Quantitative finance

Interviews preparation

Vivien Tisserand

Abstract

This is a summary of interview questions found on the internet, books, *etc.*, along with more in-depth digressions related to quantitative finance. It is a mixed of applied mathematics and computer science.
I am not found of brainteasers, they are a poor way to assess for a candidate's ability to be an asset for the team. This work smoothly transitioned to a sort of *vademecum* in applied mathematics: through several questions, it goes through different techniques that are easy to forget with time. I myself refer to it quite often when I forget about how to write the Lagrangian in a constrained optimization problem, or the general solution of a second-order differential equation...

Contents

Chapter 1

Probability

1.1 Correlated bivariate distribution

Let (X, Y) follow a bivariate normal standard distribution with correlation ρ. Find the expectation:

$$\mathbb{E}[\operatorname{sgn}(X)\operatorname{sgn}(Y)].$$

We are interested in the joint distribution of X and Y:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right).$$

To see what happens here, we can compare the density contour of this distribution with the independent case. The covariance matrix is symmetric thus diagonalizable. We can find its eigenvalues and its eigenvectors (through classic computations or noticing this is a circulant matrix). With $P = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$,

$$\Sigma = P \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix} P^{-1}.$$

This gives us the shape of the correlated distribution. Qualitatively, we can say that, as ρ defines how rotated and squished the distribution is, the bigger ρ, the higher the probabiliy of X and Y being the same sign.

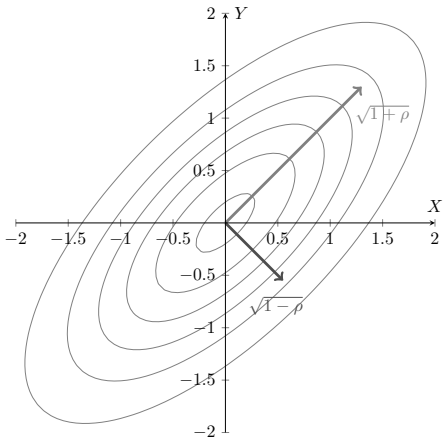

Figure 1.1: Density contours of a bivariate normal law, with $\rho = 0.7$

Back to our problem: the random variable $\operatorname{sgn}(X)\operatorname{sgn}(Y)$ takes values in the set $\{-1, 1\}$. Thus, to get its expectancy, we can compute these discrete probabilities:

$$\begin{aligned}\mathbb{E}[\operatorname{sgn}(X)\operatorname{sgn}(Y)] &= 1 \times \mathbb{P}(\operatorname{sgn}(X)\operatorname{sgn}(Y) = 1) - 1 \times \mathbb{P}(\operatorname{sgn}(X)\operatorname{sgn}(Y) = -1) \\ &= 1 \times \mathbb{P}(\operatorname{sgn}(X)\operatorname{sgn}(Y) = 1) - 1 \times (1 - \mathbb{P}(\operatorname{sgn}(X)\operatorname{sgn}(Y) = 1)) \\ &= 2\mathbb{P}(\operatorname{sgn}(X)\operatorname{sgn}(Y) = 1) - 1.\end{aligned}$$

Using the symmetry of the distribution,

$$\mathbb{P}(\operatorname{sgn}(X)\operatorname{sgn}(Y) = 1) = \mathbb{P}(X > 0, Y > 0) + \mathbb{P}(X < 0, Y < 0) = 2\mathbb{P}(X > 0, Y > 0),$$

thus the only thing we need to compute is $\mathbb{P}(X > 0, Y > 0)$.

If

$$\begin{pmatrix} U \\ V \end{pmatrix} = \Sigma^{-1/2} \begin{pmatrix} X \\ Y \end{pmatrix},$$

then (U, V) follows an independent bivariate normal standard distribution. Inverting Σ we get:

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

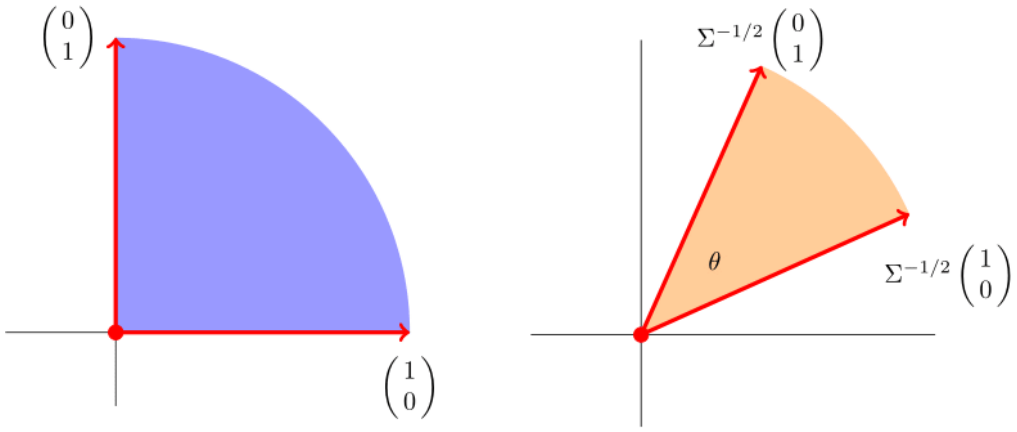

Figure 1.2: Area of the event $X > 0, Y > 0$ for $\rho = 0$ (left) and $\rho \neq 0$ (right). From here.

Then, there exists a $\theta \in [0, 2\pi]$ such that $\mathbb{P}(X > 0, Y > 0) = \dfrac{\theta}{2\pi}$. This θ verifies

$$\cos\theta = \frac{\langle u, v \rangle}{\|u\|\|v\|}.$$

with $u = \Sigma^{-1/2} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $v = \Sigma^{-1/2} \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

$$\begin{aligned}\langle u, v \rangle &= (1\ 0)\,\Sigma^{-1}(0\ 1)^T = -\rho/(1-\rho^2) \\ \|u\|^2 &= (1\ 0)\,\Sigma^{-1}(1\ 0)^T = 1/(1-\rho^2) \\ \|v\|^2 &= (0\ 1)\,\Sigma^{-1}(0\ 1)^T = 1/(1-\rho^2)\end{aligned}$$

so that $\cos(\theta) = -\rho$. Putting it all together gives

$$\mathbb{P}(X > 0, Y > 0) = \frac{\arccos(-\rho)}{2\pi}.$$

Finally,

$$\boxed{\mathbb{E}[\operatorname{sgn}(X)\operatorname{sgn}(Y)] = \frac{2\arccos(-\rho)}{\pi} - 1.}$$

Note that if $\rho = 0$, we have $\mathbb{E}[\operatorname{sgn}(X)\operatorname{sgn}(Y)] = 0$; it converges to 1 as $\rho \longrightarrow 1$ and to -1 as $\rho \longrightarrow -1$, which gives us confidence in our answer.

1.2 The coupons collector

> A chocolate company launches a marketing campaign: for each chocolate bar you buy, you get one collectible card out of a set of n possible cards. We can assume the card are uniformly distributed among the chocolate bars.
> How many chocolate bars should you buy to complete the collection?

Well, at least n, even if we are very lucky.
The first bar we open will yield to a new card. For the second bar, we have a probability $\frac{1}{n}$ to get the same card we already have, thus $\frac{n-1}{n}$ to get a new card. This follows a geometric law: the expectation for such an event is $\frac{n}{n-1}$. And so on, decreasing the probability for each new card we acquire.
The total expectancy will be the sum of all of these individual processes:

$$\mathbb{E}[N] = \sum_{k=0}^{n-1} \frac{n}{n-k},$$

with N the random variable that counts the number of chocolate bars eaten to get the full collection.

We realize that we are actually dealing with the harmonic sum $H_n = \sum_{k=1}^{n} \frac{1}{k}$, which can be squeezed between two integrals to get the equivalent: $H_n \sim_{n\to+\infty} \log(n)$.
Thus $N \sim_{n\to+\infty} n\log(n)$.

To give a confidence interval around the number of chocolate bars we should buy, let's pull up some concentration inequalities.
We still deal with the sum of independent geometric variables so the variance is easy to compute:

$$\begin{aligned}
\operatorname{Var}[N] &= \sum_{k=1}^{n} \operatorname{Var}[N_i] \\
&= \sum_{k=1}^{n} \left(1 - \frac{n-k+1}{n}\right) \left(\frac{n}{n-k+1}\right)^2 \\
&= n \sum_{k=1}^{n} \frac{k-1}{(n-k+1)^2} \sim_{n\to+\infty} n\frac{\pi^2}{6}.
\end{aligned}$$

Applying Chebychev's inequality, we get:

$$\mathbb{P}(|\mathbb{E}[N] - N| \geq k\sigma) \leq \frac{1}{k^2}.$$

Some other inequalities could be used to raffinate this result: Chernoff bounds, Vysochanskij–Petunin inequality, etc.

