

Contents

I	Introduction	1
1	Introduction To Advanced Algorithmic Trading	3
1.1	The Hunt for Alpha	3
1.2	Why Time Series Analysis, Bayesian Statistics and Machine Learning?	3
1.2.1	Bayesian Statistics	4
1.2.2	Time Series Analysis	4
1.2.3	Machine Learning	5
1.3	How Is The Book Laid Out?	6
1.4	Required Technical Background	6
1.4.1	Mathematics	7
1.4.2	Programming	7
1.5	How Does This Book Differ From "Successful Algorithmic Trading"?	8
1.6	Software Installation	8
1.6.1	Installing Python	8
1.6.2	Installing R	9
1.7	QSTrader Backtesting Simulation Software	9
1.7.1	Alternatives	10
1.8	Where to Get Help	10
II	Bayesian Statistics	11
2	Introduction to Bayesian Statistics	13
2.1	What is Bayesian Statistics?	13
2.1.1	Frequentist vs Bayesian Examples	14
2.2	Applying Bayes' Rule for Bayesian Inference	17
2.3	Coin-Flipping Example	18
3	Bayesian Inference of a Binomial Proportion	23
3.1	The Bayesian Approach	24
3.2	Assumptions of the Approach	24
3.3	Recalling Bayes' Rule	25
3.4	The Likelihood Function	25
3.4.1	Bernoulli Distribution	25
3.4.2	Bernoulli Likelihood Function	26
3.4.3	Multiple Flips of the Coin	27
3.5	Quantifying our Prior Beliefs	27

3.5.1	Beta Distribution	27
3.5.2	Why Is A Beta Prior Conjugate to the Bernoulli Likelihood?	30
3.5.3	Multiple Ways to Specify a Beta Prior	30
3.6	Using Bayes' Rule to Calculate a Posterior	31
4	Markov Chain Monte Carlo	35
4.1	Bayesian Inference Goals	35
4.2	Why Markov Chain Monte Carlo?	36
4.2.1	Markov Chain Monte Carlo Algorithms	36
4.3	The Metropolis Algorithm	37
4.4	Introducing PyMC3	38
4.5	Inferring a Binomial Proportion with Markov Chain Monte Carlo	38
4.5.1	Inferring a Binomial Proportion with Conjugate Priors Recap	39
4.5.2	Inferring a Binomial Proportion with PyMC3	39
4.6	Bibliographic Note	45
5	Bayesian Linear Regression	47
5.1	Frequentist Linear Regression	47
5.2	Bayesian Linear Regression	48
5.3	Bayesian Linear Regression with PyMC3	49
5.3.1	What are Generalised Linear Models?	49
5.3.2	Simulating Data and Fitting the Model with PyMC3	50
5.4	Bibliographic Note	55
5.5	Full Code	56
6	Bayesian Stochastic Volatility Model	59
6.1	Stochastic Volatility	59
6.2	Bayesian Stochastic Volatility	60
6.3	PyMC3 Implementation	63
6.3.1	Obtaining the Price History	63
6.3.2	Model Specification in PyMC3	65
6.3.3	Fitting the Model with NUTS	65
6.4	Full Code	67
III	Time Series Analysis	71
7	Introduction to Time Series Analysis	73
7.1	What is Time Series Analysis?	73
7.2	How Can We Apply Time Series Analysis in Quantitative Finance?	74
7.3	Time Series Analysis Software	74
7.4	Time Series Analysis Roadmap	75
7.5	How Does This Relate to Other Statistical Tools?	76
8	Serial Correlation	77
8.1	Expectation, Variance and Covariance	77
8.1.1	Example: Sample Covariance in R	78

