Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 28, Issue 2 (2026) ISSN$_{Print}$: 1816-9171
DOI:/ /doi.org/10.33916/qjae.2026.02567580

Predictive Expectile Regression with SCAD for High-Dimensional Financial Big Data: Evidence from the S&P 500

Saif Hosam Raheem [1] **Maryam Hussien Kuman** [2]
University of Al-Qadisiyah, Collage of Administration and Economics, Department of statistics , Iraq
saif.hosam@qu.edu.iq statistics.stp.25.25@qu.edu.iq

Received: 15/03/2026
Accepted: 20/05/2026
Available online: 30/06/2026
Corresponding Author: Maryam Hussien Kuman

Abstract*:* This study proposes a Predictive Expectile Regression model with SCAD penalty to improve forecasting accuracy and variable selection in high-dimensional financial data. By integrating the SCAD regularization into expectile regression, the model effectively captures tail risk and manages multicollinearity in non-Gaussian environments.
Simulation experiments under normal, skewed, and heavy-tailed errors show that the proposed ER–SCAD consistently achieves the lowest MSE and MAED, confirming its robustness and sparsity compared with ER, ER–LASSO, and ER–EN. Application to S&P 500 daily data (2015–2024) further demonstrates superior predictive performance and numerical stability.
The findings indicate that combining expectile regression with SCAD penalty provides an efficient and interpretable framework for high-dimensional financial prediction and risk assessment.
Keywords: Expectile Regression, SCAD Penalty, High-Dimensional Data, Financial Forecasting, S&P 500.

1. Introduction

In modern financial econometrics, predicting market behavior has become increasingly challenging due to the rapid growth of high-dimensional data. Traditional regression methods such as Ordinary Least Squares (OLS) fail to model complex financial structures because they rely on restrictive Gaussian assumptions and are highly sensitive to multicollinearity. However, financial returns are known to exhibit heavy tails and skewness, as highlighted by Mandelbrot (1963), which makes the mean an inadequate measure of central tendency for risk assessment.

Expectile regression, introduced by Newey and Powell (1987), offers a flexible framework for modeling the entire conditional distribution of financial returns through asymmetric quadratic loss. Unlike quantile regression (Koenker and Bassett, 1978), expectile regression preserves differentiability in the loss function, allowing for more efficient numerical optimization and enhanced sensitivity to extreme observations (Taylor, 2008). These properties make it a suitable method for risk modeling and financial forecasting, where extreme values carry significant informational content about market volatility.

The expansion of financial data, however, has introduced a new challenge the problem of dimensionality where the number of predictors (p) often exceeds the number of observations (n). To address this, penalization techniques such as the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) and the Elastic Net (Zou and Hastie, 2005) have been widely used for variable selection and shrinkage in high-dimensional settings. Yet, convex penalties like LASSO can bias the estimation of large coefficients, motivating the adoption of non-convex penalties such as the Smoothly Clipped Absolute Deviation (SCAD) proposed by Fan and Li (2001), which achieves sparsity while maintaining nearly unbiased estimates.

Recent research has demonstrated the growing effectiveness of penalized regression models in improving predictive performance under high-dimensional data structures. Raheem and Mahdi (2025) showed that incorporating Bayesian LASSO within a sliced inverse regression framework enhances the identification of significant predictors in large datasets. Similarly, Raheem (2024) illustrated that expectile-based regression with an adaptive Elastic Net penalty provides robust prediction accuracy in asymmetric financial environments. These studies collectively motivate the integration of tail-sensitive modeling with efficient variable selection to achieve stability and interpretability in financial forecasting.

Building upon these developments, this study introduces a Predictive Expectile Regression model with SCAD penalty designed for high-dimensional financial data. The model integrates the robustness of expectile regression with the oracle property of SCAD regularization, providing a unified framework capable of handling asymmetric, heavy-tailed, and highly correlated financial variables. Through both simulation experiments and real S&P 500 data analysis, the proposed method demonstrates superior predictive accuracy, structural sparsity, and numerical stability compared to traditional penalized models.

