h)$ depends only on h. A strictly stationary process is *ergodic* if every shift-invariant event has probability 0 or 1.

Ergodicity is what licenses the replacement of population expectations by time averages; it is the minimal condition under which a single realized price path is informative about the law that generated it.

Definition 2 (Strong mixing)**.** For $-\infty \le a \le b \le \infty$ let $\mathcal{F}_a^b = \sigma(r_t : a \le t \le b)$. The *strong ($\alpha$-)mixing* coefficient is

$$\alpha(n) \;=\; \sup_{t} \sup_{A \in \mathcal{F}_{-\infty}^{t},\, B \in \mathcal{F}_{t+n}^{\infty}} \big|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)\big|.$$

The process is *strongly mixing* if $\alpha(n) \to 0$ as $n \to \infty$.

Mixing quantifies asymptotic independence between the distant past and future and is the workhorse dependence condition for limit theory in econometrics (White, 2001). We will also use the martingale structure of errors.

Definition 3 (Optimal forecast and innovation)**.** Under squared-error loss the *optimal one-step forecast* of r_{t+1} given $\mathcal{F}_t$ is the conditional mean $\mu_t := \mathbb{E}[r_{t+1} \mid \mathcal{F}_t]$, and the *innovation* is $\varepsilon_{t+1} := r_{t+1} - \mu_t$. By construction $\mathbb{E}[\varepsilon_{t+1} \mid \mathcal{F}_t] = 0$, so $\{\varepsilon_t\}$ is a *martingale-difference sequence* (MDS) with respect to $\{\mathcal{F}_t\}$.

Proposition 1 (Optimality of the conditional mean)**.** *For any $\mathcal{F}_t$-measurable predictor g_t with $\mathbb{E}[g_t^2] < \infty$,*

$$\mathbb{E}\big[(r_{t+1} - g_t)^2\big] = \mathbb{E}\big[(r_{t+1} - \mu_t)^2\big] + \mathbb{E}\big[(\mu_t - g_t)^2\big] \;\ge\; \mathbb{E}\big[(r_{t+1} - \mu_t)^2\big],$$

with equality iff $g_t = \mu_t$ almost surely.

Proof. Write $r_{t+1} - g_t = (r_{t+1} - \mu_t) + (\mu_t - g_t) = \varepsilon_{t+1} + (\mu_t - g_t)$. Since $\mu_t - g_t$ is $\mathcal{F}_t$-measurable and $\mathbb{E}[\varepsilon_{t+1} \mid \mathcal{F}_t] = 0$, the cross term vanishes: $\mathbb{E}[\varepsilon_{t+1}(\mu_t - g_t)] = \mathbb{E}[(\mu_t - g_t)\mathbb{E}[\varepsilon_{t+1} \mid \mathcal{F}_t]] = 0$. Expanding the square gives the identity; the inequality and equality condition are immediate. □

Proposition 1 is elementary but frames everything that follows: the term $\mathbb{E}[(r_{t+1} - \mu_t)^2] = \mathbb{E}[\varepsilon_{t+1}^2]$ is *irreducible*, and all statistical effort is spent estimating μ_t — driving down the second, *estimable* term $\mathbb{E}[(\mu_t - g_t)^2]$.

3 The signal-to-noise obstacle

3.1 Predictive risk and the predictability ceiling

Suppose a modeler uses a forecast $\hat{\mu}_t = \hat{g}(\boldsymbol{x}_t; \hat{\theta}_n)$ with parameter estimated on n observations. Adding and subtracting μ_t and the population-optimal $g^*(\boldsymbol{x}_t)$ within the model class, Proposition

1 extends to the standard *noise / approximation / estimation* decomposition

$$\underbrace{\mathbb{E}[(r_{t+1}-\hat{\mu}_t)^2]}_{\text{predictive risk}} = \underbrace{\mathbb{E}[\varepsilon_{t+1}^2]}_{\text{irreducible}} + \underbrace{\mathbb{E}[(\mu_t - g^*(\boldsymbol{x}_t))^2]}_{\text{approximation (bias)}} + \underbrace{\mathbb{E}[(g^*(\boldsymbol{x}_t)-\hat{\mu}_t)^2]}_{\text{estimation (variance)}}. \tag{1}$$

Define the *population predictive* R^2 as $R^2_* = \mathrm{Var}(\mu_t)/\mathrm{Var}(r_{t+1})$. For daily equity returns R^2_* is tiny; empirically the out-of-sample R^2 relative to the historical mean is at best a fraction of a percent at monthly frequencies (Welch and Goyal, 2008; Campbell and Thompson, 2008). The following makes precise why a tiny R^2_* is so punishing.

