Mathematical Documentation and Trading Strategies

Quantitative Trading Research Platform

PRECOG Quant Task Project

February 2026

Contents

1 Executive Summary

This document provides comprehensive mathematical documentation for a quantitative trading research framework. The platform implements a complete algorithmic trading pipeline encompassing:

- Data acquisition and feature engineering using advanced signal processing
- Statistical arbitrage via cointegration analysis
- Machine learning-based alpha and risk prediction
- Systematic backtesting with realistic transaction cost modeling

The framework processes 100 anonymized assets and develops systematic trading strategies based on mathematical models and statistical tests.

2 Data Cleaning and Feature Engineering

2.1 Technical Indicators

2.1.1 Relative Strength Index (RSI)

The RSI is a momentum oscillator measuring the speed and magnitude of price changes over a rolling window of $n = 14$ days.

Price Change:

$$\Delta_t = \text{Close}_t - \text{Close}_{t-1} \tag{1}$$

Gain and Loss Separation:

$$\text{Gain}_t = \max(\Delta_t, 0) \tag{2}$$

$$\text{Loss}_t = \max(-\Delta_t, 0) \tag{3}$$

Average Gain and Loss (14-day SMA):

$$\text{AvgGain}_t = \frac{1}{14} \sum_{i=t-13}^{t} \text{Gain}_i \tag{4}$$

$$\text{AvgLoss}_t = \frac{1}{14} \sum_{i=t-13}^{t} \text{Loss}_i \tag{5}$$

Relative Strength:

$$\text{RS}_t = \frac{\text{AvgGain}_t}{\text{AvgLoss}_t} \tag{6}$$

RSI Calculation:

$$\text{RSI}_t = 100 - \frac{100}{1 + \text{RS}_t} \tag{7}$$

The RSI oscillates between 0 and 100, with values above 70 traditionally indicating overbought conditions and values below 30 indicating oversold conditions.

2.1.2 Rogers-Satchell Volatility

The Rogers-Satchell (RS) volatility estimator utilizes intraday high-low-open-close data to provide an unbiased estimate of volatility without assuming zero drift.

Per-Period RS Variance:

$$\text{RS}_t = \ln\left(\frac{H_t}{O_t}\right) \cdot \ln\left(\frac{H_t}{C_t}\right) + \ln\left(\frac{L_t}{O_t}\right) \cdot \ln\left(\frac{L_t}{C_t}\right) \tag{8}$$

where:

- H_t = High price at time t
- L_t = Low price at time t
- O_t = Open price at time t
- C_t = Close price at time t

Rolling Volatility (14-day window):

$$\text{RS\_Vol}_t = \sqrt{\frac{1}{14} \sum_{i=t-13}^{t} \text{RS}_i} \tag{9}$$

2.2 Kalman Filter for Signal Smoothing

A one-dimensional Kalman filter is applied to smooth noisy technical indicators while preserving signal quality. The filter uses process noise covariance $Q = 0.01$ and measurement noise covariance $R = 1.0$.

State Prediction:

$$\hat{x}_{t|t-1} = \hat{x}_{t-1|t-1} \tag{10}$$

$$P_{t|t-1} = P_{t-1|t-1} + Q \tag{11}$$

Kalman Gain:

$$K_t = \frac{P_{t|t-1}}{P_{t|t-1} + R} \tag{12}$$

State Update:

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t(z_t - \hat{x}_{t|t-1}) \tag{13}$$

$$P_{t|t} = (1 - K_t) \cdot P_{t|t-1} \tag{14}$$

where:

- $\hat{x}_{t|t}$ = Filtered state estimate at time t
- z_t = Measurement (observed value) at time t
- $P_{t|t}$ = Error covariance estimate at time t
- K_t = Kalman gain at time t

2.3 Engineered Feature Set

The following features are engineered for each asset:

Feature	**Formula**	**Window**
ret_14d	$\frac{\text{Close}_t}{\text{Close}_{t-14}} - 1$	14 days
RSI_kalman	$\mathcal{K}(\text{RSI}_{14})$	-
ret_14d_kalman	$\mathcal{K}(\text{ret}_{14d})$	-
RS_vol_kalman	$\mathcal{K}(\text{RS\_Vol})$	-
RSI_slope	$\Delta(\text{RSI\_kalman})$	1st diff
RSI_accel	$\Delta^2(\text{RSI\_kalman})$	2nd diff
risk_adj_mom	$\frac{\text{ret\_14d\_kalman}}{\text{RS\_vol\_kalman}}$	-
vol_z_14	$\frac{\text{Volume}_t - \mu_{14}(\text{Volume})}{\sigma_{14}(\text{Volume})}$	14 days

Table 1: Engineered Feature Set

where $\mathcal{K}(\cdot)$ denotes the Kalman filter operator, Δ denotes first difference, and Δ^2 denotes second difference.

3 Similarity Analysis and Cointegration

3.1 Correlation-Based Asset Screening

Assets are screened for high correlation to identify potential trading pairs:

Pearson Correlation Coefficient:

$$\rho(r_i, r_j) = \frac{\mathrm{Cov}(r_i, r_j)}{\sigma_{r_i} \cdot \sigma_{r_j}} \tag{15}$$

where r_i and r_j are return series for assets i and j.

Screening Criterion:

$$\rho(r_i, r_j) > 0.90 \tag{16}$$

3.2 Cointegration Testing

For asset pairs passing the correlation screen, cointegration is tested to identify mean-reverting relationships.

3.2.1 Log-Price Construction

$$S_t^{\log} = \ln(S_t) \tag{17}$$

where S_t is the price of the asset at time t.

3.2.2 Hedge Ratio Estimation

The hedge ratio β is estimated via ordinary least squares (OLS) regression:

Linear Regression Model:

$$S_{1,t}^{\log} = \alpha + \beta \cdot S_{2,t}^{\log} + \epsilon_t \tag{18}$$

OLS Solution:

$$\beta = \frac{\mathrm{Cov}(S_1^{\log}, S_2^{\log})}{\mathrm{Var}(S_2^{\log})} \tag{19}$$

3.2.3 Spread Construction

The cointegrated spread is constructed as:

$$\mathrm{Spread}_t = S_{1,t}^{\log} - \beta \cdot S_{2,t}^{\log} \tag{20}$$

Under cointegration, this spread is stationary and mean-reverting.

3.2.4 Augmented Dickey-Fuller (ADF) Test

The ADF test is used to verify stationarity of the spread:

Test Regression:

$$\Delta\mathrm{Spread}_t = \alpha + \gamma \cdot \mathrm{Spread}_{t-1} + \sum_{i=1}^{p} \delta_i \Delta\mathrm{Spread}_{t-i} + \epsilon_t \tag{21}$$

Null Hypothesis: $H_0 : \gamma = 0$ (unit root exists, non-stationary)

Alternative Hypothesis: $H_1 : \gamma < 0$ (stationary)
Significance Level: $\alpha = 0.05$ (5% significance)

A p-value below 0.05 provides evidence to reject the null hypothesis and conclude that the spread is stationary.

3.3 Distance Metrics

For clustering and similarity analysis, correlation-based distance is used:

$$d(i, j) = 1 - |\rho(r_i, r_j)| \tag{22}$$

This distance metric ranges from 0 (perfect correlation) to 2 (perfect negative correlation).

4 Machine Learning Models

4.1 Walk-Forward Cross-Validation

To prevent look-ahead bias, a walk-forward validation scheme is employed:

- **Number of Splits:** 5 time-series splits
- **Method:** TimeSeriesSplit (scikit-learn)
- **Property:** Each test set uses only data following the training set chronologically

4.2 Target Variable Construction

4.2.1 Forward 14-Day Return (Alpha Target)

$$y_t^{\text{return}} = \frac{\text{Close}_{t+14}}{\text{Close}_t} - 1 \tag{23}$$

This represents the 14-day forward return, which the model aims to predict.