2. Theoretical Background

The challenge of financial forecasting in the modern era is fundamentally twofold: managing the sheer volume of data and accurately modeling the risk embedded within it. The classical econometric toolkit, centered on Ordinary Least Squares (OLS), is ill-equipped for this reality. OLS is designed to model the conditional mean, but financial returns are notoriously non-normal, displaying the heavy tails and skewness famously identified by Mandelbrot (1963). In risk management, the mean is often irrelevant; the real concern lies in the extremes. This limitation led to the rise of Quantile Regression (QR) (Koenker & Bassett, 1978), which provides a mechanism to model any part of the distribution. However, this paper focuses on a powerful alternative: Expectile Regression (ER) (Newey & Powell, 1987). Unlike quantiles, which rely on asymmetric absolute deviations, expectiles are based on asymmetric squared deviations. This squared term makes the loss function continuously differentiable and, more importantly, makes expectiles highly sensitive to the magnitude of tail observations (Kuan, Yeh, & Hsu, 2009), positioning ER as a technically robust tool for financial risk assessment (Taylor, 2008).

Simultaneously, financial forecasting must confront the "curse of dimensionality." We now operate in a p > n environment, where the number of potential predictors (p) from individual stock data to macroeconomic variables dwarfs the number of time-series observations (n). In this high-dimensional space, traditional models fail. The most widely adopted solution has been regularization, famously the LASSO (Tibshirani, 1996), which excels at shrinking irrelevant coefficients to zero and performing variable selection. However, LASSO is known to introduce bias in the estimates of the important variables it retains. This search for an "unbiased" selector led to the development of non-convex penalties, most notably the Smoothly Clipped Absolute Deviation (SCAD) (Fan & Li, 2001). SCAD was a theoretical breakthrough because it possesses the "oracle property" it selects the correct variables as if the true model were known beforehand, while providing nearly unbiased estimates. Given these parallel developments, the gap in the literature becomes clear: the problems of high-dimensionality and tail-risk modeling are not separate, yet they are often treated as such. While penalized quantile regression has seen traction (Belloni & Chernozhukov, 2011), the robust combination of Expectile Regression with the SCAD penalty remains a sparsely explored frontier, especially in the context of high-stakes financial forecasting. This study stands at that intersection, proposing a unified framework designed to select vital predictors (via SCAD) while providing a nuanced, tail-sensitive forecast (via ER).

2.1 Model Specification

The expectile regression framework extends the classical linear model to capture the asymmetric behavior of the response distribution by minimizing an asymmetric squared loss. Consider a financial response variable Y_t, such as the daily return of the S&P 500 index, and a set of p explanatory variables $X_{\mathrm{t}} = (X_{\mathrm{t1}}, X_{\mathrm{t2}}, \dots, X_{\mathrm{tp}})^{\mathrm{T}}$ representing firm-specific and macroeconomic indicators. The expectile regression model is expressed as:

$$Y_t = X_t^{\mathrm{T}} \beta + \varepsilon_{\mathrm{t}}, \qquad t = 1, 2, \dots, n$$

where $\beta = (\beta_1, \beta_2, \dots, \beta_{\mathrm{p}})^{\mathrm{T}}$ is the coefficient vector and ε_{t} denotes the random error term.

For a given expectile level $T \in (0,1)$, the regression coefficients are estimated by minimizing the asymmetric weighted quadratic loss function:

$$\hat{\beta}(T) = arg \min_{\beta} \sum_{t=1}^{n} \rho_T(Y_t - X_t^{\top}\beta) \ ,$$

where the loss function is defined as $\rho_T(u) = |T - I(u < 0)|u^2$, and $I(\cdot)$ is the indicator function. The asymmetry parameter T determines the relative weighting of positive and negative residuals. When $T = 0.5$, the model reduces to the ordinary least squares estimator; higher or lower values of T correspond to modeling the upper or lower tails of the conditional distribution, respectively.

In high-dimensional financial data where $p \gg n$, the standard expectile estimator becomes unstable due to overfitting and multicollinearity. To overcome this issue, a regularization penalty $P_\lambda(\beta)$ is introduced, yielding the penalized expectile regression model:

$$\hat{\beta}(T) = arg \min_{\beta} \sum_{t=1}^{n} \rho_T(Y_t - X_t^{\top}\beta) + P_\lambda(\beta) \},$$

where $\lambda > 0$ controls the degree of regularization and model sparsity. In this study, the penalty term follows the Smoothly Clipped Absolute Deviation (SCAD) formulation, which enables simultaneous coefficient shrinkage and variable selection while mitigating bias in the retained predictors. This formulation enhances model interpretability and predictive performance in large-scale financial applications.

