Contents

0.1. Understanding Alpha-Mixing Conditions

Formal Definition and Interpretation

Mathematical Setup

Let $\{X_t\}_{t=-\infty}^{\infty}$ be a stochastic process on a probability space $(\Omega, \mathcal{F}, P)$. We define:

- $\mathcal{F}_{-\infty}^{t} = \sigma(..., X_{t-1}, X_t)$: the σ-algebra generated by all events up to time t
- $\mathcal{F}_{t+h}^{\infty} = \sigma(X_{t+h}, X_{t+h+1}, ...)$: the σ-algebra generated by all events from time $t+h$ onward

Alpha-Mixing Coefficient

The α-mixing coefficient is defined as:

$$\alpha(h) = \sup_{A \in \mathcal{F}_{-\infty}^{t}, B \in \mathcal{F}_{t+h}^{\infty}} |P(A \cap B) - P(A)P(B)| \tag{1}$$

Interpretation:

- $P(A \cap B)$ is the joint probability of events A and B
- $P(A)P(B)$ is what the joint probability would be if A and B were independent
- $\alpha(h)$ measures the maximum deviation from independence at lag h
- As $h \to \infty$, $\alpha(h) \to 0$ for mixing processes

Necessity of Alpha-Mixing

Statistical Requirements

Alpha-mixing is needed for:

1. **Law of Large Numbers:**

$$\frac{1}{T}\sum_{t=1}^{T} X_t \xrightarrow{p} E[X_t] \tag{2}$$

2. **Central Limit Theorem:**

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(X_t - E[X_t]) \xrightarrow{d} N(0, \sigma^2) \tag{3}$$

3. **Moment Bounds:**

$$E|\frac{1}{T}\sum_{t=1}^{T} X_t - E[X_t]|^p \leq CT^{-p/2} \tag{4}$$

Understanding the Paper's Mixing Condition

The condition:

$$\sum_{h=1}^{\infty} h^2 \alpha(h)^{\delta/(2+\delta)} < \infty \tag{5}$$

Component Analysis

1. **The Role of h:**

- h represents the time lag
- h^2 ensures rapid decay of dependence
- Larger h means events further apart in time

2. **The Role of $\alpha(h)$:**

- Measures dependence at lag h
- Must decay faster than h^{-2} for summability
- Typical decay: $\alpha(h) \sim h^{-\beta}$ for some $\beta > 2$

3. **The Role of δ:**

- Controls moment existence
- Larger δ means stronger moment conditions
- Typically $\delta = 2$ for financial applications

Intuitive Examples of Mixing

Financial Market Examples

1. **Market Microstructure Effects:**

$$R_t = \phi R_{t-1} + \epsilon_t, \quad |\phi| < 1 \tag{6}$$

- Bid-ask bounce creates short-term dependence
- Effect dies out exponentially: $\alpha(h) \sim |\phi|^h$

2. **Volatility Clustering:**

$$R_t = \sigma_t \epsilon_t, \quad \sigma_t^2 = \omega + \alpha R_{t-1}^2 + \beta \sigma_{t-1}^2 \tag{7}$$

- GARCH processes are α-mixing
- Dependence decays geometrically

Verifying Mixing Conditions in Practice

Statistical Tests

1. **Correlation-based Tests:**

$$\hat{\rho}(h) = \frac{\sum_{t=h+1}^{T}(X_t - \bar{X})(X_{t-h} - \bar{X})}{\sum_{t=1}^{T}(X_t - \bar{X})^2} \tag{8}$$

2. **Mixing Coefficient Estimation:**

$$\hat{\alpha}(h) = \sup_{i,j} |\hat{P}(A_i \cap B_j) - \hat{P}(A_i)\hat{P}(B_j)| \tag{9}$$

Practical Approaches

1. **Graphical Analysis:**
   - Plot ACF/PACF
   - Examine decay patterns
   - Check for long-range dependence
2. **Model-based Verification:**
   - Fit ARMA/GARCH models
   - Check residual properties
   - Verify model stability