Chapter 2

Statistics

2.1 Estimating the support of an uniform law

Suppose that we have $x_1, \dots, x_n$ observations from an uniform law $X \sim \mathcal{U}[0, \theta]$, where θ is an unknown parameter that we want to estimate. Give at least two estimators for θ and compare them.

Method of moments: Having a look at the first order moment, it appears that $\mathbb{E}[X] = \theta/2$. Taking the empirical counter-party of this theoretical quantity, we have $\boxed{\hat{\theta}^{\text{MM}} = \frac{2}{n}\sum_{i=1}^{n} x_i.}$
By applying the strong law of large numbers and the continuous mapping theorem, $\hat{\theta}^{\text{MM}} \xrightarrow{a.s.} \theta$. Thus $\boxed{\text{this estimator is consistent}}$.
We want asymptotic results on the convergence of this estimator. Before using the CLT, we have to check for the existence of a second-order moment.

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_{\mathbb{R}} x^2 f(x)\,\mathrm{d}x \\
&= \int_0^{\theta} x^2 \frac{1}{\theta}\,\mathrm{d}x \\
&= \left[\frac{1}{3\theta}x^3\right]_0^{\theta} \\
&= \frac{\theta^2}{3} < +\infty
\end{aligned}$$

Thus, we have $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \dfrac{\theta^2}{12}$. So, $\mathbb{V}[2X_1] = \dfrac{\theta^2}{3}$.

By applying the central limit theorem, we have :

$$\sqrt{n}(\hat{\theta}^{\text{MM}} - \theta) \xrightarrow{(d)} \mathcal{N}\left(0, \frac{\theta^2}{3}\right).$$

We have to evaluate the risk of this estimator, that we write as the sum of the squared bias and the variance :

$$\text{MSE}(\hat{\theta}^{\text{MM}}) = \mathbb{E}[(\hat{\theta}^{\text{MM}} - \theta)^2] = \mathbb{E}[(\hat{\theta}^{\text{MM}} - \mathbb{E}[\hat{\theta}^{\text{MM}}])^2] + \mathbb{E}[\hat{\theta}^{\text{MM}} - \theta]^2 = \mathbb{V}[\hat{\theta}^{\text{MM}}] + (\mathbb{E}[\hat{\theta}^{\text{MM}}] - \theta)^2.$$

We have $\mathbb{E}[\hat{\theta}^{\mathrm{MM}}] = 0$ and $\mathbb{V}[\hat{\theta}^{\mathrm{MM}}] = \frac{1}{n^2} n \mathbb{V}[2X_1] = \frac{\theta^2}{3n}$.

Thus,

$$\mathrm{MSE}(\hat{\theta}^{\mathrm{MM}}) = \frac{\theta^2}{3n}.$$

Maximum likelihood: Let's write the likelihood of this model:

$$\begin{aligned} L((X_1, \dots, X_n), \theta) &= \prod_{i=1}^{n} f_X(X_i) \\ &= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(X_i) \\ &= \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(X_i). \end{aligned}$$

And this function is maximized by choosing the smallest θ such that all of the X_i lie in $[0, \theta]$, that is $\hat{\theta}^{\mathrm{MLE}} = \max_{1 \leq i \leq n} X_i$.

To check the consistency of this estimator, we will have a look at its convergence (in probability). Let $\theta \in \Theta$ and $\varepsilon > 0$:

$$\begin{aligned} \mathbb{P}_\theta(|\hat{\theta}^{\mathrm{MLE}} - \theta| \geq \varepsilon) &= \mathbb{P}_\theta(\hat{\theta}^{\mathrm{MLE}} \geq \theta + \varepsilon) + \mathbb{P}_\theta(\hat{\theta}^{\mathrm{MLE}} \leq \theta - \varepsilon) \\ &= 0 + \mathbb{P}_\theta(\max_{1 \leq i \leq n} X_i \leq \theta - \varepsilon) \\ &= \prod_{i=1}^{n} \mathbb{P}_\theta(X_i \leq \theta - \varepsilon) \\ &= \left(1 - \frac{\varepsilon}{\theta}\right)^n \underset{n \to +\infty}{\longrightarrow} 0. \end{aligned}$$

Thus, $\hat{\theta}^{\mathrm{MLE}} \xrightarrow{\mathbb{P}} \theta$: this estimator is consistent.