Proposition 2 (When does a predictor help out of sample?)**.** *Consider the location model $r_{t+1} = \mu + \delta\, z_t + \varepsilon_{t+1}$ with z_t standardized, $\mathrm{Var}(\varepsilon) = \sigma^2$, and $\{\varepsilon_t\}$ an MDS. Compare the conditional forecast $\hat{\mu}_t = \hat{\mu} + \hat{\delta} z_t$ (OLS on n points) to the unconditional forecast $\bar{r}$. To first order, the conditional forecast has lower expected out-of-sample MSE than the unconditional one iff*

$$\frac{\delta^2}{\sigma^2} > \frac{1}{n}, \qquad \textit{equivalently} \qquad R^2_* > \frac{1}{n+1}.$$

Proof sketch. The conditional model spends one extra degree of freedom: its estimation variance exceeds the unconditional model's by σ^2/n to leading order (the variance of $\hat{\delta} z_t$ averaged over z_t), while its systematic gain is the explained variance δ^2. The conditional forecast wins exactly when the explained variance exceeds the extra estimation variance, $\delta^2 > \sigma^2/n$. Dividing by $\mathrm{Var}(r) = \delta^2 + \sigma^2$ gives the R^2_* form. □

Remark 1. Proposition 2 is the rigorous content of the Campbell and Thompson (2008) dictum: a real but minuscule signal is only *usable* once the sample is large enough that $1/n$ falls below R^2_*. With $R^2_* \approx 0.5\%$ one needs $n \gtrsim 200$ effective observations merely to break even, before any model misspecification or multiple-testing penalty. Estimation error, not the absence of signal, is the binding constraint.

3.2 Finite-sample bias: the predictive-regression problem

A second, subtler estimation problem afflicts the very regressions used to detect predictability. Many predictors (dividend yields, valuation ratios) are highly persistent and their innovations are correlated with returns.

Proposition 3 (Stambaugh bias)**.** *Consider the predictive system*

$$r_{t+1} = \alpha + \beta x_t + u_{t+1}, \qquad x_{t+1} = \theta + \rho\, x_t + v_{t+1},$$

with $\mathbb{E}[(u_{t+1}, v_{t+1}) \mid \mathcal{F}_t] = \mathbf{0}$ and $\mathrm{Cov}(u_{t+1}, v_{t+1}) = \sigma_{uv}$. Then the OLS estimator $\hat{\beta}$ inherits the

finite-sample bias of $\hat{\rho}$:

$$\mathbb{E}[\hat{\beta} - \beta] \;=\; \frac{\sigma_{uv}}{\sigma_v^2}\,\mathbb{E}[\hat{\rho} - \rho] \;\approx\; \frac{\sigma_{uv}}{\sigma_v^2}\Big(-\frac{1+3\rho}{n}\Big),$$

using the Stambaugh (1999) expansion (the Kendall bias of the AR(1) coefficient).

Because $\sigma_{uv} < 0$ for typical valuation-ratio predictors and ρ is near one, the bias inflates $\hat{\beta}$ away from zero, manufacturing apparent predictability. The lesson is methodological: in-sample t-statistics on persistent predictors overstate significance, and inference must be bias-corrected or based on the out-of-sample criterion of Proposition 2.

4 Forecasting consistency and inference under dependence

We now ask when a forecasting procedure is *consistent* — its forecasts (or the parameters indexing them) converge to their population targets — and how to attach *valid* confidence statements to comparisons of forecasts. Both require limit theory for dependent sequences.

Definition 4 (Consistency of a forecasting procedure)**.** Let $\hat{\theta}_n$ be the estimator underlying $\hat{\mu}_t = g(\boldsymbol{x}_t; \hat{\theta}_n)$, with population target $\theta_0 = \arg\min_\theta \mathbb{E}[(r_{t+1} - g(\boldsymbol{x}_t;\theta))^2]$. The procedure is *(weakly) consistent* if $\hat{\theta}_n \xrightarrow{p} \theta_0$, and *predictively consistent* if $\mathbb{E}[(g(\boldsymbol{x}_t;\hat{\theta}_n) - g(\boldsymbol{x}_t;\theta_0))^2] \to 0$.

Assumption 1 (Regular dependent design)**.** $\{(r_t, \boldsymbol{x}_t)\}$ is strictly stationary and ergodic; $g(\boldsymbol{x};\theta)$ is continuous in θ on a compact Θ with θ_0 interior; $\mathbb{E}\sup_\theta (r_{t+1} - g(\boldsymbol{x}_t;\theta))^2 < \infty$; and the process is strongly mixing with $\sum_n \alpha(n)^{1-2/\nu} < \infty$ for some $\nu > 2$ with $\mathbb{E}|r_t|^\nu < \infty$ and $\mathbb{E}\|\boldsymbol{x}_t\|^\nu < \infty$.

