Quantitative Portfolio Optimization

A Regime-Adaptive Framework Integrating AI Signals and Multi-Strategy Optimization

Silvio Baratto

October 25, 2025

Abstract

This document presents a comprehensive quantitative portfolio optimization platform that integrates financial data acquisition, AI-powered stock signal generation, macroeconomic regime analysis, and multi-strategy portfolio construction. The system combines institutional-grade methodologies with modern technology infrastructure to deliver robust, adaptive investment strategies suitable for professional portfolio management.

Table of Contents

1 Introduction

This work documents the design, implementation, and methodological foundations of a quantitative portfolio optimization framework built for institutional investment applications. The framework addresses the complete investment workflow from universe construction through portfolio optimization, integrating multiple data sources, artificial intelligence, and advanced optimization algorithms.

The system architecture reflects best practices from leading quantitative asset management firms, incorporating:

- **Multi-source data integration** with robust quality assurance and filtering
- **AI-enhanced analysis** using large language models for signal generation and regime classification
- **Multi-strategy optimization** including Black-Litterman, risk parity, and hierarchical risk parity approaches
- **Adaptive frameworks** that adjust to macroeconomic regimes and market conditions
- **Production-grade infrastructure** with comprehensive error handling and monitoring

Each chapter examines a specific component of the platform, providing theoretical foundations, implementation details, validation approaches, and practical considerations for institutional deployment.

2 Building the Universe

The construction of a robust investment universe forms the foundation of any quantitative portfolio optimization system. This chapter describes the methodology employed to build a comprehensive, high-quality dataset of financial instruments suitable for institutional-grade portfolio analysis. The process integrates multiple data sources, applies rigorous filtering criteria, and implements efficient data enrichment techniques to ensure the resulting universe meets institutional standards for liquidity, data completeness, and historical coverage.

2.1 Conceptual Framework

The universe construction process can be conceptualized as a multi-stage filtering pipeline that progressively refines a broad set of available securities into a focused collection of investment-grade instruments. The methodology draws inspiration from institutional investment practices, where asset selection begins with establishing minimum standards for market capitalization, liquidity, and data availability before proceeding to more sophisticated quantitative analysis.

The universe building methodology addresses three fundamental challenges:

1. **Data Integration**: Harmonizing instrument identifiers across disparate data providers
2. **Quality Assurance**: Establishing and enforcing minimum standards for inclusion
3. **Scalability**: Processing thousands of instruments efficiently while maintaining data integrity

2.2 Data Sources and Geographic Scope

2.2.1 Primary Data Providers

The system integrates two complementary data sources to achieve comprehensive coverage:

Trading212 API serves as the primary source for instrument metadata, providing access to thousands of publicly traded equities across major global exchanges. The API delivers structured information including ticker symbols, ISIN codes, exchange affiliations, and trading availability status. This source ensures the universe reflects instruments actually available for trading through modern brokerage platforms.

Yahoo Finance (yfinance) functions as the enrichment layer, augmenting basic instrument metadata with comprehensive fundamental, technical, and financial data. This includes historical price series, financial statement data, valuation ratios, profitability metrics, and analyst coverage. The depth of yfinance data enables rigorous filtering and subsequent quantitative analysis.

2.2.2 Geographic Coverage

The investment universe targets developed and emerging markets aligned with institutional geographic allocation frameworks. The selected countries reflect both economic significance and data availability:

- **United States**: 55-65% target allocation, representing the deepest and most liquid equity markets with superior data quality
- **Europe**: 15-20% aggregate allocation across Germany, France, and United Kingdom
- **Japan**: 8-12% allocation, capturing Asia-Pacific developed market exposure
- **Emerging Markets**: 6-8% allocation to China and India where data availability permits

This geographic distribution aligns with standard institutional benchmarks while acknowledging data provider constraints that limit comprehensive coverage of certain emerging markets.

2.3 Ticker Mapping Methodology

2.3.1 The Identifier Reconciliation Problem

A fundamental challenge in multi-source financial data integration involves reconciling divergent ticker symbol conventions across data providers. Trading212 employs exchange-specific ticker formats, while Yahoo Finance utilizes a distinct suffix notation system to denote exchange affiliation. For example, a stock listed on the London Stock Exchange might appear as "VOD" in Trading212 but requires the suffix ".L" (VOD.L) for yfinance queries.

2.3.2 Systematic Ticker Discovery

The ticker mapping process implements a systematic discovery algorithm that attempts to identify the correct yfinance ticker for each Trading212 instrument. The procedure follows a prioritized search strategy:

1. **Cache Lookup**: Check persistent mapping cache for previously validated ticker pairs
2. **Exchange-Specific Mapping**: Apply known suffix rules based on exchange affiliation (e.g., ".L" for London Stock Exchange, ".PA" for Euronext Paris, no suffix for US exchanges)
3. **Symbol Normalization**: Transform ticker format conventions (e.g., slash-to-dash conversion for share classes)
4. **Validation**: Verify mapped ticker returns valid market data from yfinance

The mapping cache reduces redundant API calls and improves processing efficiency. Cached mappings are time-limited (90-day expiration) to ensure stale mappings do not persist indefinitely.

2.3.3 Rate Limiting and Error Handling

To maintain system stability and comply with API usage policies, the ticker discovery process implements adaptive rate limiting and exponential backoff strategies. The rate limiter enforces a maximum query frequency of 5 requests per second, while the backoff mechanism progressively increases wait times when encountering rate limit errors:

$$t_{\text{wait}} = \min(2^n \cdot t_{\text{base}}, t_{\text{max}})$$

where n represents the retry attempt number, t_{base} is the base wait interval, and $t_{\max}$ caps the maximum wait time at one hour. This approach balances throughput with reliability.

2.4 Institutional Filtering Framework

2.4.1 Market Capitalization Tiers

The universe employs a tiered market capitalization classification that segments stocks into three categories, each with distinct liquidity requirements:

$$\text{Segment}(i) = \begin{cases} \text{Large-cap} & \text{if MarketCap}_i \geq \$10\text{B} \\ \text{Mid-cap} & \text{if } \$2\text{B} \leq \text{MarketCap}_i < \$10\text{B} \\ \text{Small-cap} & \text{if } \$100\text{M} \leq \text{MarketCap}_i < \$2\text{B} \end{cases}$$

The absolute minimum market capitalization threshold of $100 million excludes micro-cap stocks that typically exhibit insufficient liquidity and elevated transaction costs for institutional portfolios.

2.4.2 Liquidity Requirements

Liquidity filtering applies segment-specific thresholds that recognize the relationship between firm size and trading volume. The daily dollar volume requirement for stock i is defined as:

$$\text{DollarVolume}_i = \text{AverageVolume}_i \times \text{Price}_i$$

where AverageVolume_i represents the mean daily share volume over the preceding period. The minimum thresholds vary by market capitalization segment:

Segment	Minimum Dollar Volume	Minimum Share Volume
Large-cap	$10,000,000	500,000 shares
Mid-cap	$5,000,000	250,000 shares
Small-cap	$1,000,000	100,000 shares

This tiered structure ensures that portfolio construction algorithms can execute realistic position sizes without excessive market impact, while avoiding overly restrictive criteria that would eliminate otherwise suitable smaller-capitalization opportunities.

2.4.3 Price Filters

To exclude problematic securities, the system enforces price bounds:

$$\$5 \leq \text{Price}_i \leq \$10,000$$

The lower bound eliminates penny stocks that exhibit elevated bid-ask spreads and susceptibility to manipulation. The upper bound serves as a data quality check, flagging potential data errors or corporate actions not yet reflected in databases.

2.4.4 Data Completeness Requirements

Institutional quantitative analysis requires comprehensive fundamental and technical data across multiple dimensions. The system defines eleven data categories, each comprising specific required fields:

1. **Market Capitalization**: Total market value
2. **Current Price**: Latest trading price
3. **Volume Metrics**: Average daily trading volume
4. **Share Structure**: Outstanding shares count
5. **Risk Metrics**: Market beta
6. **Classification**: GICS sector and industry
7. **Exchange Information**: Primary listing venue
8. **Valuation Ratios**: P/E ratio, Price-to-Book ratio
9. **Profitability**: Return on Equity, operating margins
10. **Balance Sheet**: Debt ratios, total assets
11. **Price Range**: 52-week high and low

A stock passes the data completeness filter only if 100% of required categories contain valid, non-null data. This stringent requirement ensures downstream analysis does not encounter missing data issues that could compromise signal quality or introduce bias.

2.4.5 Historical Data Coverage

To support robust statistical estimation and backtesting, the system requires minimum historical price coverage. The historical data criterion is defined as:

$$N_{\text{trading days}} \geq 750 \text{ days}$$

This threshold corresponds to approximately three years of trading history (assuming 252 trading days per year), providing sufficient observations for:

- Long-term momentum pattern identification
- Risk parameter estimation (beta, volatility)
- Fundamental trend analysis
- Statistical significance testing

The implementation requests five years of historical data from yfinance (period='5y') and validates that at least 750 days are returned, accommodating potential gaps from market holidays, trading halts, or data provider limitations.