In order to estimate the risk of this estimator, we have to look at the law that the maximum of n independent uniform laws follows. This is done by looking at the cumulative distribution function. Let $x \in [0, \theta]$:

$$\begin{aligned} \mathbb{P}_\theta(X_{(n)} \leq x) &= \mathbb{P}_\theta\left(\bigcap_{i=1}^{n} X_i \leq x\right) \\ &= \prod_{i=1}^{n} \mathbb{P}_\theta(X_i \leq x) \\ &= \left(\frac{x}{\theta}\right)^n. \end{aligned}$$

Thus,

$$F_{X_{(n)}} = \begin{cases} 0 & \text{if } x < 0 \\ \left(\frac{x}{\theta}\right)^n & \text{if } 0 \leq x \leq \theta \\ 1 & \text{if } x > \theta \end{cases}$$

This cdf as smooth as we need to take its derivative: that will be the density we were looking for:

$$f_{X_{(n)}}(x) = n\frac{x^{n-1}}{\theta^n}\mathbb{1}_{[0,\theta]}(x)$$

Let's compute the bias and the variance.

$$\mathbb{E}[\hat{\theta}^{\text{MLE}}] = \int_{\mathbb{R}} x f_{X_{(n)}}(x)\,\mathrm{d}x = \int_0^\theta \frac{n}{\theta^n}x^n\,\mathrm{d}x = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n}{n+1}\theta.$$

Then, the bias is : $B(\hat{\theta}^{\text{MLE}}) = \dfrac{n}{n+1}\theta - \theta = -\dfrac{1}{n+1}\theta \neq 0$. We can introduce a corrected estimator that we will consider too : $\hat{\theta}^{\text{MLE}}_{\text{corr}} = \dfrac{n+1}{n}\hat{\theta}^{\text{MLE}}$, such that $\mathbb{E}[\hat{\theta}^{\text{MLE}}_{\text{corr}}] = \theta$: an unbiased estimator.

Then, we have

$$\mathbb{E}[(\hat{\theta}^{\text{MLE}})^2] = \int_{\mathbb{R}} x^2 f_{X_{(n)}}(x)\,\mathrm{d}x = \int_0^\theta \frac{n}{\theta^n}x^{n+1}\,\mathrm{d}x = \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\theta = \frac{n}{n+2}\theta^2.$$

And

$$\text{MSE}(\hat{\theta}^{\text{MLE}}) = \mathbb{E}[(\hat{\theta}^{\text{MLE}} - \theta)^2] = \mathbb{E}[(\hat{\theta}^{\text{MLE}})^2] - 2\theta\mathbb{E}[\hat{\theta}^{\text{MLE}}] + \theta^2$$

Thus,

$$\text{MSE}(\hat{\theta}^{\text{MLE}}) = \frac{n}{n+2}\theta^2 - 2\frac{n}{n+1}\theta^2 + \theta^2 = \frac{2\theta^2}{(n+1)(n+2)}.$$

And

$$\text{MSE}(\hat{\theta}^{\text{MLE}}_{\text{corr}}) = \left(\frac{n+1}{n}\right)^2\mathbb{E}[(\hat{\theta}^{\text{MLE}})^2] - 2\frac{n+1}{n}\theta\mathbb{E}[(\hat{\theta}^{\text{MLE}} + \theta^2 = \frac{\theta^2}{n(n+1)}.$$

Maximum a posteriori: We write the likelihood of the model in terms of θ :

$$L((X_1,\ldots,X_n),\theta) = \frac{1}{\theta^n}\prod_{i=1}^n \mathbb{1}_{[0,\theta]}(X_i) = \frac{1}{\theta^n}\mathbb{1}_{[X_{(n)},=\infty[}(\theta).$$

Remark : the set such that $L(.,\theta) > 0$ is $[0,\theta]$: it depends on θ, thus the model is not regular. Keep that in mind when dealing with Fisher information for instance.