Theorem 1 (Consistency and asymptotic normality of the forecast estimator)**.** *Under Assumption 1 and standard identification/smoothness conditions, the least-squares estimator satisfies $\hat{\theta}_n \xrightarrow{p} \theta_0$ and*

$$\sqrt{n}\,(\hat{\theta}_n - \theta_0) \;\xrightarrow{d}\; \mathcal{N}\big(\mathbf{0},\; \boldsymbol{H}^{-1}\boldsymbol{S}\,\boldsymbol{H}^{-1}\big), \qquad \boldsymbol{S} = \sum_{h=-\infty}^{\infty} \mathrm{Cov}(\boldsymbol{s}_t, \boldsymbol{s}_{t+h}),$$

where $\boldsymbol{s}_t$ is the score (estimating-equation) summand and $\boldsymbol{H}$ its expected Jacobian. The long-run variance $\boldsymbol{S}$ *accounts for serial dependence.*

This is a standard consequence of an ergodic-theorem law of large numbers and a central limit theorem for mixing/martingale-difference sequences (White, 2001; for the functional version, Billingsley, 1968). The content for practice is the sandwich form: ignoring the off-diagonal covariances in $\boldsymbol{S}$ underestimates sampling variability whenever errors are autocorrelated, as they are for overlapping or volatility-clustered returns.

Theorem 2 (HAC consistency)**.** *Let $\widehat{\boldsymbol{S}} = \sum_{|h|\le B_n} K(h/B_n)\,\widehat{\boldsymbol{\Gamma}}(h)$ be a kernel long-run-variance estimator with sample autocovariances $\widehat{\boldsymbol{\Gamma}}(h)$, a kernel K (e.g. Bartlett, as in Newey and West,*

1987), and bandwidth $B_n \to \infty$, $B_n/n \to 0$. Under regularity and moment conditions, $\widehat{\boldsymbol{S}} \xrightarrow{p} \boldsymbol{S}$. The data-driven bandwidth and optimal-kernel theory are due to Andrews (1991).

4.1 Comparing forecasts

Consistency of an individual procedure is not the practical question; the practical question is whether one forecast *beats* another out of sample, with a valid p-value. Let $d_t = \ell(e_t^{(1)}) - \ell(e_t^{(2)})$ be the loss differential of two forecasts and $\bar{d} = \frac{1}{m}\sum_t d_t$ over m out-of-sample points.

Theorem 3 (Diebold–Mariano / West)**.** *If $\{d_t\}$ is covariance-stationary with long-run variance $S_d = \sum_h \operatorname{Cov}(d_t, d_{t+h}) \in (0,\infty)$, then $\sqrt{m}\,\bar{d}\,/\,\widehat{S}_d^{1/2} \xrightarrow{d} \mathcal{N}(0,1)$ under $H_0 : \mathbb{E}[d_t] = 0$, with $\widehat{S}_d$ a HAC estimator. Diebold and Mariano (1995) establish the test for given forecasts; West (1996) extends the asymptotics to account for estimation error in the forecasting models; Clark and West (2007) adjust the statistic for* nested *models, where the standard test is undersized.*

Theorem 4 (Conditional predictive ability)**.** *Giacomini and White (2006) test the hypothesis of equal* conditional *predictive ability, $\mathbb{E}[d_{t+1} \mid \mathcal{F}_t] = 0$, via the moment condition $\mathbb{E}[\boldsymbol{h}_t d_{t+1}] = \mathbf{0}$ for instruments $\boldsymbol{h}_t \in \mathcal{F}_t$. Under a fixed estimation window and mixing, the Wald statistic $m\,\overline{(\boldsymbol{h}d)}^{\top}\widehat{\boldsymbol{S}}^{-1}\overline{(\boldsymbol{h}d)} \xrightarrow{d} \chi^2_q$. This both tests and* guides *forecast selection, since the instruments reveal* when *one model dominates.*

When several candidate forecasts are compared at once, single-comparison p-values are invalid; one needs the family-wise corrections of Section 7 (White, 2000; Hansen, 2005; Hansen et al., 2011). All of this machinery, however, presumes the moment conditions ($\mathbb{E}|r|^{\nu} < \infty$, $S_d < \infty$) that heavy tails may violate — the subject of the next section.