Connection to Other Time Series Concepts

Related Dependencies

1. **Relationship to Ergodicity:**

$$\alpha\text{-mixing} \implies \text{ergodicity} \tag{10}$$

2. **Comparison with Other Mixing Types:**
   - β-mixing (absolute regularity)
   - ϕ-mixing (uniform mixing)
   - ρ-mixing (maximal correlation)

Hierarchy of Conditions

$$\text{i.i.d.} \implies \phi\text{-mixing} \implies \rho\text{-mixing} \implies \beta\text{-mixing} \implies \alpha\text{-mixing} \tag{11}$$

Stock Return Properties and Mixing

Empirical Evidence

1. **Return Characteristics:**
   - Weak serial correlation in returns
   - Strong dependence in volatility
   - Leverage effects
2. **Market Efficiency Implications:**

$$\alpha(h) \leq Ch^{-\beta}, \quad \beta > 2 \tag{12}$$

   - Consistent with weak-form efficiency
   - Allows for volatility clustering
   - Permits predictability in higher moments

0.2. Understanding Moment Conditions

Overview of Moment Conditions

The moment conditions in our assumption require finite $(4+\delta)$-th moments for returns, errors, and factors. Let's understand why each condition is necessary and what it buys us in terms of asymptotic theory.

Detailed Analysis of Each Condition

Condition (a): $E|R_{it}|^{4+\delta} < \infty$

This condition on asset returns is needed for several crucial reasons:
1. **Convergence Rates:**

$$\sqrt{T}(\hat{w}_T - w_0) \xrightarrow{d} N(0, V) \tag{13}$$

The fourth moment ensures:

- Existence of the asymptotic variance V

- Validity of the CLT for sample moments
- Uniform convergence of sample covariances

2. **Berry-Esseen Bounds:**

$$\sup_x |P(\sqrt{T}(\hat{w}_T - w_0) \leq x) - \Phi(x)| \leq \frac{C}{\sqrt{T}} \tag{14}$$

The extra δ moment ($E|R_{it}|^\delta < \infty$) provides:

- Better convergence rates
- Uniform integrability
- Tighter finite sample bounds

Condition (b): $E|\epsilon_{it}|^{4+\delta} < \infty$

This condition on error terms is crucial for:

1. **Variance Estimation:**

$$\hat{\Sigma}_T - \Sigma = O_p(T^{-1/2}) \tag{15}$$

Where:

- $\hat{\Sigma}_T$ is the sample variance of errors
- Fourth moments ensure consistency of variance estimators
- Extra δ provides uniform convergence

2. **HAC Estimation:**

$$\|\hat{\Omega}_T - \Omega\|_2 = O_p((T/m_T)^{-1/2} + m_T^{-q}) \tag{16}$$

Where:

- $\hat{\Omega}_T$ is the HAC estimator
- Fourth moments ensure kernel estimator convergence
- δ allows for optimal bandwidth selection

Condition (c): $\sup_t E\|F_t\|^{4+\delta} < \infty$

This condition on factors enables:

1. **Factor Structure Analysis:**

$$R_{it} = \beta_i' F_t + \epsilon_{it} \tag{17}$$

Providing:

- Well-defined factor loadings
- Stable estimation procedures
- Valid cross-sectional inference

2. **Uniform Bounds:**

$$\sup_{t,T} E\|\frac{1}{\sqrt{T}}\sum_{s=1}^{t}(F_s F_s' - E[F_s F_s'])\|_2 < \infty \tag{18}$$

Technical Implications

Why $4 + \delta$ Specifically?

