The Quant Interview Cheat Sheet

Written by the Team at QuantGuide.io

Contents

1 Preface

1.1 Introductory Remarks

Quant interviews are hard. And for good reason, too – at the intersection of computer science, statistics, and finance lies the most lucrative job out of university, with graduate opportunities commanding up to $600,000 per year in compensation from base salary, sign-on bonus, and performance bonus (and internships paying over $150 per hour).

Naturally, every year is more competitive than the last, and sometimes we even wonder if we'd still be able to pass the interviews at the top trading firms today. We're uncompromising believers that you do not have to be a genius to land a role in quant and that there exists a path of persistent preparation and tenacious thoughtfulness that will lead to an offer. We carefully thought about what resources would have helped us back then that would have made our lives less stressful when preparing for interviews, so we authored this sheet in hopes that it can make your studying and interviewing processes easier.

1.2 How to Use This

This document is meant to assist you in your preparation for interviews. It will refresh you on fundamental ideas that universally show up in Quantitative Trading (QT) and Research (QR) interviews. Sections labelled **QT** and **QR** are to represent the type of interview that the content is relevant for. Note that this document only consists of important results without derivation/proof. It's imperative that you properly practice questions and review the concepts listed here, as this is not a substitute for studying.

– The Team at QuantGuide.io

2 Probability

2.1 Key Distributions

Name	Modeling Intuition	PMF/PDF	MGF	μ	σ^2
Bernoulli	Toss a coin, 1 if heads, else 0, coin lands heads with probability p	$f(t;p) = \begin{cases} p & \text{if } t = 1 \\ 1-p & \text{if } t = 0 \end{cases}$	$pe^\theta + (1-p)$	p	$p(1-p)$
Binomial	Toss a coin n times, probability of t heads, coin lands heads with probability p	$f(t;n,p) = \binom{n}{t}p^t(1-p)^{n-t}$	$\left[pe^\theta + (1-p)\right]^n$	np	$np(1-p)$
Geometric	Probability of tossing coin t times until heads, coin lands heads with probability p	$f(t;p) = p(1-p)^{t-1}$	$\frac{pe^\theta}{1-(1-p)e^\theta}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson	Probability of t occurrences within a fixed time interval or space; parameter λ represents average number of occurrences	$f(t;\lambda) = \frac{\lambda^t e^{-\lambda}}{t!}$	$e^{\lambda(e^\theta-1)}$	λ	λ
Exponential	Probability distribution of time between events in a Poisson process occurring with rate λ	$f(t;\lambda) = \lambda e^{-\lambda t}\ \mathbf{1}_{t\geq 0}$	$\frac{\lambda}{\lambda-\theta}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Uniform	Uniform on a line	$f(t;a,b) = \frac{1}{b-a}\ \mathbf{1}_{t\in[a,b]}$	$\frac{e^{b\theta}-e^{a\theta}}{\theta(b-a)}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	Standard univariate normal, z-score transform	$f(t) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$e^{\mu\theta+\frac{1}{2}\sigma^2\theta^2}$	μ	σ^2

2.2 Formulas and Laws

Common Relationships Between Distributions (QT/QR)

1. $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ IID $\implies \displaystyle\sum_{i=1}^{n} X_i \sim \text{Binom}(n,p)$

2. $X_1, \ldots, X_r \sim \text{Geom}(p)$ IID $\implies \displaystyle\sum_{i=1}^{r} X_i \sim \text{NegBinom}(r,p)$

3. $X_i \sim \text{Poisson}(\lambda_i)$, $1 \leq i \leq n$ independent $\implies \displaystyle\sum_{i=1}^{n} X_i \sim \text{Poisson}\left(\sum_{i=1}^{n} \lambda_i\right)$

4. $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$ IID $\implies \displaystyle\sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda^{-1})$ (Shape-Scale parameterization)

5. $X_i \sim N(\mu, \sigma^2)$, $1 \leq i \leq n$ independent $\implies \displaystyle\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$

6. If $X \sim \text{Unif}(0,1)$ and $F(x)$ is an invertible CDF, then $Y = F^{-1}(X)$ has CDF $F(x)$

Conditional Probability, Bayes, and Law of Total Probability (QT/QR)

Consider events $A_1, \ldots, A_n$ which form a partition of the sample space as well as event B. Then,

$$\mathbb{P}(A_1 \,|\, B) = \frac{\mathbb{P}(A_1 \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1)}{\sum_{i=1}^{n} \mathbb{P}(B \cap A_i)} = \frac{\mathbb{P}(B \,|\, A_1)\mathbb{P}(A_1)}{\sum_{i=1}^{n} \mathbb{P}(B \,|\, A_i)\mathbb{P}(A_i)}$$

Law of Total Expectation and Variance (QT/QR)

For two random variables X, Y defined on the same sample space,

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X \,|\, Y]] \overset{\text{discrete } Y}{=} \sum_{i=1}^{\infty} \mathbb{P}(Y = y_i)\, \mathbb{E}[X \,|\, Y = y_i] \overset{\text{continuous } Y}{=} \int_{\mathbb{R}} \mathbb{E}[X \mid Y = y] f_Y(y) dy \\ \text{Var}(X) &= \text{Var}(\mathbb{E}[X \mid Y]) + \mathbb{E}[\text{Var}(X \mid Y)], \text{ where } \text{Var}(X \mid Y) = \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2 \mid Y] \end{aligned}$$

Intuitively, the Law of Total Expectation says that if we ”average over all averages” of X obtained by some information about Y, we obtain the true average. Similarly, the Law of Total Variance says that the true variance comes from two sources: between samples (the first term) and within samples (the second term).

