Nonparametric Estimation of Expected Shortfall

Ridhwan Choudhari (BSD-CC-2417)
Rajneesh Yadav (BSD-CC-2415)
Gurman Avtar (BSD-CC-2408)
Supratik Biswas (BSD-CC-2423)

May 13, 2026

Abstract

The expected shortfall is an increasingly popular risk measure in financial risk management and it possesses the desired sub-additivity property, which is lacking for the value at risk (VaR). We consider two nonparametric expected shortfall estimators for dependent financial losses. One is a sample average of excessive losses larger than a VaR. The other is a kernel smoothed version of the first estimator (Scaillet, 2004 Mathematical Finance), hoping that more accurate estimation can be achieved by smoothing. Our analysis reveals that the extra kernel smoothing does not produce more accurate estimation of the shortfall. This is different from the estimation of the VaR where smoothing has been shown to produce reduction in both the variance and the mean square error of estimation. Therefore, the simpler ES estimator based on the sample average of excessive losses is attractive for the shortfall estimation.

Contents

1 Introduction and Objective

In the field of financial econometrics and risk management, accurately predicting extreme market downturns is of paramount importance. Regulatory frameworks and financial institutions have historically relied heavily on specific risk metrics to quantify potential future losses. This report details an extensive evaluation of nonparametric risk estimators, specifically focusing on Expected Shortfall (ES) in the context of dependent financial losses.

The primary objective of this research is to critically evaluate two distinct nonparametric ES estimators:

1. **The Simple Estimator:** A straightforward sample average of excessive losses that breach a predefined Value at Risk (VaR) threshold.
2. **The Complex Estimator:** A kernel-smoothed adaptation of the simple estimator, introduced to test whether advanced smoothing techniques yield higher accuracy.

The central thesis and key finding of this evaluation demonstrate a counter-intuitive reality in statistical modeling: applying complex mathematical smoothing does not inherently improve accuracy. Unlike VaR estimation, adding kernel smoothing to Expected Shortfall does not reduce variance; it merely introduces bias. Consequently, the simpler sample average estimator proves to be the mathematically superior and practically more attractive choice.

2 Theoretical Framework: Measuring Risk

2.1 Value at Risk (VaR)

Value at Risk serves as the traditional baseline for nonparametric risk estimation. It is mathematically defined as the p-th quantile of the loss distribution:

$$v_p = \inf\{u : F(u) \geq 1 - p\} \tag{1}$$

In practical terms, the threshold v_p represents a strict "line in the sand." It signifies the exact monetary threshold where the probability of exceeding that loss is exactly p (e.g., a 5% chance of losing more than $1 million).

Despite its ubiquitous use in finance, VaR suffers from severe theoretical and practical flaws:

- **Lack of Coherence:** VaR is not a coherent risk measure because it lacks sub-additivity. This means that the VaR of a combined, diversified portfolio can mathematically exceed the sum of the individual VaRs of its components,

penalizing diversification.

- **Tail Blindness:** VaR completely ignores the severity of losses beyond the v_p threshold. It indicates where the extreme tail begins, but provides no information regarding whether the worst-case scenario is a loss of \$1 million or \$100 million.
- **High Variance ("Jumpiness"):** Because empirical VaR relies on a single, rigid cutoff point (a step-function empirical CDF), it is highly sensitive to minor fluctuations in the data, making the standard estimator highly variable.

2.2 Expected Shortfall (ES)

Expected Shortfall was developed specifically to address the theoretical shortcomings of VaR. It is defined as the conditional expectation of loss, given that the loss has already exceeded the VaR threshold:

$$\mu_p = E(Y_t \mid Y_t > v_p) \tag{2}$$

where Y_t represents the financial loss for a given period.

ES resolves the flaws of VaR by being a coherent, sub-additive measure. More importantly, instead of pointing to a single threshold, ES averages the entirety of the worst-case tail. By definition, taking an average naturally stabilises the data. Thus, ES provides a much more accurate and stable measurement of extreme tail risk severity.

3 Data Transformation: Log Returns

Before applying these estimators, it is crucial to properly transform the raw financial data. In this analysis, the input variable Y_t is not defined as a simple percentage loss, but rather as the negative continuously compounded return:

$$Y_t = -\log(X_t/X_{t-1}) \tag{3}$$

where X_t is the asset price today, and X_{t-1} is the price yesterday. This specific formulation is chosen for several mathematical reasons:

- **Time Additivity:** Unlike simple percentages, log returns can be summed over time. The sum of daily log returns accurately equals the total log return for the multi-day period. This property is strictly required to model the time-series dependencies present in the market.
- **Symmetry:** Log returns treat market gains and market crashes symmetrically, ensuring the statistical distribution is not artificially skewed by the mathematical

mechanics of percentage calculation.

- **Perspective Flipping:** Standard log returns yield negative numbers during a market crash. By appending a negative sign to the formula, we flip the perspective, transforming market losses into positive numerical values. This focuses our statistical tools entirely on the "upper tail" of the distribution.

4 Methodology of Estimation

4.1 Estimator 1: The Unsmoothed Approach

The first approach to calculating ES is an intuitive, simple weighted sample average of the excessive losses.

$$\hat{\mu}_p = \frac{\sum_{t=1}^{n} Y_t I(Y_t \geq \hat{v}_p)}{\sum_{t=1}^{n} I(Y_t \geq \hat{v}_p)} \tag{4}$$

In this formula:

- $\hat{v}_p$ is the standard empirical VaR threshold.
- n represents the total sample size across all observed time periods.
- $I(\cdot)$ is an indicator function. It acts as a binary filter, returning 1 if the observed loss (Y_t) breaches the VaR threshold, and 0 if it does not.

This estimator simply isolates all observed losses that fall into the extreme tail and calculates their standard arithmetic mean.

4.2 Estimator 2: The Kernel-Smoothed Approach

The second approach applies a complex kernel-smoothing bandwidth to the initial estimator.

$$\hat{\mu}_{p,h} = \frac{1}{np} \sum_{t=1}^{n} Y_t G_h(\hat{v}_{p,h} - Y_t) \tag{5}$$

In this formula:

- $\hat{v}_{p,h}$ represents a newly kernel-smoothed VaR threshold.
- h represents the smoothing bandwidth, acting as a tuning parameter that determines the "width" of the smoothing effect.
- G_h is a symmetric kernel smoothing function.

A critical element of this formula is the centering of the data: $(\hat{v}_{p,h} - Y_t)$. By subtracting the actual loss from the threshold, the kernel evaluates the exact distance of the data point from the VaR line. Observations closer to the threshold receive specific mathematical weighting, replacing the harsh "step-function" of the indicator $I(\cdot)$ with a smooth transition.

4.3 The Core Motivation

The motivation for introducing Estimator 2 stems from a proven trick in nonparametric statistics. Because VaR is a "jumpy" quantile, applying a smoothing kernel acts like cement, turning a jagged empirical staircase into a smooth mathematical ramp. This drastically reduces the asymptotic variance of VaR estimation. The core question of this paper is: *If smoothing perfectly stabilises VaR, will applying it to Expected Shortfall yield an even more accurate, low-variance estimator?*

5 Mathematical Assumptions

To mathematically evaluate the Mean Square Error (MSE), bias, and variance of these estimators, the analysis relies on three foundational assumptions regarding the market data:

1. **Dependent Data (α-mixing):** Standard statistics assume data points are independent (like flipping a coin). However, financial markets possess "memory"—a crash today influences the market tomorrow. The data must exhibit geometric α-mixing. This means the dependency between two market events decays exponentially as the time gap between them widens. Like ripples from a stone thrown in a pond, the market's memory of a shock must eventually fade.