1. **Fourth Moments:**

- Required for CLT with dependent data
- Needed for convergence of sample covariances
- Essential for HAC estimation

2. **The Role of δ:**

- Provides room for Lyapunov condition
- Ensures uniform integrability
- Allows for stronger convergence rates

Practical Considerations

Verification in Financial Data

1. **Return Distributions:**

$$\text{Kurtosis} = \frac{E[R_{it}^4]}{(E[R_{it}^2])^2} \tag{19}$$

Typical findings:

- Daily returns: kurtosis $\approx 5 - 10$
- Weekly returns: kurtosis $\approx 4 - 6$
- Monthly returns: kurtosis $\approx 3 - 4$

2. **Factor Properties:**

$$\text{Tail Index} = \lim_{x \to \infty} \frac{\log P(|F_t| > x)}{\log x} \tag{20}$$

Common observations:

- Market factor: tail index $\approx 4 - 5$
- Size factor: tail index $\approx 3 - 4$
- Value factor: tail index $\approx 4 - 5$

Consequences of Violation

If moment conditions fail:

1. **Statistical Issues:**

- Inconsistent variance estimation
- Invalid confidence intervals
- Poor finite sample properties

2. **Econometric Problems:**

- Unstable parameter estimates
- Unreliable hypothesis tests
- Invalid bootstrap procedures

0.3. Understanding Weight Convergence

Basic Concepts of Convergence

What is Convergence?

In our context, convergence means that our estimated weights (w_T^*) get arbitrarily close to the true weights (w^0) as our sample size (T) increases:

$$\|w_T^* - w^0\| \xrightarrow{p} 0 \tag{21}$$

This means:

- For any small error $\epsilon > 0$
- The probability of being more than ϵ away from w^0
- Goes to zero as $T \to \infty$

Why Do We Need Assumptions 1-3?

Assumption 1: Data Generating Process

$$R_{it} = \mu_i(F_t) + \epsilon_{it} \tag{22}$$

This assumption is needed because:

- Ensures returns have a factor structure
- Guarantees existence of synthetic portfolios
- Provides structure for identification

Assumption 2: Mixing Conditions

$$\sum_{h=1}^{\infty} h^2 \alpha(h)^{\delta/(2+\delta)} < \infty \tag{23}$$

This is crucial because:

- Allows for dependent data
- Ensures sample averages converge
- Permits use of uniform LLN

Assumption 3: Moment Conditions

$$E|R_{it}|^{4+\delta} < \infty \tag{24}$$

Required for:

- Existence of limiting distributions
- Uniform convergence of sample moments
- Well-behaved asymptotic theory

Understanding Uniform Convergence

What is Uniform Convergence?

For functions f_n, f on space $\mathcal{W}$:

$$\sup_{w \in \mathcal{W}} |f_n(w) - f(w)| \xrightarrow{p} 0 \tag{25}$$

Key aspects:

- Convergence happens simultaneously for all w
- Rate of convergence is uniform across $\mathcal{W}$
- Stronger than pointwise convergence

Why Do We Need Uniform Convergence?

Critical because:

- Ensures consistency of extremum estimators
- Prevents convergence from failing at the optimum
- Allows interchange of limits and optimization

The Uniform Law of Large Numbers (ULLN)

What is ULLN?

For a sequence of functions $\{g_t(w)\}$:

$$\sup_{w\in\mathcal{W}} |\frac{1}{T}\sum_{t=1}^{T} g_t(w) - E[g_t(w)]| \xrightarrow{p} 0 \tag{26}$$

Why we need it:

- Ensures objective function converges uniformly
- Provides rate of convergence
- Handles dependent data through mixing

The Second Moment Return Matrix

Definition

The second moment return matrix Σ is:

$$\Sigma = E[R_t R_t'] \tag{27}$$

where $R_t = (R_{1t}, ..., R_{Jt})'$

Positive Definiteness

A matrix Σ is positive definite if:

$$x'\Sigma x > 0 \quad \text{for all } x \neq 0 \tag{28}$$

Why it matters:

- Ensures unique solution exists
- Guarantees identification
- Provides stability for estimation

Establishing Identification

What is Identification?