Covariance and Correlation (QT/QR)

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \qquad \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

Covariance and correlation are measurements of linear association of X and Y. An example of uncorrelated but not independent random variables are Z and Z^2, where $Z \sim N(0, 1)$.

Properties of Expectation, Variance, and Covariance (QT/QR)

Let a, b, c, and d be real constants and X and Y be random variables with finite mean and variance. Then all of the following hold:

1. $\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$
2. $\text{Var}(aX + b) = a^2\text{Var}(X)$
3. $\text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y)$
4. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$
5. If X and Y are independent and have finite mean, then X and Y are uncorrelated.
6. $\text{Corr}(aX + b, cY + d) = \text{sign}(ac)\text{Corr}(X, Y)$
7. $\text{Cov}(X, X) = \text{Var}(X)$
8. The correlation and covariance matrices are both positive semidefinite.
9. $|\text{Corr}(X, Y)| \leq 1$

Tail Sum and Integral (QT/QR)

If X is a non-negative random variable, then

$$\mathbb{E}[X] \overset{\text{continuous } X}{=} \int_0^{\infty} \mathbb{P}(X > t)\, dt \overset{\text{integer-valued } X}{=} \sum_{t=0}^{\infty} \mathbb{P}(X > t)$$

Law of the Unconscious Statistician (LOTUS) (QT/QR)

Let X be a random variable and $g(x)$ be a function. Then we have

$$\mathbb{E}[g(X)] \overset{\text{continuous } X}{=} \int_{\mathbb{R}} g(x) f_X(x)\, dx \overset{\text{discrete } X}{=} \sum_{k \in \text{Supp}(X)} g(k)\mathbb{P}[X = k]$$

Central Limit Theorem (QT/QR)
Suppose that $X_1, X_2, \ldots, X_n$ are IID random variables with finite mean μ and variance σ^2. Furthermore, suppose that n is large. Define $S_n = X_1 + \cdots + X_n$ and $\overline{X}_n = \frac{S_n}{n}$. Then we have that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$$

This intuitively says that sums of large amounts of IID random variables approach a standard normal distribution when re-scaled to have mean 0 and variance 1. This is useful to approximate sums whose exact distribution is not known.

Law of Large Numbers (QT/QR)
Suppose that $X_1, X_2, \ldots$ are IID random variables with finite mean μ. Define $\overline{X}_n = \dfrac{X_1 + \cdots + X_n}{n}$. With probability 1, we have that

$$\lim_{n\to\infty} \overline{X}_n = \mu$$

This intuitively says that the sample average approaches the true average as the sample size grows large. This theorem is the basis of Monte Carlo Sampling, as we are attempting to find the true average by averaging the results of simulations.

2.3 Markov Chains

An $n \times n$ transition matrix $\boldsymbol{P}$ is defined as follows for a discrete state space with n states $\mathcal{X} = \{x_1, x_2, \ldots x_n\}$, where P_{ij} is the probability of transitioning from state x_i to state x_j. Each entry in P is within $[0,1]$, and the sum of entries for each row must total 1. Note that not all transition matrices are stationary.

Stationary Distribution (QR)
Solve for row vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$, where π_i is the stationary probability of the i-th state.

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}, \quad \boldsymbol{\pi}\mathbf{1} = 1$$

Expected Time to State (Example)
Suppose we have a 3×3 transition matrix $\boldsymbol{P}$, and we wish to find the expected time to reach x_3 from x_1. We can solve the following system of equations for μ_1, where μ_i denotes the expected time to reach state x_3 starting from x_i:

$$\mu_1 = 1 + \boldsymbol{P}_{11}\mu_1 + \boldsymbol{P}_{12}\mu_2 + \boldsymbol{P}_{13}$$
$$\mu_2 = 1 + \boldsymbol{P}_{21}\mu_1 + \boldsymbol{P}_{22}\mu_2 + \boldsymbol{P}_{23}$$