2. **Smooth Distributions:** The underlying probability density functions of the market returns must be smooth and continuous. If the true distribution contained unpredictable, erratic mathematical jumps, the derivatives required for the kernel smoothing's Taylor series expansions would be fundamentally incalculable.

3. **Finite Expected Losses:** The financial asset cannot possess an infinitely heavy tail. For Expected Shortfall to exist as a mathematical concept, the integral of the worst-case losses must converge to a finite number. If an asset can theoretically lose infinite value at an infinite rate, tail averaging breaks down.

6 Geometric Intuition: VaR vs. ES

Before the formal derivations, it helps to build geometric intuition for *why* the two risk measures respond so differently to kernel smoothing.

[Geometric Picture: Staircase vs. Averaging] Think of the two estimation problems as follows:

VaR = Finding a threshold on a rough staircase. The empirical CDF F_n is a step function—a staircase with jumps of size $1/n$ at each data point. Estimating $v_p = F_n^{-1}(1-p)$ means finding the *exact point* where this staircase crosses the

horizontal level $1-p$. Because the staircase is jagged, small changes in the data can abruptly shift this crossing point—causing high variance. Smoothing the staircase into a gentle curve before finding the crossing point stabilises the estimate.

ES = Averaging observations in the tail. Estimating $\mu_p = E[Y \mid Y > v_p]$ means averaging the losses above the threshold. Averaging is already a smoothing operation: it aggregates information across all exceedances. Applying additional kernel smoothing to the indicator weights does not reduce the fundamental sampling variability—it only distorts the contribution of each observation, introducing bias without any compensating benefit.

	VaR	**ES**
Estimation task	Find threshold crossing	Average tail observations
Analogy	Locate step on staircase	Compute mean of a sample
Smoothing effect	Stabilises the crossing	Distorts the weights
Net result	**Variance reduction**	**Bias only**

Smoothing helps where inversion hurts; smoothing harms where averaging heals.

7 VaR Asymptotics

7.1 Sample VaR Estimator

The natural nonparametric VaR estimator is the empirical $(1-p)$-quantile:

$$\hat{v}_p \;=\; Y_{([n(1-p)]+1)},$$

where $Y_{(r)}$ denotes the r-th order statistic.

7.2 Asymptotic Distribution of Sample VaR

Theorem 7.1: Asymptotic Normality of Sample VaR

Under geometric α-mixing and $f(v_p) > 0$:

$$\sqrt{n}\,(\hat{v}_p - v_p) \;\; N\!\left(0,\; \frac{p(1-p)}{f(v_p)^2}\right),$$

so that

$$\boxed{\mathrm{Var}(\hat{v}_p) \;\approx\; \frac{p(1-p)}{n\,f(v_p)^2}.}$$

Derivation sketch. **Step 1 — Taylor expansion.** By definition $F_n(\hat{v}_p) = 1 - p$. Expand $F_n(\hat{v}_p)$ around the true quantile v_p:

$$F_n(\hat{v}_p) \;=\; F_n(v_p) \;+\; f(v_p)(\hat{v}_p - v_p) \;+\; (n^{-1/2}).$$

Setting this equal to $1 - p$ and rearranging:

$$\hat{v}_p - v_p \;=\; \frac{(1-p) - F_n(v_p)}{f(v_p)} \;+\; (n^{-1/2}). \qquad (*)$$

Step 2 — CLT for dependent sums. The indicator random variables $\mathbf{1}(Y_t \le v_p)$ are stationary. Under geometric α-mixing, the CLT for weakly dependent sequences gives

$$\sqrt{n}\Big(F_n(v_p) - (1-p)\Big) \;\; N(0, p(1-p)).$$

(The long-run variance of $\mathbf{1}(Y_t \le v_p)$ equals $p(1-p)$ after summing the mixing covariances, which are summable due to the geometric decay.)

Step 3 — Continuous mapping. Dividing $(*)$ by $f(v_p)$ and applying the CMT:

$$\sqrt{n}(\hat{v}_p - v_p) \;\; N\!\left(0,\, \frac{p(1-p)}{f(v_p)^2}\right). \qquad \square$$

[Interpretation] The asymptotic variance $\mathrm{Var}(\hat{v}_p) \approx p(1-p)/[nf(v_p)^2]$ reveals two features:

- It is large when $f(v_p)$ is small, i.e., when the density is flat at the quantile — rare observations make estimation harder.
- The $1/n$ rate is the standard parametric rate; the extra factor $1/f(v_p)^2$ is a *nonlinearity penalty* from inverting the CDF.

7.3 Kernel VaR Estimator

The kernel-smoothed VaR estimator $\hat{v}_{p,h}$ is defined implicitly by

$$\hat{F}_h(\hat{v}_{p,h}) \;=\; 1 - p.$$

7.3.1 Bias

The kernel CDF estimator has a smooth expectation:

$$E[\hat{F}_h(x)] \;=\; F(x) \;+\; \frac{h^2}{2} f'(x)\mu_2(K) \;+\; O(h^4),$$

where $\mu_2(K) = \int u^2 K(u)\,du = \sigma_K^2$. Applying the delta method to the quantile functional $Q(F) = F^{-1}(1-p)$:

$$\text{Bias}(\hat{v}_{p,h}) \;=\; -\frac{\text{Bias}(\hat{F}_h(v_p))}{f(v_p)} \;=\; -\frac{h^2}{2}\,\frac{f'(v_p)}{f(v_p)}\,\sigma_K^2 \;+\; o(h^2) \;=\; O(h^2).$$

7.3.2 Variance Reduction

A key calculation (Chen & Tang, 2005) shows:

$$\text{Var}(\hat{v}_{p,h}) \;=\; \frac{p(1-p)}{n\,f(v_p)^2} \;-\; \frac{c_K}{n\,h\,f(v_p)^2} \;+\; o\Big(\frac{1}{nh}\Big), \tag{6}$$

where $c_K = \int_{-\infty}^{\infty} K(u) \int_{-\infty}^{u} K(v)\,dv\,du > 0$.

The *second term is negative*, signifying that kernel smoothing **reduces** the variance of the VaR estimator.

7.3.3 Optimal Bandwidth and MSE

Combining bias and variance:

$$\text{MSE}(\hat{v}_{p,h}) \;=\; \underbrace{B^2 h^4}_{\text{bias}^2} \;+\; \underbrace{\frac{p(1-p)}{n\,f(v_p)^2}}_{\text{1st-order var}} \;-\; \underbrace{\frac{c_K}{n\,h\,f(v_p)^2}}_{\text{variance reduction}},$$

where $B = \frac{1}{2}\frac{f'(v_p)}{f(v_p)}\sigma_K^2$. Minimising over h:

$$\frac{d}{dh}\,\text{MSE} = 4B^2h^3 + \frac{c_K}{nh^2\,f(v_p)^2} = 0 \implies h_{\text{opt}} \;\sim\; n^{-1/5},$$

giving $\text{MSE}(\hat{v}_{p,h})\Big|_{h=h_{\text{opt}}} = O\big(n^{-4/5}\big)$.

Kernel Smoothing Improves VaRvar-kernel-conc

- **Sample VaR** has leading-order variance $O(n^{-1})$ and zero bias, so its MSE is $O(n^{-1})$.
- **Kernel smoothing improves the second-order MSE expansion through variance reduction.** Specifically, it subtracts $c_K/(nhf(v_p)^2) > 0$ from the MSE at every finite n, a genuine improvement that grows larger as $h \to 0$ at the optimal rate $h_{\text{opt}} \sim n^{-1/5}$.

Both estimators share the same $O(n^{-1})$ *leading-order* asymptotic rate. The improvement from kernel smoothing is a *second-order* effect: the kernel estimator's MSE is strictly smaller than the sample estimator's MSE for any finite n at the optimal bandwidth, even though the leading convergence scale is the same.