1. **Flat prior:**

   We apply the definition for a Bayesian estimator with a prior density π_0 :

$$\begin{aligned}\hat{\theta}^{\mathrm{B}} &= \frac{\displaystyle\int_{\Theta} \theta L(x,\theta)\pi_0(\theta)\,\mathrm{d}\lambda(\theta)}{\displaystyle\int_{\Theta} L(x,\theta)\pi_0(\theta)\,\mathrm{d}\lambda(\theta)} \\ &= \frac{\displaystyle\int_{X_{(n)}}^{+\infty} \theta^{-n+1}\,\mathrm{d}\theta}{\displaystyle\int_{X_{(n)}}^{+\infty} \theta^{-n}\,\mathrm{d}\theta} \\ &= \frac{n-1}{n-2} X_{(n)}.\end{aligned}$$

Bias and MSE are not computed there for sanity reasons.

2. **Jeffreys prior:**

   The density function of this prior is proportional to the squareroot of the determinant of the Fisher information matrix.

   Thus we need to compute this quantity for this model, with n observations (as it is not regular, $I_n \neq nI_1$) :

$$I_n(\theta) = \mathbb{E}\left[\frac{\partial \log L_n(\theta)}{\partial\theta}^2\right]$$

   We have $I_n(\theta) = \mathbb{E}[(-n/\theta)^2] = \dfrac{n^2}{\theta^2}$

   (If we had taken the expectancy of the second-order derivative of the log-likelihood, we would not have had the same result has the model is not regular.)

   This gives us the noninformative prior (Jeffreys) : $\pi_0(\theta) \propto \theta^{-1}$.

2.2 Building a statistical test

Let's assume you have a batch of a hundred observations (numbers). There are two hypotheses and one is true :

- H_0: these observations are independent draws from a Gaussian $\mathcal{N}(0, 1/18)$,
- H_1: each observation has been obtained by averaging 6 uniforms $\mathcal{U}([-1,1])$ random variables.

How would you find out which scenario is true.

Taken from @adad8m on Twitter.

Chapter 3

Machine learning

3.1 Linear regression

We are interested in the basic linear regression model where given observations $(x_i, y_i)_{1 \leq i \leq n}$ we want to build the model

$$Y = \alpha + \beta X + \varepsilon.$$

Explain the underlying assumptions in the model and derive estimators for the coefficients.

The underlying assumptions of the linear model are the following:

- No perfect multicollinearity between the explanatory variables, otherwise the parameter β is not identifiable. This is the **full-rank** assumption.
- Independence of errors. Generalized least squares can handle correlated errors.
- **Homoscedasticity**, or constant variance, which can be tested on the residuals. If there is heteroscedasticity, the Gauss-Markov theorem doesn't apply, thus the estimators derived are not the Best Linear Unbiased Estimators (BLUE). It can be corrected thanks to a weighted least squares approach, or a logarithmization of the data.
- **Exogeneity**.

We have to minimize the Euclidean distance between the predicted values by the model for Y, $\hat{Y}$ and the actual values. This can be written as :

$$(\hat{\alpha}, \hat{\beta}) \in \text{argmin}_{(\alpha,\beta)\in\mathbb{R}^2} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

Replacing with the model, we have to find parameters that minimize $f(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$. We will write the first order conditions and check by computing the Hessian that we are actually looking at a minimum.

$$\begin{cases} \dfrac{\partial f}{\partial \alpha}(\alpha, \beta) & = 0 \\ \dfrac{\partial f}{\partial \beta}(\alpha, \beta) & = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum_i (y_i - \alpha - \beta x_i) & = 0 \\ -2\sum_i x_i (y_i - \alpha - \beta x_i) & = 0 \end{cases}$$

The first line give $\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$ while the second line can be written as the following :

$$\begin{aligned}\sum_i x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) &= \sum_i y_i x_i - (\overline{y} - \hat{\beta}\overline{x})x_i - \hat{\beta}x_i^2 \\ &= \sum_i y_i - \overline{y} + \hat{\beta}(\overline{x} - x_i).\end{aligned}$$