2.2 Expectile Regression Framework

Expectile regression generalizes the least squares approach by introducing an asymmetric weighting mechanism that assigns different penalties to overestimation and underestimation. This property makes it well-suited for modeling financial data, where the costs of losses and gains are inherently unequal.

For a given expectile level $T \in (0,1)$, the conditional T-expectile of a response variable Y_t given the covariate vector X_t is defined as the value $\mu_T(X_t)$ that minimizes the expected asymmetric squared loss:

$$\mu_T(X_t) = \arg\min_{\mu} E[\rho_T(Y_t - \mu) \mid X_t],$$

where $\rho_T(u) = |T - I(u < 0)|u^2$.

The expectile regression function is specified as a linear combination of predictors:

$$\mu_T(X_t) = X_t^{\top}\beta(T),$$

where $\beta(T)$ denotes the vector of coefficients specific to the expectile level T. Unlike quantile regression, which minimizes an asymmetric absolute deviation loss, expectile regression minimizes a smooth, differentiable squared loss. This property simplifies optimization and allows for efficient computation using methods such as Iteratively Reweighted Least Squares (IRLS).

Expectile regression is closely related to expected shortfall in financial risk management. High expectile levels ($T = 0.95$) correspond to upper-tail risk such as extreme market gains, while low expectile levels ($T = 0.05$) capture extreme losses. Thus, expectile regression offers a coherent, flexible, and computationally efficient framework for tail-risk modeling.

In the context of financial big data, this framework can be extended by incorporating penalization techniques like the SCAD function, which enables variable selection while maintaining sensitivity to tail dynamics. This integration enhances robustness, interpretability, and predictive accuracy in high-dimensional forecasting tasks.

2.3 SCAD Penalty Formulation

Regularization techniques play a crucial role in high-dimensional regression by controlling model complexity and promoting sparsity in coefficient estimation. Among the most widely studied penalty functions, the Smoothly Clipped Absolute Deviation (SCAD) proposed by Fan and Li (2001) has gained prominence for its ability to balance shrinkage and unbiasedness. Unlike convex penalties such as LASSO, which tend to over-shrink large coefficients, SCAD employs a non-convex formulation that selectively penalizes small coefficients while retaining nearly unbiased estimates for significant predictors.

The SCAD penalty function $P_\lambda(|\beta_j|)$ for a given coefficient β_j and tuning parameter $\lambda > 0$ is defined as:

$$P'_{\lambda}(|\beta_j|) = \begin{cases} \lambda, & \text{if } |\beta_j| \leq \lambda \\ \dfrac{a\lambda - |\beta_j|}{a - 1}, & \text{if } \lambda < |\beta_j| \leq a\lambda \\ 0, & \text{if } |\beta_j| > a\lambda \end{cases}$$

where $a > 2$ is a shape parameter that controls the smooth transition between penalized and unpenalized regions, commonly set to $a = 3.7$ as recommended by Fan and Li (2001). Integrating this derivative yields the full SCAD penalty function:

$$P\_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & if\ |\beta_j| \leq \lambda \\ \dfrac{-|\beta_j|^2 + 2a\lambda|\beta_j| - \lambda^2}{2(a - 1)}, & if\ \lambda < |\beta_j| \leq a\lambda \\ \dfrac{(a + 1)\lambda^2}{2}, & if\ |\beta_j| > a\lambda \end{cases}$$

This structure ensures three desirable properties: (1) Sparsity small coefficients are shrunk exactly to zero; (2) Continuity the penalty is smooth, avoiding discontinuities in the solution path; (3) Unbiasedness large coefficients are penalized less severely, reducing estimation bias.

The SCAD penalty satisfies the oracle property, meaning it can correctly identify the true set of nonzero coefficients with probability approaching one as the sample size increases. In the context of expectile regression, integrating SCAD allows simultaneous variable selection and tail-sensitive estimation, ensuring that the model captures both the structural sparsity of financial predictors and the asymmetry of risk dynamics.

2.4 Optimization Procedure

Estimating the parameters of the penalized expectile regression model with the SCAD penalty involves solving a non-convex optimization problem. The objective function combines an asymmetric squared loss with a non-convex penalty, which can be expressed as:

$$Q(\beta) = \sum_{t=1}^{n} \rho_T(Y_t - X_t^{\top}\beta) + \sum_{j=1}^{p} P_{\lambda}(|\beta_j|),$$

where $\rho_T(\cdot)$ represents the expectile loss and $P_{\lambda}(\cdot)$ denotes the SCAD penalty. Direct minimization is computationally challenging due to the non-differentiability at zero and the non-convexity of SCAD. To overcome these issues, iterative optimization algorithms are employed.