Gambler's Ruin/Random Walks (QT/QR)
Suppose that you and a friend start with \$$a$ and \$$b$, respectively, where $a, b > 0$ are integers. Each round, a fair coin in flipped. You receive \$1 from your friend if it appears heads, while you must pay \$1 to your friend if it appears tails. The game ends once one of the players has no money left. The probability that you are ruined i.e. have no money left is

$$\frac{a}{a+b}$$

If the coin has probability p of appearing heads, define $\varrho = \dfrac{p}{1-p}$. The probability that you are ruined is

$$\frac{1-\varrho^b}{1-\varrho^{a+b}}$$

3 Statistical Learning

3.1 Linear Regression

Simple Linear Regression (QR)
Suppose we have a variable X and another variable Y which we assume has some sort of linear relationship with X, and suppose we have m samples in a data set. Specifically, we assume

$$Y = \beta_0 + \beta_1 X + \epsilon,\ \ \mathbb{E}[\epsilon] = 0,\ \ \mathrm{Var}(\epsilon) = \sigma^2,\ \ \epsilon \text{ and } X \text{ independent}$$

The best estimates for β_0, β_1 that minimize the residual sum of squares $\left(RSS = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2\right)$ are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \approx \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)},\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},\quad \text{and}$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{m}(x_i - \bar{x})^2},\quad \mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2 \sum_{i=1}^{m} x_i^2}{m\sum_{i=1}^{m}(x_i - \bar{x})^2},\quad \hat{\sigma}^2 = \frac{1}{m-2}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

OLS Assumptions: Independent observations, constant variance (homoscedasticity), and no multicollinearity

Terms for Model Assessment and Inference (QR)

Term	Math	Intuition
RSS (Residual Sum of Squares)	$\sum_{i=1}^{m} (y_i - \hat{y}_i)^2$	Measures the amount of variability that is left unexplained after performing the regression
SE (Standard Error)	$\frac{\sigma}{\sqrt{m}} \approx \frac{\hat{\sigma}}{\sqrt{m}}$	Used in hypothesis testing on the coefficients; $t = \frac{\hat{\beta}}{SE(\hat{\beta})}$, distributed as t_{m-p-1}
RSE (Residual Standard Error)	$\sqrt{\frac{\mathrm{RSS}}{m-2}} = \sqrt{\frac{\mathrm{RSS}}{m-p-1}}$	Absolute measure of lack of fit of model to the data, generalized for p predictors
TSS (Total Sum of Squares)	$\sum_{i=1}^{m} (y_i - \bar{y})^2$	Total variance in Y
R^2	$1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$	Proportion of variance in Y explained by X
Adjusted R^2	$1 - \frac{\mathrm{RSS}/(m-p-1)}{\mathrm{TSS}/(m-1)}$	Similar to R^2, adjusted for p predictors and m examples
$\mathrm{Corr}(X,Y)$	$\frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i-\bar{y})^2}}$	Sample correlation between X, Y

Multiple Linear Regression (QR)
Our model assumes a linear relationship between Y and $X_1, X_2, \ldots, X_p$.

$$Y = \epsilon + \beta_0 + \sum_{i=1}^{n} \beta_i X_i,\ \ \mathbb{E}[\epsilon] = 0,\ \ \mathrm{Var}(\epsilon) = \sigma^2$$

We can express the above in matrix form. Denote $\mathbf{X}$ as an $m \times (p+1)$ matrix with each row an input vector with 1 in the first position corresponding to β_0. Further, let $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\intercal$, and let $\mathbf{y}$ denote the m-vector of outputs corresponding to $\mathbf{X}$. Then we have that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\mathbb{E}[\epsilon_i] = 0$, and $\mathrm{Var}(\epsilon_i) = \sigma^2$.

We can compute the unbiased estimator $\hat{\boldsymbol{\beta}}$ as follows:

$$\mathrm{RSS} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \Rightarrow \frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}^\intercal(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^\intercal\mathbf{X})^{-1}\sigma^2,\quad \hat{\sigma}^2 = \frac{1}{m-p-1}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

F Distribution for Subset Selection (QR)

1. Is a subset of $X_1, X_2, \ldots, X_p$ useful in predicting Y?
We may conduct the hypothesis test by computing the F-statistic defined below, where subscripts with 1 denote relation to the larger model.

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(m - p_1 - 1)}$$

The F statistic corresponds to the $F_{p_1 - p_0, m - p_1 - 1}$ distribution.

2. What subset of $X_1, X_2, \ldots, X_p$ helps explain Y?
The F-ratio is defined as follows, where k denotes the number of predictors currently in the model, RSS_0 and RSS_1 are the measures of error for the original and new (appropriately fitted) model, respectively.

$$F = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_0/(m - k - 2)}$$

Forward selection greedily adds predictors into the model that produce the largest value of F, whereas backward selection sequentially deletes predictors producing the smallest value of F at each stage. Stopping condition is an F-ratio $\geq$ $(100 - \alpha)$th percentile for some α, typically 5 or 10, of the $F_{1, N-k-2}$ distribution.

Coefficient Shrinkage Methods (QR)

Previously, we aimed to minimize $\mathrm{RSS} = \sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$. Regularization methods such as ridge and lasso shrink regression coefficients towards zero, which can significantly reduce coefficient estimates' variance.