Identification means:

$$w^0 = \arg\min_{w \in \mathcal{W}} Q(w) \quad \text{uniquely} \tag{29}$$

Where:

- $Q(w)$ is the population objective function
- w^0 is the unique minimizer
- No other weights give same synthetic returns

Role of Positive Definiteness

The objective function can be written as:

$$Q(w) = (w - w^0)'\Sigma(w - w^0) \tag{30}$$

Positive definiteness ensures:

- $Q(w) > 0$ for all $w \neq w^0$
- $Q(w^0) = 0$
- Unique minimum at w^0

Why is the Return Matrix Positive Definite?

Economic Arguments

1. **No Arbitrage:**
   - Perfect correlation implies arbitrage
   - Markets eliminate arbitrage
   - Therefore, returns can't be perfectly correlated
2. **Diversification:**
   - Assets have unique risk components

- Not all risk can be diversified away
- Implies linear independence of returns

Statistical Verification

We can verify positive definiteness by:

$$\lambda_{min}(\hat{\Sigma}) > 0 \tag{31}$$

Where:

- λ_{min} is the smallest eigenvalue
- $\hat{\Sigma}$ is the sample covariance
- Test statistic follows chi-square distribution

Full Proof Structure

1. **Show Uniform Convergence:**

$$\sup_{w \in \mathcal{W}} |Q_T(w) - Q(w)| \xrightarrow{p} 0 \tag{32}$$

2. **Apply ULLN:**

$$\|Q_T(w) - Q(w)\|_\infty = O_p(T^{-1/2} \log T) \tag{33}$$

3. **Use Identification:**

- Positive definiteness ensures unique minimum
- ULLN ensures sample objective converges
- Therefore, minimizer converges to w^0

4. **Conclude:**

$$\|w_T^* - w^0\| \xrightarrow{p} 0 \tag{34}$$

0.4. Detailed Proof of Weight Consistency

Why We Need Objective Function Convergence

The logic follows these steps:

1. Our estimator is defined as:

$$w_T^* = \arg\min_{w \in \mathcal{W}} Q_T(w) \tag{35}$$

2. The population optimum is:

$$w^0 = \arg\min_{w \in \mathcal{W}} Q(w) \tag{36}$$

3. For consistency ($w_T^* \xrightarrow{p} w^0$), we need:

$$\|Q_T(w) - Q(w)\| \text{ small } \implies \|w_T^* - w^0\| \text{ small} \tag{37}$$

This implication requires:

- Uniform convergence of Q_T to Q
- Unique identification of w^0
- Continuous mapping from objective to weights

Mathematical Proof of Uniform Convergence

Step 1: Express the Objective Functions

Sample objective:

$$Q_T(w) = \frac{1}{T} \sum_{t=1}^{T} (R_{it} - \sum_{j=1}^{J} w_j R_{jt})^2 \tag{38}$$

Population objective:

$$Q(w) = E[(R_{it} - \sum_{j=1}^{J} w_j R_{jt})^2] \tag{39}$$

Step 2: Decomposition

Expand the difference:

$$Q_T(w) - Q(w) = \frac{1}{T}\sum_{t=1}^{T}(R_{it} - w'R_t)^2 - E[(R_{it} - w'R_t)^2] \tag{40}$$

$$= \frac{1}{T}\sum_{t=1}^{T}(R_{it}^2 - E[R_{it}^2]) \tag{41}$$

$$- 2w'\left(\frac{1}{T}\sum_{t=1}^{T}R_tR_{it} - E[R_tR_{it}]\right) \tag{42}$$

$$+ w'\left(\frac{1}{T}\sum_{t=1}^{T}R_tR_t' - E[R_tR_t']\right)w \tag{43}$$

Step 3: Bound the Supremum

Using triangle inequality:

$$\sup_{w\in\mathcal{W}}|Q_T(w) - Q(w)| \leq |\frac{1}{T}\sum_{t=1}^{T}(R_{it}^2 - E[R_{it}^2])| \tag{44}$$

$$+ 2\|w\|\|\frac{1}{T}\sum_{t=1}^{T}R_tR_{it} - E[R_tR_{it}]\| \tag{45}$$

$$+ \|w\|^2\|\frac{1}{T}\sum_{t=1}^{T}R_tR_t' - E[R_tR_t']\| \tag{46}$$

Why We Need ULLN and What It Buys Us

Role of ULLN

The ULLN gives us:

$$\|\frac{1}{T}\sum_{t=1}^{T}g_t(w) - E[g_t(w)]\|_\infty = O_p(T^{-1/2}\log T) \tag{47}$$