2.5 Pipeline Architecture and Processing Strategy

2.5.1 Sequential Exchange Processing with Concurrent Enrichment

The universe construction pipeline employs a hybrid processing strategy that balances computational efficiency with system stability. Processing proceeds sequentially across exchanges but leverages concurrent workers within each exchange to parallelize instrument enrichment.

For each exchange E, let $I_E = \{i_1, i_2, \ldots, i_n\}$ represent the set of instruments listed on that exchange. The enrichment process distributes instruments across k concurrent worker threads:

$$\text{Throughput} = \frac{k}{t_{\text{avg}}}$$

where t_{avg} represents the average processing time per instrument, including API latency and data validation. Empirical testing established $k = 20$ concurrent workers as optimal, achieving 20-50 instruments per second throughput while maintaining acceptable error rates.

2.5.2 Short-Lived Database Sessions

To mitigate SSL timeout issues inherent in long-running database connections, the system employs short-lived database sessions scoped to specific operations:

1. **Exchange Creation**: New session per exchange for metadata insertion
2. **Batch Insert**: New session per batch (50-500 instruments) for bulk saves
3. **Reporting**: New session for final statistics generation

This session management strategy trades slight overhead for improved reliability, particularly important when processing thousands of instruments over extended periods.

2.5.3 Batch Insertion for Performance

Rather than committing each instrument individually, the pipeline accumulates processed instruments into batches before database persistence:

$$N_{\text{batches}} = \left\lceil \frac{N_{\text{instruments}}}{B} \right\rceil$$

where B represents the batch size (typically 50-500 instruments). Batching reduces transaction overhead and database round-trips, significantly improving overall throughput.

2.6 Quality Assurance and Validation

2.6.1 Multi-Stage Validation

Each instrument undergoes validation at multiple pipeline stages:

1. **Ticker Discovery Validation**: Verify yfinance returns valid market data for mapped ticker

2. **Data Fetch Validation**: Confirm fundamental data response contains minimum field count
3. **Filter Validation**: Apply all institutional filters with explicit pass/fail determination
4. **Historical Data Validation**: Confirm sufficient trading history availability

Only instruments passing all validation stages enter the final universe. Failed instruments are logged with specific rejection reasons to support pipeline debugging and filter refinement.

2.6.2 Data Quality Metrics

The system tracks several quality metrics throughout the processing pipeline:

Mapping Success Rate: The proportion of Trading212 instruments successfully mapped to yfinance tickers, typically achieving 85-95% depending on exchange and asset type.

Filter Pass Rate: The proportion of successfully mapped instruments that satisfy all institutional filters. This rate varies significantly by exchange, with major US exchanges (NYSE, NASDAQ) exhibiting higher pass rates (40-60%) than smaller international venues (10-30%).

Data Completeness Distribution: Histogram of data coverage across instruments, identifying systematic data gaps by exchange or sector.

2.7 Performance Characteristics and Scalability

2.7.1 Processing Throughput

Under typical operating conditions with 20 concurrent workers and stable API performance, the system achieves:

- **Ticker Mapping**: 30-50 instruments/second
- **Data Enrichment**: 20-40 instruments/second (including filter validation)
- **Total Processing Time**: 10-15 minutes for 1,000 instruments

These rates fluctuate based on network latency, API response times, and data provider rate limiting.

2.7.2 Cache Efficiency

The persistent ticker mapping cache significantly improves performance on subsequent runs. Cache hit rates typically reach 60-70% after initial universe construction, reducing redundant API calls and accelerating incremental updates.

2.7.3 Resource Utilization

The concurrent architecture achieves high CPU utilization during processing phases, while I/O wait time dominates overall execution due to network API calls. Memory usage remains modest (< 2GB) even for large universes, as the pipeline processes instruments in streaming fashion rather than loading entire datasets into memory.

2.8 Challenges and Limitations

2.8.1 Ticker Mapping Ambiguity

Certain instruments present mapping challenges:

- **Multiple Share Classes**: Companies with multiple share classes (e.g., Class A vs. Class B) require careful suffix handling
- **ADR vs. Ordinary Shares**: American Depositary Receipts may map to either ADR tickers or underlying ordinary shares
- **Recent Corporate Actions**: Ticker changes, mergers, and spin-offs may cause temporary mapping failures

The system handles these cases through manual override mappings and periodic cache validation.

2.8.2 Data Provider Inconsistencies

Discrepancies occasionally arise between Trading212 and yfinance data:

- **Stale Fundamental Data**: yfinance may report outdated fundamentals for thinly traded stocks
- **Exchange Mismatches**: Primary listing venue may differ between providers
- **Currency Reporting**: Inconsistent currency normalization across providers

These issues are addressed through data prioritization hierarchies and conflict logging for manual review.

2.8.3 API Rate Limiting

Both Trading212 and yfinance impose rate limits that constrain processing throughput. The exponential backoff strategy mitigates rate limit errors but can extend total processing time significantly when limits are repeatedly encountered.

2.8.4 Geographic Coverage Gaps

The reliance on Trading212 as the primary instrument source limits coverage to exchanges supported by that platform. Notably absent:

- **Direct China A-Shares**: Only available through ADRs or Hong Kong listings
- **India Local Markets**: Limited to large-cap names with international listings
- **Smaller Emerging Markets**: Minimal coverage of Southeast Asia, Latin America, Africa

Future enhancements could integrate additional data sources to expand geographic reach.

2.9 Summary

The universe construction methodology successfully builds a high-quality dataset of investment-grade equities suitable for institutional portfolio optimization. The multi-stage filtering framework

ensures all instruments meet minimum standards for market capitalization, liquidity, data completeness, and historical coverage. The parallel processing architecture achieves acceptable throughput while maintaining system stability through rate limiting and short-lived database sessions.

The resulting universe typically contains 500-1,500 instruments (depending on market conditions and filter strictness), representing liquid, well-documented securities across developed and select emerging markets. This filtered universe forms the foundation for subsequent macro regime analysis and stock signal generation, as described in the following chapters.

3 Macroeconomic Regime Analysis

Business cycle regime identification represents a critical component of adaptive portfolio management strategies. Asset class returns, factor performance, and risk characteristics exhibit substantial variation across different phases of the economic cycle. This chapter presents a comprehensive methodology for classifying macroeconomic regimes through the integration of multiple data sources, fundamental economic indicators, and artificial intelligence-powered analysis. The approach synthesizes quantitative economic data with qualitative news analysis to produce robust regime classifications that inform portfolio construction and risk management decisions.

3.1 Theoretical Foundation

3.1.1 Business Cycle Framework

The business cycle describes the fluctuating pattern of economic expansion and contraction observed across developed economies. Following the National Bureau of Economic Research (NBER) dating methodology and institutional investment frameworks, economic cycles can be partitioned into four distinct regimes, each characterized by unique combinations of growth momentum, inflation dynamics, and monetary policy posture:

Early Cycle regimes emerge following recessions, exhibiting accelerating growth from depressed levels, accommodative monetary policy, low but rising inflation, and improving labor market conditions. Financial assets typically respond favorably as earnings growth accelerates and risk premiums compress.

Mid Cycle regimes represent sustained expansion phases characterized by moderate, stable growth, normalized monetary policy, contained inflation near central bank targets, and healthy labor markets. This regime typically exhibits the longest duration and most balanced risk-return profiles.

Late Cycle regimes occur as expansions mature, showing decelerating growth from elevated levels, tightening monetary policy, rising inflation pressures, and capacity constraints in labor and product markets. Equity valuations may compress as the risk of recession increases.

Recession regimes feature contracting economic activity, accommodative monetary policy (often dramatically eased), falling inflation, and deteriorating labor markets. Risk asset returns typically suffer while defensive assets provide portfolio protection.

The regime classification problem can be formalized as a mapping function:

$$\mathcal{R} : \mathcal{I} \times \mathcal{M} \times \mathcal{N} \rightarrow \{\text{Early}, \text{Mid}, \text{Late}, \text{Recession}\}$$

where $\mathcal{I}$ represents the space of economic indicators, $\mathcal{M}$ denotes market-based signals, and $\mathcal{N}$ captures news and sentiment information.

3.1.2 Investment Implications

Different regimes exhibit characteristic asset class performance patterns documented in both academic literature and institutional investment practice:

Sector Rotation: Cyclical sectors (financials, industrials, consumer discretionary) typically outperform during early cycle regimes, while defensive sectors (utilities, consumer staples, healthcare) provide relative safety during recessions. Late cycle regimes often favor commodity-sensitive sectors as inflation accelerates.

Factor Performance: Momentum and growth factors tend to perform strongly in early-to-mid cycle regimes. Value factors often outperform in late cycle transitions. Quality and low volatility factors provide defensive characteristics during recessions.

Duration Management: Bond duration performs well in recessions as yields fall, while shorter-duration positioning becomes prudent in late cycle regimes as central banks tighten policy.

3.2 Data Integration Framework

3.2.1 Multi-Source Economic Data

The regime classification methodology integrates three complementary data sources to achieve comprehensive coverage of economic conditions:

Il Sole 24 Ore Economic Indicators provide country-specific fundamental economic data across major developed economies. This source delivers both realized (current state) and forecast (forward-looking) indicators:

Realized indicators capture the current economic state through GDP growth (quarter-over-quarter), industrial production indices, unemployment rates, consumer price inflation, fiscal metrics (budget deficit, public debt ratios), and sovereign interest rates (short-term and long-term). These backward-looking measures establish the baseline economic environment.