Ridge Regression
Objective function to be minimized, for hyperparameter λ:

$$\sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\intercal (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^\intercal \boldsymbol{\beta}$$

$$\Rightarrow \hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}^\intercal \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\intercal \mathbf{y}$$

As $\lambda \to \infty$, the impact of the shrinkage penalty grows, and ridge regression coefficient estimates will approach zero. Note that we don't want to shrink the intercept, since it is, conceptually, a measure of the center of the response when predictors are at zero. Note that ridge regression will shrink all coefficients towards zero but will not set any of them exactly to zero, reducing model interpretability for large p. Lasso overcomes this disadvantage.

Lasso Regression
Objective function to be minimized, for hyperparameter λ, yielding sparse models:

$$\sum_{i=1}^{m} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$

Bias-Variance Tradeoff (QR)

We have a relationship $Y = f(X) + \epsilon$, where $\mathbb{E}[\epsilon] = 0, \mathrm{Var}(\epsilon) = \sigma^2$. Suppose we have learned a function $\hat{f}(\cdot)$ to approximate $f(\cdot)$ via a training set D. Then, the expected prediction error over a test set x can be decomposed.

$$\mathrm{EPE}(x) = \mathbb{E}\left[\left(f(x) - \hat{f}(x) \right)^2 \middle| X = x \right] = \sigma^2 + \left(\mathrm{Bias}\left[\hat{f}(x) \right] \right)^2 + \mathrm{Var}\left[\hat{f}(x) \right]$$

$$\mathrm{Bias}\left[\hat{f}(x) \right] := \mathbb{E}_D \left[\hat{f}(x) - f(x) \right]$$

3.2 Classification

k-Nearest Neighbors (QR)
Unlike other classification methods, k-nearest neighbors is a non-parametric method where we classify data points based on the data points closest to it. We first calculate the distance between the new observation and each training point. We then identify the k training observations that are closest. We then classify the new observation point as the most common class within the k observations.

Logistic Regression (QR)
Suppose we now want to model our Y as a binary output for given X. We can now model the probability of X belonging in a particular category. Here, we will assign X to group 1 if $\mathbb{P}(Y = 1 \mid X = x) > 0.5$ and 0 otherwise. This is done via the *logistic* function:

$$f(z) = \frac{1}{1 + e^{-z}}$$

Logistic regression is also known as a linear classification model. This is due to the fact that we use a linear regression as the input into the logistic function, which then normalizes it into a probability. In other words, $z = \beta_0 + \beta_1 x$.

To find β, we can use the method of maximum likelihood, where the probabilities for each x_i is denoted by $f(\hat{\beta}_0 + \hat{\beta}_1 x_i)$. Unlike in linear regression, there is no closed form solution and we must use iterative methods to determine $\hat{\beta}$. Multiple logistic regression is done in the same manner, but now we have to calculate additional values of β.

Generative Models (QR)
Logistic regression models $\mathbb{P}(Y = k \mid X = x)$ directly. In generative models, we can use Bayes' Theorem to obtain an alternative formula for the probability:

$$\begin{aligned}\mathbb{P}(Y = k \mid X = x) &= \frac{\mathbb{P}(X = x \mid Y = k)\mathbb{P}(Y = k)}{\mathbb{P}(X = x)} \\ &= \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} \pi_l f_l(x)}\end{aligned}$$

Generative models involve calculating these values individually, before combining them to indirectly obtain the desired probability, hence, generative models.

Discriminant Analysis (QR)
Here, we assume that the data is normal. In other words, $f_k(x) \sim \mathcal{N}(\mu_k, \sigma_k^2)$. We can obtain estimates of μ_k and σ_k^2 using standard maximum likelihood estimates. The prior can be estimated by using the proportion of training data that is in each class.

When we assuming that $\sigma_1 = \sigma_2 = \cdots = \sigma_k$, this gives us the result for linear discriminant analysis. When we ease these assumptions, and allow unique σ_k, we obtain the result for quadratic discriminant analysis. Above, we assumed $f_k(x)$ to be univariate. However, linear and quadratic discriminant analysis also apply when using multivariate Gaussians.

Naive Bayes (QR)
Naive Bayes makes a strong (yet different from discriminant analysis) assumption about the data distribution. We make thee assumption that within each class, the p predictors are independent from each other. This gives the following statement:

$$f_k(x) = f_{k1}(x_1) f_{k2}(x_2) \dots f_{kp}(x_p)$$

This greatly simplifies computation as it assumes no dependence between the independent variables. In discriminant analysis, we assumed that the data follows a normal distribution, which may have some non-diagonal covariance structure.

3.3 Tree Methods

Decision Trees (QR)
Tree methods can be used for both classification and regression. We will start with a singular decision tree. A decision trees will make a sequence of if-else statements, which eventually leads to some group in the classification scenario, or some value in the regression scenario. In general, decision trees are very easy to interpret and can handle non-quantitative predictors well. On the other side, decision trees can overfit to the data (low bias) and have high variance. Furthermore, the accuracy isn't the greatest. These problems can be remedied with bagging or boosting.

Bagging (QR)
Bagging is a method that is used to reduce the variance of a decision tree. In other words, from a single dataset, we can keep training decision trees with replacement (bootstrapping), and then average all the trees to obtain a model with lower variance (aggregation).