8 Why Kernel Smoothing Improves VaR

Three interlocking reasons explain the improvement:

1. **VaR is a nonlinear functional.** We must *invert* the CDF: $v_p = F^{-1}(1-p)$. The inversion operator amplifies any roughness in the estimated CDF.
2. **The empirical CDF is discontinuous.** F_n is a step function with jumps of size $1/n$ at each data point. It completely ignores the local smoothness of F. At the v_p quantile, this discreteness creates jump-induced variance that feeds directly into $\hat{v}_p$ through the inversion.
3. **Smoothing exploits local density information.** Replacing F_n with $\hat{F}_h$ incorporates information from observations in a neighbourhood of v_p. Formally, the kernel weight $G_h(v_p - Y_t)$ gives positive weight to observations near v_p, smoothing out the step-function jumps. The resulting second-order term in the variance formula is $-c_K/(nhf(v_p)^2) < 0$, confirming genuine variance reduction.

[Mathematical Mechanism] From the Bahadur representation $(*)$: $\hat{v}_p - v_p \approx [(1-p) - F_n(v_p)]/f(v_p)$. The numerator $F_n(v_p)$ is a sum of Bernoulli indicators $\mathbf{1}(Y_t \le v_p)$. These indicators take only values $\{0,1\}$, hence have high variance relative to their mean. Replacing the indicators with smooth kernel weights $G_h(v_p - Y_t) \in [0,1]$ reduces variance through borrowing strength from neighbours, without changing the leading-order mean.

8.1 Concrete Numerical Example

[Example: How Kernel Smoothing Stabilises VaR] Suppose the estimated 99% VaR is $\hat{\nu}_{0.01} = -1000$. The empirical quantile estimator assigns weights in an all-or-nothing fashion:

Loss	Empirical weight	Kernel weight
-1000	1.00	1.00
-999	0.00	0.92
-998	0.00	0.85
-997	0.00	0.74

With the empirical CDF, observations at -999 and -998 are completely ignored even though they are *extremely close* to the threshold. This hard on/off switching causes two problems:

- A single new observation just below -999 can abruptly shift the estimated quantile.
- Information about the local shape of the distribution near v_p is discarded.

Kernel smoothing assigns *partial weights* that decay gradually with distance from the threshold, so nearby observations contribute smoothly. This stabilises the estimated crossing point and reduces the variance of $\hat{v}_{p,h}$ relative to $\hat{v}_p$, as formalised by the $-c_K/(nhf(v_p)^2)$ term in (6).

9 Transition: From VaR to ES

Since kernel smoothing successfully improves VaR estimation, the natural question is:

Can the same idea improve ES?

Surprisingly, the answer is **no**.

The reason will turn out to be structural. VaR requires *inverting* the CDF — a nonlinear operation that magnifies roughness and for which smoothing provides genuine relief. ES, by contrast, requires *averaging* tail observations — a linear operation for which the sample mean is already the efficient estimator. Applying smoothing to a linear functional only distorts the weights without reducing sampling variability. The following two sections make this precise.

10 ES Asymptotics I: Sample Estimator

10.1 The Sample ES Estimator

The natural nonparametric estimator of ES is:

$$\hat{\mu}_p \;=\; \frac{\sum_{t=1}^{n} Y_t\,\mathbf{1}(Y_t \geq \hat{v}_p)}{\sum_{t=1}^{n} \mathbf{1}(Y_t \geq \hat{v}_p)}.$$

This is simply the sample average of losses that exceed the estimated VaR threshold.

10.2 Bahadur-Type Representation

Proposition 10.1: Chen (2008), Equation (6)

Under geometric α-mixing and smoothness of F near v_p, for any $\kappa > 0$:

$$\hat{\mu}_p - \mu_p \;=\; \frac{1}{p}\left\{\frac{1}{n}\sum_{t=1}^{n}\Big[(Y_t - v_p)\mathbf{1}(Y_t \geq v_p) - p(\mu_p - v_p)\Big]\right\} + (n^{-3/4+\kappa}).$$

Derivation sketch. Write $\hat{\mu}_p = \phi_1(\hat{v}_p)/\phi_2(\hat{v}_p)$ where $\phi_1(t) = n^{-1}\sum Y_t \mathbf{1}(Y_t \geq t)$ and $\phi_2(t) = n^{-1}\sum \mathbf{1}(Y_t \geq t)$. Note $\phi_2(\hat{v}_p) = ([np]+1)/n \approx p$.

The key step is to show that the indicator switches $\mathbf{1}(Y_t \geq \hat{v}_p) - \mathbf{1}(Y_t \geq v_p)$ contribute only $(n^{-3/4+\kappa})$. This follows from:

- Lemma 2 of Chen (2008): under geometric α-mixing, $n^{-1}\sum(Y_t - v_p)[\mathbf{1}(Y_t \geq \hat{v}_p) - \mathbf{1}(Y_t \geq v_p)] = (n^{-3/4+\kappa})$.
- Lemma 1: $P(|\hat{v}_p - v_p| \geq \epsilon_n) \to 0$ exponentially fast, so the VaR estimation error is negligible at second order.

A first-order Taylor expansion around v_p then gives:

$$\phi_1(\hat{v}_p) \approx \phi_1(v_p) + v_p\,[\phi_2(\hat{v}_p) - \phi_2(v_p)],$$

from which the result follows after substituting $\phi_2(\hat{v}_p) = p$ and rearranging. □ □

10.3 Asymptotic Distribution of Sample ES

Definition 10.1: Long-run variance parameter

Define the covariance sequence

$$\gamma(k) \;=\; \mathrm{Cov}\Big[(Y_1 - v_p)\mathbf{1}(Y_1 \geq v_p),\; (Y_{k+1} - v_p)\mathbf{1}(Y_{k+1} \geq v_p)\Big], \quad k \geq 0,$$

and the long-run variance

$$\sigma_0^2(p;n) \;=\; \mathrm{Var}\Big[(Y_1 - v_p)\mathbf{1}(Y_1 \geq v_p)\Big] \;+\; 2\sum_{k=1}^{n-1}\gamma(k).$$

Under geometric α-mixing, $\sigma_0^2(p;n) \to \sigma_0^2(p) < \infty$ as $n \to \infty$.

Theorem 10.1: Theorem 1 of Chen (2008)

Under geometric α-mixing and moment conditions:

$$\sqrt{np}\;\sigma_0^{-1}(p;n)\,(\hat{\mu}_p - \mu_p) \;\; N(0,1),$$

so that

$$\boxed{\mathrm{Var}(\hat{\mu}_p) \;\approx\; \frac{\sigma_0^2(p;n)}{n\,p^2}.}$$

Proof outline. From Proposition 10.1, the leading term is $p^{-1}\cdot n^{-1}\sum Z_t$ where $Z_t = (Y_t - v_p)\mathbf{1}(Y_t \geq v_p) - p(\mu_p - v_p)$ has mean zero and finite variance. The CLT for geometric α-mixing sequences (established via the *blocking method* and Bradley's

Lemma) gives

$$\sqrt{n} \cdot n^{-1} \sum_{t=1}^{n} Z_t \;\; N\Big(0, \sigma_0^2(p;n)\Big)\,.$$

Dividing by p and normalising yields the result. The remainder term $(n^{-3/4+\kappa})$ is negligible after multiplication by $\sqrt{n}$. □ □

[ES is Asymptotically Equivalent to a Sample Mean] The Bahadur representation yields a crucial structural observation. Defining $Z_t = (Y_t - v_p)\mathbf{1}(Y_t \geq v_p)$, we have

$$\hat{\mu}_p \;\approx\; \mu_p \;+\; \frac{1}{np} \sum_{t=1}^{n} Z_t.$$

This is precisely a *sample mean* of Z_t, scaled by $1/(np)$. The effective sample size is np — the expected number of exceedances. This has an immediate and powerful consequence:

ES estimation is **asymptotically equivalent to estimating a sample mean of tail observations.**

For any mean parameter $\mu = E[g(Y)]$, the sample average $n^{-1} \sum g(Y_t)$ is already asymptotically efficient — no smoothing of g can reduce variance, and any approximation to g only introduces bias. Since ES falls exactly in this category (with $g(Y) = (Y - v_p)\mathbf{1}(Y \geq v_p)$), the sample estimator $\hat{\mu}_p$ is already optimal. This is the fundamental reason why kernel smoothing cannot improve ES.