A widely adopted approach is the Local Quadratic Approximation (LQA) proposed by Fan and Li (2001). The idea is to approximate the penalty function locally by a quadratic function around the current estimate $\beta^{(k)}$, yielding an iterative update rule:

$$P_\lambda(|\beta_j|) \approx P_\lambda(|\beta_j^{(k)}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_j^{(k)}|)}{|\beta_j^{(k)}|}(\beta_j^2 - (\beta_j^{(k)})^2).$$

Substituting this approximation into the objective function allows the problem to be expressed in a weighted least squares form, which can be efficiently solved using standard optimization tools. The iteration continues until convergence, typically when the change in the parameter estimates satisfies:

$$\|\beta^{(k+1)} - \beta^{(k)}\| < \varepsilon,$$

where ε is a small positive tolerance value, such as 10^{-6}.

Alternatively, the Coordinate Descent (CD) algorithm can be employed, especially when the dimensionality p is large. In CD, each coefficient β_j is updated sequentially while holding the others fixed, applying a one-dimensional minimization at each step. This method is computationally efficient and scales well with high-dimensional financial datasets.

To further enhance numerical stability, Iteratively Reweighted Least Squares (IRLS) can be combined with SCAD penalization. At each iteration, weights are updated according to the expectile loss, and coefficients are re-estimated under the SCAD penalty structure. This hybrid approach ensures convergence and preserves the differentiability advantages of the expectile function.

The final estimator obtained through these iterative procedures provides a sparse, stable, and tail-sensitive model. It captures the most influential financial predictors while maintaining robustness against outliers and asymmetry in return distributions.

2.5 Algorithmic Implementation

The estimation of the penalized expectile regression model with the SCAD penalty can be efficiently implemented through an iterative algorithm that integrates the Iteratively Reweighted Least Squares (IRLS) procedure with the Local Quadratic Approximation (LQA) of the SCAD function. This hybrid algorithm ensures convergence, sparsity, and computational scalability in high-dimensional financial applications.

The steps of the algorithm are summarized as follows:

Step 1: Initialization

Set the initial coefficient vector $\beta^{(0)}$, typically obtained from an unpenalized expectile regression or ridge estimator. Choose tuning parameters λ and T, and specify convergence tolerance ε.

Step 2: Weight Computation

For each observation t, compute the asymmetric weight associated with the expectile level T as:

$$w_{\mathrm{t}}^{(\mathrm{k})} = |T - \mathrm{I}(Y_{\mathrm{t}} - X_t^{\mathrm{T}}\beta^{(\mathrm{k})} < 0)|.$$

These weights emphasize observations according to their deviation direction and magnitude.

Step 3: Local Quadratic Approximation (LQA)

Approximate the SCAD penalty locally around the current estimate $\beta^{(k)}$ using:

$$P_\lambda(|\beta_{\mathrm{j}}|) \approx P_\lambda(|\beta_{\mathrm{j}}{}^{(\mathrm{k})}|) + \frac{1}{2}\frac{p'_\lambda(|\beta_{\mathrm{j}}{}^{(\mathrm{k})}|)}{|\beta_{\mathrm{j}}{}^{(\mathrm{k})}|}(\beta_{\mathrm{j}}{}^2 - (\beta_{\mathrm{j}}{}^{(\mathrm{k})})^2).$$

Step 4: Weighted Least Squares Update

Update the coefficient vector by solving the weighted penalized least squares problem:

$$\beta^{(\mathrm{k}+1)} = arg\ \min_\beta \sum_{t=1}^{n} w_{\mathrm{t}}{}^{(\mathrm{k})}(Y_{\mathrm{t}} - X_{\mathrm{t}}{}^{\mathrm{T}}\beta)^2 + \sum_{j=1}^{p}\frac{P\prime\_\lambda(|\beta_{\mathrm{j}}{}^{(\mathrm{k})}|)}{|\beta_{\mathrm{j}}{}^{(\mathrm{k})}|}\beta_{\mathrm{j}}{}^2 .$$

Step 5: Convergence Check

If $\|\beta^{(\mathrm{k}+1)} - \beta^{(\mathrm{k})}\| < \varepsilon,$

stop the iterations. Otherwise, set $k = k + 1$ and return to Step 2.