Due to bootstrapping, not all data points will be used when training. On average, each tree will only use about 2/3 of the data. The remaining data is referred to OOB (out-of-bag) observations, which can be used to estimate out-of-sample test error.

Random Forests (QR)
Random Forests essentially use bagging, but with a small modification. In bagging, we may have high correlation between features and thus our trees can all be highly correlated. In random forests, we decorrelate the trees by forcing the trees to consider only a small subset of the features, m, rather than all of them. When we set $m = p$, we obtain bagged trees.

Boosting (QR)
In bagging and random forests, we reduced variance by using an ensemble of decision trees. Here, we focus on reducing bias by sequentially growing decision trees. In addition, we train the model on a modified version of the data, rather than a bootstrapped dataset. More specifically, we train the model on the residuals.

First, we will train the data on the original dataset. Then, we calculate the residuals and re-train a tree on the residuals. We then add the models together and repeat. Boosting is additive, focusing on reducing the error with each sequentially grown tree.

3.4 Neural Networks

Neural Networks (QR)
Neural networks are a very powerful way of introducing non-linearity to both regression and classification problems. This is done through the usage of activation functions. Let us denote an activation function as $f(x)$. A neural network does the following:

$$y = f_n(f_{n-1}(\dots f_1(Wx + b)))$$

We call W, the weights, and b the bias, and x is the input, which can range from a scalar value to a vector. The f_n subscript denotes that the activation functions between layers don't need to be the same. Common activation functions are ReLU, softmax, or tanh. One can notice that without activation functions, the result simplifies to that of a linear regression, where W and b are the βs. Similarly, a logistic regression is one of the most basic versions of a neural network, where the activation function is the sigmoid function.

One can see that there can be an extremely large amount of parameters. These parameters are trained through gradient descent and backpropagation. Many different losses can be used − which can explain the versatility of neural networks. A downside is they require a large amount of data to be effective. They are also not necessarily the quickest and lack interpretability. All 3 statements are reasons why neural networks are not that widely used in finance, and many go towards simpler models, like regression for their needs. However, we can see the strength of neural networks, especially with the sudden popularity of large-language models.

4 Linear Algebra

4.1 Matrix Basics

Fundamental Knowledge (QT/QR)
Let A and B be square $n \times n$ matrices. Then all of the following hold:

$$\cos(\theta) = \frac{x^\intercal y}{\|x\|\|y\|} \qquad (AB)^\intercal = B^\intercal A^\intercal \qquad (AB)^{-1} = B^{-1}A^{-1} \qquad A^{-1}A = AA^{-1} = I \quad \text{rank}(A) + \text{null}(A) = n$$

$$Av = \lambda v \;\; \Rightarrow \;\; (A - \lambda I)v = 0 \;\; \Rightarrow \;\; \det(A - \lambda I) = 0 \qquad \det(A) = \frac{1}{\det(A^{-1})} \qquad \det(A) = \det(A^T)$$

$$\det(AB) = \det(A)\det(B) \qquad \det(cA) = c^n \det(A) \qquad \det(A) = \prod_{i=1}^{n} \lambda_i \qquad \text{trace}(A) = \sum_{i=1}^{n} A_{ii} = \sum_{i=1}^{n} \lambda_i$$

Nonsingular Matrices (QR)
A nonsingular matrix is invertible. A $(n \times n)$ is nonsingular if and only if any (and therefore all) of the following hold:

1. Columns of A span $\mathbb{R}^n$, or equivalently, $\text{rank}(A) = \dim(\text{range}(A)) = n$
2. $A^\intercal$ is nonsingular
3. $\det(A) \neq 0$
4. $Ax = 0$ has only the trivial solution $x = 0$; $\dim(\text{nul}(A)) = 0$

Note that if $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, then $A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$. Larger inverses may be found via Gauss-Jordan Elimination:

$$[A \mid I] \quad \overset{\text{elementary row operations}}{\Rightarrow} \quad [I \mid A^{-1}]$$

2D Rotation Matrices (QR)
2D Rotation matrices by θ radians counter-clockwise about the origin are matrices in the form $R_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$.

Orthogonal Matrices (QR)
Orthogonal matrices (unitary matrices in the reals) are square with orthonormal row and column vectors. They are nonsingular and satisfy $Q^T = Q^{-1}$. Orthogonal matrices can be intepreted as rotation matrices.

Idempotent Matrices (QR)
Idempotent matrices are square matrices satisfying $A^2 = A$. In other words, the effect of applying the linear transformation A twice is the same as applying it once. Projection matrices are Idempotent.

Positive Semi-definite Matrices (QR)
Covariance and correlation matrices are always positive semi-definite and positive definite if there is no perfect linear dependence among random variables. Each of the following conditions is a necessary and sufficient condition for A to be positive semi-definite/definite:

Positive Semi-Definite	Positive Definite
$z^\intercal Az \geq 0$ for all column vectors z	$z^\intercal Az > 0$ for all nonzero column vectors z
All eigenvalues are nonnegative	All eigenvalues are positive
All upper left/lower right submatrices have nonnegative determinants	All upper left/lower right submatrices have positive determinants

Note that if A has negative diagonal elements, then A cannot be positive semi-definite.