8.2	Correlation	80
8.2.1	Example: Sample Correlation in R	80
8.3	Stationarity in Time Series	80
8.4	Serial Correlation	82
8.5	The Correlogram	83
8.5.1	Example 1 - Fixed Linear Trend	84
8.5.2	Example 2 - Repeated Sequence	84
8.6	Next Steps	86
9	Random Walks and White Noise Models	87
9.1	Time Series Modelling Process	87
9.2	Backward Shift and Difference Operators	88
9.3	White Noise	89
9.3.1	Second-Order Properties	89
9.3.2	Correlogram	90
9.4	Random Walk	91
9.4.1	Second-Order Properties	91
9.4.2	Correlogram	92
9.4.3	Fitting Random Walk Models to Financial Data	93
10	Autoregressive Moving Average Models	97
10.1	How Will We Proceed?	98
10.2	Strictly Stationary	98
10.3	Akaike Information Criterion	99
10.4	Autoregressive (AR) Models of order p	99
10.4.1	Rationale	99
10.4.2	Stationarity for Autoregressive Processes	100
10.4.3	Second Order Properties	101
10.4.4	Simulations and Correlograms	101
10.4.5	Financial Data	105
10.5	Moving Average (MA) Models of order q	111
10.5.1	Rationale	112
10.5.2	Definition	112
10.5.3	Second Order Properties	112
10.5.4	Simulations and Correlograms	113
10.5.5	Financial Data	117
10.5.6	Next Steps	124
10.6	Autogressive Moving Average (ARMA) Models of order p, q	124
10.6.1	Bayesian Information Criterion	124
10.6.2	Ljung-Box Test	125
10.6.3	Rationale	125
10.6.4	Definition	126
10.6.5	Simulations and Correlograms	126
10.6.6	Choosing the Best ARMA(p,q) Model	130
10.6.7	Financial Data	132

10.7	Next Steps	134
11	Autoregressive Integrated Moving Average and Conditional Heteroskedastic Models	135
11.1	Quick Recap	135
11.2	Autoregressive Integrated Moving Average (ARIMA) Models of order p, d, q	136
11.2.1	Rationale	136
11.2.2	Definitions	136
11.2.3	Simulation, Correlogram and Model Fitting	137
11.2.4	Financial Data and Prediction	139
11.2.5	Next Steps	144
11.3	Volatility	144
11.4	Conditional Heteroskedasticity	144
11.5	Autoregressive Conditional Heteroskedastic Models	145
11.5.1	ARCH Definition	145
11.5.2	Why Does This Model Volatility?	146
11.5.3	When Is It Appropriate To Apply ARCH(1)?	146
11.5.4	ARCH(p) Models	146
11.6	Generalised Autoregressive Conditional Heteroskedastic Models	147
11.6.1	GARCH Definition	147
11.6.2	Simulations, Correlograms and Model Fittings	147
11.6.3	Financial Data	150
11.7	Next Steps	153
12	Cointegrated Time Series	155
12.1	Mean Reversion Trading Strategies	155
12.2	Cointegration	156
12.3	Unit Root Tests	156
12.3.1	Augmented Dickey-Fuller Test	157
12.3.2	Phillips-Perron Test	157
12.3.3	Phillips-Ouliaris Test	157
12.3.4	Difficulties with Unit Root Tests	157
12.4	Simulated Cointegrated Time Series with R	157
12.5	Cointegrated Augmented Dickey Fuller Test	162
12.6	CADF on Simulated Data	163
12.7	CADF on Financial Data	164
12.7.1	EWA and EWC	165
12.7.2	RDS-A and RDS-B	168
12.8	Full Code	171
12.9	Johansen Test	174
12.9.1	Johansen Test on Simulated Data	175
12.9.2	Johansen Test on Financial Data	178
12.9.3	Full Code	181
13	State Space Models and the Kalman Filter	185
13.1	Linear State-Space Model	186

13.2	The Kalman Filter	187
13.2.1	A Bayesian Approach	188
13.2.2	Prediction	189
13.3	Dynamic Hedge Ratio Between ETF Pairs Using the Kalman Filter	191
13.3.1	Linear Regression via the Kalman Filter	191
13.3.2	Applying the Kalman Filter to a Pair of ETFs	192
13.3.3	TLT and ETF	192
13.3.4	Scatterplot of ETF Prices	192
13.3.5	Time-Varying Slope and Intercept	193
13.4	Next Steps	196
13.5	Bibliographic Note	196
13.6	Full Code	196
14	Hidden Markov Models	201
14.1	Markov Models	202
14.1.1	Markov Model Mathematical Specification	203
14.2	Hidden Markov Models	204
14.2.1	Hidden Markov Model Mathematical Specification	205
14.2.2	Filtering of Hidden Markov Models	205
14.3	Regime Detection with Hidden Markov Models	206
14.3.1	Market Regimes	207
14.3.2	Simulated Data	207
14.3.3	Financial Data	210
14.4	Next Steps	213
14.5	Bibliographic Note	213
14.6	Full Code	213
IV	Statistical Machine Learning	217
15	Introduction to Machine Learning	219
15.1	What is Machine Learning?	219
15.2	Machine Learning Domains	220
15.2.1	Supervised Learning	220
15.2.2	Unsupervised Learning	220
15.2.3	Reinforcement Learning	220
15.3	Machine Learning Techniques	220
15.3.1	Linear Regression	221
15.3.2	Linear Classification	221
15.3.3	Tree-Based Methods	221
15.3.4	Support Vector Machines	221
15.3.5	Artificial Neural Networks and Deep Learning	221
15.3.6	Bayesian Networks	221
15.3.7	Clustering	221
15.3.8	Dimensionality Reduction	221
15.4	Machine Learning Applications	222