5 Heavy tails

That asset returns are not Gaussian has been known since Mandelbrot (1963) and Fama (1965); the modern catalogue of "stylized facts" — heavy tails, volatility clustering, near-zero autocorrelation of returns alongside strong autocorrelation of squared returns — is given by Cont (2001). Heavy tails are not a cosmetic nuisance: they determine which moments exist, and therefore which estimators and limit theorems are even valid.

5.1 Regular variation and the tail index

Definition 5 (Regular variation)**.** A measurable $L : (0,\infty) \to (0,\infty)$ is *slowly varying* if $L(\lambda x)/L(x) \to 1$ as $x \to \infty$ for every $\lambda > 0$. A random variable X has a *regularly varying right tail with index $\alpha > 0$* if

$$\mathbb{P}(X > x) \;=\; x^{-\alpha} L(x), \qquad x \to \infty,$$

for some slowly varying L. We write $X \in \mathrm{RV}(-\alpha)$. The parameter α is the *tail index*; $\xi := 1/\alpha$ is the *extreme-value index*.

Power-law tails are the canonical heavy tails: they decay polynomially, not exponentially, so large moves are far more probable than under a normal law. The single most consequential fact is the exact correspondence between the tail index and the existence of moments.

Proposition 4 (Tail index and moments)**.** *If $X \in \mathrm{RV}(-\alpha)$ with $\alpha > 0$, then*

$$\mathbb{E}|X|^p < \infty \quad \textit{for } 0 \le p < \alpha, \qquad \textit{and} \qquad \mathbb{E}|X|^p = \infty \quad \textit{for } p > \alpha.$$

Proof idea. By the layer-cake representation $\mathbb{E}|X|^p = \int_0^\infty p\, x^{p-1} \mathbb{P}(|X| > x)\, dx$. With $\mathbb{P}(|X| > x) \sim x^{-\alpha} L(x)$, the integrand behaves like $p\, x^{p-1-\alpha} L(x)$ at infinity, which is integrable iff $p - 1 - \alpha < -1$, i.e. $p < \alpha$ (slowly varying factors do not affect the threshold; see Embrechts et al., 1997, Resnick, 2007). □

Empirically, daily equity-return tail indices typically fall in the range $\alpha \approx 3$–5 (Cont, 2001). By Proposition 4 this means the *variance exists* ($\alpha > 2$) but the *fourth moment may not* ($\alpha < 4$): the population kurtosis can be infinite, so the *sample* kurtosis does not converge and is a near-useless statistic, and any estimator whose asymptotics rely on a finite fourth moment (much variance-of-variance inference, naive GARCH standard errors) stands on unstable ground.

5.2 Extreme-value limit laws

Regular variation is precisely the condition under which sample maxima and threshold exceedances have non-degenerate limits.

Theorem 5 (Fisher–Tippett–Gnedenko)**.** *Let $X_1, \dots, X_n$ be i.i.d. with maximum $M_n = \max_i X_i$. If there exist $a_n > 0, b_n \in \mathbb{R}$ and a non-degenerate G with $\mathbb{P}((M_n - b_n)/a_n \le x) \to G(x)$, then G is, up to location-scale, a* generalized extreme value *(GEV) law $G_\xi(x) = \exp\{-(1+\xi x)^{-1/\xi}\}$ (for $1 + \xi x > 0$), with shape $\xi \in \mathbb{R}$. A distribution with $X \in \mathrm{RV}(-\alpha)$ lies in the maximum domain of attraction of the Fréchet case $\xi = 1/\alpha > 0$.*

The result originates with Fisher–Tippett and Gnedenko (1943); see de Haan and Ferreira (2006) for the modern treatment. Its statistical counterpart, the basis of the *peaks-over-threshold* (POT) method, characterizes the distribution of exceedances above a high threshold.

Theorem 6 (Pickands–Balkema–de Haan)**.** *X lies in a max-domain of attraction iff, for a suitable positive function $\beta(u)$,*

$$\lim_{u \uparrow x_F} \sup_{0 \le y < x_F - u} \left| \mathbb{P}(X - u \le y \mid X > u) - H_{\xi, \beta(u)}(y) \right| = 0,$$

where $H_{\xi,\beta}(y) = 1 - (1 + \xi y/\beta)^{-1/\xi}$ is the generalized Pareto distribution *(GPD) and x_F is the right endpoint.*

Balkema and de Haan (1974) and Pickands (1975) established this convergence: above a high enough threshold the excess distribution is approximately GPD, with the same shape ξ as the GEV limit. This is what makes tail estimation feasible from a finite sample — one need only model the exceedances.