4.2.2 Forward 14-Day Volatility (Risk Target)

$$y_t^{\text{risk}} = \sigma_{14}(\text{PctChange}_{t\to t+14}) \tag{24}$$

where σ_{14} denotes the standard deviation over the 14-day forward window.

4.3 XGBoost Model Configuration

Two separate XGBoost regressors are trained:

1. **Alpha Model:** Predicts forward returns
2. **Risk Model:** Predicts forward volatility

Hyperparameters:

- `n_estimators`: 500 (number of boosting rounds)
- `learning_rate`: 0.01 (shrinkage to prevent overfitting)
- `max_depth`: 3 (maximum tree depth)
- `subsample`: 0.7 (row sampling ratio)
- `colsample_bytree`: 0.7 (feature sampling ratio)
- `objective`: `reg:squarederror` (L2 loss)
- `n_jobs`: -1 (parallel processing)

4.4 Loss Function

XGBoost minimizes the squared error loss:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{k=1}^{K}\Omega(f_k) \tag{25}$$

where:

- y_i = True target value
- $\hat{y}_i$ = Predicted value
- $\Omega(f_k)$ = Regularization term for tree k
- K = Number of trees

4.5 Signal Engineering

4.5.1 Raw Score Construction

The raw trading score combines predicted return and predicted risk:

$$\text{RawScore}_t = \frac{\text{pred\_return}_t}{\text{pred\_risk}_t + \epsilon} \tag{26}$$

where $\epsilon = 10^{-6}$ prevents division by zero.

4.5.2 Signal Smoothing

A 3-day moving average smooths the raw scores:

$$\text{TradeScore}_t = \frac{1}{3}\sum_{i=0}^{2}\text{RawScore}_{t-i} \tag{27}$$

4.5.3 Inverse Volatility Weighting

Position weights are inversely proportional to predicted risk:

$$w_t^{\text{inv\_vol}} = \frac{1}{\text{pred\_risk}_t + \epsilon_w} \tag{28}$$

where $\epsilon_w = 10^{-4}$ ensures numerical stability.

5 Backtesting Framework

5.1 Portfolio Construction

5.1.1 Asset Selection

At each rebalancing date:

1. Rank all assets by `TradeScore`
2. Select top $N = 10$ assets with score above threshold $\tau = 1.0$
3. Apply sticky rank logic to reduce turnover

5.1.2 Position Sizing

Step 1 - Normalize Inverse Volatility Weights:

$$w_i = \frac{w_i^{\text{inv\_vol}}}{\sum_{j \in \text{TopN}} w_j^{\text{inv\_vol}}} \tag{29}$$

Step 2 - Calculate Position Values:

$$\text{PositionValue}_i = \text{PortfolioValue} \times w_i \tag{30}$$

Step 3 - Determine Shares:

$$\text{Shares}_i = \left\lfloor \frac{\text{PositionValue}_i}{\text{Price}_i} \right\rfloor \tag{31}$$

5.2 Transaction Cost Modeling

Cost Components:

- Bid-Ask Spread: 10 basis points (bps) $= 0.001 = 0.1\%$
- Slippage: Implicit in spread
- Commission: Variable (implementation-dependent)

Transaction Cost per Trade:

$$\text{Cost}_{\text{trade}} = |\text{TradeValue}| \times c \tag{32}$$

where $c = 0.001$ (10 bps).
Total Daily Costs:

$$\text{Cost}_t = \sum_{i \in \text{Trades}} |\text{TargetValue}_i - \text{CurrentValue}_i| \times c \tag{33}$$

5.3 Return Calculations

5.3.1 Daily Portfolio Return

$$r_t = \frac{\text{PV}_t - \text{PV}_{t-1} - \text{Cost}_t}{\text{PV}_{t-1}} \tag{34}$$

where PV_t is the portfolio value at time t.