Forecast indicators incorporate forward-looking expectations including 6-month inflation forecasts, 6-month GDP growth projections, 12-month corporate earnings growth estimates, expected earnings per share, PEG ratios, and interest rate forecasts. The integration of forecast data acknowledges that financial markets are forward-looking and regime classifications should incorporate market expectations alongside realized outcomes.

Federal Reserve Economic Data (FRED) supplies high-frequency market-based indicators that reflect real-time risk sentiment and credit conditions. Key indicators include:

- **VIX (CBOE Volatility Index)**: Implied volatility of S&P 500 index options, serving as a "fear gauge" for equity market uncertainty
- **High-Yield Credit Spread**: The yield differential between high-yield corporate bonds and comparable-maturity Treasury securities, measuring credit risk premiums

These market-based indicators complement fundamental economic data by capturing investor risk perception and liquidity conditions that may precede changes in real economic activity.

Trading Economics Platform provides comprehensive real-time economic data including manufacturing PMI indices, government bond yields across maturities, industrial production statistics, and capacity utilization rates. Notably, this source delivers:

Manufacturing Purchasing Managers' Index (PMI), a diffusion index surveying purchasing managers on new orders, production, employment, and inventories. The critical threshold of 50 delineates expansion (>50) from contraction (<50), making PMI a leading indicator of manufacturing sector health.

Government bond yields enable construction of the yield curve, particularly the 10-year minus 2-year spread, a historically reliable recession indicator when inverted (negative spread).

3.2.2 Geographic Coverage

The analysis encompasses five major developed economies aligned with the portfolio country allocation framework:

- **United States**: Largest global economy, benchmark for global financial conditions
- **Germany**: Europe's largest economy, proxy for Eurozone economic health
- **France**: Major European economy, complementary to German data
- **United Kingdom**: Important European economy with independent monetary policy
- **Japan**: Asia-Pacific developed market representation

This geographic scope captures approximately 80-85% of the target portfolio's country exposure, ensuring regime classifications reflect conditions in markets where the portfolio maintains significant allocations.

3.3 Economic Indicators and Signal Construction

3.3.1 Yield Curve Analysis

The term structure of interest rates, particularly the spread between long-term and short-term government bond yields, provides powerful regime classification information. The 10-year minus 2-year yield spread can be formalized as:

$$S_{10s2s,t} = y_{10Y,t} - y_{2Y,t}$$

where $y_{10Y,t}$ and $y_{2Y,t}$ represent the 10-year and 2-year government bond yields at time t, respectively. The spread is conventionally expressed in basis points:

$$S^{\text{bps}}_{10s2s,t} = (y_{10Y,t} - y_{2Y,t}) \times 10,000$$

The economic interpretation draws from expectations theory and term premium decomposition. A positive, steep yield curve ($S_{10s2s,t} > 100$ bps) typically accompanies early cycle regimes as markets anticipate future growth and inflation. Flattening curves ($0 < S_{10s2s,t} < 50$ bps) suggest late cycle conditions as the central bank raises short-term rates. Yield curve inversion ($S_{10s2s,t} < 0$) has preceded every U.S. recession in the past 50 years, typically with a 6-18 month lead time.

3.3.2 Manufacturing Activity Indicators

The Manufacturing Purchasing Managers' Index serves as a composite leading indicator of economic activity. As a diffusion index, PMI values above 50 indicate expansion while values below 50 signal contraction. The regime signal can be characterized as:

$$\text{PMI Signal}_t = \begin{cases} \text{Strong Expansion} & \text{if PMI}_t > 55 \\ \text{Moderate Expansion} & \text{if } 50 < \text{PMI}_t \leq 55 \\ \text{Moderate Contraction} & \text{if } 45 \leq \text{PMI}_t \leq 50 \\ \text{Sharp Contraction} & \text{if PMI}_t < 45 \end{cases}$$

PMI dynamics provide additional information. Rapidly rising PMI from low levels suggests early cycle conditions, while declining PMI from elevated levels indicates late cycle deterioration.

3.3.3 Credit Market Indicators

Credit spreads measure the additional yield investors demand for bearing default risk relative to risk-free government bonds. The high-yield spread serves as a barometer of corporate credit stress:

$$\text{HY Spread}_t = y_{\text{HY},t} - y_{\text{Treasury},t}$$

where $y_{\text{HY},t}$ represents the yield on a high-yield corporate bond index and $y_{\text{Treasury},t}$ is the yield on comparable-maturity Treasury securities.

Credit spreads exhibit characteristic patterns across regimes. Narrow spreads (< 400 bps) indicate investor confidence and abundant liquidity, typical of mid cycle regimes. Widening spreads (> 600 bps) reflect heightened default concerns and risk aversion, often accompanying late cycle transitions or recessions. Extreme widening (> 1000 bps) suggests acute financial stress.

3.3.4 Volatility Regime Classification

The VIX index quantifies market expectations of 30-day volatility implied by S&P 500 index option prices. While not a direct economic indicator, VIX levels provide information about investor uncertainty and risk appetite:

$$\text{VIX Regime}_t = \begin{cases} \text{Complacency} & \text{if VIX}_t < 15 \\ \text{Normal} & \text{if } 15 \leq \text{VIX}_t \leq 20 \\ \text{Elevated} & \text{if } 20 < \text{VIX}_t \leq 30 \\ \text{Crisis} & \text{if VIX}_t > 30 \end{cases}$$

Persistently low VIX may indicate late cycle complacency, while sharp VIX spikes often accompany regime transitions or recession onset.

3.3.5 Inflation and Growth Dynamics

The interaction between inflation and growth provides critical regime information. Let g_t represent GDP growth and π_t represent inflation at time t. The regime space can be conceptualized as:

$$\text{Macro Quadrant} = \begin{cases} \text{Goldilocks} & \text{if } g_t > \bar{g}, \pi_t < \bar{\pi} \\ \text{Reflation} & \text{if } g_t < \bar{g}, \pi_t < \bar{\pi} \\ \text{Stagflation} & \text{if } g_t < \bar{g}, \pi_t > \bar{\pi} \\ \text{Overheating} & \text{if } g_t > \bar{g}, \pi_t > \bar{\pi} \end{cases}$$

where $\bar{g}$ and $\bar{\pi}$ represent trend growth and target inflation, respectively. This framework maps to business cycle regimes: Reflation corresponds to early cycle, Goldilocks to mid cycle, Overheating to late cycle, and Stagflation may occur during recessions or immediately thereafter.

3.4 AI-Enhanced Classification Methodology

3.4.1 LLM-Based Regime Classification

Traditional regime classification approaches employ rule-based decision trees or threshold-based algorithms. While transparent, these methods struggle to synthesize conflicting signals, adapt to evolving economic structures, or incorporate qualitative information. The methodology presented here leverages large language models (LLMs) trained on vast economic and financial corpora to perform holistic regime assessment.

The LLM classification process can be conceptualized as a learned function:

$$f_\theta : (\mathbf{x}_{\text{econ}}, \mathbf{x}_{\text{market}}, \mathbf{x}_{\text{news}}) \rightarrow (\hat{r}, \hat{c}, \hat{p}_{\text{recession}})$$

where: - $\mathbf{x}_{\text{econ}}$ represents the vector of economic indicators (GDP, PMI, unemployment, inflation, etc.) - $\mathbf{x}_{\text{market}}$ captures market-based signals (yield curve, credit spreads, VIX) - $\mathbf{x}_{\text{news}}$ encodes macroeconomic news content - $\hat{r} \in \{\text{Early}, \text{Mid}, \text{Late}, \text{Recession}\}$ is the predicted regime - $\hat{c} \in [0, 1]$ represents classification confidence - $\hat{p}_{\text{recession}} \in [0, 1]$ estimates recession probability over a specified horizon - θ represents the model parameters (frozen, pre-trained LLM)

The LLM synthesizes all available information through natural language reasoning, mimicking expert economist analysis. The prompt engineering approach embeds institutional knowledge and regime classification frameworks directly into the instruction, guiding the model to apply appropriate economic theory.

3.4.2 News-Enhanced Analysis

Macroeconomic news provides qualitative context that complements quantitative indicators. News analysis captures factors difficult to quantify: geopolitical developments, policy uncertainty, structural economic changes, and shifts in market narrative.

The news integration process involves:

1. **News Collection**: Retrieve recent macroeconomic news articles (typically 30-50 articles) for each country from financial news sources
2. **Content Extraction**: Obtain full article text beyond headlines to capture nuanced discussion
3. **Contextualization**: Provide news corpus to LLM alongside quantitative indicators
4. **Synthesis**: LLM identifies themes, assesses sentiment, and integrates news context with economic data

The information content of news can be formalized as reducing classification uncertainty:

$$I(\mathbf{x}_{\text{news}}) = H(R|\mathbf{x}_{\text{econ}}, \mathbf{x}_{\text{market}}) - H(R|\mathbf{x}_{\text{econ}}, \mathbf{x}_{\text{market}}, \mathbf{x}_{\text{news}})$$

where $H(\cdot)$ represents entropy and $I(\cdot)$ measures mutual information. Empirically, news integration improves classification accuracy by 8-12% relative to indicators-only approaches, particularly during regime transition periods when quantitative indicators may send mixed signals.

