
Prerequisites for Adaptive Portfolio Construction and Regime Detection

A self-contained primer covering Hidden Markov Models, the EM algorithm, sparse PCA, CVaR optimisation, and mean-reversion analysis.

Melvin Alvarez-Caradu
Independent Researcher, HFThot Research Lab
<https://hfthot-lab.eu>
June 2026 — v1.0

DISTRIBUTION: REGISTERED USERS — FREE TIER

This document provides a self-contained mathematical primer for readers approaching the HFThot Research Lab white paper “Adaptive Portfolio Construction with Regime Detection and Execution Optimisation” (WP3). We develop, from first principles, the essential mathematical machinery: Hidden Markov Models and the Baum–Welch algorithm, spectral clustering and model selection via BIC, Principal Component Analysis (classical and sparse), convex optimisation for portfolio construction, Conditional Value-at-Risk and its LP reformulation, and mean-reversion modelling via the Ornstein–Uhlenbeck process. Each section contains full derivations, worked examples, and exercises.

Prerequisites

Required background for this primer: Linear algebra (eigenvalues, SVD, positive definite matrices), probability theory (Bayes’ theorem, conditional distributions, maximum likelihood), basic optimisation (Lagrange multipliers, KKT conditions), and elementary statistics (mean, variance, covariance). No prior knowledge of HMMs, portfolio theory, or risk measures is assumed.

Contents

1	Probability and Statistics Review	3
1.1	Sample Spaces and Events	3
1.2	Conditional Probability and Bayes’ Theorem	3
1.3	Common Probability Distributions	4
1.3.1	Discrete Distributions	4
1.3.2	Continuous Distributions	4
1.4	Maximum Likelihood Estimation	4
1.5	The Law of Large Numbers and Central Limit Theorem	5
1.6	Exercises	5
2	Markov Chains	7
2.1	Definitions and Notation	7
2.2	Application: Market Regimes	7
2.3	Exercises	7
3	Hidden Markov Models	9
3.1	Model Specification	9
3.2	The Three Fundamental Problems	9
3.3	The Forward–Backward Algorithm	9
3.4	The Viterbi Algorithm (Decoding)	10
3.5	Exercises	10
4	The Baum–Welch Algorithm (EM for HMMs)	12
4.1	The Expectation–Maximisation Framework	12
4.2	Baum–Welch Update Equations	12
4.3	Model Selection: BIC	12
4.4	Exercises	13

5	Principal Component Analysis	14
5.1	Classical PCA	14
5.2	Singular Value Decomposition (SVD)	14
5.3	Sparse PCA	14
5.4	Exercises	15
6	Fundamentals of Optimisation	16
6.1	Unconstrained Optimisation	16
6.2	Gradient Descent	16
6.3	Constrained Optimisation: Lagrange Multipliers	16
6.4	Convex Functions and Sets	17
6.5	Exercises	17
7	Convex Optimisation for Portfolios	19
7.1	Portfolio Optimisation: Setup	19
7.2	Mean–Variance (Markowitz)	19
7.3	Risk Parity	19
7.4	The KKT Conditions	19
7.5	Exercises	20
8	Conditional Value-at-Risk	21
8.1	Value-at-Risk (VaR)	21
8.2	CVaR: Definition and Properties	21
8.3	Rockafellar–Uryasev LP Reformulation	21
8.4	Portfolio CVaR Minimisation as LP	22
8.5	Exercises	22
9	Introduction to Stochastic Differential Equations	24
9.1	From Random Walks to Brownian Motion	24
9.2	What Is a Stochastic Differential Equation?	24
9.3	Itô’s Formula: The Chain Rule for SDEs	25
9.4	The Euler–Maruyama Method	25
9.5	The Milstein Method (Higher Order)	25
9.6	Exercises	26
10	Mean-Reversion and the Ornstein–Uhlenbeck Process	27
10.1	Definition	27
10.2	Distributional Properties	27
10.3	Half-Life of Mean Reversion	27
10.4	Estimation via OLS Regression	28
10.5	Exercises	28
11	Regime-Conditional Portfolio Construction	29
11.1	The Adaptive Framework	29
11.2	Regime Transition Detection	29
11.3	Ablation Study Design	29
11.4	Exercises	30
12	Summary and Connections to WP3	31
	References	32

1 Probability and Statistics Review

This section is designed for readers who need a refresher on the probability and statistics concepts used throughout this primer. Readers comfortable with Bayes' theorem, maximum likelihood, and common distributions may skip to Section 2.

1.1 Sample Spaces and Events

Definition 1.1 (Probability Space). A *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ consists of:

- Ω : the *sample space* (set of all possible outcomes);
- $\mathcal{F}$: a σ -algebra of *events* (measurable subsets of Ω);
- $\mathbb{P}$: a *probability measure* with $\mathbb{P}(\Omega) = 1$.

Example 1.1 (Coin Toss). For a fair coin tossed twice: $\Omega = \{HH, HT, TH, TT\}$, $\mathcal{F} = 2^\Omega$ (all subsets), $\mathbb{P}(\{HH\}) = \mathbb{P}(\{HT\}) = \mathbb{P}(\{TH\}) = \mathbb{P}(\{TT\}) = 1/4$. The event “at least one head” is $A = \{HH, HT, TH\}$ with $\mathbb{P}(A) = 3/4$.

1.2 Conditional Probability and Bayes' Theorem

Definition 1.2 (Conditional Probability). For events A, B with $\mathbb{P}(B) > 0$:

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \quad (1)$$

Theorem 1.1 (Bayes' Theorem). For a partition $\{B_1, \dots, B_K\}$ of Ω with $\mathbb{P}(B_k) > 0$:

$$\mathbb{P}(B_k | A) = \frac{\mathbb{P}(A | B_k)\mathbb{P}(B_k)}{\sum_{j=1}^K \mathbb{P}(A | B_j)\mathbb{P}(B_j)}. \quad (2)$$

Proof. By the definition of conditional probability, $\mathbb{P}(B_k | A) = \mathbb{P}(A \cap B_k)/\mathbb{P}(A)$. Using $\mathbb{P}(A \cap B_k) = \mathbb{P}(A | B_k)\mathbb{P}(B_k)$ and the law of total probability $\mathbb{P}(A) = \sum_j \mathbb{P}(A | B_j)\mathbb{P}(B_j)$ gives (2). $\square$

Example 1.2 (Medical Test — Worked). A disease has prevalence 1%. A test has sensitivity 95% and specificity 90%. If a patient tests positive, what is $\mathbb{P}(\text{disease} | \text{positive})$?

Solution. Let D = disease, $+$ = positive test.

$$\mathbb{P}(D | +) = \frac{\mathbb{P}(+ | D)\mathbb{P}(D)}{\mathbb{P}(+ | D)\mathbb{P}(D) + \mathbb{P}(+ | D^c)\mathbb{P}(D^c)} \quad (3)$$

$$= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.10 \times 0.99} = \frac{0.0095}{0.0095 + 0.099} = \frac{0.0095}{0.1085} \approx 8.8\%. \quad (4)$$

Despite an accurate test, the low prevalence means most positives are false! This is why *prior probability* matters — a key insight for Bayesian inference.

Key Insight

Bayes' theorem is the mathematical foundation for regime detection in HMMs: the “prior” is the transition probability from the previous state, and the “likelihood” is the emission probability of the observed data given each state. The “posterior” is the updated belief

about which regime we are in.

1.3 Common Probability Distributions

1.3.1 Discrete Distributions

Definition 1.3 (Bernoulli and Binomial). $X \sim \text{Bernoulli}(p)$: $\mathbb{P}(X = 1) = p$, $\mathbb{P}(X = 0) = 1 - p$.
 $Y \sim \text{Binomial}(n, p)$: $\mathbb{P}(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$. $\mathbb{E}[Y] = np$, $\text{Var}(Y) = np(1 - p)$.

Definition 1.4 (Poisson Distribution). $X \sim \text{Poisson}(\lambda)$: $\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$. $\mathbb{E}[X] = \text{Var}(X) = \lambda$. Used to model rare events (e.g., number of large price jumps per day).

1.3.2 Continuous Distributions

Definition 1.5 (Normal (Gaussian) Distribution). $X \sim \mathcal{N}(\mu, \sigma^2)$ has density:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right). \quad (5)$$

The standard normal ($\mu = 0, \sigma = 1$) CDF is denoted $\Phi(x)$ and its density $\phi(x)$.

Definition 1.6 (Multivariate Normal). $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$ positive definite:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right). \quad (6)$$

Example 1.3 (Bivariate Normal — Worked). Let $(X, Y) \sim \mathcal{N}(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix})$. The conditional distribution is:

$$Y \mid X = x \sim \mathcal{N}(\rho x, 1 - \rho^2).$$

For $\rho = 0.8$ and $X = 2$: $Y \mid X = 2 \sim \mathcal{N}(1.6, 0.36)$. This shows that knowing X narrows our uncertainty about Y — the conditional variance $1 - \rho^2 = 0.36$ is much smaller than the marginal variance 1.

1.4 Maximum Likelihood Estimation

Definition 1.7 (Maximum Likelihood Estimator). Given i.i.d. observations $x_1, \dots, x_n$ from a distribution $f(x \mid \theta)$, the *maximum likelihood estimator* (MLE) is:

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^n f(x_i \mid \theta) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n \log f(x_i \mid \theta). \quad (7)$$

Example 1.4 (MLE for Gaussian Mean — Worked). For $x_1, \dots, x_n \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known:

$$\ell(\mu) = \sum_{i=1}^n \log f(x_i \mid \mu) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2. \quad (8)$$

Setting $\ell'(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0$ gives:

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}. \quad (9)$$

The MLE for the Gaussian mean is simply the sample mean.

Practical Application

Connection to HMMs: The Baum–Welch algorithm (Section 4) is an EM algorithm that finds MLEs for HMM parameters. The M-step updates (29)–(30) are *weighted* MLEs where the weights $\gamma_t(k)$ are the posterior probabilities of each hidden state.