5.3.2 Cumulative Return

$$R_{\text{total}} = \frac{\text{PV}_{\text{final}}}{\text{PV}_{\text{initial}}} - 1 \tag{35}$$

5.3.3 Annualized Return

$$R_{\text{annual}} = (1 + R_{\text{total}})^{\frac{252}{N_{\text{days}}}} - 1 \tag{36}$$

where N_{days} is the number of trading days in the backtest period.

5.4 Risk Metrics

5.4.1 Sharpe Ratio

The annualized Sharpe ratio measures risk-adjusted returns:

$$\text{Sharpe} = \frac{\mu_r - r_f}{\sigma_r} \times \sqrt{252} \tag{37}$$

where:

- μ_r = Mean daily return
- r_f = Risk-free rate (typically assumed to be 0 for simplicity)
- σ_r = Standard deviation of daily returns
- 252 = Annualization factor (trading days per year)

5.4.2 Sortino Ratio

The Sortino ratio penalizes only downside volatility:

$$\text{Sortino} = \frac{\mu_r - r_{\text{target}}}{\sigma_{\text{downside}}} \times \sqrt{252} \tag{38}$$

where:

$$\sigma_{\text{downside}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \min(r_i - r_{\text{target}}, 0)^2} \tag{39}$$

5.4.3 Maximum Drawdown

The maximum drawdown measures the largest peak-to-trough decline:

$$\text{MDD} = \min_{t \in [0,T]} \left(\frac{\text{PV}_t}{\max_{s \in [0,t]} \text{PV}_s} - 1 \right) \tag{40}$$

5.4.4 Average Drawdown

$$\text{AvgDD} = \frac{1}{N_{\text{dd}}} \sum_{i=1}^{N_{\text{dd}}} \text{DD}_i \tag{41}$$

where N_{dd} is the number of periods in drawdown.

5.5 Performance Attribution

5.5.1 Transaction Cost Impact

$$\text{CostImpact} = R_{\text{no\_costs}} - R_{\text{with\_costs}} \tag{42}$$

5.5.2 Benchmark Comparison

Outperformance:

$$\text{Outperf} = R_{\text{strategy}} - R_{\text{benchmark}} \tag{43}$$

Information Ratio:

$$\text{IR} = \frac{\mu_{r_s - r_b}}{\sigma_{r_s - r_b}} \times \sqrt{252} \tag{44}$$

where r_s and r_b are strategy and benchmark returns respectively.

6 Statistical Tests and Hypothesis Testing

6.1 Augmented Dickey-Fuller (ADF) Test

Purpose: Test for unit root (non-stationarity)
Null Hypothesis: H_0: Series has a unit root (non-stationary)
Alternative Hypothesis: H_1: Series is stationary
Test Statistic:

$$\text{ADF} = \frac{\hat{\gamma}}{\text{SE}(\hat{\gamma})} \tag{45}$$

Decision Rule: Reject H_0 if p-value < 0.05

6.2 Johansen Cointegration Test

Purpose: Test for cointegration between multiple time series
Null Hypothesis: H_0: No cointegrating relationship
Alternative Hypothesis: H_1: At least one cointegrating vector exists
Test Statistics:

- Trace statistic
- Maximum eigenvalue statistic

6.3 Correlation Hypothesis Test

Null Hypothesis: H_0: $\rho = 0$ (no linear relationship)
Test Statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{46}$$

where r is the sample correlation and n is the sample size.

6.4 Sharpe Ratio Significance Test

Null Hypothesis: H_0: Strategy Sharpe = Benchmark Sharpe
Test Statistic (under independence):

$$z = \frac{\text{SR}_{\text{strategy}} - \text{SR}_{\text{benchmark}}}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{47}$$

7 Strategy Summary

7.1 Overall Trading Strategy

The implemented strategy is a **long-only quantitative equity strategy** based on machine learning predictions of future returns and risks.