4.2 Matrix Decompositions

Diagonalizable Matrices (QR)
A is diagonalizable if and only if it has linearly independent eigenvectors, or equivalently, if the geometric multiplicity and the algebraic multiplicity of all the eigenvalues agree. A special case of this is if A has n distinct eigenvalues. Suppose we have eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding eigenvectors $v_1, \ldots, v_n$. Then

$$A = XDX^{-1}, \;\; X = \begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}, \;\; D = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{bmatrix}$$

Intuitively, this says that we can find a basis consisting of the eigenvectors of A. Useful for computing large powers of A, as $A^n = XD^nX^{-1}$. An important example is A being real and symmetric implies A is diagonalizable.

Singular Value Decomposition (QR)
SVD is powerful in low-rank approximations of matrices. Unlike eigenvalue decomposition, SVD uses two unique bases (left/right singular vectors). For orthogonal matrices $U(m \times m), V(n \times n)$ and diagonal matrix $\Sigma(m \times n)$ with nonnegative diagonal entries in nonincreasing order, we can write any $m \times n$ matrix A as:

$$A = U\Sigma V^{\intercal}$$

Intuitively, this says that we can express A as a diagonal matrix with suitable choices of (orthogonal) bases.

QR Decomposition (QR)
For nonsingular A, we can write $A = QR$, where Q is orthogonal and R is an upper triangular matrix with positive diagonal elements. QR decomposition assists in increasing the efficiency of solving $Ax = b$ for nonsingular A:

$$Ax = b \Rightarrow QRx = b \Rightarrow Rx = Q^{-1}b = Q^{\intercal}b$$

QR decomposition is very useful in efficiently solving large numerical systems and inversion of matrices. Furthermore, it is also used in least-squares when our data is not full rank.

LU and Cholesky Decompositions (QR)
For nonsingular A, we can write $A = LU$, where L is a lower and U is an upper triangular matrix. This decomposition assists in solving $Ax = b$ as well as computing the determinant:

$$\det(A) = \det(L)\det(U) = \prod_{i=1}^{n} L_{ii} \prod_{j=1}^{n} U_{jj}$$

If A is symmetric positive definite, then A can be expressed as $A = R^{\intercal}R$ via Cholesky decomposition, where R is an upper triangular matrix with positive diagonal entries. Cholesky decomposition is essentially LU decomposition with $L = U^{\intercal}$. These decompositions are both useful for solving large linear systems.

Projections (QR)
Fix a vector $v \in \mathbb{R}^n$. The projection of $x \in \mathbb{R}^n$ onto v is given by

$$\text{proj}_v(x) = P_v x = \frac{vv^T}{||v||^2}x = \frac{x \cdot v}{||v||^2}v$$

More generally, if $S = \text{Span}\{v_1, \ldots, v_k\} \subseteq \mathbb{R}^n$ has orthogonal basis $\{v_1, \ldots, v_k\}$, then the projection of $x \in \mathbb{R}^n$ onto S is given by

$$\text{proj}_S(x) = \sum_{i=1}^{k} \frac{x \cdot v_i}{||v_i||^2} v_i$$

The main property is that $\text{proj}_S(x) \in S$ and $x - \text{proj}_S(x)$ is orthogonal to any $s \in S$. Linear Regression can be viewed as a projection of our observed data onto the subspace formed by the span of the collected data.

5 Calculus

Differentiation (QT/QR)
At all points x where the functions and the derivatives are defined,

$$\frac{d}{dx}(x^n) = nx^{n-1} \quad \frac{d}{dx}\sin(x) = \cos(x) \quad \frac{d}{dx}\cos(x) = -\sin(x) \quad \frac{d}{dx}\tan(x) = \sec^2(x)$$

$$\frac{d}{dx}\sec(x) = \sec(x)\tan(x) \quad \frac{d}{dx}\csc(x) = -\csc(x)\cot(x) \quad \frac{d}{dx}\cot(x) = -\csc^2(x)$$

$$\frac{d}{dx}\arcsin(x) = \frac{1}{\sqrt{1-x^2}} \quad \frac{d}{dx}\arctan(x) = \frac{1}{1+x^2} \quad \frac{d}{dx}\operatorname{arcsec}(x) = \frac{1}{|x|\sqrt{1-x^2}}$$

$$\frac{d}{dx}(e^x) = e^x \quad \frac{d}{dx}(f(x) \pm g(x)) = f'(x) \pm g'(x) \quad \frac{d}{dx}(f(x)g(x)) = f'(x)g(x) + g'(x)f(x)$$

$$\frac{d}{dx}(\ln(x)) = \frac{1}{x} \quad \frac{d}{dx}f(g(x)) = f'(g(x))g'(x) \quad \frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2}$$

$$\frac{d}{dx}\left(f(x)^{g(x)}\right) = f(x)^{g(x)}\left[g'(x)\ln(f(x)) + g(x)\cdot\frac{f'(x)}{f(x)}\right] \quad \frac{d}{dx}(x^x) = x^x(\ln(x)+1)$$