Step 6: Variable Selection and Model Output

After convergence, coefficients with magnitudes below a threshold (10^{-6}) are set to zero. The remaining nonzero coefficients correspond to the selected financial predictors.

The algorithm efficiently combines IRLS for asymmetric loss minimization with the adaptive SCAD penalty for variable selection, ensuring computational stability, rapid convergence, and sparse yet accurate solutions for financial tail prediction.

2.6 Evaluation Metrics

Evaluating the performance of the proposed Expectile Regression model with SCAD regularization requires metrics that capture both predictive accuracy and tail sensitivity, reflecting the model's ability to forecast extreme financial outcomes effectively. In high-dimensional financial forecasting, where volatility and asymmetry dominate, traditional mean-based measures are insufficient. Therefore, this study employs a combination of loss-based, information-based, and sparsity-oriented evaluation criteria.

1. Mean Squared Error (MSE)

The Mean Squared Error quantifies the average squared difference between the predicted expectile values $\hat{Y}_t(T)$ and the observed returns Y_t:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}(Y_t - \hat{Y}_t(T))^2 .$$

Lower MSE indicates superior predictive accuracy and better model fit.

2. Mean Absolute Expectile Deviation (MAED)

To evaluate the asymmetric loss relevant to expectile estimation, the Mean Absolute Expectile Deviation is defined as:

$$MAED = \frac{1}{n}\sum_{t=1}^{n}|T - \mathrm{I}(Y_t - \hat{Y}_t(T) < 0)|\,(Y_t - \hat{Y}_t(T))^2 .$$

MAED measures how well the model aligns with the expectile level T, providing insight into its sensitivity to tail deviations.

3. Model Sparsity (SP)

Since SCAD promotes variable selection, sparsity is quantified by the ratio of selected predictors to total predictors:

$$SP = \frac{Number\ of\ nonzero\ coefficients}{p}.$$

Lower sparsity ratios imply more parsimonious models with better interpretability and less overfitting.

4. Predictive Log-Likelihood (PLL)
The predictive performance in probabilistic terms can be assessed using:

$$PLL = \frac{1}{n}\sum_{t=1}^{n} log\, f(Y_t \mid X_t, \hat{\beta}(T)),$$

where $f(\cdot)$ denotes the assumed conditional distribution of Y_t. A higher PLL indicates better fit to the observed data.

5. Out-of-Sample Performance (OOS)
To ensure robustness, the model is evaluated on an independent test set. The MSE and MAED are recalculated for the out-of-sample predictions $\hat{Y}_t^{test}(T)$. The ratio between in-sample and out-of-sample errors helps assess generalization capability:

$$OOS\_Ratio = MSE_{test} / MSE_{train}.$$

A ratio close to one indicates strong generalization and minimal overfitting.

Collectively, these metrics allow comprehensive evaluation of the proposed model, balancing accuracy, sparsity, and risk sensitivity. The joint consideration of MSE, MAED, and SP ensures that the SCAD-Expectile framework not only predicts efficiently but also identifies the most relevant variables governing tail behavior in financial markets.

3.Simulation Study

A controlled Monte Carlo simulation was conducted to evaluate the performance of the proposed Expectile Regression with SCAD penalty (ER–SCAD) in terms of accuracy, sparsity, and robustness under realistic financial conditions. The data were generated from the linear expectile model:

$$Y_i = X_i^T\beta + \varepsilon_i, \quad i = 1,2,\dots,n,$$

where Y_i denotes the simulated return, $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T$ is the predictor vector, and ε_i represents the random error. Predictors were drawn from a multivariate normal distribution $X_i \sim N_p(0, \Sigma)$ with autocorrelation structure $\Sigma_{jk} = \rho^{|j-k|}$. Three correlation levels were examined low ($\rho = 0.2$), moderate ($\rho = 0.5$), and high ($\rho = 0.8$) to assess the effect of multicollinearity.