Thus

$$\begin{aligned}&\sum_i x_i(y_i - \hat{\alpha} - \hat{\beta}x_i) = 0 \qquad \Leftrightarrow \sum_i y_i - \overline{y} + \hat{\beta}(\overline{x} - x_i) = 0 \\ &\Leftrightarrow \hat{\beta} = \frac{\sum_i y_i - \overline{y}}{\sum_i x_i - \overline{x}} \\ &\Leftrightarrow \hat{\beta} = \frac{\sum_i (y_i - \overline{y})(x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} \\ &\Leftrightarrow \boxed{\hat{\beta} = \frac{\widehat{\mathrm{Cov}}(X, Y)}{\widehat{\mathrm{Var}}(X)}.}\end{aligned}$$

3.2 LASSO estimator

We consider a penalized regression, with the ℓ_1 norm. The minimization problem now is:

$$\min_{\beta \in \mathbb{R}^p} \quad \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

where λ is an hyperparameter.

This is called the LASSO regression. Provide an analytical expression of the solution. For simplicity of the computation, we will assume that X is orthonormal (that means $X^T X = I_p$).

We first expand the objective function:

$$\begin{aligned}\|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 &= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta + \lambda \sum_{i=1}^p |\beta_i| \\ &= Y^T Y + \sum_{i=1}^p -2\hat{\beta}_i^{\mathrm{OLS}}\beta_i + \beta_i^2 + \lambda|\beta_i|\end{aligned}$$

For the optimization we can get rid of the first term that doesn't depend on β, and then, as the objective function is separable,

$$\min_{\beta \in \mathbb{R}^p} \quad \sum_{i=1}^p -2\hat{\beta}_i^{\mathrm{OLS}}\beta_i + \beta_i^2 + \lambda|\beta_i| = \sum_{i=1}^p \min_{\beta_i \in \mathbb{R}} -2\hat{\beta}_i^{\mathrm{OLS}}\beta_i + \beta_i^2 + \lambda|\beta_i|$$

we can just minimize each $f_i \colon x \mapsto -2\hat{\beta}_i^{\mathrm{OLS}}\beta_i + \beta_i^2 + \lambda|\beta_i|$. The first order condition is enough to find a minimum (we have a polynomial in β with a positive quadratic coefficient) and

$$f_i'(x) = \begin{cases} 2x - (2\hat{\beta}_i - \lambda) & \text{if } x < 0 \\ 2x - (2\hat{\beta}_i + \lambda) & \text{if } x > 0 \end{cases}$$

We then have to break the analysis in three cases : $\hat{\beta}_i < -\lambda/2$, $\hat{\beta}_i \in [-\lambda/2, \lambda/2]$ and $\hat{\beta}_i > -\lambda/2$. This gives us the maxima, respectively $x = \hat{\beta}_i + \lambda/2$, $x = 0$ and $x = \hat{\beta}_i - \lambda/2$. Overall, we find the following closed-form expression:

$$\hat{\beta}_i^{\text{LASSO}} = \text{sign}(\hat{\beta}_i)(|\hat{\beta}_i| - \lambda/2)_+$$

LASSO regression is often used as a **feature selection** tool: with the ℓ_1 penalization term, some of the smaller β_i's are set to 0. Adjusting the hyperparameter λ is a way to select more or less features.

There exist others penalized regression :

- *Ridge regression*, with a penalization on the ℓ_2 norm.
- *Elastic net*, that combines both penalizations.

3.3 Bayes classifier

Let's put ourselves under a **binary classification** setup: we have a distribution – or dataset – $(X, Y) \in \mathbb{R} \times \{0, 1\}$ of features / target and want to build a classifier h.
In particular, we are looking for a classifier that minimizes the misclassification error $\mathbb{P}(h(X) \neq Y|X)$. This optimal classifier is called **Bayes predictor**, derive its expression.

Let's compute the risk for a classifier h:

$$\begin{aligned}
\mathbb{P}(h(X) \neq Y|X) &= 1 - \mathbb{P}(h(X) = Y|X) \\
&= 1 - \mathbb{E}\left[\mathbb{1}\{h(X) = Y\}|X\right] \\
&= 1 - \mathbb{E}\left[\mathbb{1}\{h(X) = 0, Y = 0\}|X\right] - \mathbb{E}\left[\mathbb{1}\{h(X) = 1, Y = 1\}|X\right] \\
&= 1 - \mathbb{E}\left[\mathbb{1}\{h(X) = 0\}\mathbb{1}\{Y = 0\}|X\right] - \mathbb{E}\left[\mathbb{1}\{h(X) = 1\}\mathbb{1}\{Y = 1\}|X\right] \\
&= 1 - \mathbb{1}\{h(X) = 0\}\mathbb{E}\left[\mathbb{1}\{Y = 0\}|X\right] - \mathbb{1}\{h(X) = 1\}\mathbb{E}\left[\mathbb{1}\{Y = 0\}|X\right] \\
&\text{as } h(X) \text{ depends only on } X \\
&= 1 - \mathbb{1}\{h(X) = 0\}(1 - \eta(X)) - \mathbb{1}\{h(X) = 1\}\eta(X) \\
&\text{with } \eta(X) = \mathbb{P}(Y = 1|X) \\
&= \eta(X) - (2\eta(X) - 1)\mathbb{1}\{h(X) = 1\}.
\end{aligned}$$

Consider g^* the classifier that minimizes the theoretical risk. By definition, $\mathbb{P}(h(X) \neq Y|X) - \mathbb{P}(g^*(X) \neq Y|X) \geq 0$, thus $(2\eta(X) - 1)\mathbb{1}\{g^*(X) = 1\} - \mathbb{1}\{h(X) = 1\} \geq 0$.

Separating cases:

- If $g^*(X) = 1$, the previous equation yields $\eta(X) \geq 1/2$,
- While If $g^*(X) = 0$, twe must have $\eta(X) < 1/2$.

Indeed, the optimal Bayes predictor is $\boxed{g^*(X) = \mathbb{1}\{\eta(X) \geq 1/2\},}$ where $\eta(X) = \mathbb{P}(Y = 1|X)$.

In the case where we have k classes, the Bayes classifier is:

$$h(X) = k, \quad k \in \arg\max_k \mathbb{P}(Y = k|X = x).$$

Chapter 4

Stochastic calculus

4.1 Recurrence of a diffusion process

Considers a Brownian diffusion process $(X_t)_{t\in[0,T]}$ solution of the following stochastic differential equation (SDE):

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t, \qquad X_0 = x.$$

where b, σ are continuous functions and $(W_t)_{t\in[0,T]}$ denotes a Brownian motion defined on a filtered probability space.

Under which conditions is this diffusion recurrent? Positive recurrent?

Application: discuss the recurrence of

1. the Brownian motion,
2. the Ornstein-Uhlenbeck process: $dX_t = -\mu(X_t - m)dt + \sigma dW_t$.

Recall that a process is called **recurrent** when leaving from any state it almost surely comes back to any other state of the diffusion ($\mathbb{P}_y(T_x < +\infty) = 1$ where T_x is the first return time $\inf\{t \geq 0 : X_t = x\}$ and $\mathbb{P}_y(\cdot)$ states that we are leaving from point y), **recurrent positive** when it does so in a finite time ($\mathbb{E}_y[T_x] < \infty$).

The infinitesimal operator of this diffusion is written:

$$\mathcal{A}f(x) = b(x)\frac{\partial f}{\partial x} + \frac{1}{2}\sigma^2(x)\frac{\partial^2 f}{\partial x^2}.$$

With an arbitrary $x_0 \in \mathbb{R}$, the scale function[1] $S(x) := \displaystyle\int_{x_0}^{x} \exp\left\{-2\int_{x_0}^{y} \frac{b(z)}{\sigma^2(z)}\,dz\right\} dy$ is a solution of $\mathcal{A}f(x) = 0$,[2] thus provides a recurrence criterion: the process X is recurrent if and only if $S(+\infty) = +\infty$ and $S(-\infty) = -\infty$.

[1] the scale function is one of the characteristics associated to a diffusion process, along with the speed measure and killed measure.

[2] this makes $(S(X_t))_{t\geq 0}$ a martingale by the way.

Chapter 5

Financial mathematics

5.1 Dupire formula for local volatility

Explain the motivation behind local volatility and prove Dupire formula.

The Black-Scholes model is really convenient as it has few parameters, yields closed-form vanilla prices and is widely known. However, the assumption of constant volatility doesn't match the observed market prices.
Indeed, looking at the implied volatility surface we observe smile / skew / smirk accross all asset classes. In equities, the volatility smile appeared after the 1987 crisis and reflected this premium buyers were ready to pay to hedge against the downside risk.
In order to build new pricing models that reflected this stylized fact, a maturity and strike dependent volatility was introduced: going from a constant σ accros all instruments to $\sigma(t, S_t)$.