7.1.1 Strategy Components

1. **Universe:** 100 anonymized assets
2. **Signal Generation:** XGBoost models predict 14-day forward returns and volatility
3. **Position Sizing:** Inverse volatility weighting (risk parity approach)
4. **Portfolio Construction:** Top 10 assets by risk-adjusted score
5. **Rebalancing:** Daily with sticky rank to reduce turnover
6. **Risk Management:** Diversification + volatility targeting

7.1.2 Key Strategy Parameters

Parameter	Value	Purpose
Initial Capital	\$1,000,000	Starting portfolio value
Transaction Cost	10 bps	Round-trip trading cost
Rebalance Frequency	Daily	Portfolio adjustment
Top N Assets	10	Number of positions
Score Threshold	1.0	Minimum signal strength
Smoothing Window	3 days	Signal noise reduction
Sticky Rank Buffer	Top 15	Turnover reduction

Table 2: Strategy Parameters

7.2 Alternative Strategies Explored

7.2.1 Pairs Trading Strategy

Based on cointegration analysis in the similarity-checking notebook:

Entry Conditions:

$$\text{Enter Long Spread if: } \text{Spread}_t < \mu_{\text{spread}} - 2\sigma_{\text{spread}} \tag{48}$$

$$\text{Enter Short Spread if: } \text{Spread}_t > \mu_{\text{spread}} + 2\sigma_{\text{spread}} \tag{49}$$

Exit Conditions:

$$\text{Exit if: } |\text{Spread}_t - \mu_{\text{spread}}| < 0.5\sigma_{\text{spread}} \tag{50}$$

This mean-reversion strategy trades the spread when it deviates significantly from its long-term mean.

8 Model Inputs and Configuration Summary

9 Performance Metrics and Expected Results

9.1 Target Performance Metrics

Based on backtesting over a 2-year period:

9.2 Strategy Viability Criteria

The strategy is considered viable for potential live trading if:

1. Sharpe Ratio > 0 AND $R_{\text{total}} > 0$
2. $R_{\text{strategy}} > R_{\text{benchmark}}$ (outperforms equal-weight)
3. $|\text{MDD}| < 20\%$ (manageable drawdown)
4. Positive Sharpe ratio after transaction costs
5. Statistically significant outperformance (t-test or bootstrap)

10 Mathematical Notation Reference

10.1 General Notation

- t = Time index
- i, j = Asset indices
- n, N = Sample size or number of observations
- μ = Mean
- σ = Standard deviation
- ρ = Correlation coefficient
- ϵ = Error term or small constant
- $\mathcal{K}(\cdot)$ = Kalman filter operator
- Δ = First difference operator
- ln = Natural logarithm

10.2 Financial Variables

- S_t = Asset price at time t
- r_t = Return at time t
- σ_t = Volatility at time t
- w_i = Weight of asset i in portfolio
- PV_t = Portfolio value at time t
- H_t, L_t, O_t, C_t = High, Low, Open, Close prices

10.3 Model Components

- $\hat{y}$ = Predicted value
- β = Regression coefficient / hedge ratio
- α = Intercept / significance level
- K_t = Kalman gain
- Q, R = Process and measurement noise covariances
- f_k = Tree/function in ensemble model

11 References and Further Reading

11.1 Technical Indicators

- Wilder, J. W. (1978). *New Concepts in Technical Trading Systems.* Trend Research.
- Rogers, L. C. G., & Satchell, S. E. (1991). Estimating Variance From High, Low and Closing Prices. *The Annals of Applied Probability*, 1(4), 504-512.

11.2 Kalman Filtering

- Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1), 35-45.
- Labbe, R. (2015). *Kalman and Bayesian Filters in Python.* GitHub: rlabbe/filterpy.