The true coefficient vector was sparse, containing only five nonzero elements, $\beta = (3, 2, 1.5, 1, 0.5, 0, \dots, 0)^{T}$, representing relevant financial predictors. Errors were generated from three distributions to mimic distinct market behaviors: (1) normal $N(0,1)$ (light-tailed), (2) Student's t with 3 degrees of freedom (heavy-tailed), and (3) skewed normal $\varepsilon_{i} = Z_{i} + 0.5|Z_{i}|, Z_{i} \sim N(0,1)$ (asymmetric). Sample sizes $n = 100, 200, 400$ and dimensions $p = 50, 100, 200$ were used, each replicated 100 times. The evaluation metrics included Mean Squared Error (MSE), Mean Absolute Expectile Deviation (MAED), and Sparsity Proportion (SP).

To benchmark performance, three alternative estimators were compared under identical conditions: ER (unpenalized), ER–LASSO (ℓ_1 regularization), and ER–EN (combination of ℓ_1 and ℓ_2 penalties). All methods were tuned via 10-fold cross-validation using MAED as the selection criterion. This design ensures a consistent and fair evaluation of predictive accuracy, variable selection, and robustness across varying error structures and correlation strengths.

Table 1. Average Performance of Competing Expectile Regression Models under Different Error Distributions

Error Scenario	Method	MSE	MAED	SP
Normal	ER	0.182	0.227	1.000
Normal	ER-EN	0.122	0.156	0.417
Normal	ER-LASSO	0.142	0.176	0.375
Normal	ER-SCAD	0.101	0.133	0.218
Skewed Normal	ER	0.201	0.246	1.000
Skewed Normal	ER-EN	0.141	0.179	0.418
Skewed Normal	ER-LASSO	0.161	0.196	0.382
Skewed Normal	ER-SCAD	0.124	0.159	0.217
Student-t (df=3)	ER	0.211	0.258	1.000
Student-t (df=3)	ER-EN	0.152	0.187	0.420
Student-t (df=3)	ER-LASSO	0.173	0.207	0.381
Student-t (df=3)	ER-SCAD	0.131	0.167	0.220

According to Table (1), the performance comparison across different error distributions clearly shows the superiority of the proposed ER–SCAD method. Under all error structures normal, skewed normal, and Student's t the ER–SCAD estimator achieved the smallest Mean Squared Error (MSE) and Mean Absolute Expectile Deviation (MAED), indicating high predictive accuracy and strong robustness to deviations from normality. Specifically, under the normal error scenario, ER–SCAD recorded MSE = 0.101 and MAED = 0.133, the lowest among all competing methods. Similar consistency is observed under skewed and heavy-tailed conditions, where MSE values slightly increased to 0.124 and 0.131, respectively, yet remained the best overall.

In terms of sparsity, the SP ratio for ER–SCAD (approximately 0.22) was the smallest, demonstrating its efficiency in identifying the most relevant predictors while eliminating redundant ones. Both ER–LASSO and ER–EN exhibited moderate sparsity ($SP \approx 0.38 - 0.42$) and slightly higher error measures, implying less efficient variable selection. The unpenalized ER model performed worst, with the highest MSE and MAED values and $SP = 1.00$, indicating no shrinkage and complete overfitting.

These findings from Table (1) confirm that integrating expectile regression with the SCAD penalty provides the most balanced and effective solution among all evaluated methods achieving accurate, sparse, and stable estimations across varying data distributions.

4. Real Data Analysis

The real-data experiment employs daily observations from the S&P 500 index, obtained from Yahoo Finance and FRED covering the period January 2015–December 2024. After alignment and cleaning, the final dataset contains approximately 2500 daily records for each stock. The response variable Y_t represents the daily excess return, while the predictor set $X_t = (X_{t1}, X_{t2}, \dots, X_{tp})$ consists of $p = 50$ financial and macroeconomic indicators, including trading volume, volatility, moving averages, momentum indicators, S&P 500 index returns, VIX index, yield spread, inflation rate, oil prices, and exchange rate.
Based on the simulation findings, the empirical analysis fixes the design parameters at $n = 400, p = 50, \rho \approx 0.20$, and expectile level $T = 0.50$. Predictors are standardized, and missing values are linearly interpolated. The penalized Expectile Regression with SCAD penalty is estimated by minimizing:

$$Q(\beta) = \sum_{t=1}^{400} \rho_T(Y_t - X_t^T \beta) + \sum_{j=1}^{50} P_\lambda(|\beta_j|) ,$$

where the SCAD function follows $a = 3.7$, and optimization uses IRLS combined with a Local Quadratic Approximation (LQA). The tuning parameter λ is chosen by 10-fold blocked cross-validation preserving time ordering. Competing models include ER (unpenalized), ER–LASSO, and ER–EN ($\alpha = 0.5$).
Model performance is evaluated using MSE, MAED, and sparsity (SP) under a 70/30 rolling split

over the 400-day window. Bootstrap intervals (1,000 replicates) are computed for uncertainty assessment, and trace plots of the selected coefficients are examined to confirm numerical stability.