5.3 Heavy tails from dynamics, and their effect on dependence estimates

Heavy tails need not be assumed in the innovations; they are *generated* by volatility dynamics. The ARCH/GARCH family (Engle, 1982; Bollerslev, 1986) produces unconditionally heavy-tailed, volatility-clustered returns even from light-tailed (e.g. Gaussian) shocks, because the stationary distribution of a GARCH process is regularly varying. A critical consequence for *predictability testing* is that heavy tails distort the sampling behavior of the sample autocorrelations themselves.

Theorem 7 (Sample ACF under heavy tails)**.** *Davis and Mikosch (1998) show that for heavy-tailed, ARCH-type processes the sample autocorrelations $\hat{\rho}(h)$ can converge at rates* slower *than $\sqrt{n}$ and to* non-Gaussian *limits expressible through stable laws, depending on which moments are finite. In particular, when the fourth moment is infinite, the usual $\pm 1.96/\sqrt{n}$ confidence bands for autocorrelations are not valid.*

Theorem 7 is sobering: the very diagnostic used to claim "returns are (un)predictable from their own past" has nonstandard sampling theory under the tails returns actually exhibit. Robust or resampling-based confidence statements (Section 7) are therefore not optional refinements but requirements.

6 Estimation under heavy tails

6.1 Tail-index estimation

The central estimation target is the tail index α (or $\xi = 1/\alpha$).

Definition 6 (Hill estimator)**.** Let $X_{(1)} \geq X_{(2)} \geq \cdots \geq X_{(n)}$ be the order statistics (descending) of a positive sample, and fix $k \in \{1, \ldots, n-1\}$. The *Hill estimator* (Hill, 1975) of $\xi = 1/\alpha$ is

$$\hat{\xi}^{\text{Hill}}_{k,n} = \frac{1}{k} \sum_{i=1}^{k} \big(\log X_{(i)} - \log X_{(k+1)}\big), \qquad \hat{\alpha} = 1/\hat{\xi}^{\text{Hill}}_{k,n}.$$

Theorem 8 (Consistency and asymptotic normality of the Hill estimator)**.** *Suppose $X \in \text{RV}(-\alpha)$. If $k = k_n \to \infty$ with $k_n/n \to 0$, then $\hat{\xi}^{\text{Hill}}_{k,n} \xrightarrow{p} \xi$. Under a second-order regular-variation refinement controlling the bias of L, and a corresponding growth restriction on k_n,*

$$\sqrt{k_n}\,(\hat{\xi}^{\text{Hill}}_{k,n} - \xi) \xrightarrow{d} \mathcal{N}(b,\ \xi^2),$$

with an asymptotic bias b that vanishes when k_n grows slowly enough (Hall, 1982; de Haan and Ferreira, 2006).

Theorem 8 exposes the defining tension of tail estimation: the variance ξ^2/k falls as k grows, but the bias rises as one reaches deeper into the (non-power-law) body of the distribution. Choosing k is a bias–variance problem with no universally optimal solution; practitioners inspect the "Hill plot" $k \mapsto \hat{\xi}_{k,n}$ for a stable region, or use adaptive, mean-squared-error-optimal selection rules. The POT alternative (Theorem 6) fits a GPD to exceedances above a threshold u by maximum likelihood, trading the choice of k for the choice of u.

6.2 Conditional EVT: filter, then estimate the tail

Returns are not i.i.d., so applying EVT to raw returns conflates volatility clustering with tail thickness. McNeil and Frey (2000) propose a two-stage *conditional EVT*: (i) fit a GARCH-type model and extract approximately i.i.d. standardized residuals $\hat{z}_t = r_t/\hat{\sigma}_t$; (ii) apply POT/Hill tail estimation to $\{\hat{z}_t\}$. Conditional quantiles (hence Value-at-Risk and expected shortfall) then combine the forecast $\hat{\sigma}_{t+1}$ with the estimated residual tail. This cleanly separates the two sources of large losses — a high *current* volatility versus a heavy *residual* tail — and is the state of the art for conditional tail-risk estimation.

6.3 Robust and quantile-based estimation

Because sample moments are fragile under heavy tails (Proposition 4), location and dispersion are better summarized by *quantiles*, which exist regardless of moments. Quantile regression (Koenker, 2005) estimates conditional quantiles $Q_\tau(r_{t+1} \mid \boldsymbol{x}_t)$ directly by minimizing the asymmetric "pinball" loss $\rho_\tau(e) = e(\tau - \mathbf{1}\{e < 0\})$, providing a moment-free description of the predictive distribution that is exactly what interval forecasting (next section) requires. For volatility specifically, *realized measures* built from high-frequency data (Andersen et al., 2003) furnish consistent, model-free variance estimates; the HAR model (Corsi, 2009) parsimoniously captures their long-memory dynamics, and bipower variation (Barndorff-Nielsen and Shephard, 2006) separates the continuous and jump contributions that heavy tails partly reflect. Tail risk estimated this way can itself be a priced state variable (Kelly and Jiang, 2014).