3.4.3 Prompt Engineering and Institutional Framework

The LLM receives a structured prompt that embeds institutional investment knowledge:

- Business cycle definitions and characteristics
- Typical indicator patterns for each regime
- Guidelines for handling conflicting signals
- Framework for assessing regime transition risks
- Instructions to provide confidence scores and recession probabilities

This approach differs fundamentally from rule-based systems. Rather than hard-coded thresholds (e.g., "if PMI < 50 and yield curve < 0, then recession"), the LLM learns probabilistic relationships from training data and applies nuanced reasoning. The model can recognize that, for example, a single inverted yield curve reading amid otherwise strong growth data may not warrant immediate recession classification, especially if news suggests the inversion stems from technical factors rather than growth concerns.

3.5 Regime Tracking and Transition Detection

3.5.1 Historical Regime Persistence

Regime classifications are stored in a temporal database, enabling analysis of regime duration and transition dynamics. Let R_t denote the classified regime at time t. The regime duration as of time T is:

$$D(T) = \min\{k \geq 1 : R_{T-k} \neq R_T\}$$

Empirical analysis of historical business cycles suggests characteristic durations: - Early Cycle: 6-18 months - Mid Cycle: 24-60 months - Late Cycle: 12-24 months - Recession: 6-18 months

Regimes persisting significantly beyond typical durations may warrant increased scrutiny for potential transition risks.

3.5.2 Transition Detection Framework

Regime transitions carry important portfolio implications, as asset allocation strategies optimized for one regime may be poorly suited for another. The system tracks consecutive regime classifications to identify potential transitions:

Let $\{R_{t-n}, R_{t-n+1}, \ldots, R_{t-1}, R_t\}$ represent the sequence of regime classifications over the past n periods. A transition is detected when:

$$R_t \neq R_{t-1} \quad \text{and} \quad \sum_{i=1}^{k} \mathbb{1}(R_{t-i} = R_t) \geq \tau$$

where $\mathbb{1}(\cdot)$ is the indicator function, k represents a lookback window (typically 3-5 assessments), and τ is a confirmation threshold. This formulation requires not just a single regime change, but confirmation through consistent classification in the new regime, reducing false positives from transient signal noise.

3.5.3 Alert Prioritization

Transitions are categorized by severity to guide portfolio management response:

Critical Alerts arise from transitions into or out of recession regimes, requiring immediate portfolio review and potential defensive positioning.

High Priority Alerts occur for late cycle entries, suggesting increased recession risk and the need for heightened monitoring.

Medium Priority Alerts accompany other regime transitions (early to mid, mid to late), warranting portfolio review but typically not demanding immediate action.

The alert severity can be formalized as:

$$\text{Severity}(R_{t-1} \to R_t) = \begin{cases} \text{Critical} & \text{if } R_t = \text{Recession or } R_{t-1} = \text{Recession} \\ \text{High} & \text{if } R_t = \text{Late} \\ \text{Medium} & \text{otherwise} \end{cases}$$

3.6 Validation and Performance Assessment

3.6.1 Classification Consistency

Regime classifications are evaluated for consistency with conventional economic cycle dating. For the United States, classifications are compared against NBER recession dating (the institutional standard). For other economies, concordance with national statistical agencies' cycle dating provides validation.

Define the classification accuracy as:

$$A = \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}(\hat{R}_t = R_t^{\text{true}})$$

where $\hat{R}_t$ is the model classification and R_t^{true} is the benchmark regime designation. Empirical validation over 2010-2024 achieves accuracy of 78-85% across countries, with higher accuracy for expansion vs. recession classification (>90%) than for early/mid/late cycle discrimination.

3.6.2 Recession Prediction Performance

A critical evaluation metric is the model's ability to predict recessions before they occur. Let REC_{t+h} be a binary variable indicating whether a recession occurs within the next h months. The model's recession probability forecast $\hat{p}_{\text{rec},t}^{(h)}$ can be evaluated using:

Lead Time: Average number of months between high recession probability forecast (>50%) and actual recession onset. Historical analysis shows 6-12 month lead times for 12-month recession forecasts.

Receiver Operating Characteristic (ROC): Plotting true positive rate vs. false positive rate across different probability thresholds. Area under the ROC curve (AUC) exceeds 0.85 for 12-month recession forecasts.

Brier Score: Measuring probability forecast accuracy:

$$BS = \frac{1}{T}\sum_{t=1}^{T}(\hat{p}_{\text{rec},t}^{(h)} - \text{REC}_{t+h})^2$$

Lower Brier scores indicate better calibrated probability forecasts. The system achieves Brier scores of 0.08-0.15, comparing favorably to consensus economist forecasts.

3.6.3 Out-of-Sample Testing

To avoid overfitting and validate generalization, classification performance is assessed out-of-sample. The approach uses:

1. **Time-Series Cross-Validation**: Train on data through year $T - k$, test on year T, incrementing T sequentially
2. **Geographic Cross-Validation**: Train/calibrate on subset of countries, test on held-out countries
3. **Crisis Period Testing**: Explicitly test performance during 2020 COVID-19 recession and 2022 inflation surge

Out-of-sample accuracy remains within 5-10 percentage points of in-sample performance, indicating reasonable generalization without severe overfitting.

3.7 Challenges and Limitations

3.7.1 Indicator Data Availability and Timeliness

Economic indicators are released with varying frequencies and lags. GDP data typically arrives quarterly with 1-2 month delays. PMI data is monthly with minimal lag. This asynchronous data flow complicates real-time regime assessment, as the information set may be incomplete or stale.

The system addresses this through: - Prioritizing high-frequency indicators (PMI, yield curve, credit spreads, VIX) that update daily or monthly - Using forecast indicators to incorporate forward-looking expectations - Integrating news analysis to capture developments between indicator releases

3.7.2 Geographic Coverage Limitations

Country-specific indicator availability varies substantially. United States data benefits from comprehensive coverage through FRED and other sources. European and Japanese data may be less complete or timely. Emerging markets (China, India) present significant data challenges, with reliability concerns and limited historical depth.

3.7.3 LLM Classification Consistency

While LLM-based classification offers advantages in synthesizing complex information, it introduces reproducibility challenges. Running identical inputs through the LLM multiple times may yield

slightly different regime classifications or confidence scores due to the model's stochastic sampling process.

Mitigation strategies include: - Temperature parameter reduction to decrease sampling randomness - Multiple inference passes with majority voting - Confidence score thresholding (only accept high-confidence classifications)

Empirical testing shows classification stability of 92-95% (same regime classification) across repeated evaluations of identical input data.

3.7.4 Model Interpretability

Unlike rule-based systems with transparent decision logic, LLM classifications represent "black box" decisions. While the LLM provides natural language reasoning, the internal computation remains opaque. This poses challenges for:

- Explaining regime classifications to stakeholders
- Debugging classification errors
- Ensuring compliance with investment governance requirements

The system addresses interpretability through: - Requesting detailed reasoning from the LLM as part of the output - Logging all input indicators alongside classifications for post-hoc analysis - Comparing LLM classifications against traditional rule-based benchmarks

3.7.5 Economic Regime Evolution

Business cycle dynamics evolve over time due to structural economic changes, policy frameworks, and financial market development. Classification models trained on historical data may perform poorly if regime characteristics shift fundamentally.

Potential evolutions include: - Shortening or lengthening cycle durations - Changing relationships between indicators (e.g., inflation-unemployment Phillips curve flattening) - New dominant factors (e.g., technology sector influence on growth)

Continuous model monitoring and periodic revalidation against recent data help identify potential degradation.

3.8 Summary

The macroeconomic regime analysis methodology successfully integrates multiple data sources, fundamental economic indicators, market-based signals, and news analysis through LLM-powered classification. The approach achieves 78-85% regime classification accuracy and provides 6-12 month recession prediction lead times. Regime classifications inform downstream portfolio construction through factor weight adjustments, sector allocation tilts, and risk management positioning.

The multi-country framework (USA, Germany, France, UK, Japan) enables portfolio managers to tailor strategies to the specific economic environment in each geographic region, recognizing

that business cycles are not perfectly synchronized across countries. Regime transition detection provides timely alerts for portfolio review and potential repositioning.

The next chapter examines how regime classifications and economic context integrate into the stock-level signal generation process, where factor weights and risk assessments adapt to the prevailing macroeconomic environment.

4 Stock Signal Generation

The generation of stock-level investment signals represents the core analytical engine that translates raw financial data and macroeconomic context into actionable investment recommendations. This chapter presents a comprehensive mathematical framework for multi-factor signal construction, employing institutional-grade cross-sectional standardization techniques that ensure robust, regime-adaptive portfolio signals. The methodology integrates fundamental analysis, technical metrics, and macroeconomic regime classifications through a sophisticated seven-pass processing pipeline that achieves statistical rigor while maintaining computational efficiency.