1.5 The Law of Large Numbers and Central Limit Theorem

Theorem 1.2 (Strong Law of Large Numbers). *If $X_1, X_2, \dots$ are i.i.d. with $\mathbb{E}[|X_1|] < \infty$, then:*

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{a.s.} \mathbb{E}[X_1] \quad \text{as } n \rightarrow \infty. \quad (10)$$

Theorem 1.3 (Central Limit Theorem). *If $X_1, X_2, \dots$ are i.i.d. with $\mathbb{E}[X_1] = \mu$ and $\text{Var}(X_1) = \sigma^2 < \infty$, then:*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } n \rightarrow \infty. \quad (11)$$

Example 1.5 (CLT for Portfolio Returns — Worked). A portfolio has 100 equally weighted assets with mean daily return $\mu = 0.05\%$ and standard deviation $\sigma = 2\%$ (assumed independent for simplicity). The portfolio return is $\bar{R} = \frac{1}{100} \sum_{i=1}^{100} R_i$. By CLT: $\bar{R} \approx \mathcal{N}(0.05\%, (2\%/\sqrt{100})^2) = \mathcal{N}(0.05\%, 0.04\%)$. The portfolio std is 0.2% vs. individual asset std of 2% — diversification reduces risk by a factor of $\sqrt{n} = 10$.

1.6 Exercises

- Bayes' theorem.** A trading signal has $\mathbb{P}(\text{signal} \mid \text{up day}) = 0.7$ and $\mathbb{P}(\text{signal} \mid \text{down day}) = 0.3$. Historical data shows $\mathbb{P}(\text{up day}) = 0.55$. Compute $\mathbb{P}(\text{up day} \mid \text{signal})$.
- MLE.** Derive the MLE for the variance σ^2 of a Gaussian with known mean $\mu = 0$, given observations $x_1, \dots, x_n$.
- Central Limit Theorem.** A stock returns have $\mu = 0.08\%$ daily and $\sigma = 1.5\%$ daily. Using CLT, approximate the probability that the average return over 252 trading days exceeds 0.15%.
- Conditional distributions.** If (X, Y) is bivariate normal with $\mu_X = 0$, $\mu_Y = 1$, $\sigma_X = 1$, $\sigma_Y = 2$, $\rho = -0.5$, find $\mathbb{E}[Y \mid X = 1]$ and $\text{Var}(Y \mid X = 1)$.

Solutions

Solution to Exercise 1. By Bayes' theorem:

$$\mathbb{P}(\text{up} \mid \text{signal}) = \frac{0.7 \times 0.55}{0.7 \times 0.55 + 0.3 \times 0.45} = \frac{0.385}{0.385 + 0.135} = \frac{0.385}{0.52} \approx 0.7404. \quad (12)$$

The signal increases the probability of an up day from 55% to $\approx 74\%$. $\square$

Solution to Exercise 2. The log-likelihood is: $\ell(\sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i x_i^2$. Taking the derivative: $\ell'(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \sum_i x_i^2 = 0$, giving $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$. $\square$

Solution to Exercise 3. Let $\bar{R}_{252}$ be the average daily return over 252 days. By CLT: $\bar{R}_{252} \approx \mathcal{N}(0.08\%, (1.5\%/\sqrt{252})^2)$. Standard error: $\text{SE} = 1.5\%/\sqrt{252} \approx 0.0945\%$. $\mathbb{P}(\bar{R}_{252} > 0.15\%) = \mathbb{P}(Z > \frac{0.15 - 0.08}{0.0945}) = \mathbb{P}(Z > 0.741) \approx 22.9\%$. $\square$

Solution to Exercise 4. For bivariate normal, $\mathbb{E}[Y \mid X = x] = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$ and $\text{Var}(Y \mid X) = \sigma_Y^2(1 - \rho^2)$. So $\mathbb{E}[Y \mid X = 1] = 1 + (-0.5) \cdot \frac{2}{1} \cdot (1 - 0) = 1 - 1 = 0$ and $\text{Var}(Y \mid X = 1) = 4(1 - 0.25) = 3$. $\square$

2 Markov Chains

2.1 Definitions and Notation

Definition 2.1 (Discrete-Time Markov Chain). A *discrete-time Markov chain* (DTMC) is a sequence of random variables $(Z_t)_{t=0,1,2,\dots}$ taking values in a finite state space $\{1, \dots, K\}$ such that:

$$\mathbb{P}(Z_{t+1} = j \mid Z_t = i, Z_{t-1}, \dots, Z_0) = \mathbb{P}(Z_{t+1} = j \mid Z_t = i) = a_{ij}. \quad (13)$$

The matrix $A = (a_{ij})_{i,j=1}^K$ is the *transition matrix*, with $a_{ij} \geq 0$ and $\sum_j a_{ij} = 1$.

Definition 2.2 (Stationary Distribution). A probability vector $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ is a *stationary distribution* if $\boldsymbol{\pi}^\top A = \boldsymbol{\pi}^\top$, i.e., $\boldsymbol{\pi}$ is a left eigenvector of A with eigenvalue 1.

Theorem 2.1 (Perron–Frobenius). *If A is irreducible and aperiodic, then the stationary distribution exists, is unique, and $A^n \rightarrow \mathbf{1}\boldsymbol{\pi}^\top$ as $n \rightarrow \infty$.*

2.2 Application: Market Regimes

In financial markets, regime-switching models posit that returns are generated by different distributions depending on the current “regime” (e.g., bull, bear, crisis). The regime sequence (Z_t) follows a Markov chain, but regimes are *not directly observable* — we only see asset returns.

Key Insight

Market regimes (bull, bear, crisis) are *hidden states* that drive observable returns. Hidden Markov Models provide a principled framework for inferring these states from data, leading to adaptive portfolio strategies that respond to regime changes.

2.3 Exercises

1. For the transition matrix $A = \begin{pmatrix} 0.95 & 0.05 \\ 0.10 & 0.90 \end{pmatrix}$ (bull $\leftrightarrow$ bear), compute the stationary distribution.
2. Show that the expected sojourn time in state i is $1/(1 - a_{ii})$.
3. Simulate a 3-state Markov chain with given A and verify that the empirical distribution converges to $\boldsymbol{\pi}$.
4. **Multinomial transition.** For a 3-state chain with transition matrix $A = \begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.1 & 0.8 & 0.1 \\ 0.05 & 0.15 & 0.8 \end{pmatrix}$, compute A^2 and interpret what it means for 2-step transitions.
5. **Absorbing states.** Modify the 2-state chain so that “bear” is absorbing ($a_{22} = 1$). What happens to A^n as $n \rightarrow \infty$? What is the economic interpretation?

Solutions

Solution to Exercise 1. The stationary distribution satisfies $\boldsymbol{\pi}^\top A = \boldsymbol{\pi}^\top$, $\pi_1 + \pi_2 = 1$. Writing out: $\pi_1 = 0.95\pi_1 + 0.10\pi_2 \Rightarrow 0.05\pi_1 = 0.10\pi_2 \Rightarrow \pi_1 = 2\pi_2$. With $\pi_1 + \pi_2 = 1$: $\pi_1 = 2/3$, $\pi_2 = 1/3$. The market spends $\approx 67\%$ of the time in bull and $\approx 33\%$ in bear. $\square$

Solution to Exercise 2. Starting from state i , the sojourn time T_i (steps spent before leaving) is geometric: $\mathbb{P}(T_i = k) = a_{ii}^{k-1}(1 - a_{ii})$ for $k \geq 1$. The expectation of a geometric distribution with parameter $p = 1 - a_{ii}$ is $\mathbb{E}[T_i] = 1/(1 - a_{ii})$.

For the bull market with $a_{11} = 0.95$: $\mathbb{E}[T_1] = 1/0.05 = 20$ days. For bore with $a_{22} = 0.90$: $\mathbb{E}[T_2] = 1/0.10 = 10$ days. Bull periods are on average twice as long as bear periods, consistent with $\pi_1/\pi_2 = 2$. $\square$

Solution to Exercise 4 (Multinomial). $A^2 = A \cdot A$. The (i, j) entry of A^2 gives $\mathbb{P}(Z_{t+2} = j \mid Z_t = i)$. By Chapman–Kolmogorov: $(A^2)_{ij} = \sum_k a_{ik}a_{kj}$. For example, $\mathbb{P}(\text{bull} \rightarrow \text{crisis in 2 steps}) = 0.9 \times 0.05 + 0.05 \times 0.1 + 0.05 \times 0.8 = 0.09$. $\square$

3 Hidden Markov Models

3.1 Model Specification

Definition 3.1 (Hidden Markov Model). A *Hidden Markov Model* (HMM) is defined by:

- **Hidden states:** $(Z_t)_{t=1}^T$ with $Z_t \in \{1, \dots, K\}$, following a Markov chain with transition matrix A and initial distribution $\boldsymbol{\pi}$;
- **Observations:** $(Y_t)_{t=1}^T$ with $Y_t \in \mathbb{R}^d$, conditionally independent given Z :

$$Y_t \mid Z_t = k \sim f_k(y_t \mid \boldsymbol{\theta}_k), \quad (14)$$

where f_k is the *emission distribution* for state k .

The full parameter set is $\boldsymbol{\Theta} = (A, \boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K)$.

For financial applications, the emission distributions are typically Gaussian:

$$Y_t \mid Z_t = k \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k), \quad (15)$$

so $\boldsymbol{\theta}_k = (\boldsymbol{\mu}_k, \Sigma_k)$.

3.2 The Three Fundamental Problems

1. **Evaluation:** Given $\boldsymbol{\Theta}$ and observations $\mathbf{Y} = (Y_1, \dots, Y_T)$, compute $\mathbb{P}(\mathbf{Y} \mid \boldsymbol{\Theta})$.
2. **Decoding:** Find the most likely state sequence $\hat{\mathbf{Z}} = \operatorname{argmax}_{\mathbf{Z}} \mathbb{P}(\mathbf{Z} \mid \mathbf{Y}, \boldsymbol{\Theta})$.
3. **Learning:** Find $\hat{\boldsymbol{\Theta}} = \operatorname{argmax}_{\boldsymbol{\Theta}} \mathbb{P}(\mathbf{Y} \mid \boldsymbol{\Theta})$.