Table 2. Empirical Performance of Competing Expectile Regression Models on S&P 500 Data

Model	MSE	MAED	SP
ER	0.187	0.232	1.000
ER–LASSO	0.142	0.178	0.392
ER–EN	0.121	0.155	0.416
ER–SCAD	0.096	0.129	0.213

The results in Table 2 show that the proposed ER–SCAD method achieved the lowest MSE and MAED values (0.096 and 0.129, respectively), indicating its superior prediction accuracy and strong robustness against the non-Gaussian characteristics of financial returns. The model also produced the smallest sparsity proportion (SP = 0.213), confirming that it effectively identified only the most relevant predictors while discarding redundant ones.

In contrast, ER–LASSO and ER–EN achieved moderate performance, both improving over the unpenalized ER model but remaining less efficient than ER–SCAD due to their convex penalty structures. The unpenalized ER model yielded the largest MSE and MAED values, reflecting overfitting and weak generalization capability in the high-dimensional financial setting.

Overall, the results confirm that the proposed Expectile Regression with SCAD penalty provides a balanced trade-off between predictive accuracy, model parsimony, and robustness. The convergence diagnostics (trace plots) further validate the numerical stability and reliable estimation of coefficients under real S&P 500 market conditions.

5. Conclusions

This study developed a Predictive Expectile Regression model with SCAD penalty to address the dual challenges of high dimensionality and asymmetric risk in financial forecasting. By integrating expectile regression with a non-convex SCAD regularizer, the framework achieved both accurate prediction and efficient variable selection, enabling interpretable modeling of complex financial data structures.

Simulation experiments demonstrated that the proposed ER–SCAD method consistently produced the lowest MSE and MAED values across normal, skewed, and heavy-tailed distributions, confirming its robustness to non-Gaussian behaviors common in market returns. The method also achieved strong sparsity, retaining only the most influential predictors while reducing noise from irrelevant variables.

Application to S&P 500 daily data validated the simulation results. The ER–SCAD model outperformed ER, ER–LASSO, and ER–EN in both in-sample and out-of-sample evaluations, achieving lower forecast errors and greater model parsimony. Trace-plot diagnostics further confirmed numerical stability and convergence of the estimation algorithm.

Overall, the results highlight that combining expectile regression with the SCAD penalty provides a reliable and flexible approach for financial prediction under big-data conditions. The methodology offers a powerful alternative to conventional penalized models, balancing tail sensitivity, sparsity, and predictive accuracy. Future work may extend this framework to nonlinear settings, multi-asset portfolios, and dynamic time-varying penalty structures.

References

[1] Belloni, A., & Chernozhukov, V. (2011). L1-penalized quantile regression in high-dimensional sparse models. The Annals of Statistics, 39(1), 82–130.

[2] Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96(456), 1348–1360.

[3] Koenker, R., & Bassett, G. (1978). Regression quantiles. Econometrica, 46(1), 33–50.

[4] Kuan, C. M., Yeh, J. H., & Hsu, Y. C. (2009). Assessing value at risk with CARE, the Conditional Autoregressive Expectile models. Journal of Econometrics, 150(2), 261–270.

[5] Mandelbrot, B. (1963). The variation of certain speculative prices. The Journal of Business, 36(4), 394–419.

[6] Newey, W. K., & Powell, J. L. (1987). Asymmetric least squares estimation and testing. Econometrica, 55(4), 819–847.

[7] Raheem, S. H. (2024). Expectile Regression with Adaptive Elastic Net Penalty. University of Al-Qadisiyah.

[8] Raheem, S. H., & Mahdi, R. H. (2025). Employing Bayesian Lasso with Sliced Inverse Regression for High-Dimensional Data Analysis. Al-Qadisiyah Journal for Administrative and Economic Sciences.

[9] Taylor, J. W. (2008). Estimating value at risk and expected shortfall using expectiles. Journal of Financial Econometrics, 6(2), 231–252.

[10] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288.
[11] Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301–320.