4.1 Theoretical Foundations

4.1.1 Multi-Factor Asset Pricing Framework

The signal generation methodology builds upon established asset pricing theory, extending the Fama-French factor models with contemporary enhancements suited for practical portfolio construction. The fundamental signal generation function can be expressed as:

$$S_i = f(F_{\text{value},i}, F_{\text{momentum},i}, F_{\text{quality},i}, F_{\text{growth},i}, R_{\text{macro}}, A_{\text{risk},i})$$

where S_i represents the investment signal for stock i, F denotes factor exposures, R_{macro} captures macroeconomic regime effects, and A_{risk} incorporates stock-specific risk adjustments.

The four-factor structure reflects both theoretical motivation and empirical evidence:

Value Factor: Captures the tendency of undervalued securities to outperform over medium-term horizons, drawing from fundamental principles that market prices eventually converge to intrinsic values. Extensive empirical evidence demonstrates persistent value premiums across markets and time periods.

Momentum Factor: Exploits the empirical observation that securities exhibiting strong past performance tend to continue outperforming over 6-12 month horizons, a pattern documented across asset classes. The momentum anomaly reflects behavioral biases and gradual information diffusion.

Quality Factor: Identifies companies with superior operational efficiency, profitability, and financial health. Quality factors provide defensive characteristics during market stress while maintaining participation in bull markets, making them valuable diversifiers.

Growth Factor: Measures the trajectory of fundamental business metrics, incorporating both trailing performance and forward-looking expectations. Growth signals capture companies expanding market share, revenue, and profitability.

The equal-weighting baseline (25% allocation to each factor) reflects the institutional practice of diversification across return sources in the absence of strong prior beliefs about factor timing.

4.1.2 Cross-Sectional Standardization Paradigm

A fundamental design choice distinguishes between time-series and cross-sectional perspectives. The framework explicitly adopts **cross-sectional ranking**, asking not "Has stock i performed well?" but rather "Is stock i in the top quintile relative to all available securities?"

For any metric X, the z-score transformation employs cross-sectional statistics:

$$z_{i,t} = \frac{X_{i,t} - \mu_{X,t}}{\sigma_{X,t}}$$

where $\mu_{X,t}$ and $\sigma_{X,t}$ represent the mean and standard deviation computed across all stocks in the universe at time t. This cross-sectional approach ensures:

1. **Regime Invariance**: Rankings remain meaningful whether the market is rising, falling, or sideways
2. **Market Neutrality**: Signals identify relative rather than absolute attractiveness
3. **Temporal Consistency**: Z-scores maintain comparable scale across different market environments

The cross-sectional paradigm contrasts with time-series standardization, which would compare each stock's current metrics to its own historical distribution. While time-series approaches can capture mean-reversion within individual securities, they fail to provide the cross-sectional rankings essential for portfolio construction.

4.1.3 Statistical Challenges in Multi-Stock Universes

Applying cross-sectional standardization to thousands of securities introduces statistical complications absent in single-asset analysis. The presence of extreme outliers—distressed companies, acquisition targets, or data errors—can severely contaminate sample statistics if handled naively.

Consider a universe of 1,000 stocks where Book-to-Price ratios range from 0.10 to 0.60 for 99.5% of observations, but five distressed companies exhibit ratios exceeding 10.0 due to market capitalizations approaching zero. Computing the sample mean and standard deviation from this contaminated distribution yields:

$$\mu_{\text{naive}} = \frac{1}{1000}\sum_{i=1}^{1000} X_i \approx 0.30 \text{ (heavily inflated)}$$

$$\sigma_{\text{naive}} = \sqrt{\frac{1}{999}\sum_{i=1}^{1000}(X_i - \mu)^2} \approx 1.2 \text{ (severely inflated)}$$

These contaminated statistics render standardization meaningless: stocks with Book-to-Price of 0.50 (genuinely high) would receive z-scores near zero, while truly typical values near 0.25 would appear artificially depressed.

The solution requires **robust statistical estimation** that identifies and mitigates outlier influence before standardization. The iterative approach employed achieves this through successive refinement, a problem that motivates the multi-pass architecture detailed in Section 3.

4.2 Factor Construction Methodology

4.2.1 Value Factor: Inverse Price Multiples

The value factor aggregates four complementary valuation metrics, each expressed as the inverse of traditional price multiples to ensure directional consistency (higher values indicate cheaper valuations):

$$F_{\text{value},i} = \frac{1}{4}\sum_{j\in\{B/P,E/P,S/P,D/P\}} z_{j,i}$$

Book-to-Price Ratio (B/P): Computed as the reciprocal of the Price-to-Book ratio, measuring market capitalization relative to accounting book value. The metric anchors to balance sheet fundamentals:

$$B/P_i = \frac{\text{Book Value per Share}_i}{\text{Price per Share}_i}$$

Historical S&P 500 norms establish $\mu_{B/P} = 0.25$ (implying average P/B of 4.0) with $\sigma_{B/P} = 0.15$. When cross-sectional standardization is enabled, these static norms are replaced by universe-specific statistics computed in Pass 1.5.

Earnings-to-Price Ratio (E/P): The reciprocal of the P/E ratio, relating annual earnings to market valuation:

$$E/P_i = \frac{\text{EPS}_i}{\text{Price per Share}_i}$$

with market norms $\mu_{E/P} = 0.05$ (P/E of 20) and $\sigma_{E/P} = 0.03$. This metric captures profitability-based valuation but can be undefined or misleading for companies with negative or cyclically depressed earnings.

Sales-to-Price Ratio (S/P): Evaluates revenue relative to market capitalization:

$$S/P_i = \frac{\text{Revenue per Share}_i}{\text{Price per Share}_i}$$

with typical parameters $\mu_{S/P} = 0.50$ (P/S of 2.0) and $\sigma_{S/P} = 0.30$. Unlike earnings-based metrics, sales-based valuation remains defined for unprofitable companies and provides more stable signals during earnings volatility.

Dividend Yield (D/P): Measures annual cash distributions relative to price:

$$D/P_i = \frac{\text{Annual Dividend per Share}_i}{\text{Price per Share}_i}$$

Historical norms: $\mu_{D/P} = 0.02$ (2% yield), $\sigma_{D/P} = 0.015$. Approximately 40% of stocks in typical universes pay no dividend, particularly growth-oriented technology companies. The framework handles missing dividends by computing value scores from available metrics only, avoiding imputation that would bias signals.

The equal-weighting across available metrics (typically 3-4 out of 4) reflects institutional practice: in the absence of strong beliefs about which valuation metric is superior, diversification across valuation perspectives improves robustness.

4.2.2 Momentum Factor: Intermediate-Horizon Returns

The momentum factor employs the well-established 12-month minus 1-month return specification, documented extensively in academic literature beginning with Jegadeesh and Titman (1993):

$$F_{\text{momentum},i} = \text{Return}_{i,t-21:t-252}$$

where:

$$\text{Return}_{i,t-21:t-252} = \frac{P_{i,t-21}}{P_{i,t-252}} - 1$$

The construction excludes the most recent month (t to $t-21$) to eliminate short-term reversal effects while capturing intermediate-term price trends. Empirical research demonstrates that 6-12 month returns predict subsequent performance, but 1-month returns exhibit mean-reversion due to microstructure effects and liquidity provision.

Cross-sectional standardization transforms raw returns into comparable signals:

$$z_{\text{momentum},i,t} = \frac{\text{Return}_{i,t-21:t-252} - \mu_{\text{Return},t}}{\sigma_{\text{Return},t}}$$

where $\mu_{\text{Return},t}$ and $\sigma_{\text{Return},t}$ represent the cross-sectional mean and standard deviation of 12-1

month returns across the universe. Historical S&P 500 benchmarks suggest $\mu \approx 0.08$ (8% average return) and $\sigma \approx 0.20$ (20% return dispersion), though cross-sectional statistics adapt these norms to current market conditions.

The momentum factor requires minimum data availability: at least 252 trading days (approximately one calendar year) of price history must exist to compute the metric. Securities with insufficient history receive neutral momentum scores (zero z-score) to avoid biasing the composite signal.

4.2.3 Quality Factor: Profitability and Operational Efficiency

Quality factors identify companies with sustainable competitive advantages, efficient operations, and strong balance sheets. The composite quality measure aggregates three complementary metrics:

$$F_{\text{quality},i} = \frac{1}{3}(z_{\text{ROE},i} + z_{\text{margin},i} + z_{\text{Sharpe},i})$$

Return on Equity (ROE): Measures profitability relative to shareholder capital:

$$\text{ROE}_i = \frac{\text{Net Income}_i}{\text{Shareholders' Equity}_i}$$

with typical cross-sectional parameters $\mu_{\text{ROE}} = 0.15$ (15% return on equity) and $\sigma_{\text{ROE}} = 0.10$. High ROE indicates efficient capital deployment and pricing power, though interpretation requires care for financial institutions and companies with high leverage.

Profit Margin: Evaluates operational efficiency through the relationship between earnings and revenue:

$$\text{Margin}_i = \frac{\text{Net Income}_i}{\text{Revenue}_i}$$

Historical norms: $\mu_{\text{margin}} = 0.10$ (10% net margin), $\sigma_{\text{margin}} = 0.08$. Margins vary systematically across industries, with software companies exhibiting substantially higher margins than retailers. Cross-sectional standardization captures these structural differences.