3.3 The Forward–Backward Algorithm

Definition 3.2 (Forward Variable).

$$\alpha_t(k) = \mathbb{P}(Y_1, \dots, Y_t, Z_t = k \mid \boldsymbol{\Theta}). \quad (16)$$

Algorithm Forward Algorithm

Initialisation ($t = 1$):

$$\alpha_1(k) = \pi_k \cdot f_k(Y_1), \quad k = 1, \dots, K. \quad (17)$$

Recursion ($t = 2, \dots, T$):

$$\alpha_t(k) = f_k(Y_t) \sum_{j=1}^K \alpha_{t-1}(j) a_{jk}. \quad (18)$$

Termination:

$$\mathbb{P}(\mathbf{Y} \mid \boldsymbol{\Theta}) = \sum_{k=1}^K \alpha_T(k). \quad (19)$$

Definition 3.3 (Backward Variable).

$$\beta_t(k) = \mathbb{P}(Y_{t+1}, \dots, Y_T \mid Z_t = k, \boldsymbol{\Theta}). \quad (20)$$

Algorithm Backward Algorithm**Initialisation** ($t = T$): $\beta_T(k) = 1$ for all k .**Recursion** ($t = T - 1, \dots, 1$):

$$\beta_t(k) = \sum_{j=1}^K a_{kj} f_j(Y_{t+1}) \beta_{t+1}(j). \quad (21)$$

The posterior state probabilities combine forward and backward:

$$\gamma_t(k) = \mathbb{P}(Z_t = k \mid \mathbf{Y}, \boldsymbol{\Theta}) = \frac{\alpha_t(k) \beta_t(k)}{\sum_j \alpha_t(j) \beta_t(j)}. \quad (22)$$

3.4 The Viterbi Algorithm (Decoding)**Algorithm Viterbi Algorithm**

Working in log-space for numerical stability:

Initialisation:

$$\delta_1(k) = \log \pi_k + \log f_k(Y_1), \quad \psi_1(k) = 0. \quad (23)$$

Recursion ($t = 2, \dots, T$):

$$\delta_t(k) = \max_{1 \leq j \leq K} [\delta_{t-1}(j) + \log a_{jk}] + \log f_k(Y_t), \quad (24)$$

$$\psi_t(k) = \operatorname{argmax}_{1 \leq j \leq K} [\delta_{t-1}(j) + \log a_{jk}]. \quad (25)$$

Backtracking:

$$\hat{Z}_T = \operatorname{argmax}_k \delta_T(k), \quad \hat{Z}_t = \psi_{t+1}(\hat{Z}_{t+1}). \quad (26)$$

Complexity: $O(TK^2)$ time, $O(TK)$ space.

3.5 Exercises

1. Implement the forward algorithm and verify on a 2-state HMM with known parameters that $\sum_k \alpha_T(k)$ matches the brute-force computation (summing over all K^T state sequences).
2. Show that the forward algorithm has complexity $O(TK^2)$.
3. Implement the Viterbi algorithm and compare the decoded state sequence with the true hidden states on simulated data.
4. **Scaling.** The forward variables $\alpha_t(k)$ can underflow for long sequences. Show that using log-scaled forward variables $\hat{\alpha}_t(k) = \alpha_t(k)/c_t$ where $c_t = \sum_k \alpha_t(k)$ avoids this, and that $\log \mathbb{P}(\mathbf{Y}) = \sum_t \log c_t$.
5. **Posterior decoding.** An alternative to Viterbi is posterior decoding: $\hat{Z}_t = \operatorname{argmax}_k \gamma_t(k)$ independently at each t . Give an example where Viterbi and posterior decoding give different answers. Which is more appropriate for regime detection?

Solutions

Solution to Exercise 1. The brute-force computation sums over all K^T state sequences: $\mathbb{P}(\mathbf{Y}) = \sum_{z_1} \dots \sum_{z_T} \pi_{z_1} f_{z_1}(y_1) \prod_{t=2}^T a_{z_{t-1}z_t} f_{z_t}(y_t)$. For $K = 2, T = 3$ this is $2^3 = 8$ terms. The forward

algorithm computes the same quantity in $O(TK^2) = O(3 \cdot 4) = 12$ operations. Both must give $\sum_k \alpha_T(k)$. $\square$

Solution to Exercise 2. The forward recursion (18) at time t requires: for each state k (K choices), sum over $j = 1, \dots, K$ predecessors, each requiring one multiplication and addition: $O(K)$ per state. Total per time step: $O(K \cdot K) = O(K^2)$. Over T steps: $O(TK^2)$. The brute-force method is $O(K^T T)$, exponentially worse. $\square$

Solution to Exercise 4 (Scaling). Define $c_t = \sum_k \alpha_t(k)$ and $\hat{\alpha}_t(k) = \alpha_t(k)/c_t$. The scaled forward recursion becomes: $\hat{\alpha}_t(k) = \frac{f_k(Y_t) \sum_j \hat{\alpha}_{t-1}(j) a_{jk}}{c_t}$ where c_t is chosen so $\sum_k \hat{\alpha}_t(k) = 1$. Since $\mathbb{P}(\mathbf{Y}) = \sum_k \alpha_T(k) = \prod_{t=1}^T c_t$, we get $\log \mathbb{P}(\mathbf{Y}) = \sum_{t=1}^T \log c_t$, which avoids underflow. $\square$

4 The Baum–Welch Algorithm (EM for HMMS)

4.1 The Expectation–Maximisation Framework

Definition 4.1 (EM Algorithm). The *Expectation–Maximisation* (EM) algorithm maximises the log-likelihood $\ell(\Theta) = \log \mathbb{P}(\mathbf{Y} \mid \Theta)$ iteratively:

1. **E-step:** Compute $Q(\Theta \mid \Theta^{\text{old}}) = \mathbb{E}_{\mathbf{Z} \mid \mathbf{Y}, \Theta^{\text{old}}} [\log \mathbb{P}(\mathbf{Y}, \mathbf{Z} \mid \Theta)]$.
2. **M-step:** Update $\Theta^{\text{new}} = \operatorname{argmax}_{\Theta} Q(\Theta \mid \Theta^{\text{old}})$.

Theorem 4.1 (EM Monotonicity). *Each EM iteration increases the log-likelihood: $\ell(\Theta^{\text{new}}) \geq \ell(\Theta^{\text{old}})$, with equality if and only if $\Theta^{\text{new}} = \Theta^{\text{old}}$.*

Sketch. Using Jensen’s inequality on the KL divergence between the posterior $\mathbb{P}(\mathbf{Z} \mid \mathbf{Y}, \Theta^{\text{old}})$ and $\mathbb{P}(\mathbf{Z} \mid \mathbf{Y}, \Theta)$, we obtain: $\ell(\Theta) \geq Q(\Theta \mid \Theta^{\text{old}}) + H(q^{\text{old}})$, where H is the entropy. Since M-step maximises Q , the bound gives the result. $\square$

4.2 Baum–Welch Update Equations

For Gaussian emission HMMS, define the transition posterior:

$$\xi_t(i, j) = \mathbb{P}(Z_t = i, Z_{t+1} = j \mid \mathbf{Y}, \Theta) = \frac{\alpha_t(i) a_{ij} f_j(Y_{t+1}) \beta_{t+1}(j)}{\sum_{i', j'} \alpha_t(i') a_{i'j'} f_{j'}(Y_{t+1}) \beta_{t+1}(j')}. \quad (27)$$

The M-step updates are:

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)}, \quad (28)$$

$$\hat{\mu}_k = \frac{\sum_{t=1}^T \gamma_t(k) Y_t}{\sum_{t=1}^T \gamma_t(k)}, \quad (29)$$

$$\hat{\Sigma}_k = \frac{\sum_{t=1}^T \gamma_t(k) (Y_t - \hat{\mu}_k)(Y_t - \hat{\mu}_k)^\top}{\sum_{t=1}^T \gamma_t(k)}, \quad (30)$$

$$\hat{\pi}_k = \gamma_1(k). \quad (31)$$

Key Insight

The Baum–Welch algorithm is guaranteed to converge to a *local* maximum of the likelihood. In practice, multiple random initialisations are used, and the best solution (by log-likelihood) is selected. For financial data, the learned states often correspond to interpretable market regimes.

4.3 Model Selection: BIC

Definition 4.2 (Bayesian Information Criterion).

$$\text{BIC} = -2\ell(\hat{\Theta}) + p \log T, \quad (32)$$

where p is the number of free parameters and T the number of observations. Lower BIC indicates a better model.

For a K -state Gaussian HMM with d -dimensional observations:

$$p = K^2 - 1 + K(d + \frac{d(d+1)}{2}). \quad (33)$$

Practical Application

WP3 Result: BIC analysis on 10 years of daily equity returns selects $K = 3$ states (bull, consolidation, crisis). $K = 2$ underfits (misses the consolidation regime), while $K \geq 4$ overfits and reduces out-of-sample Sharpe by ~ 0.1 .

4.4 Exercises

1. Derive the M-step update (29) by taking the derivative of Q with respect to $\boldsymbol{\mu}_k$ and setting to zero.
2. Implement the full Baum–Welch algorithm for a 3-state, 1D Gaussian HMM. Verify on simulated data that the estimated parameters converge to the true values.
3. Compute the BIC for $K = 2, 3, 4, 5$ on a simulated dataset from a 3-state model and verify that BIC correctly selects $K = 3$.
4. **EM monotonicity.** Show why EM cannot decrease the likelihood by expanding the log-likelihood as $\ell(\Theta) = Q(\Theta|\Theta^{\text{old}}) + H(\Theta|\Theta^{\text{old}})$ using the KL divergence.
5. **Initialisation strategy.** Explain why K -means clustering on the observations is a sensible initialisation for Baum–Welch. What are the alternatives?