15.4.1	Forecasting and Prediction	222
15.4.2	Natural Language Processing	222
15.4.3	Factor Models	222
15.4.4	Image Classification	222
15.4.5	Model Accuracy	223
15.4.6	Parametric and Non-Parametric Models	223
15.4.7	Statistical Framework for Machine Learning Domains	224
16	Supervised Learning	225
16.1	Mathematical Framework	225
16.2	Classification	226
16.3	Regression	226
16.3.1	Financial Example	227
16.4	Training	227
17	Linear Regression	229
17.1	Linear Regression	229
17.2	Probabilistic Interpretation	230
17.2.1	Basis Function Expansion	232
17.3	Maximum Likelihood Estimation	232
17.3.1	Likelihood and Negative Log Likelihood	233
17.3.2	Ordinary Least Squares	234
17.4	Simulated Data Example with Scikit-Learn	235
17.5	Full Code	238
17.6	Bibliographic Note	241
18	Tree-Based Methods	243
18.1	Decision Trees - Mathematical Overview	243
18.2	Decision Trees for Regression	244
18.2.1	Creating a Regression Tree and Making Predictions	245
18.2.2	Pruning The Tree	246
18.3	Decision Trees for Classification	247
18.3.1	Classification Error Rate/Hit Rate	247
18.3.2	Gini Index	247
18.3.3	Cross-Entropy/Deviance	247
18.4	Advantages and Disadvantages of Decision Trees	248
18.4.1	Advantages	248
18.4.2	Disadvantages	248
18.5	Ensemble Methods	248
18.5.1	The Bootstrap	248
18.5.2	Bootstrap Aggregation	249
18.5.3	Random Forests	250
18.5.4	Boosting	250
18.5.5	Python Scikit-Learn Implementation	251
18.6	Bibliographic Note	257
18.7	Full Code	257

19 Support Vector Machines	261
19.1 Motivation for Support Vector Machines	261
19.2 Advantages and Disadvantages of SVMs	262
19.2.1 Advantages	262
19.2.2 Disadvantages	262
19.3 Linear Separating Hyperplanes	263
19.4 Classification	265
19.5 Deriving the Classifier	266
19.6 Constructing the Maximal Margin Classifier	267
19.7 Support Vector Classifiers	268
19.8 Support Vector Machines	271
19.8.1 Bibliographic Notes	273
20 Model Selection and Cross-Validation	275
20.1 Bias-Variance Trade-Off	275
20.1.1 Machine Learning Models	275
20.1.2 Model Selection	276
20.1.3 The Bias-Variance Tradeoff	278
20.2 Cross-Validation	281
20.2.1 Overview of Cross-Validation	282
20.2.2 Forecasting Example	282
20.2.3 Validation Set Approach	283
20.2.4 k-Fold Cross Validation	284
20.2.5 Python Implementation	285
20.2.6 k-Fold Cross Validation	289
20.2.7 Full Python Code	292
21 Unsupervised Learning	301
21.1 High Dimensional Data	302
21.2 Mathematical Overview of Unsupervised Learning	302
21.3 Unsupervised Learning Algorithms	303
21.3.1 Dimensionality Reduction	303
21.3.2 Clustering	303
21.4 Bibliographic Note	304
22 Clustering Methods	305
22.1 K-Means Clustering	305
22.1.1 The Algorithm	306
22.1.2 Issues	307
22.1.3 Simulated Data	308
22.1.4 OHLC Clustering	310
22.2 Bibliographic Note	318
22.3 Full Code	318
23 Natural Language Processing	325
23.1 Overview	325