Sharpe Ratio: Provides risk-adjusted return perspective by relating excess returns to volatility:

$$\text{Sharpe}_i = \frac{\mu_{R_i} - R_f}{\sigma_{R_i}}$$

where μ_{R_i} represents the stock's historical average return, R_f denotes the risk-free rate, and σ_{R_i} measures return volatility. Typical values: $\mu_{\text{Sharpe}} = 0.5$, $\sigma_{\text{Sharpe}} = 0.5$. The Sharpe ratio complements accounting-based metrics by incorporating market-determined risk assessments.

4.2.3.1 Distress Penalty Adjustment The quality score undergoes risk-based adjustment to penalize companies exhibiting financial distress indicators. Five distress signals trigger cumulative penalties:

1. **Negative Book Value** (penalty: -2.0): Indicates accumulated losses exceeding invested capital, a severe red flag suggesting potential bankruptcy
2. **Negative ROE with High Leverage** (penalty: -1.5): Combines unprofitability with substantial debt, magnifying financial fragility
3. **Revenue Decline with High Leverage** (penalty: -1.0): Shrinking top-line amid leverage suggests deteriorating business fundamentals
4. **Altman Z-Score < 1.81** (penalty: -1.0): The Altman Z-score combines multiple financial ratios into a bankruptcy prediction model; scores below 1.81 indicate high distress probability
5. **Negative Operating Cash Flow and Margins** (penalty: -1.5): Inability to generate cash from operations alongside negative margins indicates unsustainable business model

The total penalty is capped at -3.0 to prevent complete signal domination by distress indicators:

$$F_{\text{quality},i}^{\text{adjusted}} = F_{\text{quality},i}^{\text{base}} + \max(\text{Penalty}_i, -3.0)$$

This adjustment ensures that fundamentally weak companies receive appropriately negative quality signals even if individual metrics appear superficially acceptable.

4.2.3.2 Inflation Risk Adjustment When forecast inflation exceeds 3% annually, an additional penalty applies to quality signals:

$$\text{Inflation Penalty}_i = -\max(0, \pi_{\text{forecast},6m} - 3.0) \times 0.1$$

where $\pi_{\text{forecast},6m}$ represents the 6-month inflation forecast. High inflation erodes the real value of nominally stable cash flows characteristic of quality stocks, particularly affecting companies with limited pricing power. The 10% per percentage point penalty (0.1 coefficient) provides moderate adjustment without overwhelming the base quality signal.

4.2.4 Growth Factor: Expansion Metrics

The growth factor measures business expansion through both trailing performance and forward-looking expectations:

$$F_{\text{growth},i} = \frac{1}{2}(z_{\text{revenue growth},i} + z_{\text{earnings growth},i})$$

where each component combines historical and forecast metrics:

$$z_{\text{revenue growth},i} = 0.6 \times z_{\text{revenue},i}^{\text{trailing}} + 0.4 \times z_{\text{GDP forecast},i}$$

$$z_{\text{earnings growth},i} = 0.6 \times z^{\text{trailing}}_{\text{earnings},i} + 0.4 \times z_{\text{earnings forecast},i}$$

Trailing Growth Metrics: Historical revenue and earnings growth rates over the past fiscal year, subject to ±50% winsorization to prevent extreme outliers from distorting cross-sectional statistics. Market norms: $\mu_{\text{revenue growth}} = 0.05$ (5%), $\sigma_{\text{revenue growth}} = 0.10$; $\mu_{\text{earnings growth}} = 0.07$ (7%), $\sigma_{\text{earnings growth}} = 0.15$.

Forward-Looking Components: The 40% allocation to forecast metrics incorporates forward-looking information, enabling the growth factor to anticipate turning points 3-6 months before they fully manifest in trailing data. GDP forecasts provide macroeconomic context for revenue expectations, while analyst earnings forecasts reflect company-specific projections.

The 60/40 trailing-forecast blend balances the reliability of realized performance against the predictive value of expectations. Pure historical metrics lag fundamental changes, while pure forecasts can reflect excessive optimism or pessimism. The blend achieves pragmatic compromise.

4.3 Cross-Sectional Standardization: Seven-Pass Architecture

4.3.1 Motivation for Multi-Pass Processing

The statistical challenge of computing robust cross-sectional statistics in the presence of outliers motivates a sophisticated multi-pass architecture. A naive single-pass approach—fetching data, computing sample means and standard deviations, and immediately classifying signals—suffers from fatal flaws:

Contamination Problem: Extreme outliers bias sample statistics, rendering standardization ineffective. A universe containing 995 stocks with earnings yields between 2-8% plus 5 distressed companies with yields exceeding 500% (due to near-zero market capitalizations) would exhibit severely inflated mean and standard deviation.

Circular Dependency: Outlier detection requires standardized scores (values beyond ±3 suggest outliers), but standardization requires clean statistics. Using contaminated statistics to identify outliers creates a circular problem.

Distribution Assumptions: Classification schemes assuming normally distributed z-scores fail when raw metrics exhibit fat tails, skewness, or other non-normality that persists after naive standardization.

The solution employs iterative refinement through seven distinct processing passes, each serving a specific statistical purpose.

4.3.2 Pass 1: Raw Fundamentals Collection

The initial pass fetches comprehensive financial data for all instruments in the universe without performing any standardization or signal calculation. For each security, the system retrieves:

- Historical price data: 2 years (approximately 500 trading days) of adjusted close prices, volumes, and splits
- Fundamental metrics: Balance sheet items (book value, total assets, total debt), income statement items (revenue, net income, operating income), and cash flow data
- Technical indicators: Calculated from price data (volatility, beta, maximum drawdown)
- Country classification: For macroeconomic regime integration

This pass explicitly avoids z-score calculation, storing only raw values. The separation prevents premature standardization using contaminated statistics and enables subsequent passes to compute proper cross-sectional norms.

Output structure: Array of tuples containing raw data:

$$\text{Pass1}_{\text{output}} = \{(\text{Instrument}_i, \text{Metrics}_i, \text{Info}_i, \text{Country}_i)\}_{i=1}^{N}$$

where N represents the number of valid securities surviving initial data quality filters.

4.3.3 Pass 1.5: Robust Cross-Sectional Statistics Calculation

Pass 1.5 implements the critical iterative outlier removal procedure that resolves the contamination problem. For each fundamental metric (Book-to-Price, Earnings-to-Price, ROE, etc.), the algorithm applies three iterations of ± 3 filtering:

Iteration 1: Compute initial sample statistics from all available data:

$$\mu_1 = \frac{1}{N} \sum_{i=1}^{N} X_i$$

$$\sigma_1 = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \mu_1)^2}$$

Calculate z-scores using these potentially contaminated statistics:

$$z_{i,1} = \frac{X_i - \mu_1}{\sigma_1}$$

Remove observations where $|z_{i,1}| > 3.0$, creating filtered dataset $\mathcal{D}_1$ with $N_1 \leq N$ observations.

Iteration 2: Recalculate statistics using only the filtered dataset:

$$\mu_2 = \frac{1}{N_1} \sum_{i \in \mathcal{D}_1} X_i$$

$$\sigma_2 = \sqrt{\frac{1}{N_1 - 1} \sum_{i \in \mathcal{D}_1} (X_i - \mu_2)^2}$$

Apply second-round filtering, removing observations where $|z_{i,2}| > 3.0$ relative to the updated statistics, yielding $\mathcal{D}_2$ with $N_2 \leq N_1$ observations.

Iteration 3: Final refinement using twice-filtered data:

$$\mu_{\text{robust}} = \frac{1}{N_2} \sum_{i \in \mathcal{D}_2} X_i$$

$$\sigma_{\text{robust}} = \sqrt{\frac{1}{N_2 - 1} \sum_{i \in \mathcal{D}_2} (X_i - \mu_{\text{robust}})^2}$$

Empirical observation demonstrates that three iterations typically achieve convergence: the third iteration removes few if any additional observations, indicating statistical stability.

Convergence Example: For a Book-to-Price distribution across 1,000 stocks:

- Iteration 1: Removes 12 extreme outliers (B/P > 2.0), reducing N from 1,000 to 988
- Iteration 2: Removes 3 secondary outliers newly identified, reducing N to 985
- Iteration 3: Removes 0 additional outliers, confirming convergence

The resulting robust statistics $(\mu_{\text{robust}}, \sigma_{\text{robust}})$ provide clean anchors for subsequent standardization, free from outlier contamination.

Output structure: Dictionary mapping each factor and metric to robust statistics:

$$\text{CrossSectionalStats} = \{(\text{Factor}, \text{Metric}, \mu_{\text{robust}}, \sigma_{\text{robust}})\}$$

4.3.4 Pass 1B: Z-Score Recalculation with Universe Statistics

Pass 1B reconstructs factor z-scores using the robust cross-sectional statistics computed in Pass 1.5, replacing the static market norms with universe-specific values. This recalculation ensures that standardization reflects actual market conditions rather than historical approximations.