Solutions

Solution to Exercise 1. We need $\frac{\partial Q}{\partial \boldsymbol{\mu}_k} = 0$ where $Q = \sum_t \sum_k \gamma_t(k) [-\frac{1}{2}(Y_t - \boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(Y_t - \boldsymbol{\mu}_k) + \text{const}]$. Taking the derivative w.r.t. $\boldsymbol{\mu}_k$: $\sum_t \gamma_t(k) \Sigma_k^{-1}(Y_t - \boldsymbol{\mu}_k) = 0 \Rightarrow \sum_t \gamma_t(k) Y_t = \boldsymbol{\mu}_k \sum_t \gamma_t(k)$, giving $\hat{\boldsymbol{\mu}}_k = \frac{\sum_t \gamma_t(k) Y_t}{\sum_t \gamma_t(k)}$. This is a *weighted average* of observations, weighted by the posterior probability of being in state k . It generalises the standard MLE $\bar{x}$ to the latent variable setting. $\square$

Solution to Exercise 3. BIC for each K : fit the model, compute $-2\ell(\hat{\Theta}) + p \log T$. For a 1D model with K states: $p = K^2 - 1 + 2K$ (transition entries + means + variances). With $T = 1000$ observations simulated from a 3-state model: $K = 2$: BIC $\approx$ large (underfits, ℓ is poor); $K = 3$: BIC $\approx$ smallest (correct model); $K = 4, 5$: BIC increases (penalty term $p \log T$ grows faster than ℓ improves). BIC correctly identifies $K = 3$ because additional states capture noise rather than structure. $\square$

Solution to Exercise 4 (EM monotonicity). Write $\ell(\Theta) = \log \mathbb{P}(\mathbf{Y}|\Theta)$. Using any distribution $q(\mathbf{Z})$:

$$\ell(\Theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{\mathbb{P}(\mathbf{Y}, \mathbf{Z}|\Theta)}{q(\mathbf{Z})} - \sum_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{\mathbb{P}(\mathbf{Z}|\mathbf{Y}, \Theta)}{q(\mathbf{Z})} \quad (34)$$

$$= \underbrace{Q(\Theta|\Theta^{\text{old}})}_{\text{ELBO}} + \underbrace{H(q) + D_{\text{KL}}(q||p(\mathbf{Z}|\mathbf{Y}, \Theta))}_{\geq 0}. \quad (35)$$

The E-step sets $q = \mathbb{P}(\mathbf{Z}|\mathbf{Y}, \Theta^{\text{old}})$, making $D_{\text{KL}} = 0$. The M-step maximises Q over Θ , which can only increase the ELBO. Since $D_{\text{KL}} \geq 0$, $\ell(\Theta^{\text{new}}) \geq \text{ELBO}^{\text{new}} \geq \text{ELBO}^{\text{old}} = \ell(\Theta^{\text{old}})$. $\square$

5 Principal Component Analysis

5.1 Classical PCA

Definition 5.1 (PCA). Given n observations $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, PCA finds the directions of maximum variance:

$$\mathbf{w}_1 = \operatorname{argmax}_{\|\mathbf{w}\|=1} \operatorname{Var}(\mathbf{w}^\top \mathbf{X}) = \operatorname{argmax}_{\|\mathbf{w}\|=1} \mathbf{w}^\top \hat{\Sigma} \mathbf{w}, \quad (36)$$

where $\hat{\Sigma}$ is the sample covariance matrix.

Theorem 5.1 (PCA Solution). *The principal components are the eigenvectors of $\hat{\Sigma}$, ordered by decreasing eigenvalue:*

$$\hat{\Sigma} \mathbf{w}_k = \lambda_k \mathbf{w}_k, \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0. \quad (37)$$

The proportion of variance explained by the first m components is:

$$\frac{\sum_{k=1}^m \lambda_k}{\sum_{k=1}^d \lambda_k}. \quad (38)$$

5.2 Singular Value Decomposition (SVD)

Theorem 5.2 (SVD). *Any matrix $X \in \mathbb{R}^{n \times d}$ admits a decomposition:*

$$X = U \Sigma V^\top, \quad (39)$$

where $U \in \mathbb{R}^{n \times r}$ has orthonormal columns, $\Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$ with $\sigma_1 \geq \dots \geq \sigma_r > 0$, and $V \in \mathbb{R}^{d \times r}$ has orthonormal columns. The columns of V are the principal components, and $\lambda_k = \sigma_k^2 / (n - 1)$.

5.3 Sparse PCA

Classical PCA loadings are dense: every asset has a nonzero weight in every component. For interpretability and robustness, we seek *sparse* loadings.

Definition 5.2 (Sparse PCA). The *sparse PCA* problem adds an ℓ_1 penalty:

$$\mathbf{w}_k = \operatorname{argmax}_{\|\mathbf{w}\|_2=1} (\mathbf{w}^\top \hat{\Sigma} \mathbf{w} - \lambda \|\mathbf{w}\|_1). \quad (40)$$

Proposition 5.1 (Thresholding Approximation). *A simple approximation to sparse PCA: compute classical PCA, then apply hard thresholding:*

$$\tilde{w}_{kj} = w_{kj} \cdot \mathbb{I}(|w_{kj}| > \tau), \quad (41)$$

followed by renormalisation. This is equivalent to sparse PCA in the orthogonal design case.

Practical Application

WP3 Application (Box–Tao Decomposition): Sparse PCA decomposes equity returns into interpretable factors:

- PC1 (market factor): loads on all assets $\Rightarrow$ systematic risk;
- PC2 (sector rotation): sparse loadings separating sectors;

- PC3 (dispersion): loads on high-vol assets only.

The sparse loadings are *regime-dependent*: the decomposition is recomputed in each HMM regime.

5.4 Exercises

1. Show that maximising $\mathbf{w}^\top \hat{\Sigma} \mathbf{w}$ subject to $\|\mathbf{w}\| = 1$ gives the eigenvector of $\hat{\Sigma}$ with largest eigenvalue.
2. Implement PCA via SVD on a portfolio of 10 equity returns and compute the proportion of variance explained by each component.
3. Implement the thresholding approximation to sparse PCA and compare the resulting loadings with classical PCA.
4. **Scree plot.** Explain how to use a scree plot (eigenvalues vs. component index) to choose the number of components m . What does a “knee” in the plot indicate?

Solutions

Solution to Exercise 1. Use Lagrange multipliers. Maximise $L(\mathbf{w}, \lambda) = \mathbf{w}^\top \hat{\Sigma} \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1)$. Setting $\nabla_{\mathbf{w}} L = 2\hat{\Sigma} \mathbf{w} - 2\lambda \mathbf{w} = 0$ gives $\hat{\Sigma} \mathbf{w} = \lambda \mathbf{w}$, i.e., $\mathbf{w}$ must be an eigenvector. The objective value is $\mathbf{w}^\top \hat{\Sigma} \mathbf{w} = \lambda \mathbf{w}^\top \mathbf{w} = \lambda$. To maximise, choose the eigenvector with the *largest* eigenvalue λ_1 . The second PC maximises variance subject to orthogonality to $\mathbf{w}_1$, leading to the second eigenvector, and so on. $\square$

Solution to Exercise 2. Given the centred data matrix $X \in \mathbb{R}^{n \times d}$ with $\hat{\Sigma} = \frac{1}{n-1} X^\top X$, compute the SVD $X = U \Sigma_{\text{svd}} V^\top$. The columns of V are the principal component directions, and $\lambda_k = \sigma_k^2 / (n-1)$. PVE for first m components: $\sum_{k=1}^m \sigma_k^2 / \sum_{k=1}^d \sigma_k^2$. Typically, for equity returns, PC1 (market factor) explains 40–60% of variance. $\square$

Solution to Exercise 4 (Scree plot). A scree plot displays eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d$ on the y -axis versus index k on the x -axis. The “knee” (elbow) is where eigenvalues transition from large to small. Components before the knee capture signal; those after capture noise. Choose m at the knee. For equity data with 50 stocks, the knee typically appears at $m = 3$ –5 (market, sector, momentum factors). $\square$

6 Fundamentals of Optimisation

This section introduces the basic concepts of optimisation theory for readers who have not taken a formal course. These foundations are essential for understanding portfolio construction (Section 7) and CVaR optimisation (Section 8).

6.1 Unconstrained Optimisation

Definition 6.1 (Local and Global Minima). A point $\mathbf{x}^*$ is a *local minimum* of $f : \mathbb{R}^d \rightarrow \mathbb{R}$ if there exists $\epsilon > 0$ such that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all $\|\mathbf{x} - \mathbf{x}^*\| < \epsilon$. It is a *global minimum* if the inequality holds for all $\mathbf{x} \in \mathbb{R}^d$.

Theorem 6.1 (First-Order Necessary Condition). *If $\mathbf{x}^*$ is a local minimum of a differentiable function f , then:*

$$\nabla f(\mathbf{x}^*) = \mathbf{0}. \quad (42)$$

Points satisfying this are called stationary points or critical points.

Theorem 6.2 (Second-Order Sufficient Condition). *If $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and the Hessian $H_f(\mathbf{x}^*) = (\frac{\partial^2 f}{\partial x_i \partial x_j})$ is positive definite (all eigenvalues > 0), then $\mathbf{x}^*$ is a strict local minimum.*

Example 6.1 (Minimising a Quadratic — Worked). Let $f(x, y) = x^2 + 2y^2 - 2xy + 4x - 6y$.

$$\nabla f = (2x - 2y + 4, 4y - 2x - 6) = (0, 0) \quad (43)$$

$$\Rightarrow 2x - 2y = -4, \quad -2x + 4y = 6. \quad (44)$$

Adding: $2y = 2 \Rightarrow y = 1$, then $x = -1$. Hessian: $H = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}$. Eigenvalues: $\lambda = 3 \pm \sqrt{5}$, both positive $\Rightarrow (x^*, y^*) = (-1, 1)$ is a strict local (and global) minimum.

6.2 Gradient Descent

Algorithm Gradient Descent

Starting from $\mathbf{x}_0$, iterate:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \eta_k \nabla f(\mathbf{x}_k), \quad (45)$$

where $\eta_k > 0$ is the *step size* (or *learning rate*).

Proposition 6.1 (Convergence for Convex Functions). *If f is convex with L -Lipschitz gradient and $\eta_k = 1/L$, then:*

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L \|\mathbf{x}_0 - \mathbf{x}^*\|^2}{2k}. \quad (46)$$

This is $O(1/k)$ convergence — 100 steps reduce the error by 100×.

6.3 Constrained Optimisation: Lagrange Multipliers

Theorem 6.3 (Method of Lagrange Multipliers). *To minimise $f(\mathbf{x})$ subject to $h(\mathbf{x}) = 0$, define the Lagrangian:*

$$\mathcal{L}(\mathbf{x}, \nu) = f(\mathbf{x}) + \nu h(\mathbf{x}). \quad (47)$$

At the optimum: $\nabla_{\mathbf{x}} \mathcal{L} = 0$ and $h(\mathbf{x}) = 0$.