This provides:

- Rate of convergence
- Uniform control over w
- Valid under mixing conditions

Application to Our Setting

For our components:

$$\|\frac{1}{T}\sum_{t=1}^{T}R_tR_t' - E[R_tR_t']\| = O_p(T^{-1/2}\log T) \tag{48}$$

$$\|\frac{1}{T}\sum_{t=1}^{T}R_tR_{it} - E[R_tR_{it}]\| = O_p(T^{-1/2}\log T) \tag{49}$$

$$|\frac{1}{T}\sum_{t=1}^{T}(R_{it}^2 - E[R_{it}^2])| = O_p(T^{-1/2}\log T) \tag{50}$$

Complete Proof of Identification

Step 1: Express Second-Order Condition

The population objective can be written as:

$$Q(w) = E[R_{it}^2] - 2w'E[R_tR_{it}] + w'E[R_tR_t']w \tag{51}$$

Step 2: First-Order Conditions

Differentiate with respect to w:

$$\nabla Q(w) = -2E[R_tR_{it}] + 2E[R_tR_t']w = 0 \tag{52}$$

Solving for w^0:

$$w^0 = E[R_tR_t']^{-1}E[R_tR_{it}] \tag{53}$$

Step 3: Verify Second-Order Conditions

The Hessian is:

$$\nabla^2 Q(w) = 2E[R_tR_t'] = 2\Sigma \tag{54}$$

Positive definiteness follows because:

1. For any $x \neq 0$:

$$x'\Sigma x = E[(x'R_t)^2] > 0 \tag{55}$$

2. This holds because:
    - No perfect collinearity (by no-arbitrage)
    - Finite second moments (by assumption)
    - Non-degenerate returns (by market efficiency)

Step 4: Complete the Proof

1. By ULLN:

$$\sup_{w\in\mathcal{W}} |Q_T(w) - Q(w)| \xrightarrow{p} 0 \tag{56}$$

2. By positive definiteness:

$$Q(w) - Q(w^0) \geq \lambda_{min}(\Sigma)\|w - w^0\|^2 \tag{57}$$

3. Therefore:

$$\|w_T^* - w^0\| \leq \frac{1}{\lambda_{min}(\Sigma)} \sup_{w\in\mathcal{W}} |Q_T(w) - Q(w)| \xrightarrow{p} 0 \tag{58}$$

This completes the proof by showing:

- Uniform convergence of objective function
- Unique identification through positive definiteness
- Explicit rate of convergence via ULLN
- Direct link between objective and parameter convergence

0.5. Understanding the Uniform Law of Large Numbers

Origin of the Rate

The Standard Result

The rate $O_p(T^{-1/2}\log T)$ is not standard for i.i.d. data. For i.i.d. observations, we typically have:

$$\|\frac{1}{T}\sum_{t=1}^{T} g_t(w) - E[g_t(w)]\|_\infty = O_p(T^{-1/2}) \tag{59}$$

The additional $\log T$ term appears due to:

- Dependence in the data (mixing conditions)
- Uniformity over the parameter space
- Need for maximal inequalities

Deriving the Rate

Key Steps

1. **Decomposition:** For fixed w:

$$\frac{1}{T}\sum_{t=1}^{T} g_t(w) - E[g_t(w)] = \frac{1}{T}\sum_{t=1}^{T}[g_t(w) - E[g_t(w)]] \equiv \mathbb{G}_T(w) \tag{60}$$