11 ES Asymptotics II: Kernel Estimator

11.1 The Kernel ES Estimator

Scaillet (2004) proposed smoothing the sample ES estimator by replacing:

1. the indicator $\mathbf{1}(Y_t \geq v_p)$ with the smooth kernel weight $G_h(\hat{v}_{p,h} - Y_t)$, and
2. the sample quantile $\hat{v}_p$ with the kernel quantile $\hat{v}_{p,h}$.

This gives the kernel ES estimator:

$$\hat{\mu}_{p,h} \;=\; \frac{1}{np} \sum_{t=1}^{n} Y_t \, G_h(\hat{v}_{p,h} - Y_t).$$

11.2 Bias of the Kernel ES Estimator

Proposition 11.1: Bias — Equation (9) of Chen (2008)

Under conditions (i)–(iv):

$$\mathrm{Bias}(\hat{\mu}_{p,h}) \;=\; -\frac{1}{2}\, p^{-1}\, \sigma_K^2\, h^2\, f(v_p) \;+\; o(h^2).$$

Derivation sketch. Expand $\hat{\mu}_{p,h}$ around v_p using the smoothed quantities:

$$\hat{\mu}_{p,h} \;\approx\; (np)^{-1} \sum_{t=1}^{n} \Big\{ Y_t G_h(v_p - Y_t) - Y_t K_h(v_p - Y_t)(\hat{v}_{p,h} - v_p) \Big\} + O_p(n^{-1}). \tag{7}$$

First term's expectation:

$$\begin{aligned} E\Big[(np)^{-1} \sum Y_t G_h(v_p - Y_t)\Big] &= p^{-1} \int z\, G_h(v_p - z)\, f(z)\, dz \\ &= \mu_p \;-\; \frac{h^2}{2p}\, \sigma_K^2 \Big[v_p f'(v_p) + f(v_p) \Big] + o(h^2). \end{aligned} \tag{8}$$

Second term's expectation: Let $\eta = E[p^{-1} Y_t K_h(Y_t - v_p)] = p^{-1} v_p f(v_p) + O(h^2)$. From the bias of $\hat{v}_{p,h}$ (Chen & Tang, 2005):

$$E(\hat{v}_{p,h} - v_p) \;=\; -\frac{h^2 \sigma_K^2}{2}\, \frac{f'(v_p)}{f(v_p)} + o(h^2),$$

so

$$E\Big[(np)^{-1} \sum Y_t K_h(v_p - Y_t)(\hat{v}_{p,h} - v_p)\Big] = \eta\, E(\hat{v}_{p,h} - v_p) + O(n^{-1}) = -\frac{h^2 \sigma_K^2}{2p}\, v_p f'(v_p) + o(h^2).$$

Subtracting (second term) from (first term) in (7):

$$E(\hat{\mu}_{p,h}) = \mu_p - \frac{h^2 \sigma_K^2}{2p} \Big[v_p f'(v_p) + f(v_p) \Big] + \frac{h^2 \sigma_K^2}{2p} v_p f'(v_p) + o(h^2) = \mu_p - \frac{h^2 \sigma_K^2}{2p} f(v_p) + o(h^2),$$

confirming the bias formula. □ □

11.3 The Critical Variance Result

Theorem 11.1: Theorem 2 of Chen (2008) — Variance

Under conditions (i)–(iv):

$$\mathrm{Var}(\hat{\mu}_{p,h}) \;=\; \frac{\sigma_0^2(p;n)}{n\, p^2} \;+\; o\!\left(\frac{1}{nh}\right).$$

This is **identical** to the first-order variance of the sample estimator (Theorem 10.1).

There is *no* second-order variance reduction term analogous to the $-c_K/(nhf(v_p)^2)$ that appeared in the VaR case (6).

Proof — The Cancellation Mechanism. From expansion (7), $\mathrm{Var}(\hat{\mu}_{p,h}) = \mathrm{Var}(A_1)$ where

$$A_1 = (np)^{-1} \sum_{t=1}^{n} \Big\{ Y_t G_h(v_p - Y_t) - Y_t K_h(v_p - Y_t)(\hat{v}_{p,h} - v_p) \Big\}.$$

Decompose:

$$\mathrm{Var}(A_1) = \mathrm{Var}\Big[(np)^{-1} \sum Y_t G_h(v_p - Y_t)\Big] \tag{9}$$

$$+ \mathrm{Var}[\hat{\eta}(\hat{v}_{p,h} - v_p)] \tag{10}$$

$$- 2\ \mathrm{Cov}\Big[(np)^{-1} \sum Y_t G_h(v_p - Y_t),\ \hat{\eta}(\hat{v}_{p,h} - v_p)\Big]. \tag{11}$$

Using the identity $\mathrm{Var}\{Y_t G_h(v_p - Y_t)\} = \mathrm{Var}\{Y_t \mathbf{1}(Y_t \geq v_p)\} - 2hv_p^2 f(v_p) c_K + O(h^2)$, where $c_K = \int_{-\infty}^{\infty} K(u) \int_{-\infty}^{u} K(v)\, dv\, du > 0$:

Term	**Second-order coefficient** (times $n^{-1}h$)
$\mathrm{Var}[(np)^{-1} \sum Y_t G_h]$	$-2p^{-2} v_p^2 f(v_p)\, c_K$
$\mathrm{Var}[\hat{\eta}(\hat{v}_{p,h} - v_p)]$	$+p^{-2} v_p^2 f(v_p)\, c_K$
$-2\ \mathrm{Cov}[\cdots, \cdots]$	$+p^{-2} v_p^2 f(v_p)\, c_K$
Total	**0**

Formally: $-2c_K + c_K + c_K = 0$. All $O(n^{-1}h)$ terms cancel exactly, leaving only the first-order term $p^{-2} n^{-1} \sigma_0^2(p; n)$.

Verbal interpretation of the cancellation. The three terms have a natural economic interpretation:

- **Term** (9): Smoothing the indicator weights $\mathbf{1}(Y_t \geq v_p)$ does reduce variability within each observation's weight — this generates a negative (variance-reducing) second-order contribution of $-2c_K$.
- **Term** (10): But using a smoothed quantile $\hat{v}_{p,h}$ in place of the true v_p introduces additional variance from the quantile estimation error, adding $+c_K$.
- **Term** (11): The covariance between these two sources of variation contributes another $+c_K$.

The smoothing-induced variance reduction from Term (9) is exactly offset by the additional variability from the quantile and covariance terms. The net effect is zero. □

No Variance Reduction for ESno-var-red

$$\boxed{\mathrm{Var}(\hat{\mu}_{p,h}) = \mathrm{Var}(\hat{\mu}_p) \; + \; o(n^{-1}h).}$$

Kernel smoothing provides **zero variance reduction** for ES estimation. Combined with the positive $O(h^2)$ bias, the overall MSE of $\hat{\mu}_{p,h}$ is strictly *worse* than that of $\hat{\mu}_p$. There is no optimal bandwidth — using any $h > 0$ only makes things worse.