Integration (QT/QR)
Disregarding the $+C$ on all the integrals,

$$\int x^n dx = \frac{x^{n+1}}{n+1},\ n \neq -1 \quad \int \sin(x)dx = -\cos(x) \quad \int \cos(x)dx = \sin(x) \quad \int \tan(x)dx = -\ln|\cos(x)|$$

$$\int \sec(x)dx = \ln|\sec(x) + \tan(x)| \quad \int \csc(x)dx = \ln|\csc(x) - \cot(x)| \quad \int \cot(x)dx = \ln|\sin(x)|$$

$$\int \frac{1}{\sqrt{1-x^2}}dx = \arcsin(x) \quad \int \frac{1}{1+x^2}dx = \arctan(x) \quad \int \frac{1}{|x|\sqrt{1-x^2}}dx = \operatorname{arcsec}(x)$$

$$\int e^x dx = e^x \quad \int \frac{1}{x}dx = \ln|x| \quad \int (f(x) \pm g(x))dx = \int f(x)dx \pm \int g(x)dx$$

$$\int u(x)v'(x)dx = u(x)v(x) - \int v(x)u'(x)dx \quad \int f'(g(x))g'(x)dx = f(g(x))$$

Taylor Series (QT/QR)
Select some point $x = x_0$. If $x_0 = 0$, we have the Maclaurin series. Generally, $f(x) = \displaystyle\sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$.

Common Maclaurin series expansions:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots$$

$$\sin(x) = \sum_{n=0} \frac{(-1)^n x^{2n+1}}{(2n+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

$$\cos(x) = \sum_{n=0} \frac{(-1)^n x^{2n}}{(2n)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots$$

Common Summation Formulae (QT/QR)

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \quad \sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \quad \sum_{k=s}^{\infty} a \cdot r^k = a \cdot \frac{r^s}{1-r} \quad \sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$$

6 Finance

6.1 Options Theory

Bonds, Underlying, Forward Contracts (QT)

We can form many assets in options theory with the usage of two assets: the underlying, and the bond. The underlying is simple. This can be a stock, a future, or any non-derivative; we will denote this as S. The bond represents the risk-free rate, which is typically represented mathematically as some constant. We will denote this as B. For simplicity, we will assume the final payout of a bond at T is always \$1. However, we need to discount the value of this bond. This is known as a discount factor, which can be written as e^{-rT}, where r corresponds to the risk-free rate and T represents the time until expiry.

Using these two assets, we can form an asset that has a linear payout. This is known as a forward contract. This involves going long 1 unit of the underlying and shorting K units of the bond. The forward as a final payoff of $V_T = S_T - K$. This is the most basic example of a derivative. In general, this K is known as the strike price. Here, we have a positive payout when $S_T > K$ and a negative payout when $S_T < K$.

Vanilla Options (QT)

The most simple of options are calls and puts. A call represents the right (but not obligation) to purchase the underlying at strike K. Similarly, a put represents the right (but not obligation) to sell the underlying at strike K. Here, we see that the forward contract represents the obligation to purchase the underlying at strike K while a call option does not. Mathematically, a call option has time-T payoff $\max(S_T - K, 0)$ and a put option has time-T payoff $\max(K - S_T, 0)$.

An important concept in options pricing is the pricing of derivatives. This can be done through the method of replication. If two assets have the same final payoff, they should also have the same intermediate payoffs. Otherwise, there would be arbitrage. One can immediately see that we can replicate a forward contract with a put option and a call option. In other words, the final payoff of a forward contract is the same as 1 unit of a call and -1 unit of a put. This is known as put-call parity. We can mathematically write this as $C - P = S - K$.

Black-Scholes (QT/QR)

Black-Scholes is most important concept in all of options theory. It gives the basis to pricing mathematical options through a differential equation as well as the concepts of greeks that are commonly used in the industry. The Black-Scholes differential equation is as follows:

$$\frac{1}{2}\sigma^2 S^2 \frac{\partial^2 C}{\partial S^2} + rS\frac{\partial C}{\partial S} + \frac{\partial C}{\partial t} = rC$$

When we consider European options (exercise only occurs at expiration), a constant risk free rate, non-dividend paying stocks, and that stocks follows a Geometric Brownian Motion, we can use Black-Scholes. In practice, this is usually not the case. This model still gives a solid framework for modern options pricing, such as certain numerical methods used in practice. Here, σ^2 represents the constant volatility of the GBM, S is the price of the underlying, C is the option, and r is the constant risk-free rate.