7 Empirical confidence and prediction intervals

We now reach the practical apex: quantifying uncertainty. We distinguish a *confidence interval* for a parameter or a forecast's expected accuracy from a *prediction interval* for the next realized return; both are imperilled by the features above, and both can be repaired.

7.1 Why asymptotic intervals fail

A Wald interval $\hat{\theta} \pm z_{1-\alpha/2}\,\widehat{\mathrm{se}}$ presumes (i) asymptotic normality at rate $\sqrt{n}$ and (ii) a consistent standard error. Heavy tails threaten both: when a relevant moment is infinite, the limit may be a non-Gaussian stable law at a rate slower than $\sqrt{n}$ (cf. Theorem 7), and the plug-in standard error may not converge. Dependence compounds the problem by inflating the true sampling variance beyond the i.i.d. formula (Theorem 1). Two routes restore validity: *resampling* and *distribution-free coverage*.

7.2 Resampling under dependence and heavy tails

The i.i.d. bootstrap (Efron, 1979) is invalid for dependent data because it destroys the serial structure. Two repairs preserve it.

Definition 7 (Block and stationary bootstrap)**.** The *moving-block bootstrap* (Künsch, 1989) resamples contiguous blocks of length ℓ and concatenates them, preserving within-block dependence. The *stationary bootstrap* (Politis and Romano, 1994) uses random block lengths Geometric(p), which makes the resampled series strictly stationary; the expected block length $1/p$ plays the role of ℓ.

Theorem 9 (Validity of the block/stationary bootstrap)**.** *For a strictly stationary, strongly mixing sequence with a finite variance and a CLT for the statistic of interest, the block and stationary bootstraps consistently estimate the sampling distribution of smooth functionals (e.g. the sample mean) provided $\ell \to \infty$, $\ell/n \to 0$ (Künsch, 1989; Politis and Romano, 1994). The block length governs a bias–variance trade-off analogous to the HAC bandwidth.*

Heavy tails, however, can break even Theorem 9: when the variance is infinite the ordinary n-out-of-n bootstrap of the mean is *inconsistent*. The robust fix is *subsampling*.

Theorem 10 (Subsampling)**.** *Politis et al. (1999) show that, under stationarity and mixing alone, recomputing a statistic on overlapping subsamples of size b with $b \to \infty$, $b/n \to 0$, consistently estimates its sampling distribution* even when the limit law is non-normal and the convergence rate is unknown — *precisely the heavy-tailed, infinite-variance regime where the standard bootstrap fails. The related m-out-of-n bootstrap shares this robustness.*

Theorem 10 is the rigorous answer to "how do I get an honest confidence interval when I am not sure my fourth (or second) moment exists?": subsample, and let the data reveal the rate and the shape of the limit.

7.3 Distribution-free prediction intervals: conformal prediction

Interval *forecasts* of the next return demand coverage of an as-yet-unrealized random variable, not a parameter. *Conformal prediction* (Vovk et al., 2005; Lei et al., 2018) delivers this with finite-sample guarantees and no distributional assumption.

Theorem 11 (Finite-sample coverage of split conformal prediction)**.** *Let* $(Z_1, \ldots, Z_n, Z_{n+1})$ *be* exchangeable*, with* $Z_i = (\boldsymbol{x}_i, r_i)$*. Fit a predictor on a training split, compute nonconformity scores* $s_i = |r_i - \hat{\mu}(\boldsymbol{x}_i)|$ *on a calibration set* $\mathcal{C}$ *of size* m*, and let* $\hat{q}$ *be the* $\lceil (1-\alpha)(m+1) \rceil$*-th smallest score. Then the interval* $\widehat{C}(\boldsymbol{x}_{n+1}) = [\hat{\mu}(\boldsymbol{x}_{n+1}) \pm \hat{q}]$ *satisfies*

$$\mathbb{P}(r_{n+1} \in \widehat{C}(\boldsymbol{x}_{n+1})) \;\geq\; 1 - \alpha,$$

exactly, for every fixed n *and every distribution.*

The guarantee in Theorem 11 rests on *exchangeability*, which time series violate: returns drift and their volatility clusters. Two extensions restore approximate validity in exactly our setting.