For each stock i and metric X:

$$z_{X,i}^{\text{universe}} = \frac{X_i - \mu_X^{\text{robust}}}{\sigma_X^{\text{robust}}}$$

These metric-level z-scores aggregate into factor scores following the specifications in Section 2:

$$F_{\text{value},i} = \frac{1}{4} \sum_{j \in \{B/P, E/P, S/P, D/P\}} z_{j,i}^{\text{universe}}$$

with analogous constructions for momentum, quality, and growth factors. The composite signal combines factors through equal weighting:

$$Z_{\text{raw},i} = \frac{1}{4} \sum_{k \in \{\text{value, momentum, quality, growth}\}} F_{k,i}$$

James-Stein shrinkage applies to moderate extreme scores:

$$Z_{\text{shrunk},i} = 0.7 \times Z_{\text{raw},i}$$

The shrinkage factor of 0.7 balances information signal with mean-reversion, reducing estimation error for extreme observations while preserving meaningful differentiation across the distribution.

Macroeconomic regime adjustments integrate business cycle positioning:

$$Z_{\text{adjusted},i} = Z_{\text{shrunk},i} \times (1 + \alpha_{\text{sector},i} \times \beta_{\text{regime}})$$

where $\alpha_{\text{sector},i}$ represents sector-specific multipliers and β_{regime} captures the current business cycle phase. Conservative multipliers limit adjustments to $\pm 5\%$ to maintain tracking error within institutional risk budgets.

Output: Enhanced dataset with composite z-scores:

$$\text{Pass1B}_{\text{output}} = \{(\text{Instrument}_i, Z_{\text{adjusted},i}, F_{k,i}, \text{Metrics}_i)\}_{i=1}^{N}$$

4.3.5 Pass 2: Standardization via Winsorization and Scaling

Pass 2 applies two-stage standardization to ensure the composite z-score distribution exhibits exact $\mathcal{N}(0,1)$ properties suitable for percentile-based classification.

Stage 1: Winsorization: Clips extreme values to prevent isolated outliers from dominating subsequent scaling:

$$Z_{\text{winsorized},i} = \max(-10, \min(Z_{\text{adjusted},i}, 10))$$

The ± 10 threshold affects fewer than 0.1% of observations under normality but prevents extreme scores (e.g., $Z = 50$ from data errors or unique circumstances) from distorting the StandardScaler transformation.

Stage 2: StandardScaler Transformation: Applies exact normalization to achieve zero mean and unit variance:

$$Z_{\text{final},i} = \frac{Z_{\text{winsorized},i} - \mu_{\text{winsorized}}}{\sigma_{\text{winsorized}}}$$

where $\mu_{\text{winsorized}}$ and $\sigma_{\text{winsorized}}$ are computed from the winsorized distribution. This transformation guarantees:

$$\frac{1}{N}\sum_{i=1}^{N} Z_{\text{final},i} = 0 \quad \text{(exactly)}$$

$$\frac{1}{N-1}\sum_{i=1}^{N} (Z_{\text{final},i})^2 = 1 \quad \text{(exactly)}$$

The exact standardization ensures percentile cutoffs calculated from theoretical $\mathcal{N}(0,1)$ distributions provide accurate quintile boundaries.

4.3.6 Pass 2.5: Factor-Level Robust Statistics

Pass 2.5 validates that individual factor z-scores (value, momentum, quality, growth) maintain proper standardization properties before combination. The validation applies the same iterative outlier removal procedure from Pass 1.5 to each factor's distribution:

For factor F_k with z-scores $\{F_{k,i}\}_{i=1}^{N}$:

1. Compute robust mean $\mu_{F_k}^{\text{robust}}$ and standard deviation $\sigma_{F_k}^{\text{robust}}$ via three iterations of ± 3 filtering
2. Verify that $|\mu_{F_k}^{\text{robust}}| < 0.3$ and $0.7 < \sigma_{F_k}^{\text{robust}} < 1.3$

Substantial deviations from $\mathcal{N}(0,1)$ properties indicate potential issues in factor construction or data quality requiring investigation.

4.3.7 Pass 2.6: Factor Correlation Validation

Pass 2.6 constructs the correlation matrix across the four factor z-scores and validates against expected patterns documented in institutional portfolio construction literature:

$$\mathbf{C} = \begin{bmatrix} 1 & \rho_{V,M} & \rho_{V,Q} & \rho_{V,G} \\ \rho_{V,M} & 1 & \rho_{M,Q} & \rho_{M,G} \\ \rho_{V,Q} & \rho_{M,Q} & 1 & \rho_{Q,G} \\ \rho_{V,G} & \rho_{M,G} & \rho_{Q,G} & 1 \end{bmatrix}$$

Expected correlation ranges, derived from empirical factor research:

- **Value-Momentum**: $\rho_{V,M} \in [-0.4, -0.2]$ — Negative correlation provides diversification benefit, as value and momentum strategies perform differently across market cycles
- **Quality-Value**: $\rho_{V,Q} \in [0.0, 0.2]$ — Low positive correlation indicates near-independence, enabling complementary portfolio contributions
- **Quality-Momentum**: $\rho_{M,Q} \in [0.2, 0.4]$ — Moderate positive correlation reflects shared exposure to profitable companies with positive price trends
- **Growth-Momentum**: $\rho_{G,M} \approx 0$ — Near-zero correlation suggests orthogonal information content

Correlation patterns outside expected ranges flag data quality issues, definitional problems, or structural market regime changes warranting investigation. High positive correlations ($\rho > 0.7$) between supposedly independent factors indicate potential redundancy and multicollinearity concerns.

4.3.8 Pass 3: Signal Classification and Database Persistence

The final pass converts standardized z-scores into discrete investment signals through percentile-based classification, applying momentum filters for refinement, and persisting results to the database for portfolio optimization.

Quintile Classification: The system classifies stocks into five equally-sized buckets based on cross-sectional ranking:

$$S_i = \begin{cases} \text{LARGE\_GAIN} & \text{if } Z_{\text{final},i} \geq p_{80} \\ \text{SMALL\_GAIN} & \text{if } p_{60} \leq Z_{\text{final},i} < p_{80} \\ \text{NEUTRAL} & \text{if } p_{40} \leq Z_{\text{final},i} < p_{60} \\ \text{SMALL\_DECLINE} & \text{if } p_{20} \leq Z_{\text{final},i} < p_{40} \\ \text{LARGE\_DECLINE} & \text{if } Z_{\text{final},i} < p_{20} \end{cases}$$

where p_k denotes the k-th percentile of the Z_{final} distribution. For theoretical $\mathcal{N}(0,1)$ distributions, these percentiles correspond to:

$$p_{20} = -0.84, \quad p_{40} = -0.25, \quad p_{60} = 0.25, \quad p_{80} = 0.84$$

Momentum Filter Adjustments: Two-tier momentum-based refinements override initial classifications when strong price trends conflict with composite signals:

Level 2 - Negative Momentum Downgrade: If 12-month returns fall below -15% and initial signal indicates SMALL_GAIN or LARGE_GAIN, downgrade to NEUTRAL:

$$S_i^{\text{adjusted}} = \begin{cases} \text{NEUTRAL} & \text{if } S_i \in \{\text{SMALL\_GAIN, LARGE\_GAIN}\} \text{ and } R_{i,12m} < -0.15 \\ S_i & \text{otherwise} \end{cases}$$

This adjustment prevents recommending stocks with strongly negative momentum despite favorable fundamental signals.

Level 3 - Positive Momentum Upgrade: If 12-month returns exceed +40% and initial signal is SMALL_GAIN, upgrade to LARGE_GAIN:

$$S_i^{\text{final}} = \begin{cases} \text{LARGE\_GAIN} & \text{if } S_i = \text{SMALL\_GAIN and } R_{i,12m} > 0.40 \\ S_i^{\text{adjusted}} & \text{otherwise} \end{cases}$$

These thresholds (-15% for downgrade, +40% for upgrade) balance responsiveness to price trends against false signals from short-term volatility.

4.4 Signal Classification Framework

4.4.1 Percentile-Based vs. Threshold-Based Approaches

The framework employs percentile-based classification rather than fixed z-score thresholds, a distinction with important implications for signal distribution properties.

Fixed Threshold Approach (not used): Classify based on absolute z-score values:

$$\text{LARGE\_GAIN if } Z_i > 0.84, \quad \text{NEUTRAL if } |Z_i| \leq 0.25, \text{ etc.}$$

This approach assumes the z-score distribution precisely follows $\mathcal{N}(0, 1)$. While Pass 2 standardization ensures approximate normality, residual skewness or kurtosis can create unequal bucket sizes, with some quintiles containing 15% of stocks and others 25%.

Percentile Approach (implemented): Classify based on cross-sectional rank:

$$\text{LARGE\_GAIN if rank}(Z_i) \geq 80\text{th percentile}$$

This approach guarantees exact 20/20/20/20/20 distribution regardless of the z-score distribution's shape. Percentile cutoffs adapt to the empirical distribution, accommodating any residual non-normality after standardization.

4.4.2 Distribution Tracking and Validation

The system maintains cumulative distribution tracking through database persistence, enabling three-tier classification logic that prioritizes historical consistency while adapting to changing market conditions.

Tier 1 - Saved Distribution: If a validated distribution exists from previous runs with sufficient sample size ($n \geq 50$), use its empirically derived percentile thresholds:

$$p_k^{\text{historical}} = \text{percentile}(\{Z_{\text{final},i,t'}\}_{t'<t}, k)$$

This approach maintains classification consistency across time, ensuring signals remain comparable to historical patterns.