What is now known as the Dupire formula is the following expression for such a volatility function:

$$\sigma(t, S_t) = \frac{\frac{\partial C}{\partial T} + (r-q)K\frac{\partial C}{\partial K} + qC}{\frac{1}{2}K^2\frac{\partial^2 C}{\partial K^2}}.$$

We assume the underlying follows a Geometric brownian motion dynamic: $dS_t = \mu S_t dt + \sigma^2 S_t dW_t$, with $\mu = r - q$. We also introduce the discount factor between time t and maturity T: $D(t,T) = \exp\left(-\int_t^T r(s)ds\right)$.
The call option price can then be written as $C(K,T) = D(t,T)\mathbb{E}_{\mathbb{Q}[(S_T-K)^+]}$.

We are interested in the probability density of the underlying at maturity: $p(S,t)$. Its variations are governed by the **Fokker-Planck equation**:

$$\frac{\partial}{\partial t}p(S,t) = -\frac{\partial}{\partial S}(\mu S p(S,t)) + \frac{1}{2}\frac{\partial^2}{\partial \sigma^2}(\sigma^2 S^2 p(S,t)).$$

Let's compute the theta of a call option:

$$\frac{\partial C}{\partial T} = \frac{\partial D(t,T)}{\partial T}\int_K^{+\infty}(S-K)p(S,T-t)dS + D(t,T)\int_K^{+\infty}(S-K)\frac{\partial p(S,T-t)}{\partial T}dS.$$

Plugging in we get:

$$\begin{aligned}\Theta + rC &= D(t,T)\int_K^{+\infty}(S-K)\left[-\frac{\partial}{\partial S}(\mu S p(S,t)) + \frac{1}{2}\frac{\partial^2}{\partial \sigma^2}(\sigma^2 S^2 p(S,t))\right] \\ &= D(t,T)\left(-\mu I_1 + \frac{1}{2}I_2\right).\end{aligned}$$

We consider the first and second order derivatives with regards to the strike. Is is known that they respectively are equal to the cumulative distribution function above the strike and the probability density at maturity (the latter being the **Breeden-Litzenberger formula**). To know what quantities we should further consider, we apply integration by parts to I_1 and I_2, with the goal to get rid of integrands and fuzzy terms.

$$\begin{aligned}I_1 &= \int_K^{+\infty}(S-K)\frac{\partial}{\partial S}(Sp(S,t)) \\ &= [(S-K)p(S,t)]_{S=K}^{S=+\infty} - \int_K^{+\infty} Sp(S,t)dS \\ &= -\int_K^{+\infty} Sp(S,t)dS.\end{aligned}$$

To explicit this last line, let's rewrite the call price as $\mathrm{C} = \mathrm{S}e^{-qT}(d_1) - Ke^{-rT}(d_2)$ such that $\int_K^{+\infty} Sp(S,t)dS = \frac{1}{D(t,T)}\left(C - K\frac{\partial C}{\partial K}\right)$.

Then,

$$\begin{aligned}I_2 &= \int_K^{+\infty}(S-K)\frac{\partial^2}{\partial \sigma^2}(\sigma^2 S^2 p(S,t))dS \\ &= \left[(S-K)\frac{\partial}{\partial \sigma}(\sigma^2 S^2 p(S,t))\right]_{S=K}^{S=+\infty} - \frac{\partial}{\partial \sigma}(\sigma^2 S^2 p(S,t))dS \\ &= -\left[(\sigma^2 S^2 p(S,t)\right]_{S=K}^{S=+\infty} \\ &= \sigma^2 K^2 p(S,t) \\ &= \sigma^2 K^2 \frac{1}{D(t,T)}\frac{\partial^2 C}{\partial K^2}.\end{aligned}$$

Going back to the theta derivation, we have

$$\frac{\partial C}{\partial T} + rC = C - K\frac{\partial C}{\partial K} + \sigma^2 K^2 \frac{\partial^2 C}{\partial K^2}.$$

Rearranging the terms, we get the Dupire formula.

There also exist a probabilistic derivation of this formula, applying Itô to the payoff $(S_T - K)^+$ and taking the expectation.