Theorem 12 (Conformal beyond exchangeability and under drift)**.** *Barber et al. (2023) bound the coverage gap of weighted conformal prediction by a total-variation measure of the data's departure from exchangeability, so that coverage degrades* gracefully *and controllably rather than failing. Gibbs and Candès (2021) propose* adaptive conformal inference*, which updates the nominal level online,* $\alpha_{t+1} = \alpha_t + \eta\,(\alpha - \mathbf{1}\{r_t \notin \widehat{C}_t\})$*, and guarantees the long-run empirical coverage frequency converges to the target* $1 - \alpha$ *under* arbitrary *distribution shift.*

Conformal intervals are the cleanest available answer to the heavy-tail problem for *prediction*: because they are built from the empirical quantiles of realized errors, they inherit the true (heavy) error distribution automatically, with no moment assumptions and no parametric tail model. Pairing an adaptive conformal band with a conditional-EVT or quantile-regression point/scale forecast yields an interval that is simultaneously sharp, heavy-tail-aware, and coverage-valid under drift.

7.4 Evaluating intervals, and the multiplicity tax

A valid interval should also be *sharp* (narrow) and well-*calibrated*. Calibration of a full predictive distribution is assessed by *strictly proper scoring rules* (Gneiting and Raftery, 2007) — the logarithmic score and the continuous ranked probability score (CRPS) — which are minimized in expectation only by the true law and so cannot be gamed; quantile/interval forecasts are scored by the pinball loss (Koenker, 2005). Finally, when many forecasts, signals, or interval procedures are screened, single-comparison coverage and p-values are invalid. The reality-check and superior-predictive-ability tests (White, 2000; Hansen, 2005), the model confidence set (Hansen et al., 2011), the multiple-testing thresholds of Harvey et al. (2016), and the deflated Sharpe ratio (Bailey and López de Prado, 2014) all correct for this selection, and should wrap any search over predictive models. Heavy tails make the multiplicity tax heavier still, since they fatten the null distribution of the maximum statistic.

8 A defensible workflow

The pieces assemble into a single, internally consistent procedure for predictive inference on returns:

1. **Diagnose the design.** Test/justify stationarity and mixing; estimate the tail index $\hat{\alpha}$ (Theorem 8) to learn *which moments exist*; recall that $\hat{\alpha} < 4$ voids fourth-moment-based inference and $\hat{\alpha} < 2$ voids variance-based inference.

2. **Model the conditional location and scale robustly.** Use forecasts whose evaluation does not presume more moments than the data have; prefer quantile/realized-measure methods (§6); bias-correct persistent predictive regressions (Proposition 3).

3. **Quantify uncertainty by resampling that matches the data.** Stationary/block bootstrap when a CLT holds (Theorem 9); subsampling or m-out-of-n when moments are in doubt (Theorem 10).

4. **Issue distribution-free prediction intervals.** Adaptive conformal bands (Theorems 11–12) for valid coverage under heavy tails and drift; score them with proper rules (Gneiting and Raftery, 2007).

5. **Pay the multiplicity tax.** Correct every comparison for the size of the search (White, 2000; Hansen, 2005; Hansen et al., 2011; Harvey et al., 2016; Bailey and López de Prado, 2014).

The workflow is conservative by design: each step assumes *less* than the classical default, and the cost of that conservatism is precisely the protection against the failure modes catalogued above.

9 Open problems

1. **Non-stationarity and breaks.** The theory above is cleanest under stationarity; real return laws drift and break. Limit theory and conformal coverage under *structural change* (beyond the graceful-degradation bounds of Barber et al., 2023) remain incomplete, as does the detection of the breaks themselves in heavy-tailed data.

2. **Inference with infinite higher moments.** Default fourth-moment asymptotics are routinely invoked where $\hat{\alpha} < 4$. A practical, widely-adopted toolkit of moment-free standard errors (subsampling-based) for the statistics analysts actually report is still lacking.

3. **Optimal threshold/anchor selection.** Choosing k in Hill estimation (Theorem 8) or u in POT, and the block length in resampling, are unresolved adaptive-bandwidth problems with finite-sample stakes.

4. **High dimension meets heavy tails.** Machine-learning predictors (Gu et al., 2020) operate in high dimension; their estimation error (§3) and the validity of their uncertainty quantification under regularly varying features and errors are open at the intersection of two literatures.

5. **Conformal under strong dependence.** Sharpening conformal coverage for strongly dependent, volatility-clustered series — rather than treating dependence only as a perturbation of exchangeability — is an active frontier.

6. **Estimation error in decisions.** Propagating heavy-tailed predictive *distributions* (not just point forecasts) and their estimation error into portfolio choice, with valid coverage on the realized objective, closes the loop from inference to action and remains largely open.