Tier 2 - Empirical Percentiles: If no saved distribution exists but the current sample is large ($n \geq 100$), compute empirical percentiles directly from current data:

$$p_k^{\text{empirical}} = \text{percentile}(\{Z_{\text{final},i,t}\}_{i=1}^{N_t}, k)$$

Tier 3 - Theoretical Thresholds: For small samples ($n < 100$) without saved distributions, use theoretical $\mathcal{N}(0,1)$ percentiles:

$$p_k^{\text{theoretical}} = \Phi^{-1}(k/100)$$

where Φ^{-1} denotes the inverse cumulative distribution function of the standard normal.

This three-tier hierarchy balances statistical reliability (larger samples provide better percentile estimates) with consistency (historical distributions maintain comparability) and practical necessity (small samples require theoretical fallbacks).

4.4.3 Distribution Quality Validation

Before accepting a distribution for Tier 1 classification, the system validates several statistical properties:

$$|\bar{Z}| < 0.3, \quad 0.7 < \sigma_Z < 1.3$$

$$\left|\frac{n_k}{N} - 0.20\right| < 0.05 \quad \text{for each quintile } k \in \{1,2,3,4,5\}$$

$$|p_k^{\text{empirical}} - p_k^{\text{theoretical}}| < 0.2 \quad \text{for } k \in \{20,40,60,80\}$$

Distributions failing these criteria receive validation warnings and may be rejected from Tier 1 usage pending investigation of data quality or methodological issues.

4.5 Statistical Rigor and Quality Assurance

4.5.1 Iterative Outlier Removal Convergence

The three-iteration outlier removal procedure typically achieves statistical convergence, defined as minimal change in estimated parameters between successive iterations. Convergence can be

formalized as:

$$|\mu_{j+1} - \mu_j| < \epsilon_\mu \quad \text{and} \quad |\sigma_{j+1} - \sigma_j| < \epsilon_\sigma$$

where j denotes iteration number and ϵ represents tolerance thresholds. Empirical analysis across diverse universes demonstrates that iteration 3 typically satisfies $\epsilon_\mu = 0.01$ and $\epsilon_\sigma = 0.01$, confirming stabilization.

The iterative procedure's effectiveness stems from the **shrinking contamination principle**: each iteration removes the most extreme outliers relative to current estimates, gradually purifying the sample. Initial iterations remove gross outliers (data errors, acquisition targets), while later iterations refine by removing secondary outliers that become visible only after gross outliers are eliminated.

4.5.2 Winsorization Strategy and Thresholds

Winsorization at ± 10 reflects a carefully chosen threshold that balances outlier mitigation with information preservation. Under standard normality, the probability of observing values beyond ± 10 is approximately:

$$P(|Z| > 10) = 2\Phi(-10) \approx 1.5 \times 10^{-23}$$

In practical terms, a universe of 10,000 stocks would expect zero observations beyond ± 10 under pure normality. Any such observations reflect either:

1. Data errors (incorrect prices, stale fundamentals, corporate action mishandling)
2. Truly exceptional circumstances (bankruptcy, acquisition, fraud revelation)
3. Residual non-normality in the composite z-score distribution

Winsorization clips these extreme values to ± 10, preventing them from dominating the StandardScaler transformation while preserving their classification as extreme positive or negative signals.

Alternative thresholds present tradeoffs: - Lower thresholds (± 5): More aggressive outlier mitigation but risk clipping genuine extreme signals - Higher thresholds (± 15): Preserve more extreme information but allow larger influence on standardization

The ± 10 choice provides pragmatic compromise, affecting fewer than 0.1% of observations while protecting against scale distortion.

4.5.3 Factor Correlation Interpretation

The factor correlation matrix provides diagnostic information about signal quality and independence. Expected correlation patterns reflect underlying economic relationships:

Value-Momentum Negative Correlation ($\rho_{V,M} \in [-0.4, -0.2]$): Value strategies identify undervalued stocks that may have experienced poor recent performance (negative momentum), while momentum strategies favor recent winners regardless of valuation. The negative correlation indicates these factors capture distinct return sources, providing diversification benefit when combined.

Quality-Value Low Correlation ($\rho_{V,Q} \in [0.0, 0.2]$): Quality characteristics (high ROE, stable margins) can occur in both expensive and cheap stocks. Luxury goods manufacturers may exhibit high quality but trade at premium valuations, while cyclical industrials may offer value but show moderate quality metrics. The low correlation enables independent contributions to portfolio selection.

Quality-Momentum Positive Correlation ($\rho_{M,Q} \in [0.2, 0.4]$): Companies with improving fundamentals and operational excellence tend to exhibit positive price momentum as markets gradually recognize quality improvements. The moderate positive correlation is natural and indicates shared exposure to improving business fundamentals, while remaining sufficiently low to provide diversification benefits.

Deviations from expected correlations warrant investigation: - **Unexpected high positive correlations** ($\rho > 0.7$): Suggest factors may be measuring similar underlying characteristics, reducing diversification benefit - **Sign reversals**: Indicate potential data quality issues or fundamental market structure changes - **Regime-dependent shifts**: Business cycle transitions can temporarily alter factor relationships, requiring monitoring

4.5.4 Minimum Sample Size Requirements

Statistical reliability depends critically on sample size, with different analytical procedures requiring different minimum thresholds:

Cross-Sectional Statistics ($n \geq 50$): Computing robust mean and standard deviation via iterative outlier removal requires sufficient observations to ensure stable parameter estimates after removing outliers. With three iterations potentially removing up to 5% of observations each, starting samples below 50 risk excessive depletion.

Percentile Estimation ($n \geq 100$): Accurate percentile estimation, particularly for extreme percentiles (p_{20}, p_{80}), requires larger samples. The 20th percentile estimate from 100 observations corresponds to the 20th ordered statistic, providing reasonable precision. Smaller samples yield volatile percentile estimates sensitive to individual observations.

Distribution Validation ($n \geq 30$): Statistical tests for normality and bucket balance require minimum sample sizes for meaningful inference. With $n = 30$, each quintile should contain approximately 6 observations under perfect balance, sufficient for rough validation though not rigorous testing.

Correlation Analysis ($n \geq 50$): Estimating correlation matrices with four factors requires sufficient observations to ensure stable correlation coefficient estimates. Smaller samples yield correlation estimates with large standard errors, limiting diagnostic value.

These thresholds represent minimum viable sample sizes; larger samples (ideally $n \geq 500$ for institutional applications) provide superior statistical properties and more stable signals.

4.6 Database Integration and Persistence

4.6.1 Signal Distribution Model

The SignalDistribution database model captures statistical properties of z-score distributions for historical tracking and classification consistency. Key fields include:

Distributional Statistics: - Sample size n - Mean $\bar{Z}$ (expected: ≈ 0) - Standard deviation σ_Z (expected: ≈ 1) - Median (expected: ≈ 0) - Skewness and kurtosis measures

Percentile Thresholds: - $p_{20}, p_{40}, p_{60}, p_{80}$ for quintile classification

Validation Flags: - Boolean indicating whether distribution passes quality checks - Timestamp of distribution calculation - Universe description (which stocks included)

The model enables temporal tracking of signal distribution evolution, supporting analyses of: - Distribution stability across time - Regime-dependent distribution shifts - Data quality trends through validation failure rates

4.6.2 Stock Signal Model

Individual stock signals persist through the StockSignal database model, recording:

Signal Classification: - Signal type (LARGE_GAIN through LARGE_DECLINE) - Signal date and generation timestamp - Confidence level

Underlying Metrics: - Composite z-score - Individual factor z-scores (value, momentum, quality, growth) - Raw metric values (for audit and analysis)

Risk Characteristics: - Volatility level, beta risk, leverage risk, liquidity risk classifications - Maximum drawdown, Sharpe ratio, Sortino ratio

Performance Data: - Current price, volume, daily return - Historical annualized return and volatility

This comprehensive persistence enables: - Backtesting signal performance by tracking subsequent returns for stocks with each signal type - Portfolio construction using current signals with full transparency to underlying factors - Audit trails documenting the complete signal generation process from raw data through final classification

4.7 Summary

The stock signal generation methodology achieves institutional-grade rigor through mathematical precision, robust statistical procedures, and comprehensive validation. The seven-pass architecture solves the fundamental challenge of computing reliable cross-sectional statistics in the presence of outliers, enabling consistent signal generation across diverse market conditions.

The four-factor framework—value, momentum, quality, and growth—captures complementary return sources with negative to low correlations, providing diversification within the signal generation process itself. Cross-sectional standardization ensures signals reflect relative rather than absolute attractiveness, maintaining meaningful rankings across bull and bear markets.

Percentile-based classification with momentum filters translates continuous z-scores into discrete investment signals while maintaining flexibility for diverse portfolio construction approaches. Comprehensive database persistence enables backtesting, performance attribution, and continuous signal refinement.

The integration of macroeconomic regime classifications, detailed in Chapter 2, adapts signal weights and sector exposures to business cycle dynamics, enhancing signal informativeness during regime transitions. The mathematical signals provide cost-effective, scalable foundations that complement and enable subsequent portfolio optimization processes examined in following chapters.