Example 6.2 (Minimum Variance Portfolio — Worked). Minimise $f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\top \Sigma \mathbf{w}$ subject to $\mathbf{1}^\top \mathbf{w} = 1$.

Lagrangian: $\mathcal{L}(\mathbf{w}, \nu) = \frac{1}{2}\mathbf{w}^\top \Sigma \mathbf{w} + \nu(\mathbf{1}^\top \mathbf{w} - 1)$.

$\nabla_{\mathbf{w}} \mathcal{L} = \Sigma \mathbf{w} + \nu \mathbf{1} = 0 \Rightarrow \mathbf{w} = -\nu \Sigma^{-1} \mathbf{1}$.

From the constraint: $\mathbf{1}^\top (-\nu \Sigma^{-1} \mathbf{1}) = 1 \Rightarrow \nu = -\frac{1}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}$.

Final solution:

$$\mathbf{w}^* = \frac{\Sigma^{-1} \mathbf{1}}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}. \quad (48)$$

This is the *global minimum-variance portfolio* — a fundamental result in portfolio theory.

6.4 Convex Functions and Sets

Definition 6.2 (Convex Set). A set $\mathcal{C} \subseteq \mathbb{R}^d$ is *convex* if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $\lambda \in [0, 1]$: $\lambda \mathbf{x} + (1 - \lambda) \mathbf{y} \in \mathcal{C}$.

Definition 6.3 (Convex Function). A function $f : \mathcal{C} \rightarrow \mathbb{R}$ is *convex* if for all $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ and $\lambda \in [0, 1]$:

$$f(\lambda \mathbf{x} + (1 - \lambda) \mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda) f(\mathbf{y}). \quad (49)$$

Key Insight

The fundamental property of convex optimisation: *every local minimum is a global minimum*. This is why portfolio optimisation problems (Markowitz QP, CVaR LP) can be solved efficiently and reliably — they are convex programmes. Non-convex problems (like neural network training) may have many local minima, making them much harder.

Proposition 6.2. *The following functions are convex:*

- *Linear:* $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$;
- *Quadratic (positive semi-definite):* $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$ with $A \succeq 0$;
- *Norms:* $f(\mathbf{x}) = \|\mathbf{x}\|_p$ for $p \geq 1$;
- *Maximum:* $f(\mathbf{x}) = \max(g(\mathbf{x}), h(\mathbf{x}))$ if g, h are convex;
- *CVaR:* $\text{CVaR}_\alpha(L)$ is a convex function of the portfolio weights.

6.5 Exercises

1. **Unconstrained minimum.** Find the minimum of $f(x, y) = x^2 + xy + y^2 - 3x - 3y$ using first- and second-order conditions.
2. **Lagrange multipliers.** Minimise $f(x, y) = x + y$ subject to $x^2 + y^2 = 1$. Give the geometric interpretation.
3. **Convexity check.** Determine whether $f(x) = e^x$, $g(x) = \log x$, and $h(x) = x^3$ are convex, concave, or neither on $\mathbb{R}_{>0}$.

Solutions

Solution to Exercise 1. $\nabla f = (2x + y - 3, x + 2y - 3) = (0, 0)$. From $2x + y = 3$ and $x + 2y = 3$: subtracting gives $x - y = 0$, so $x = y = 1$. $H = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$. Eigenvalues: $3, 1 > 0 \Rightarrow$ positive definite, so $(1, 1)$ is a strict minimum with $f(1, 1) = 1 + 1 + 1 - 3 - 3 = -3$. $\square$

Solution to Exercise 2. Lagrangian: $\mathcal{L} = x + y + \lambda(x^2 + y^2 - 1)$. $\partial\mathcal{L}/\partial x = 1 + 2\lambda x = 0 \Rightarrow x = -1/(2\lambda)$. $\partial\mathcal{L}/\partial y = 1 + 2\lambda y = 0 \Rightarrow y = -1/(2\lambda)$. $x = y$ and $x^2 + y^2 = 2x^2 = 1 \Rightarrow x = \pm 1/\sqrt{2}$. Minimum: $(x, y) = (-1/\sqrt{2}, -1/\sqrt{2})$ with $f = -\sqrt{2}$. Geometrically: we find the point on the unit circle closest to the direction $(-1, -1)$. $\square$

Solution to Exercise 3. $f(x) = e^x$: $f''(x) = e^x > 0$ for all $x \Rightarrow$ **convex**. $g(x) = \log x$: $g''(x) = -1/x^2 < 0$ for $x > 0 \Rightarrow$ **concave**. $h(x) = x^3$: $h''(x) = 6x$, which changes sign at $x = 0 \Rightarrow$ **neither** on $\mathbb{R}_{>0}$ (actually convex on $(0, \infty)$ since $h'' > 0$ there; but neither on all of $\mathbb{R}$). $\square$

7 Convex Optimisation for Portfolios

7.1 Portfolio Optimisation: Setup

Definition 7.1 (Portfolio Problem). Given d assets with expected return $\boldsymbol{\mu} \in \mathbb{R}^d$ and covariance $\Sigma \in \mathbb{R}^{d \times d}$, find weights $\mathbf{w} \in \mathbb{R}^d$ that optimise a risk-return objective subject to constraints:

$$\min_{\mathbf{w}} R(\mathbf{w}) \quad \text{s.t.} \quad \mathbf{1}^\top \mathbf{w} = 1, \quad \mathbf{w} \geq 0. \quad (50)$$

7.2 Mean–Variance (Markowitz)

Definition 7.2 (Markowitz, 1952). The mean–variance problem minimises variance for a target return μ^* :

$$\min_{\mathbf{w}} \frac{1}{2} \mathbf{w}^\top \Sigma \mathbf{w} \quad \text{s.t.} \quad \mathbf{w}^\top \boldsymbol{\mu} \geq \mu^*, \quad \mathbf{1}^\top \mathbf{w} = 1, \quad \mathbf{w} \geq 0. \quad (51)$$

This is a quadratic program (QP) solvable in polynomial time.

Remark 7.1. The Markowitz framework is sensitive to estimation errors in $\boldsymbol{\mu}$ and Σ . Robust alternatives include:

- Shrinkage estimators for Σ (Ledoit–Wolf);
- Risk parity (equal risk contribution);
- CVaR optimisation (Section 8).

7.3 Risk Parity

Definition 7.3 (Risk Parity). A portfolio has *equal risk contribution* if each asset contributes equally to total portfolio risk:

$$w_i \cdot \frac{(\Sigma \mathbf{w})_i}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}} = \frac{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}}{d}, \quad \forall i. \quad (52)$$

Proposition 7.1 (Inverse-Volatility Approximation). For diagonal Σ (uncorrelated assets), risk parity gives:

$$w_i = \frac{1/\sigma_i}{\sum_j 1/\sigma_j}. \quad (53)$$

This is the “inverse volatility” weighting, commonly used as a simple risk parity proxy.

7.4 The KKT Conditions

Theorem 7.1 (Karush–Kuhn–Tucker). For the convex programme $\min f(\mathbf{x})$ s.t. $g_i(\mathbf{x}) \leq 0$, $h_j(\mathbf{x}) = 0$, the optimal $\mathbf{x}^*$ satisfies:

$$\nabla f(\mathbf{x}^*) + \sum_i \mu_i \nabla g_i(\mathbf{x}^*) + \sum_j \nu_j \nabla h_j(\mathbf{x}^*) = 0, \quad (54)$$

$$\mu_i g_i(\mathbf{x}^*) = 0, \quad \forall i, \quad (55)$$

$$\mu_i \geq 0, \quad \forall i. \quad (56)$$

7.5 Exercises

1. Derive the unconstrained Markowitz solution (no short-sale constraint) using Lagrange multipliers. Show it is linear in μ^* .
2. For $d = 3$ assets with given μ and Σ , solve the Markowitz QP numerically using `scipy.optimize.minimize`.
3. Implement risk parity via the inverse-volatility approximation and compare the Sharpe ratio with Markowitz for a portfolio of 5 equity ETFs.
4. **KKT conditions.** For the Markowitz problem with short-sale constraints ($w_i \geq 0$), write out the full KKT conditions. Show that KKT implies some assets will have zero weight (“corner solution”).

Solutions

Solution to Exercise 1 (Unconstrained Markowitz). Lagrangian with return constraint and budget constraint: $\mathcal{L} = \frac{1}{2}\mathbf{w}^\top \Sigma \mathbf{w} - \lambda(\mathbf{w}^\top \mu - \mu^*) - \gamma(\mathbf{1}^\top \mathbf{w} - 1)$. FOC: $\Sigma \mathbf{w} = \lambda \mu + \gamma \mathbf{1}$, so $\mathbf{w} = \lambda \Sigma^{-1} \mu + \gamma \Sigma^{-1} \mathbf{1}$. Denote $A = \mathbf{1}^\top \Sigma^{-1} \mu$, $B = \mu^\top \Sigma^{-1} \mu$, $C = \mathbf{1}^\top \Sigma^{-1} \mathbf{1}$. From constraints: $\lambda B + \gamma A = \mu^*$ and $\lambda A + \gamma C = 1$. Solving: $\lambda = \frac{C\mu^* - A}{BC - A^2}$, $\gamma = \frac{B - A\mu^*}{BC - A^2}$. The solution is linear in μ^* : $\mathbf{w}(\mu^*) = \mathbf{w}_0 + \mu^* \cdot \mathbf{w}_1$ where $\mathbf{w}_0, \mathbf{w}_1$ depend only on Σ and μ . This traces out the *efficient frontier* as μ^* varies. $\square$

Solution to Exercise 3. Risk parity weights: $w_i \propto 1/\sigma_i$ where $\sigma_i = \sqrt{\Sigma_{ii}}$. For $d = 5$ assets with annual volatilities (15%, 20%, 25%, 10%, 30%): $w \propto (1/0.15, 1/0.20, 1/0.25, 1/0.10, 1/0.30) = (6.67, 5.0, 4.0, 10.0, 3.33)$. Normalising: $w = (0.230, 0.172, 0.138, 0.345, 0.115)$. Risk parity overweights low-volatility assets. Compared to Markowitz, it is more robust to estimation errors in expected returns μ because it depends only on volatilities, not on uncertain mean estimates. $\square$

8 Conditional Value-at-Risk

8.1 Value-at-Risk (VaR)

Definition 8.1 (VaR). The *Value-at-Risk* at confidence level $1 - \alpha$ is the α -quantile of the loss distribution:

$$\text{VaR}_\alpha(L) = \inf\{x : \mathbb{P}(L \leq x) \geq \alpha\}. \quad (57)$$

VaR answers: “What is the maximum loss that occurs with probability α ?” For $\alpha = 5\%$: “We have 95% confidence that losses will not exceed VaR.”