10 Conclusion

The difficulty of forecasting returns is, stripped to its mathematics, a problem of estimation under two adversarial conditions: a signal so faint that estimation error dominates predictive risk (Section 3), and a noise distribution so heavy-tailed and dependent that the classical guarantees of consistency, normality, and coverage do not directly apply (Sections 4–5). The state of the art meets these conditions not with a better point forecast but with better *inference*: tail-aware estimation (Section 6), resampling that adapts to the data's true moments, and distribution-free conformal prediction valid under drift (Section 7). The recurring discipline is to assume less — fewer moments, no exchangeability, no single hypothesis — and to let the data, through subsampling and conformal calibration, supply the rates and shapes that theory cannot safely presume. An estimate of a return is worth little without an honest account of its uncertainty; producing that honest account, under the tails and dependence that financial data truly exhibit, is both the harder task and the one that matters.

References

Torben G. Andersen, Tim Bollerslev, Francis X. Diebold, and Paul Labys. Modeling and forecasting realized volatility. *Econometrica*, 71(2):579–625, 2003.

Donald W. K. Andrews. Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 59(3):817–858, 1991.

David H. Bailey and Marcos López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *Journal of Portfolio Management*, 40(5):94–107, 2014.

August A. Balkema and Laurens de Haan. Residual life time at great age. *Annals of Probability*, 2(5):792–804, 1974.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *Annals of Statistics*, 51(2):816–845, 2023.

Ole E. Barndorff-Nielsen and Neil Shephard. Econometrics of testing for jumps in financial economics using bipower variation. *Journal of Financial Econometrics*, 4(1):1–30, 2006.

Patrick Billingsley. *Convergence of Probability Measures*. John Wiley & Sons, New York, 1968.

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327, 1986.

John Y. Campbell and Samuel B. Thompson. Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531, 2008.

Todd E. Clark and Kenneth D. West. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311, 2007.

Rama Cont. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.

Fulvio Corsi. A simple approximate long-memory model of realized volatility. *Journal of Financial Econometrics*, 7(2):174–196, 2009.

Richard A. Davis and Thomas Mikosch. The sample autocorrelations of heavy-tailed processes with applications to arch. *Annals of Statistics*, 26(5):2049–2080, 1998.

Laurens de Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer, New York, 2006.

Francis X. Diebold and Roberto S. Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *Annals of Statistics*, 7(1): 1–26, 1979.

Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling Extremal Events for Insurance and Finance*. Springer, Berlin, 1997.

Robert F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007, 1982.

Eugene F. Fama. The behavior of stock-market prices. *Journal of Business*, 38(1):34–105, 1965.

Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383–417, 1970.

Raffaella Giacomini and Halbert White. Tests of conditional predictive ability. *Econometrica*, 74 (6):1545–1578, 2006.

Isaac Gibbs and Emmanuel Candès. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672, 2021.

Boris V. Gnedenko. Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3):423–453, 1943.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273, 2020.

Peter Hall. On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Society, Series B*, 44(1):37–42, 1982.

Peter Reinhard Hansen. A test for superior predictive ability. *Journal of Business & Economic Statistics*, 23(4):365–380, 2005.

Peter Reinhard Hansen, Asger Lunde, and James M. Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

Campbell R. Harvey, Yan Liu, and Heqing Zhu. . . . and the cross-section of expected returns. *Review of Financial Studies*, 29(1):5–68, 2016.

Bruce M. Hill. A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5):1163–1174, 1975.

Bryan Kelly and Hao Jiang. Tail risk and asset prices. *Review of Financial Studies*, 27(10): 2841–2871, 2014.

Roger Koenker. *Quantile Regression*. Cambridge University Press, Cambridge, 2005.

Hans R. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17(3):1217–1241, 1989.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Benoit Mandelbrot. The variation of certain speculative prices. *Journal of Business*, 36(4): 394–419, 1963.

Alexander J. McNeil and Rüdiger Frey. Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach. *Journal of Empirical Finance*, 7(3–4):271–300, 2000.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

James Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1): 119–131, 1975.

Dimitris N. Politis and Joseph P. Romano. The stationary bootstrap. *Journal of the American Statistical Association*, 89(428):1303–1313, 1994.

Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *Subsampling*. Springer, New York, 1999.

Sidney I. Resnick. *Heavy-Tail Phenomena: Probabilistic and Statistical Modeling*. Springer, New York, 2007.

Robert F. Stambaugh. Predictive regressions. *Journal of Financial Economics*, 54(3):375–421, 1999.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Ivo Welch and Amit Goyal. A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508, 2008.

Kenneth D. West. Asymptotic inference about predictive ability. *Econometrica*, 64(5):1067–1084, 1996.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

Halbert White. *Asymptotic Theory for Econometricians*. Academic Press, San Diego, revised edition, 2001.