11.3 Cointegration and Pairs Trading

- Engle, R. F., & Granger, C. W. J. (1987). Co-integration and Error Correction: Representation, Estimation, and Testing. *Econometrica*, 55(2), 251-276.
- Johansen, S. (1991). Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models. *Econometrica*, 59(6), 1551-1580.

11.4 Machine Learning for Finance

- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD.*
- De Prado, M. L. (2018). *Advances in Financial Machine Learning.* Wiley.
- De Prado, M. L. (2020). *Machine Learning for Asset Managers.* Cambridge University Press.

11.5 Portfolio Construction and Backtesting

- Sharpe, W. F. (1994). The Sharpe Ratio. *Journal of Portfolio Management*, 21(1), 49-58.
- Sortino, F. A., & Price, L. N. (1994). Performance Measurement in a Downside Risk Framework. *Journal of Investing*, 3(3), 59-64.
- Bailey, D. H., & Lopez de Prado, M. (2014). The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. *Journal of Portfolio Management*, 40(5), 94-107.

11.6 Software and Libraries

- Statsmodels: Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python.
- XGBoost Documentation: `https://xgboost.readthedocs.io/`
- PyKalman: `https://pykalman.github.io/`
- Pandas TA: `https://github.com/twopirllc/pandas-ta`

12 Conclusion

This document provides a comprehensive mathematical foundation for the quantitative trading research platform. The framework combines:

- **Signal Processing:** Kalman filtering for noise reduction
- **Technical Analysis:** RSI, Rogers-Satchell volatility, and derived features
- **Statistical Arbitrage:** Cointegration-based pairs trading
- **Machine Learning:** XGBoost for alpha and risk prediction
- **Portfolio Management:** Inverse volatility weighting and risk parity
- **Realistic Backtesting:** Transaction costs, slippage, and market impact

The mathematical rigor and systematic approach ensure that the strategy is:

1. **Reproducible:** All calculations are clearly defined
2. **Testable:** Hypotheses can be statistically validated
3. **Robust:** Walk-forward validation prevents overfitting
4. **Realistic:** Transaction costs are explicitly modeled

Future enhancements could include:

- Alternative machine learning models (LSTM, Transformer networks)
- Multi-factor risk models (Fama-French, Carhart)
- Regime detection and adaptive strategies
- Options strategies for hedging and income generation
- High-frequency microstructure analysis

Disclaimer: This documentation is for educational and research purposes only. Past performance does not guarantee future results. Trading involves substantial risk of loss. Consult a qualified financial advisor before making any investment decisions.

Component	Parameter	Value
Technical Indicators		
	RSI Window	14 days
	RS Volatility Window	14 days
	Kalman Q (Process Noise)	0.01
	Kalman R (Measurement)	1.0
Feature Engineering		
	Return Lookback	14 days
	Volume Z-score Window	14 days
	RSI Derivatives	1st & 2nd order
Cointegration Analysis		
	Correlation Threshold	0.90
	ADF Significance Level	0.05
	Log-Price Transform	Yes
Machine Learning		
	Prediction Horizon	14 days
	XGBoost Trees	500
	Learning Rate	0.01
	Max Tree Depth	3
	Subsample Ratio	0.7
	Feature Sampling	0.7
	CV Splits	5 (time-series)
Signal Processing		
	Signal Smoothing	3-day MA
	Risk Adjustment Epsilon	10^{-6}
	Weight Scaling Epsilon	10^{-4}
Backtesting		
	Rebalance Frequency	Daily
	Transaction Cost	10 bps
	Position Count	10 assets
	Initial Capital	$1,000,000
	Backtest Duration	2 years

Table 3: Complete Model Configuration

Metric	**Expected Range**	**Target**
Total Return	15-25%	¿20%
Annualized Return	7-12%	¿10%
Sharpe Ratio	0.8-1.2	¿1.0
Maximum Drawdown	-8% to -15%	¡-10%
Win Rate	45-55%	¿50%
Transaction Costs	1-2% of returns	¡1.5%

Table 4: Expected Performance Metrics