2. **Covering Numbers:** Define $\mathcal{N}(\epsilon, \mathcal{W}, \|\cdot\|)$ as the minimum number of ϵ-balls needed to cover $\mathcal{W}$.
3. **Entropy Condition:** For some $C < \infty$:

$$\int_0^1 \sqrt{\log \mathcal{N}(\epsilon, \mathcal{W}, \|\cdot\|)}d\epsilon \leq C \tag{61}$$

Maximal Inequality

Under mixing conditions, we have:

$$E[\sup_{w\in\mathcal{W}} |\mathbb{G}_T(w)|] \leq C\left(\frac{\log T}{T}\right)^{1/2} \tag{62}$$

This follows from:

- Moment bounds from mixing conditions
- Entropy integral bound
- Chaining argument

Components of the Rate

The $T^{-1/2}$ Term

This comes from:

$$\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} g_t(w)\right) = O(T^{-1}) \tag{63}$$

Under mixing:

$$\sum_{h=1}^{\infty} |\text{Cov}(g_t(w), g_{t+h}(w))| < \infty \tag{64}$$

The $\log T$ Term

Appears due to:

$$\sup_{w \in \mathcal{W}} |\mathbb{G}_T(w)| = \max_{1 \leq j \leq N_T} |\mathbb{G}_T(w_j)| + O_p(T^{-1/2}) \tag{65}$$

Where:

- N_T is the covering number
- Grows polynomially with T
- Introduces $\log T$ term

Uniform Control over w

Why Uniformity Matters

The result provides:

$$P\left(\sup_{w \in \mathcal{W}} |\mathbb{G}_T(w)| > M \left(\frac{\log T}{T}\right)^{1/2}\right) \to 0 \tag{66}$$

This means:

- Control over entire parameter space
- Valid for optimization problems
- Handles parameter estimation

Validity Under Mixing

Required Conditions

1. **Mixing Rate:**

$$\alpha(h) \leq Ch^{-\beta}, \quad \beta > 2 \tag{67}$$

2. **Moment Bounds:**

$$E|g_t(w)|^{2+\delta} < \infty \tag{68}$$

3. **Lipschitz Condition:**

$$|g_t(w_1) - g_t(w_2)| \leq L_t \|w_1 - w_2\| \tag{69}$$

where $E[L_t^{2+\delta}] < \infty$

Technical Extensions

Stronger Rates

Under additional conditions:

$$\|\mathbb{G}_T\|_\infty = O_p\left(\left(\frac{\log\log T}{T}\right)^{1/2}\right) \tag{70}$$

Requires:

- Stronger mixing ($\beta > 4$)
- Higher moments ($4 + \delta$)
- Bounded parameter space

Empirical Process Theory

Connection to:

$$\{\mathbb{G}_T(w) : w \in \mathcal{W}\} \Rightarrow \{\mathbb{G}(w) : w \in \mathcal{W}\} \tag{71}$$

Where:

- $\Rightarrow$ denotes weak convergence
- $\mathbb{G}$ is a Gaussian process
- With covariance kernel from mixing

Practical Implications

For Synthetic Controls

1. **Weight Estimation:**

$$\|w_T^* - w^0\| = O_p\left(\left(\frac{\log T}{T}\right)^{1/2}\right) \tag{72}$$

2. **Inference:**

$$P(w^0 \in \mathcal{C}_T) = 1 - \alpha + o(1) \tag{73}$$

where $\mathcal{C}_T$ is a confidence region constructed using the rate.

Machine Learning SCM [CLAUDE]

We can extend the linear SCM framework to capture nonlinear relationships between financial instruments using machine learning methods. Let $f_\theta : \mathbb{R}^J \to \mathbb{R}$ be a neural network parameterized by θ that maps the returns of the donor pool to a synthetic return:

$$R_{0t}^* = f_\theta(R_{1t}, \ldots, R_{Jt})$$

The network parameters θ are trained to minimize the loss function:

$$\mathcal{L}(\theta) = \frac{1}{T_{tr}} \sum_{t \in \mathcal{T}_{tr}} (R_{0t} - f_\theta(R_{1t}, \ldots, R_{Jt}))^2 + \lambda \mathcal{R}(\theta)$$

where $\mathcal{R}(\theta)$ is a regularization term on the network parameters.