Solving the differential equation gives a closed form for the price of an option, as a function of t, S, r, and σ. Partial derivatives can be taken to obtain the sensitivity of the option to these values.

$$\begin{aligned} \text{Delta} &: \frac{\partial C}{\partial S} \\ \text{Gamma} &: \frac{\partial^2 C}{\partial S^2} \\ \text{Theta} &: \frac{\partial C}{\partial t} \\ \text{Vega} &: \frac{\partial C}{\partial \sigma} \\ \text{Rho} &: \frac{\partial C}{\partial r} \end{aligned}$$

One of the biggest empirical downsides of Black-Scholes is the existence of a non-constant volatility. In real life, there exists a volatility skew. In other words, the volatility is not constant across different strikes. This can lead to huge inaccuracies when pricing with Black-Scholes.

6.2 Portfolio Theory

Two Asset Portfolio (QT/QR)

The theme of portfolio optimization is that diversification improves portfolio performance. Consider a two-asset portfolio, denoted asset 1 with mean μ_1 and variance σ_1^2 and asset 2 with mean μ_2 and variance σ_2^2, with correlation ρ. We weight these assets with weights $w_1 = w$ and $w_2 = 1 - w$. We have the following returns of the portfolio.

$$\mu = w\mu_1 + (1-w)\mu_2$$
$$\sigma^2 = w^2\sigma_1^2 + 2w(1-w)\rho\sigma_1\sigma_2 + (1-w)^2\sigma_2^2$$

If the assets are perfectly correlated $\rho = 1$, we have $\sigma = w\sigma_1 + (1-w)\sigma_2$. In other words, the volatility scales with the returns linearly. If we have $\rho < 1$, then we have $\sigma < w\sigma_1 + (1-w)\sigma_2$. In other words, the volatility of the portfolio decreases while the expected returns stays the same. We see that this only requires $\rho < 1$ for the benefits of diversification to appear.

In fact, if $\rho = -1$, then we can obtain a risk-free portfolio. We have the following weight: $w = \frac{\sigma_1}{\sigma_1+\sigma_2}$.

We can generalize this to larger portfolios. The portfolio variance can be written as $\sigma = \frac{1}{n}\mathbb{E}(\sigma_i^2) + \frac{n-1}{n}\mathbb{E}(\sigma_{i,j})$. As we take $n \to \infty$, the individual variances of the securities do not matter: only the covariance between the securities. Once the portfolio is large enough, the only influence of portfolio variance is the average covariance structure between all the assets. Idiosyncratic risk is the variance that can be diversified away by increasing the number of assets while systematic risk is the risk inherent to the market – it cannot be diversified.

Portfolio Optimization (QT/QR)

Given a large number of assets, we can determine the optimal portfolio via mean-variance optimization. We can imagine the portfolio existing in a subspace of mean and variance vectors. We denote w as a vector of weights for the assets and μ as vector of average returns. We solve the following optimization problem. We will minimize $w'\Sigma w$ such that $w'\mu = \mu_p$ and $w'1 = 1$. In other words, we want to minimize the variance, such that we obtain a specific return and such that we have an overall portfolio weight of 1. Solving this optimization gives the optimal portfolio, one that gives the lowest variance for a desired level of returns. As one increases the level of desired returns, the portfolio variance will also increase. Hence the term, ”high risk, high return”.

There are various ways to measure portfolio performance. Assuming investors only care about portfolio returns and variance, we can denote the Sharpe Ratio as μ_p/σ_p. This represents the level of return for a unit of variance. In general, investors are risk averse and don’t just care about portfolio return and variance. For example, investors may care more about negative variance – volatility that makes them lose money. A Sortino ratio is also a useful metric, μ_p/σ_d, where σ_d is the volatility only on ”bad” days. Another similar metric is known as the Calmar ratio, which normalizes the level of return to the maximum drawdown. In other words, μ_p/Maximum Drawdown. Many portfolio managers have different metrics, depending on the goals of themselves and their clients. Some may have the strategy of collecting fees from clients, and are willing to take more volatility, caring only about Sharpe ratio. Others may care about their reputation, and instead give great consideration to their Calmar ratio, in order to not lose clients.

Pricing Models (QT/QR)

The Capital Asset Pricing Model (CAPM) is a linear factor model that explains the returns of any asset through that of a market portfolio.

$$\mathbb{E}(r_i) = \beta\ \mathbb{E}(r_m)$$

The premium of a given asset is a scaled version of the total market. This is what typically is referred to as β in the market. Hedge funds and portfolio managers do not want β, as this means that their returns are correlated to the market. Note that the above equation does not have an intercept term, of α. This is what many portfolio managers seek: returns that are not correlated to the market. In practice, regression tests show that CAPM does not hold, though it is a good theoretical starting point. We would expect there to be no intercept term. Empirical evidence shows otherwise.

The CAPM seeks to explain returns through the market. However, there may be other factors that can explain market returns better. The Fama-French three factor model expands on the market factor by adding a size and value factor. The size factor is designed by going long small stocks and shorting large stocks while the value factor longs value stocks and shorts growth stocks. Nowadays, there are many factor models, with factors like momentum, profitability, and investment. Many times, portfolio managers compare their returns to these factors to see if they have exposure. A portfolio manager may believe that they have found α, but in reality, they have created a portfolio that is highly correlated to one of the Fama-French factors.