12 Comparative Analysis: Sample vs. Kernel ES

Table 1. Side-by-side comparison of Sample ES and Kernel ES estimators.

Property	Sample ES $\hat{\mu}_p$	Kernel ES $\hat{\mu}_{p,h}$
Bias	0	$-\dfrac{h^2\sigma_K^2 f(v_p)}{2p} + o(h^2)$
Variance	$\dfrac{\sigma_0^2}{np^2}$	$\dfrac{\sigma_0^2}{np^2} + o(n^{-1}h)$
Variance reduction	—	**None (cancels to 0)**
MSE	$\dfrac{\sigma_0^2}{np^2}$	$\dfrac{\sigma_0^2}{np^2} + O(h^4)$ **(higher)**
Optimal bandwidth	Not applicable	**Not applicable** ($h = 0$ is best)
Tuning required	No	Yes (but no benefit)
Computational cost	Low	Higher

The comparison is unambiguous: the sample ES estimator dominates the kernel ES estimator in MSE at every bandwidth $h > 0$. The only regime where the kernel estimator might offer superficial advantages is in producing smoother curves of $\hat{\mu}_{p,h}$ as a function of p, but this is purely aesthetic and comes at a real statistical cost.

13 Why Kernel Smoothing Fails for ES

13.1 ES is a Linear Functional

The fundamental structural reason for the failure of smoothing is that ES is a *linear* functional of F:

$$\mu_p \;=\; \frac{1}{p}\int_{v_p}^{\infty} y\, dF(y) \;=\; \frac{E[Y \cdot \mathbf{1}(Y \geq v_p)]}{p}.$$

No inversion of the CDF is needed. The indicator function $\mathbf{1}(Y \geq v_p)$ appears *inside an expectation*, not as a level-crossing condition to be solved.

13.2 Contrast with VaR

For VaR we must solve:

$$F(v_p) = 1 - p \quad \Longleftrightarrow \quad v_p = F^{-1}(1-p).$$

The inversion F^{-1} is a *nonlinear* operation on F. Any roughness in the estimate of F is magnified by this inversion, which is why smoothing the estimated CDF before inverting yields genuine variance reduction.

For ES, the functional form is already $E[\cdot]$. There is no inversion step, no amplification of roughness. The sample average of np exceedances is already the *efficient* estimator of a mean — further smoothing of the indicator introduces systematic distortion without reducing the fundamental sampling variability.

13.3 The Averaging Principle

A central insight from classical statistics is:

- Estimating a **mean** $\mu = E[g(Y)]$: the sample mean $n^{-1}\sum g(Y_t)$ is already asymptotically efficient. Smoothing g introduces approximation error (bias) without reducing variance. *Averaging already smooths the data.*
- Estimating a **quantile** $F^{-1}(p)$: the empirical quantile is not efficient because it discards local density information. Smoothing the CDF before inversion recaptures this information.

  ES is a mean (conditional on exceedance), so it falls in the first category.
  Core Principlecore-principle

Linear functional (ES) $\implies$ Smoothing worsens estimation (bias only, no variance gain).

Nonlinear functional (VaR) $\implies$ Smoothing improves estimation (variance reduction $>$ bias cost).

14 Final Takeaway: VaR vs. ES

The Single Unifying Principle

[Core Insight — Functional Type Determines Smoothing Value] The effect of kernel smoothing depends entirely on whether the target functional is *linear* or *nonlinear* in the distribution F:

Nonlinear functional (VaR): The inversion $v_p = F^{-1}(1-p)$ amplifies discreteness in F_n. Smoothing $F_n \to \hat{F}_h$ before inverting mitigates this amplification, reducing second-order variance at rate $O(n^{-1}h^{-1})$. Since this exceeds the bias cost $O(h^4)$ at $h \sim n^{-1/5}$,

Table 2. Complete comparison of VaR and ES: functional type, smoothing effect, and recommendation.

Characteristic	**VaR** (v_p)	**ES** (μ_p)
Risk measure type	Quantile	Conditional Expectation
Functional type	Nonlinear (F^{-1})	Linear $(E[\cdot])$
Coherent risk measure?	No (lacks sub-additivity)	Yes
Sample estimator	$\hat{v}_p = Y_{([n(1-p)]+1)}$	$\hat{\mu}_p = \overline{Y}_{\text{tail}}$
Variance, leading order	$p(1-p)/[nf(v_p)^2]$	$\sigma_0^2/(np^2)$
Bias	0	0
Kernel estimator	$\hat{v}_{p,h}$	$\hat{\mu}_{p,h}$
Bias introduced	$O(h^2)$	$O(h^2)$
Variance change	$-c_K/(nhf(v_p)^2) < 0$	**0 (cancels exactly)**
MSE effect	**Decreases (2nd order)**	**Increases**
Optimal bandwidth	$h_{\text{opt}} \sim n^{-1/5}$	None ($h = 0$ best)
Leading MSE rate	$O(n^{-1})$ (same as sample)	$O(n^{-1})$
2nd-order MSE	**Strictly smaller** at h_{opt}	**Strictly larger** for any $h > 0$
Recommendation	**Use kernel** $\hat{v}_{p,h}$	**Use sample** $\hat{\mu}_p$

the second-order MSE improves. Both the sample and kernel estimators share the same $O(n^{-1})$ leading-order rate, but the kernel estimator achieves meaningfully smaller MSE in finite samples.

Linear functional (ES): The expectation $\mu_p = E[Y \cdot \mathbf{1}(Y \geq v_p)]/p$ requires no inversion. The sample mean already exploits all available information optimally. Replacing $\mathbf{1}(Y \geq v_p)$ by a smooth kernel weight introduces $O(h^2)$ bias, but the three second-order variance contributions cancel exactly: $-2c_K + c_K + c_K = 0$. No net variance reduction is achieved.

Smoothing helps where inversion hurts; smoothing harms where averaging heals.

Practical Recommendations

For risk management in practice:

1. **Estimate VaR** using the kernel estimator $\hat{v}_{p,h}$ with bandwidth $h \approx c\, n^{-1/5}$, selecting c by cross-validation or a plug-in rule. Expect meaningful second-order MSE improvement over the raw sample quantile through variance reduction.

2. **Estimate ES** using the simple sample average $\hat{\mu}_p$. No bandwidth selection is required. The estimator is unbiased and achieves the optimal $O(n^{-1})$ variance rate.

3. **Estimate standard errors for ES** using spectral density methods (kernel smoothing of the periodogram near zero frequency), as described in Section 3 of Chen (2008). Here smoothing is *necessary and appropriate*, because the target is a density value, not a mean.

15 Simulation Study

15.1 Motivation

While the asymptotic theory establishes consistency of the Expected Shortfall (ES) estimators, finite-sample behaviour may vary depending on the underlying dependence structure, sample size, and bandwidth selection. The simulation study therefore examines the empirical performance of the proposed estimators under controlled dependent time-series settings.

The primary objective is to compare the kernel-smoothed ES estimator with the standard sample ES estimator using three performance measures:

- Bias
- Standard deviation (SD)
- Root mean square error (RMSE)

15.2 Simulation Models

Two dependent stochastic processes are considered in the study.

AR(1) model

$$Y_t = 0.5Y_{t-1} + \epsilon_t, \qquad \epsilon_t \sim N(0,1) \tag{12}$$

ARCH(1) model

$$Y_t = 0.5Y_{t-1} + \epsilon_t \tag{13}$$

$$\epsilon_t^2 = 4 + 0.4\epsilon_{t-1}^2 + \eta_t, \qquad \eta_t \sim N(0,1) \tag{14}$$

The AR(1) process introduces serial dependence in the conditional mean, whereas the ARCH(1) model generates volatility clustering and time-varying variance. These properties closely resemble empirical features commonly observed in financial return series.