Remark 8.1. VaR has serious shortcomings:

- Not *subadditive*: $\text{VaR}(A + B)$ can exceed $\text{VaR}(A) + \text{VaR}(B)$;
- Ignores the severity of tail losses beyond VaR;
- Not a *coherent risk measure* in the sense of Artzner et al. (1999).

8.2 CVaR: Definition and Properties

Definition 8.2 (Conditional Value-at-Risk). The *Conditional Value-at-Risk* (CVaR), also called *Expected Shortfall*, is:

$$\text{CVaR}_\alpha(L) = \mathbb{E}[L \mid L \geq \text{VaR}_\alpha(L)]. \quad (58)$$

Equivalently, for continuous distributions:

$$\text{CVaR}_\alpha(L) = \frac{1}{\alpha} \int_0^\alpha \text{VaR}_u(L) \, du. \quad (59)$$

Theorem 8.1 (Coherence). *CVaR is a coherent risk measure: it satisfies monotonicity, translation invariance, positive homogeneity, and subadditivity:*

$$\text{CVaR}(L_1 + L_2) \leq \text{CVaR}(L_1) + \text{CVaR}(L_2). \quad (60)$$

Key Insight

CVaR penalises diversification less than VaR (due to subadditivity) and focuses on the *expected* tail loss, not just the threshold. It is the risk measure of choice for portfolio optimisation in WP3.

8.3 Rockafellar–Uryasev LP Reformulation

Theorem 8.2 (Rockafellar & Uryasev, 2000). *CVaR can be computed by solving:*

$$\text{CVaR}_\alpha = \min_{\zeta \in \mathbb{R}} \left\{ \zeta + \frac{1}{\alpha} \mathbb{E}[(L - \zeta)^+] \right\}. \quad (61)$$

The minimiser $\zeta^* = \text{VaR}_\alpha$.

Proof. Define $F_\alpha(\zeta) = \zeta + \frac{1}{\alpha} \mathbb{E}[(L - \zeta)^+]$.

$$\begin{aligned} F_\alpha(\zeta) &= \zeta + \frac{1}{\alpha} \int_\zeta^\infty (x - \zeta) f_L(x) \, dx \\ &= \zeta + \frac{1}{\alpha} \left[\int_\zeta^\infty x f_L(x) \, dx - \zeta(1 - F_L(\zeta)) \right]. \end{aligned} \quad (62)$$

Taking the derivative: $F'_\alpha(\zeta) = 1 - \frac{1}{\alpha}(1 - F_L(\zeta)) = 0$ gives $F_L(\zeta^*) = 1 - \alpha$, so $\zeta^* = \text{VaR}_\alpha$. Substituting back gives $F_\alpha(\zeta^*) = \text{CVaR}_\alpha$. $\square$

8.4 Portfolio CVaR Minimisation as LP

For n return scenarios $\mathbf{r}_1, \dots, \mathbf{r}_n$, the CVaR-minimising portfolio is found by:

$$\min_{\mathbf{w}, \zeta, \mathbf{u}} \quad \zeta + \frac{1}{n\alpha} \sum_{s=1}^n u_s \quad (63)$$

$$\text{s.t.} \quad u_s \geq -\mathbf{w}^\top \mathbf{r}_s - \zeta, \quad s = 1, \dots, n, \quad (64)$$

$$u_s \geq 0, \quad s = 1, \dots, n, \quad (65)$$

$$\mathbf{1}^\top \mathbf{w} = 1, \quad \mathbf{w} \geq 0. \quad (66)$$

This is a *linear programme* with $d + 1 + n$ variables and $2n + 1$ constraints, solvable efficiently even for large n .

Practical Application

WP3 Implementation: In each HMM regime k , we solve a separate CVaR LP using the regime-conditional return scenarios: $\{\mathbf{r}_s : Z_s = k\}$. The portfolio is then rebalanced when the HMM posterior probability of the current regime changes.

8.5 Exercises

1. For a Gaussian loss $L \sim \mathcal{N}(\mu, \sigma^2)$, derive closed-form VaR and CVaR: $\text{VaR}_\alpha = \mu + \sigma \Phi^{-1}(1 - \alpha)$, $\text{CVaR}_\alpha = \mu + \sigma \frac{\phi(\Phi^{-1}(1 - \alpha))}{\alpha}$.
2. Implement the LP formulation and verify that ζ^* equals the empirical α -quantile of losses.
3. Compare CVaR-optimised and Markowitz portfolios on historical equity data. Which has better tail behaviour?
4. **Non-coherence of VaR.** Construct a two-asset example where $\text{VaR}_\alpha(L_1 + L_2) > \text{VaR}_\alpha(L_1) + \text{VaR}_\alpha(L_2)$, demonstrating that VaR is not subadditive.

Solutions

Solution to Exercise 1 (Gaussian VaR and CVaR). For $L \sim \mathcal{N}(\mu, \sigma^2)$, $\text{VaR}_\alpha = \inf\{x : \Phi(\frac{x - \mu}{\sigma}) \geq 1 - \alpha\} = \mu + \sigma \Phi^{-1}(1 - \alpha)$.

For CVaR: $\text{CVaR}_\alpha = \mathbb{E}[L \mid L \geq \text{VaR}_\alpha] = \mu + \sigma \mathbb{E}[Z \mid Z \geq \Phi^{-1}(1 - \alpha)]$ where $Z \sim \mathcal{N}(0, 1)$. Using the conditional expectation of a truncated normal: $\mathbb{E}[Z \mid Z \geq z_\alpha] = \frac{\phi(z_\alpha)}{1 - \Phi(z_\alpha)} = \frac{\phi(z_\alpha)}{\alpha}$ where $z_\alpha = \Phi^{-1}(1 - \alpha)$. Therefore: $\text{CVaR}_\alpha = \mu + \sigma \frac{\phi(\Phi^{-1}(1 - \alpha))}{\alpha}$.

Numerical example: For $\alpha = 5\%$, $\mu = 0$, $\sigma = 1$: $\Phi^{-1}(0.95) = 1.645$, so $\text{VaR}_{5\%} = 1.645$ and $\text{CVaR}_{5\%} = \phi(1.645)/0.05 = 0.1031/0.05 = 2.063$. CVaR exceeds VaR by $\approx 25\%$ for Gaussian losses. $\square$

Solution to Exercise 2 (LP verification). Given n scenarios with losses $\ell_1, \dots, \ell_n$, the LP minimises $\zeta + \frac{1}{n\alpha} \sum_s u_s$ with $u_s \geq \ell_s - \zeta$ and $u_s \geq 0$. At the optimum, $u_s = (\ell_s - \zeta)^+ = \max(\ell_s - \zeta, 0)$. The objective becomes $F(\zeta) = \zeta + \frac{1}{n\alpha} \sum_s (\ell_s - \zeta)^+$. Setting $F'(\zeta) = 0$: $1 - \frac{1}{n\alpha} \#\{s : \ell_s > \zeta\} = 0$, so $\#\{s : \ell_s > \zeta^*\} = n\alpha$, meaning ζ^* is the α -quantile of the empirical loss distribution, i.e., $\zeta^* = \text{VaR}_\alpha$. $\square$

Solution to Exercise 4 (Non-coherence of VaR). Consider two binary losses: $L_1 = L_2 = 0$ with probability 96% and $L_1 = L_2 = 100$ with probability 4%, independently. At $\alpha = 5\%$:

$\text{VaR}_{5\%}(L_1) = \text{VaR}_{5\%}(L_2) = 0$ (since $\mathbb{P}(L_i = 0) = 0.96 > 0.95$). For $L_1 + L_2$: $\mathbb{P}(L_1 + L_2 = 0) = 0.96^2 = 0.9216$, $\mathbb{P}(L_1 + L_2 = 100) = 2 \times 0.96 \times 0.04 = 0.0768$, $\mathbb{P}(L_1 + L_2 = 200) = 0.04^2 = 0.0016$. $\mathbb{P}(L_1 + L_2 \leq 0) = 0.9216 < 0.95$, so $\text{VaR}_{5\%}(L_1 + L_2) = 100 > 0 + 0$. Diversification *increases* VaR! This pathology does not occur with CVaR. $\square$

9 Introduction to Stochastic Differential Equations

This section provides an accessible introduction to SDEs for readers who have not taken a course in stochastic calculus. Understanding SDEs is essential for the Ornstein–Uhlenbeck process (Section 10) and the dynamic models used throughout WP3.

9.1 From Random Walks to Brownian Motion

Definition 9.1 (Symmetric Random Walk). Let $\xi_1, \xi_2, \dots$ be i.i.d. with $\mathbb{P}(\xi_i = +1) = \mathbb{P}(\xi_i = -1) = 1/2$. The *random walk* is $S_n = \sum_{i=1}^n \xi_i$.

The key insight is the *scaling limit*: as we take more frequent, smaller steps, the random walk converges to a continuous process.

Theorem 9.1 (Donsker’s Invariance Principle (Informal)). *If we rescale the random walk: $W_t^{(n)} = \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor}$, then as $n \rightarrow \infty$, $W^{(n)} \Rightarrow W$, where W is a Brownian motion.*

Definition 9.2 (Brownian Motion). A stochastic process $(W_t)_{t \geq 0}$ is a (standard) *Brownian motion* if:

- (i) $W_0 = 0$;
- (ii) $W_t - W_s \sim \mathcal{N}(0, t - s)$ for $0 \leq s < t$ (Gaussian increments);
- (iii) Increments on non-overlapping intervals are independent;
- (iv) $t \mapsto W_t$ is continuous almost surely.

Key Insight

Brownian motion is the fundamental “noise term” in continuous-time finance. Every SDE is driven by Brownian motion, and the roughness of Brownian paths (they are continuous but *nowhere differentiable*) is what makes stochastic calculus different from ordinary calculus.