23.2	Supervised Document Classification	326
23.3	Preparing a Dataset for Classification	326
23.4	Vectorisation	338
23.5	Term-Frequency Inverse Document-Frequency	338
23.6	Training the Support Vector Machine	340
23.7	Performance Metrics	342
23.8	Full Code	344
V Quantitative Trading Techniques		349
24	Introduction to QSTrader	351
24.1	Motivation for QSTrader	351
24.2	Design Considerations	352
24.3	Installation	354
25	Introductory Portfolio Strategies	355
25.1	Motivation	355
25.2	The Trading Strategies	355
25.3	Data	356
25.4	Python QSTrader Implementation	357
25.4.1	MonthlyLiquidateRebalanceStrategy	358
25.4.2	LiquidateRebalancePositionSizer	359
25.4.3	Backtest Interface	360
25.5	Strategy Results	361
25.5.1	Transaction Costs	361
25.5.2	US Equities/Bonds 60/40 ETF Portfolio	361
25.5.3	"Strategic" Weight ETF Portfolio	362
25.5.4	Equal Weight ETF Portfolio	364
25.6	Full Code	365
26	ARIMA+GARCH Trading Strategy on Stock Market Indexes Using R . . .	369
26.1	Strategy Overview	369
26.2	Strategy Implementation	370
26.3	Strategy Results	373
26.4	Full Code	376
27	Cointegration-Based Pairs Trading using QSTrader	381
27.1	The Hypothesis	381
27.2	Cointegration Tests in R	382
27.3	The Trading Strategy	384
27.4	Data	385
27.5	Python QSTrader Implementation	385
27.6	Strategy Results	391
27.6.1	Transaction Costs	391
27.6.2	Tearsheet	391
27.7	Full Code	392

28 Kalman Filter-Based Pairs Trading using QSTrader	399
28.1 The Trading Strategy	399
28.1.1 Data	401
28.2 Python QSTrader Implementation	401
28.3 Strategy Results	407
28.4 Next Steps	407
28.5 Full Code	409
29 Supervised Learning for Intraday Returns Prediction using QSTrader	415
29.1 Prediction Goals with Machine Learning	415
29.1.1 Class Imbalance	416
29.2 Building a Prediction Model on Historical Data	417
29.3 QSTrader Strategy Object	422
29.4 QSTrader Backtest Script	425
29.5 Results	428
29.6 Next Steps	428
29.7 Full Code	431
30 Sentiment Analysis via Sentdex Vendor Sentiment Data with QSTrader . . .	439
30.1 Sentiment Analysis	439
30.1.1 Sentdex API and Sample File	440
30.2 The Trading Strategy	441
30.3 Data	441
30.4 Python Implementation	443
30.4.1 Sentiment Handling with QSTrader	443
30.4.2 Sentiment Analysis Strategy Code	447
30.5 Strategy Results	450
30.5.1 Transaction Costs	450
30.5.2 Sentiment on S&P500 Tech Stocks	450
30.5.3 Sentiment on S&P500 Energy Stocks	451
30.5.4 Sentiment on S&P500 Defence Stocks	452
30.6 Full Code	453
31 Market Regime Detection with Hidden Markov Models using QSTrader . . .	457
31.1 Regime Detection with Hidden Markov Models	457
31.2 The Trading Strategy	458
31.3 Data	458
31.4 Python Implementation	459
31.4.1 Returns Calculation with QSTrader	459
31.4.2 Regime Detection Implementation	460
31.5 Strategy Results	470
31.5.1 Transaction Costs	470
31.5.2 No Regime Detection Filter	470
31.5.3 HMM Regime Detection Filter	471
31.6 Full Code	472

32 Strategy Decay	481
32.1 Calculating the Annualised Rolling Sharpe Ratio	482
32.2 Python QSTrader Implementation	483
32.3 Strategy Results	486
32.3.1 Kalman Filter Pairs Trade	486
32.3.2 Aluminum Smelting Cointegration Strategy	486
32.3.3 Sentdex Sentiment Analysis Strategy	488