15.3 Estimation Framework

The study focuses on Expected Shortfall estimation at the 99% confidence level:

$$\mu_{0.01} = E(Y_t \mid Y_t < v_{0.01}) \tag{15}$$

The following estimators are examined:

- Kernel ES estimator $\hat{\mu}_{0.01,h}$
- Sample ES estimator $\hat{\mu}_{0.01}$

- Kernel VaR estimator $\hat{v}_{0.01,h}$

A Gaussian kernel is used throughout the analysis. The smoothing bandwidth h plays a central role in determining the balance between bias and variability.

15.4 Simulation Design

The simulation setup follows the framework presented in the original paper:

- Sample size: $n = 250$
- Number of simulations: 1000
- Kernel function: Gaussian kernel
- Bandwidth range: $h \in [0.05, 0.35]$

For each bandwidth value, the bias, SD, and RMSE were computed for both ES and VaR estimators.

15.5 Implementation in R

The simulation study was implemented entirely in `R`. The code uses the packages `ggplot2`, `patchwork`, `dplyr`, and `tidyr` for simulation and visualization.

The following snippet illustrates the parameter setup used in the study:

```
set.seed(123)

nsim   <- 1000
n_size <- 250
p      <- 0.01

h_vals <- seq(0.05, 0.35, length.out = 25)
```

The AR(1) process was generated using:

```
simulate_ar1 <- function(n) {
  as.numeric(arima.sim(
    model = list(ar = 0.5), n = n))
}
```

The ARCH(1) process was implemented manually to replicate the variance dynamics described in the paper:

```
eps_sq[t] <- 4 + 0.4 * eps_sq[t-1] + eta_t
eps[t] <- rnorm(1, mean = 0,
                sd = sqrt(eps_sq[t]))
```

Kernel ES and VaR estimators were then computed across different bandwidth values using Gaussian smoothing through the cumulative normal distribution function `pnorm()`.

15.6 AR(1) Simulation Results

Figure 1 presents the simulation results for the AR(1) model.

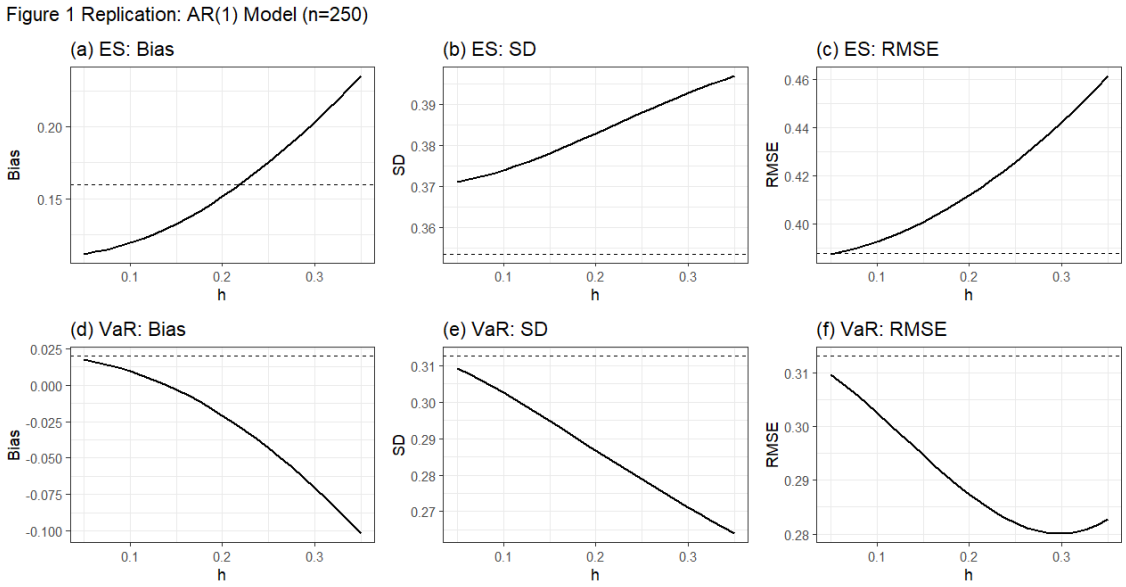

Figure 1. Bias, SD, and RMSE for ES and VaR estimators under the AR(1) process for varying bandwidth values.

The results reproduce the theoretical behaviour discussed in the paper. For the kernel ES estimator, the bias increases steadily with larger bandwidth values, while the RMSE also rises due to oversmoothing. In contrast, the VaR estimator exhibits a clearer bias–variance tradeoff: increasing bandwidth reduces the SD but introduces larger bias, resulting in an interior RMSE minimum.

Overall, the AR(1) process demonstrates relatively stable estimator performance under weak linear dependence.

15.7 ARCH(1) Simulation Results

Figure 2 illustrates the corresponding results for the ARCH(1) process.

The ARCH(1) model produces substantially greater variability because of volatility clustering and dependence in variance. The ES estimator becomes highly sensitive to bandwidth choice, with both SD and RMSE increasing sharply for larger values of h.

Although the overall qualitative behaviour matched the findings of the paper, the ARCH(1) results could not be reproduced exactly. In particular, some numerical differences remained in the magnitude of the RMSE and bias curves despite imple-