9.2 What Is a Stochastic Differential Equation?

An SDE is an equation of the form:

$$dX_t = \underbrace{b(X_t, t)}_{\text{drift}} dt + \underbrace{\sigma(X_t, t)}_{\text{diffusion}} dW_t. \quad (67)$$

The *drift* b describes the deterministic trend, and the *diffusion* σ describes the random fluctuations. The rigorous meaning is the *integral equation*:

$$X_t = X_0 + \int_0^t b(X_s, s) ds + \int_0^t \sigma(X_s, s) dW_s, \quad (68)$$

where the second integral is an *Itô integral*.

Remark 9.1 (Why “ dW_t ” is only notation). Since Brownian motion is nowhere differentiable, dW_t/dt does not exist. The notation $dX_t = b dt + \sigma dW_t$ is shorthand for the integral equation above. This distinction is important: it means standard calculus rules need modification.

9.3 Itô's Formula: The Chain Rule for SDEs

The most important tool in stochastic calculus is Itô's formula — the stochastic version of the chain rule.

Theorem 9.2 (Itô's Formula). *If X_t satisfies $dX_t = b dt + \sigma dW_t$ and $f(t, x)$ is $C^{1,2}$, then:*

$$df(t, X_t) = \left(\frac{\partial f}{\partial t} + b \frac{\partial f}{\partial x} + \frac{1}{2} \sigma^2 \frac{\partial^2 f}{\partial x^2} \right) dt + \sigma \frac{\partial f}{\partial x} dW_t. \quad (69)$$

The extra term $\frac{1}{2} \sigma^2 f_{xx}$ (absent in ordinary calculus) is called the Itô correction and arises because $(dW_t)^2 = dt$.

Example 9.1 (Geometric Brownian Motion — Worked). The stock price model $dS_t = \mu S_t dt + \sigma S_t dW_t$ is solved by:

Apply Itô's formula to $f(S) = \log S$:

$$d(\log S_t) = \frac{1}{S_t} dS_t - \frac{1}{2} \frac{1}{S_t^2} (dS_t)^2 = \left(\mu - \frac{\sigma^2}{2} \right) dt + \sigma dW_t. \quad (70)$$

Integrating: $\log S_t = \log S_0 + (\mu - \sigma^2/2)t + \sigma W_t$, so:

$$S_t = S_0 \exp \left\{ \left(\mu - \frac{\sigma^2}{2} \right) t + \sigma W_t \right\}. \quad (71)$$

Note the $-\sigma^2/2$ correction: without it, $\mathbb{E}[S_t]$ would not equal $S_0 e^{\mu t}$.

9.4 The Euler–Maruyama Method

Most SDEs cannot be solved analytically. The *Euler–Maruyama* method is the simplest numerical scheme:

Algorithm Euler–Maruyama

Given SDE $dX_t = b(X_t) dt + \sigma(X_t) dW_t$ and time step Δt :

$$X_{n+1} = X_n + b(X_n) \Delta t + \sigma(X_n) \sqrt{\Delta t} Z_n, \quad Z_n \sim \mathcal{N}(0, 1) \text{ i.i.d.} \quad (72)$$

Proposition 9.1 (Euler–Maruyama Convergence). *Under standard Lipschitz conditions on b and σ :*

- **Strong convergence:** $\mathbb{E}[|X_T - X_T^{(\Delta t)}|] = O(\sqrt{\Delta t})$ (order $\frac{1}{2}$);
- **Weak convergence:** $|\mathbb{E}[f(X_T)] - \mathbb{E}[f(X_T^{(\Delta t)})]| = O(\Delta t)$ (order 1).

Weak convergence is sufficient for computing expectations (e.g., option prices).

Example 9.2 (Simulating the OU Process — Worked). For the OU process $dX_t = \kappa(\mu - X_t) dt + \sigma dW_t$:

$$X_{n+1} = X_n + \kappa(\mu - X_n) \Delta t + \sigma \sqrt{\Delta t} Z_n. \quad (73)$$

With $\kappa = 0.1$, $\mu = 0$, $\sigma = 1$, $\Delta t = 0.01$, $X_0 = 5$:

$$X_1 = 5 + 0.1(0 - 5)(0.01) + 1 \cdot \sqrt{0.01} \cdot Z_0 = 4.995 + 0.1 Z_0. \quad (74)$$

After many steps, X will fluctuate around $\mu = 0$ with long-run variance $\sigma^2/(2\kappa) = 5$.

9.5 The Milstein Method (Higher Order)

Algorithm Milstein Method

The Milstein scheme adds a correction term for strong convergence:

$$X_{n+1} = X_n + b(X_n)\Delta t + \sigma(X_n)\sqrt{\Delta t} Z_n + \frac{1}{2}\sigma(X_n)\sigma'(X_n)((\sqrt{\Delta t} Z_n)^2 - \Delta t). \quad (75)$$

Strong convergence: $O(\Delta t)$ (order 1), vs. $O(\sqrt{\Delta t})$ for Euler–Maruyama.

Remark 9.2. For the OU process with constant diffusion ($\sigma' = 0$), the Milstein correction vanishes and reduces to Euler–Maruyama. But for GBM ($\sigma(X) = \sigma X$, $\sigma' = \sigma$) the correction matters and improves pathwise accuracy.

9.6 Exercises

1. **Random walk simulation.** Simulate a symmetric random walk with $n = 10000$ steps and rescale to $W_t^{(n)}$. Plot the result and compare with a sample Brownian motion path.
2. **Itô's formula.** Apply Itô's formula to $f(x) = x^2$ with $dX_t = \mu dt + \sigma dW_t$. Verify that $\mathbb{E}[X_t^2] = X_0^2 + (2\mu X_0 + \sigma^2)t + \mu^2 t^2$ for constant μ, σ .
3. **Euler–Maruyama convergence.** Simulate GBM ($dS = 0.05S dt + 0.2S dW$, $S_0 = 100$) using Euler–Maruyama with $\Delta t \in \{0.1, 0.01, 0.001\}$. Compare the distribution of S_T (at $T = 1$) with the exact lognormal solution.
4. **Milstein improvement.** Repeat the GBM simulation with the Milstein method and compare strong errors with Euler–Maruyama.

Solutions

Solution to Exercise 2. With $f(x) = x^2$, $f'(x) = 2x$, $f''(x) = 2$:

$$d(X_t^2) = 2X_t dX_t + \frac{1}{2} \cdot 2 \cdot (dX_t)^2 = 2X_t(\mu dt + \sigma dW_t) + \sigma^2 dt \quad (76)$$

$$= (2\mu X_t + \sigma^2) dt + 2\sigma X_t dW_t. \quad (77)$$

Taking expectations (the dW_t integral has zero mean): $\frac{d}{dt}\mathbb{E}[X_t^2] = 2\mu\mathbb{E}[X_t] + \sigma^2 = 2\mu(X_0 + \mu t) + \sigma^2$. Integrating: $\mathbb{E}[X_t^2] = X_0^2 + (2\mu X_0 + \sigma^2)t + \mu^2 t^2$. ✓ $\square$

Solution to Exercise 3. The exact GBM solution is $S_T = 100 \exp\{(0.05 - 0.02)T + 0.2W_T\} = 100 \exp\{0.03 + 0.2W_1\}$ with $W_1 \sim \mathcal{N}(0, 1)$, so $\log S_T \sim \mathcal{N}(\log 100 + 0.03, 0.04)$. For Euler–Maruyama, the weak error scales as $O(\Delta t)$: $\Delta t = 0.1$: error $\sim 10\%$; $\Delta t = 0.01$: error $\sim 1\%$; $\Delta t = 0.001$: error $\sim 0.1\%$. Verify by comparing $\mathbb{E}[\log S_T]$ and $\text{Var}(\log S_T)$ from 10000 Monte Carlo paths against the exact values. $\square$

10 Mean-Reversion and the Ornstein–Uhlenbeck Process

10.1 Definition

Definition 10.1 (Ornstein–Uhlenbeck Process). The *Ornstein–Uhlenbeck* (OU) process satisfies the SDE:

$$dX_t = \kappa(\mu - X_t) dt + \sigma dW_t, \quad (78)$$

where $\kappa > 0$ is the mean-reversion speed, μ is the long-run mean, and $\sigma > 0$ is the diffusion coefficient.

Theorem 10.1 (Explicit Solution). *The OU process has the explicit solution:*

$$X_t = \mu + (X_0 - \mu)e^{-\kappa t} + \sigma \int_0^t e^{-\kappa(t-s)} dW_s. \quad (79)$$

Proof. Apply Itô's formula to $f(t, X_t) = e^{\kappa t}(X_t - \mu)$:

$$df = e^{\kappa t}[\kappa(X_t - \mu) + \kappa(\mu - X_t)] dt + \sigma e^{\kappa t} dW_t = \sigma e^{\kappa t} dW_t. \quad (80)$$

Integrating: $e^{\kappa t}(X_t - \mu) = (X_0 - \mu) + \sigma \int_0^t e^{\kappa s} dW_s$, which gives the result after multiplying by $e^{-\kappa t}$. $\square$

10.2 Distributional Properties

Proposition 10.1. *The OU process has:*

$$\mathbb{E}[X_t] = \mu + (X_0 - \mu)e^{-\kappa t}, \quad (81)$$

$$\text{Var}(X_t) = \frac{\sigma^2}{2\kappa}(1 - e^{-2\kappa t}), \quad (82)$$

$$\lim_{t \rightarrow \infty} X_t \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\kappa}\right). \quad (83)$$

10.3 Half-Life of Mean Reversion

Definition 10.2 (Half-Life). The *half-life* of mean reversion is the time for the expected deviation from the mean to halve:

$$t_{1/2} = \frac{\ln 2}{\kappa}. \quad (84)$$

For a spread with $\kappa = 0.05$ per day: $t_{1/2} = \ln(2)/0.05 \approx 13.9$ trading days.

Key Insight

The half-life determines the optimal holding period for mean-reversion strategies. In WP3, spreads with $t_{1/2}$ between 5 and 20 days are selected as tradeable. Faster reversion (< 5 days) is captured by transaction costs; slower reversion (> 20 days) has too much noise risk.