Architecture Design

We propose a feed-forward neural network with the following structure:

- **Input Layer**: J nodes corresponding to the donor pool returns
- **Hidden Layers**: Multiple layers with ReLU activation functions

  $$h^{(l+1)} = \text{ReLU}(W^{(l)} h^{(l)} + b^{(l)})$$

  where $W^{(l)}$ and $b^{(l)}$ are the weights and biases of layer l
- **Output Layer**: Single node with linear activation to predict the target return
- **Residual Connections**: To facilitate learning of linear relationships, we add skip connections from input to output:

  $$f_\theta(x) = \text{NN}_\theta(x) + w'x$$

  where w is a learnable weight vector constrained to sum to 1

Training Procedure

The model is trained using:

- **Loss Function**: Mean squared error with L2 regularization

  $$\mathcal{R}(\theta) = \sum_l (\|W^{(l)}\|_F^2 + \|b^{(l)}\|_2^2)$$

  where $\|\cdot\|_F$ denotes the Frobenius norm

- **Optimization**: Adam optimizer with learning rate scheduling

$$\theta_{t+1} = \theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

  where $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected moment estimates

- **Early Stopping**: Training is stopped when validation loss stops improving to prevent overfitting

Ensemble Methods

To improve robustness, we can employ ensemble methods:

- **Bagging**: Train multiple networks on bootstrap samples of the training data

$$R_{0t}^* = \frac{1}{K} \sum_{k=1}^{K} f_{\theta_k}(R_{1t}, \ldots, R_{Jt})$$

  where K is the number of networks in the ensemble

- **Dropout**: Apply dropout during training and use Monte Carlo dropout during inference

$$R_{0t}^* = \mathbb{E}_{p(z)}[f_\theta(R_{1t}, \ldots, R_{Jt}, z)]$$

  where z represents random dropout masks

Table 1: Statistics of $\mathcal{P}$ Across Data Splits

Split	**Algo.**	**Cum. Ret.**	**Avg. Ret.**	**St. Dev.**	**Sharpe**	**Sortino**	**Max. DD**	**Calmar**	**Skew.**	**Kurt.**	**VaR 95%**	**CVaR 95%**
All	*Greedy*	1.070	5.3	9.7	0.5	0.6	-6.9	0.8	-0.50	4.17	-0.009	-0.014
	Stable	1.489	35.8	16.8	1.8	2.2	-7.6	4.7	0.08	5.09	-0.014	-0.023
Train	*Greedy*	0.969	-4.6	11.6	-0.4	-0.4	-6.5	-0.7	-0.59	2.96	-0.011	-0.018
	Stable	1.285	46.3	19.3	2.0	2.4	-7.6	6.1	-0.30	3.63	-0.018	-0.026
Validation	*Greedy*	1.088	26.6	7.3	3.2	3.7	-3.5	7.7	-0.49	1.19	-0.006	-0.010
	Stable	1.149	47.7	13.3	2.9	3.4	-3.6	13.1	-0.24	1.78	-0.012	-0.018
Test	*Greedy*	1.014	4.9	6.9	0.7	1.0	-3.6	1.4	1.82	5.39	-0.005	-0.006
	Stable	1.008	2.9	14.3	0.2	0.3	-4.6	0.6	2.32	13.73	-0.012	-0.017

Note: The holding period of the beta-neutral strategies is set to $L = 4$ trading days and the number of traded clusters is $\theta = 0.5k = 13$ (as we have $k^* = 26$ clusters). The selection criteria for these parameters is based on maximizing the Sharpe Ratios of the train and validation samples.