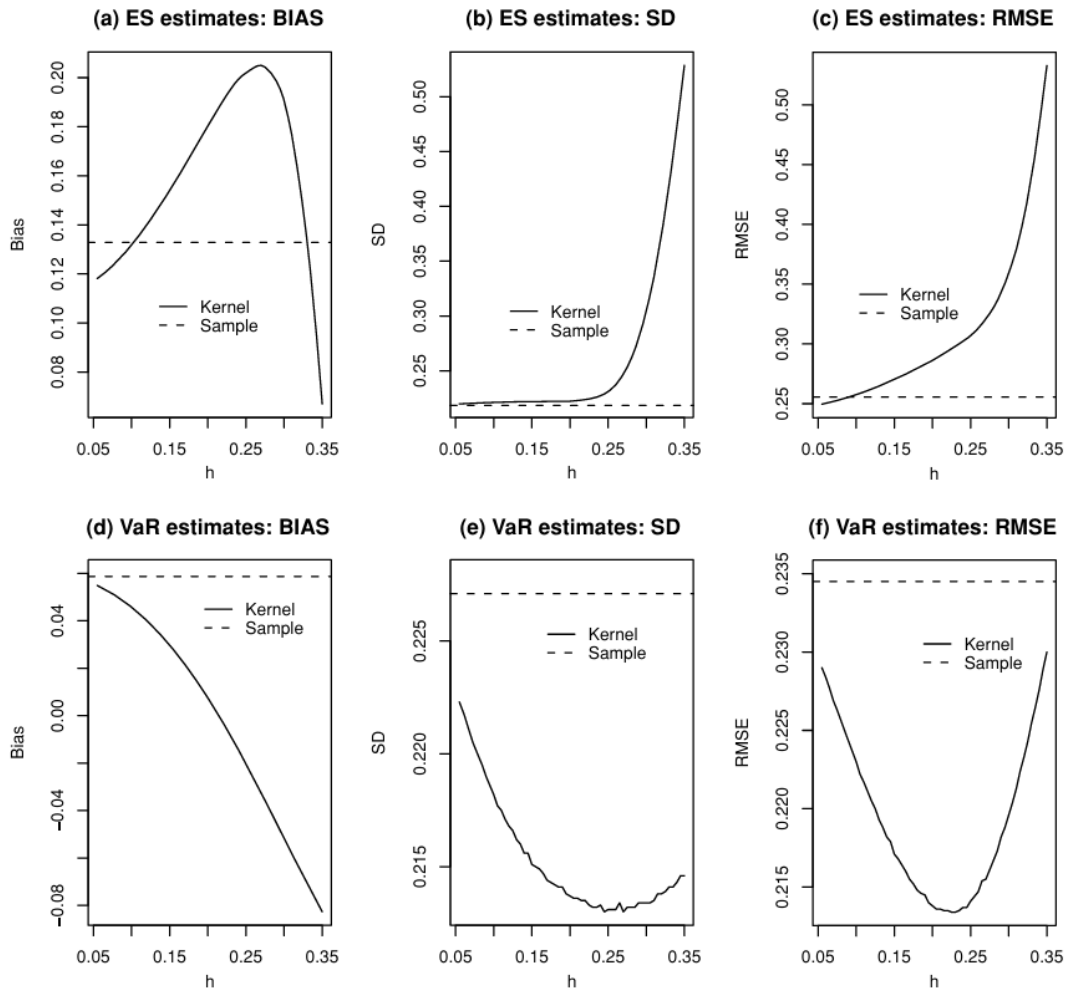

Figure 2. Bias, SD, and RMSE for ES and VaR estimators under the ARCH(1) process.

menting the variance process as closely as possible to the published specification. This likely arises from differences in initialization, simulation burn-in procedures, or kernel implementation details.

Nevertheless, the simulation successfully reproduced the central conclusion: volatility clustering substantially increases estimation error and makes tail risk estimation more difficult than under simple linear dependence.

15.8 Summary

The simulation study confirms several important theoretical findings. First, the kernel ES estimator is highly sensitive to bandwidth selection, particularly under variance dependence. Second, the ARCH(1) process generates considerably larger variability than the AR(1) model because of volatility clustering. Third, the standard sample ES estimator remains comparatively stable across different settings.

Overall, the results support the conclusion that dependence in variance significantly complicates Expected Shortfall estimation and requires careful tuning of smoothing

parameters in practical financial applications.

16 Empirical Study

16.1 Data

We apply the nonparametric ES estimation methodology to two major European equity indices: the FTSE 100 (`^FTSE`) and the DAX (`^GDAXI`). Both indices are sourced from Yahoo Finance and cover the period October 1st, 2001 to September 30th, 2003. After removing missing values arising from non-synchronous trading calendars, 491 daily log loss observations were obtained for the FTSE 100 and 496 for the DAX. The analysis is conducted over three sub-periods: Year 1 (October 2001 – September 2002), Year 2 (October 2002 – September 2003), and the full two-year period (October 2001 – September 2003).

The log loss series are defined as $Y_t = -\log(X_t/X_{t-1})$, where X_t denotes the closing index level on day t. The choice of this two-year window is deliberate: it captures a period of sustained market stress in global equity markets, making it a demanding environment in which to evaluate the behaviour of tail risk estimators.

16.2 Return Series and Autocorrelation Structure

Figure 3 displays the daily log loss series and their sample autocorrelation functions (ACF) for both indices across the full observation window.

Both return series fluctuate around zero throughout the observation window. A prominent feature of both series is volatility clustering: periods of large losses tend to be followed by further large movements, most visibly for both indices around days 150–350. This bunching of extreme observations is a well-known characteristic of financial return data and is indicative of serial dependence in the series. The ACF plots confirm this visual impression. While most autocorrelations remain within the significance bands, notable spikes are observed for the FTSE 100 at lags 2 and 18, and for the DAX at lags 5–7 and 18 – a pattern too systematic to be attributed to sampling noise alone.

16.3 Box–Pierce Test for Serial Dependence

To formally assess the presence of serial dependence in each series, we apply the Box–Pierce test. The test statistic is

$$Q = n \sum_{k=1}^{29} \hat{\rho}^2(k) \;\sim\; \chi^2_{29} \quad \text{under } H_0, \tag{16}$$

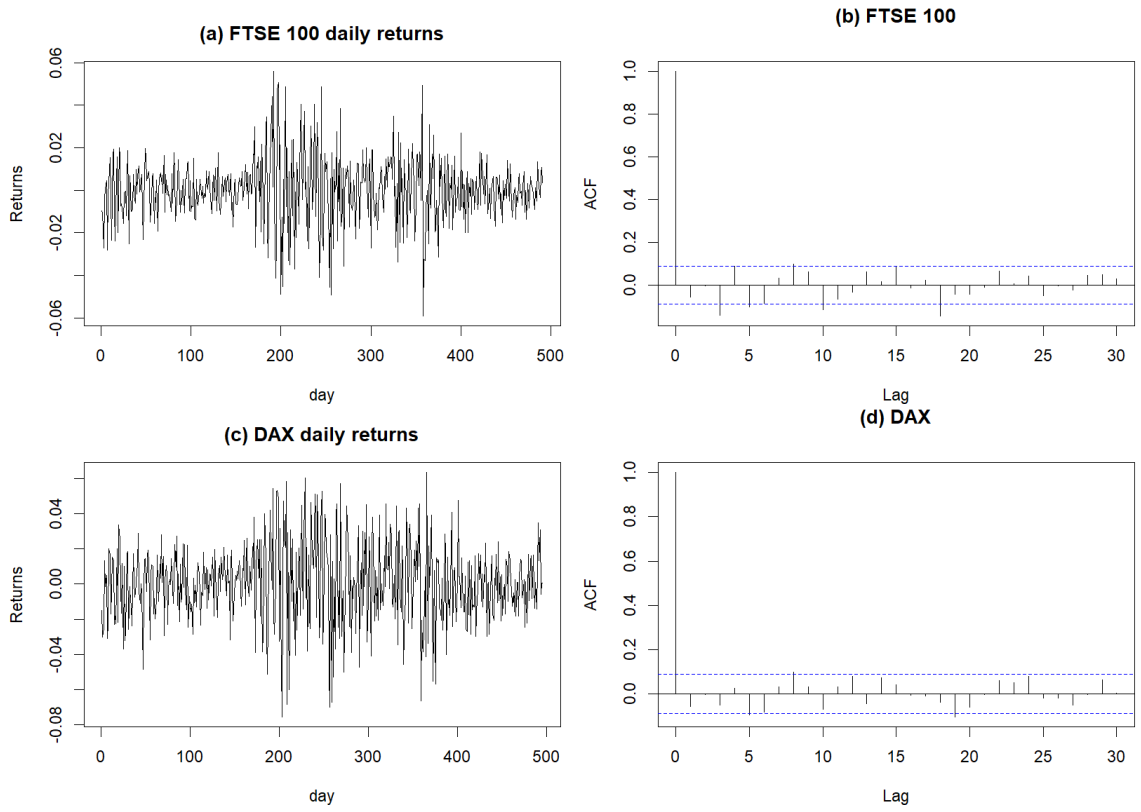

Figure 3. Daily log loss series (panels a and c) and sample autocorrelation functions (panels b and d) for FTSE 100 and DAX, October 2001 – September 2003.

where $\hat{\rho}(k)$ denotes the sample autocorrelation at lag k and H_0 is the null hypothesis of no serial autocorrelation. The results are reported in Table 3.

Series	Q statistic	p-value
FTSE 100	64.114	0.0002
DAX	44.855	0.0304

Table 3. Box–Pierce test results for FTSE 100 and DAX (29 lags).

Both p-values lie below the 5% significance level, and we therefore reject H_0 for both series. Serial dependence is statistically confirmed in both cases. The FTSE 100 exhibits particularly strong dependence (p = 0.0002), while the DAX result, though significant, sits closer to the boundary (p = 0.0304). The presence of dependence in both series means that observations cannot be treated as independent, and the asymptotic variance of the ES estimator must account for the covariance structure of the series through the long-run variance term $\sigma_0^2(p; n)$.