10.4 Estimation via OLS Regression

The OU process in discrete time satisfies:

$$X_{t+1} - X_t = \kappa(\mu - X_t)\Delta t + \sigma\sqrt{\Delta t}\varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 1). \quad (85)$$

Running OLS regression of ΔX_t on X_t :

$$\Delta X_t = a + bX_t + \varepsilon_t, \quad (86)$$

the estimates are:

$$\hat{\kappa} = -\hat{b}/\Delta t, \quad (87)$$

$$\hat{\mu} = -\hat{a}/\hat{b}, \quad (88)$$

$$\hat{t}_{1/2} = \ln(2)/\hat{\kappa} = -\ln(2)\Delta t/\hat{b}. \quad (89)$$

Practical Application

Statistical test: The estimated $\hat{b}$ should be significantly negative (ADF test with $p < 0.05$) to confirm mean reversion. Non-stationary spreads ($\hat{b} \approx 0$) are discarded.

10.5 Exercises

1. Simulate an OU process with $\kappa = 0.1$, $\mu = 0$, $\sigma = 1$ using the Euler–Maruyama method. Verify that the sample distribution converges to $\mathcal{N}(0, 5)$.
2. Estimate κ from simulated data using OLS and verify that $\hat{\kappa} \approx 0.1$.
3. Show that the autocorrelation of the OU process is $\rho(\tau) = e^{-\kappa\tau}$.
4. **ADF test interpretation.** If the ADF test on a spread gives test statistic -2.1 and critical value -2.86 (5% level), what do you conclude? What if the test statistic is -3.5 ?

Solutions

Solution to Exercise 1. Using Euler–Maruyama with $dX_t = 0.1(0 - X_t)dt + 1 \cdot dW_t$: $X_{n+1} = X_n - 0.1X_n\Delta t + \sqrt{\Delta t}Z_n$. The stationary distribution is $\mathcal{N}(\mu, \sigma^2/(2\kappa)) = \mathcal{N}(0, 1/(2 \times 0.1)) = \mathcal{N}(0, 5)$. After running 10^5 steps with $\Delta t = 0.01$, discard the first 1000 (burn-in), then compute the sample mean (≈ 0) and sample variance (≈ 5). $\square$

Solution to Exercise 2. From the discrete-time regression $\Delta X_t = a + bX_t + \varepsilon_t$: $\hat{\kappa} = -\hat{b}/\Delta t$. If $\Delta t = 1$ (daily), then $\hat{\kappa} = -\hat{b}$. With true $\kappa = 0.1$, we expect $\hat{b} \approx -0.1$. The standard error of $\hat{b}$ decreases as $O(1/\sqrt{T})$ where T is the number of observations. For $T = 1000$ daily observations, the estimate is typically accurate to $\pm 30\%$. $\square$

Solution to Exercise 3. The autocovariance of the stationary OU process is: $\text{Cov}(X_t, X_{t+\tau}) = \frac{\sigma^2}{2\kappa}e^{-\kappa|\tau|}$. The autocorrelation is $\rho(\tau) = \frac{\text{Cov}(X_t, X_{t+\tau})}{\text{Var}(X_t)} = e^{-\kappa|\tau|}$.

Derivation: Using the explicit solution: $X_{t+\tau} = \mu + (X_t - \mu)e^{-\kappa\tau} + \sigma \int_t^{t+\tau} e^{-\kappa(t+\tau-s)} dW_s$. The stochastic integral is independent of X_t , so: $\text{Cov}(X_t, X_{t+\tau}) = e^{-\kappa\tau} \text{Var}(X_t) = e^{-\kappa\tau} \frac{\sigma^2}{2\kappa}$. $\square$

Solution to Exercise 4 (ADF test). The ADF test has H_0 : unit root ($b = 0$, no mean reversion) vs. H_1 : $b < 0$ (mean-reverting). Test statistic $-2.1 > -2.86$: fail to reject H_0 at 5%. The spread is not significantly mean-reverting — do not trade it. Test statistic $-3.5 < -2.86$: reject H_0 at 5%. The spread is mean-reverting and is a candidate for a mean-reversion strategy. $\square$

11 Regime-Conditional Portfolio Construction

11.1 The Adaptive Framework

The WP3 adaptive portfolio framework combines the tools developed in previous sections into a unified pipeline:

1. **Regime Detection** (Sections 3–4): A Gaussian HMM with $K = 3$ states is fitted to daily returns via Baum–Welch. The posterior probabilities $\gamma_t(k)$ indicate the current regime.
2. **Factor Decomposition** (Section 5): Within each regime, sparse PCA extracts interpretable factor loadings. Assets are grouped by their sparse PC scores.
3. **Risk Budgeting** (Sections 7–8): The portfolio is optimised by minimising CVaR at the 5% level, using only the return scenarios from the current regime.
4. **Execution Timing** (Section 10): Mean-reversion analysis identifies optimal entry/exit timing for spread-based positions.

11.2 Regime Transition Detection

A regime transition is signalled when:

$$\gamma_t(k^*) > \gamma_{\text{threshold}} \quad \text{and} \quad k^* \neq k_{t-1}^*, \quad (90)$$

where $k^* = \operatorname{argmax}_k \gamma_t(k)$ and $\gamma_{\text{threshold}} = 0.7$.

Key Insight

The threshold $\gamma_{\text{threshold}} = 0.7$ prevents whipsawing: the model must be 70% confident of a regime change before rebalancing. This reduces turnover by $\sim 40\%$ compared to rebalancing on every argmax change.

11.3 Ablation Study Design

An *ablation study* systematically removes components to measure their marginal contribution:

Layer	Component	Marginal Sharpe
0	Equal weight baseline	0.00
+1	Markowitz mean-variance	+0.35
+2	HMM regime switching	+0.42
+3	Sparse PCA factors	+0.28
+4	CVaR replacing variance	+0.22
+5	Execution timing (OU)	+0.25

Total cumulative Sharpe improvement: $0.35 + 0.42 + 0.28 + 0.22 + 0.25 = 1.52$.

11.4 Exercises

1. Implement the full adaptive pipeline on simulated data with known regime switches and verify that the HMM detects them correctly.
2. Compare the out-of-sample Sharpe ratio of the adaptive portfolio with a static Markowitz portfolio on 10 years of equity ETF data.
3. Design your own ablation study for a 4-component trading strategy and interpret the results.

Solutions

Solution to Exercise 1. Pipeline steps: (1) Simulate 3 regimes with distinct means and variances (e.g., bull: $\mathcal{N}(0.05\%, 1\%)$, consolidation: $\mathcal{N}(0, 0.5\%)$, crisis: $\mathcal{N}(-0.1\%, 3\%)$) with known transition matrix. (2) Fit a 3-state Gaussian HMM via Baum–Welch. (3) Compute $\gamma_t(k)$ and check if $\arg\max_k \gamma_t(k)$ matches the true regime at each t . Accuracy is typically 85–95% depending on how well-separated the regimes are. The main source of error is at regime boundaries, where the posterior is uncertain. $\square$

Solution to Exercise 2. Use 10 years of daily returns for 5–10 equity ETFs (e.g., SPY, QQQ, IWM, EFA, AGG). **Static Markowitz:** Estimate $\hat{\mu}, \hat{\Sigma}$ over the full sample, compute optimal weights, apply them for the full period. **Adaptive:** Fit HMM, compute regime-conditional $\hat{\mu}_k, \hat{\Sigma}_k$, optimise CVaR in each regime, rebalance when regime changes. Typical result: adaptive Sharpe ≈ 0.8 – 1.2 vs. static ≈ 0.4 – 0.7 . The improvement comes mainly from reducing exposure during crisis regimes. $\square$

Solution to Exercise 3 (Ablation design). Example: a momentum strategy with 4 components: (0) Baseline: equal-weight; (+1) Momentum signal (12–1 month return); (+2) Volatility scaling; (+3) Sector neutralisation; (+4) Transaction cost modelling. Ablation: remove each component and measure the Sharpe drop. If removing momentum signal drops Sharpe by -0.6 but removing sector neutralisation only by -0.1 , the signal is $6\times$ more important than sector neutralisation. The ablation table should be read bottom-up (contribution) and top-down (robustness to removal). $\square$

12 Summary and Connections to WP3

This primer has covered the mathematical foundations needed for WP3:

Topic	Section	WP3 Application
Markov chains	§2	Regime dynamics
Hidden Markov Models	§3	Regime detection
Baum–Welch (EM)	§4	Parameter estimation
PCA / Sparse PCA	§5	Factor decomposition
Convex optimisation	§7	Portfolio construction
CVaR	§8	Tail risk management
OU process	§10	Mean-reversion timing
Adaptive pipeline	§11	Full framework

Key Insight

The central thesis of WP3 is that combining *regime-aware asset allocation* (HMM + CVaR) with *sparse factor models* (for interpretability) and *mean-reversion execution* (OU timing) produces a portfolio strategy that significantly outperforms static allocation. The ablation study shows that each component contributes meaningfully to the total Sharpe improvement of 1.52 over the equal-weight baseline.

References

- [1] Artzner, P., Delbaen, F., Eber, J.-M., Heath, D. (1999). *Coherent measures of risk*. Mathematical Finance, 9(3), 203–228.
- [2] Baum, L.E., Petrie, T., Soules, G., Weiss, N. (1970). *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains*. Annals of Mathematical Statistics, 41(1), 164–171.
- [3] Dempster, A.P., Laird, N.M., Rubin, D.B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*. J. Royal Statistical Society B, 39(1), 1–38.
- [4] Hamilton, J.D. (1989). *A new approach to the economic analysis of nonstationary time series and the business cycle*. Econometrica, 57(2), 357–384.
- [5] Ledoit, O., Wolf, M. (2004). *A well-conditioned estimator for large-dimensional covariance matrices*. J. Multivariate Analysis, 88(2), 365–411.
- [6] Markowitz, H. (1952). *Portfolio selection*. The Journal of Finance, 7(1), 77–91.
- [7] Rabiner, L.R. (1989). *A tutorial on hidden Markov models and selected applications in speech recognition*. Proceedings of the IEEE, 77(2), 257–286.
- [8] Rockafellar, R.T., Uryasev, S. (2000). *Optimization of conditional value-at-risk*. Journal of Risk, 2(3), 21–41.
- [9] Schwarz, G. (1978). *Estimating the dimension of a model*. Annals of Statistics, 6(2), 461–464.
- [10] Viterbi, A. (1967). *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*. IEEE Transactions on Information Theory, 13(2), 260–269.