16.4 VaR, ES and Standard Error Estimates at $p = 0.01$

Table 4 presents the kernel VaR estimate $\hat{v}_{0.01}$, the sample ES estimate $\hat{\mu}_{0.01}$, and the associated standard error (SE) obtained via spectral density estimation, for each

index and sub-period at the 99% confidence level ($p = 0.01$).

Period	VaR $\hat{v}_{0.01}$	ES $\hat{\mu}_{0.01}$	SE
FTSE 2001–2002	0.0499	0.0518	0.0051
FTSE 2002–2003	0.0349	0.0408	0.0068
FTSE 2001–2003	0.0472	0.0506	0.0025
DAX 2001–2002	0.0564	0.0575	0.0049
DAX 2002–2003	0.0513	0.0559	0.0063
DAX 2001–2003	0.0548	0.0586	0.0024

Table 4. Kernel VaR $\hat{v}_{0.01}$, sample ES $\hat{\mu}_{0.01}$, and standard errors at $p = 0.01$ for FTSE 100 and DAX across three time periods.

Which index was riskier? The DAX produced consistently higher ES and VaR estimates than the FTSE 100 across all three periods. At $p = 0.01$, the DAX ES exceeded that of the FTSE by approximately 10% in Year 1, and by as much as 37% in Year 2, indicating substantially greater tail risk in the German index throughout the observation window.

Which period was riskiest? For the FTSE 100, Year 1 was clearly the period of highest risk, with ES falling from 0.0518 to 0.0408 in Year 2 – a reduction of approximately 21%, consistent with a market that recovered meaningfully in the second year. The DAX presents a markedly different picture: its ES declined only marginally between years (from 0.0575 to 0.0559), suggesting that the German market remained under sustained stress throughout the entire two-year window rather than recovering in the second year.

Standard errors and sample size. Consistent with the asymptotic theory, the full two-year standard errors are substantially smaller than those from either individual year for both indices (FTSE: 0.0025 versus 0.0051 and 0.0068; DAX: 0.0024 versus 0.0049 and 0.0063). This confirms that larger samples yield more precise ES estimates, as the effective sample size for ES estimation grows with n.

16.5 ES Estimates with 95% Confidence Bands

Figure 4 shows the sample ES estimates $\hat{\mu}_p$ and their 95% confidence bands over 20 equally spaced values of p from 0.01 to 0.03, for all three sub-periods and both indices. The bands are constructed as $\hat{\mu}_p \pm 1.96 \times \widehat{\mathrm{SE}}$, where the standard error is obtained from spectral density estimation as described earlier.

As expected from theory, the ES estimate decreases monotonically as p increases across all panels: averaging over a less extreme portion of the tail naturally reduces the conditional expected loss. The DAX curves lie above the FTSE curves in every sub-period and at every value of p, reinforcing the conclusion that the DAX carried higher tail risk throughout. For the FTSE 100, the Year 1 curve lies noticeably

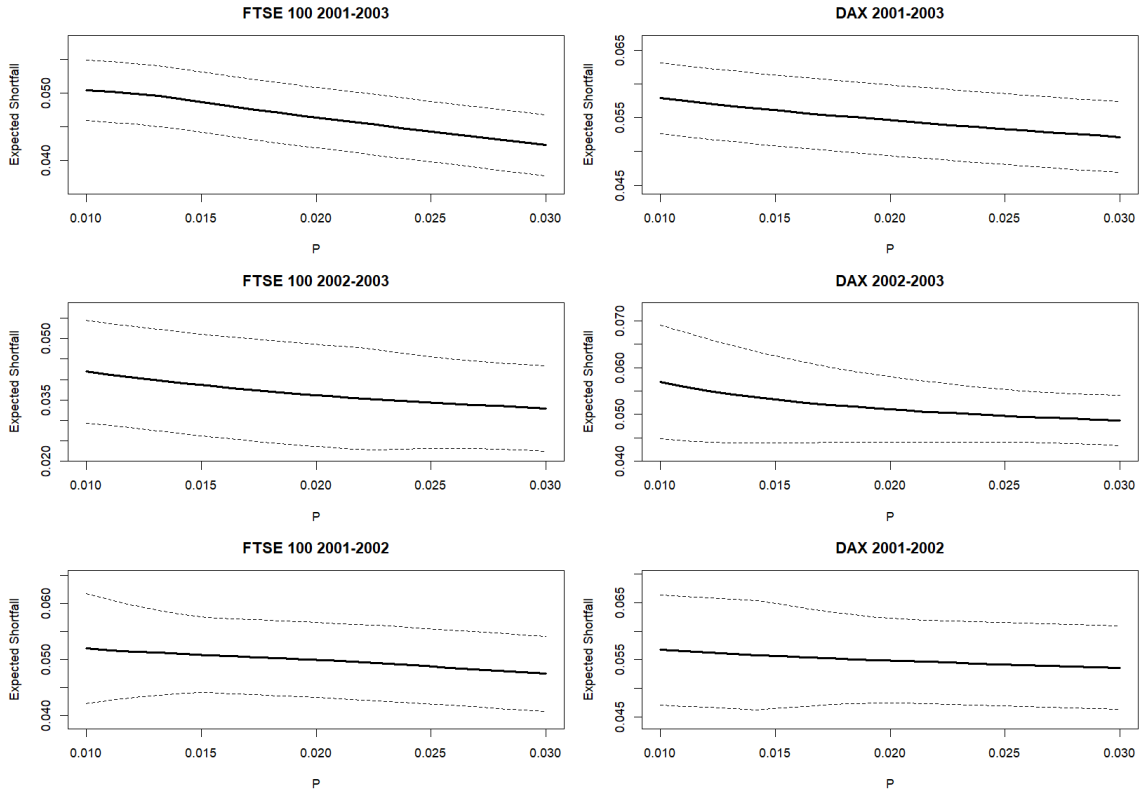

Figure 4. Sample ES estimates (solid line) with 95% confidence bands (dashed lines) for FTSE 100 (left column) and DAX (right column) across three sub-periods, for $p \in [0.01, 0.03]$.

above the Year 2 curve, reflecting the clear reduction in risk in the second year. For the DAX, the two single-year curves are much closer together, consistent with the persistent risk levels observed in Table 4.

The confidence bands narrow substantially when moving from single-year to full two-year estimates, reflecting the precision gains from larger samples. One additional observation is that the DAX Year 2 bands are visibly wider than those of DAX Year 1, despite similar sample sizes. This is consistent with the higher SE for DAX 2002–2003 in Table 4 (0.0063 versus 0.0049), and may reflect greater irregularity in the tail behaviour of the DAX during the second sub-period.

16.6 Summary

The nonparametric ES estimator produces stable, well-behaved estimates across both indices and all three time periods. The key findings are as follows. First, serial dependence is confirmed for both the FTSE 100 and the DAX, validating the use of a dependence-aware variance estimator. Second, the DAX consistently exhibited higher tail risk than the FTSE 100 across all periods and all values of p. Third, while risk declined meaningfully in the FTSE 100 between Year 1 and Year 2, the DAX showed near-persistent risk levels throughout the full observation window – a finding that distinguishes the two European markets during this period. Fourth, the standard errors decrease as expected with larger sample sizes, confirming the reliability and precision of the sample ES estimator under real-world dependent financial data.

Key References

- Chen, S. X. (2008). Nonparametric estimation of expected shortfall. *Journal of Financial Econometrics*, 6(1), 87-107.
- Scaillet, O. (2004). Nonparametric estimation and sensitivity analysis of expected shortfall. *Mathematical Finance*, 14(1), 115-129.