Quantitative finance

Interviews preparation

Vivien Tisserand

Abstract

This is a summary of interview questions found on the internet, in 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 fond 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

Bud1

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.

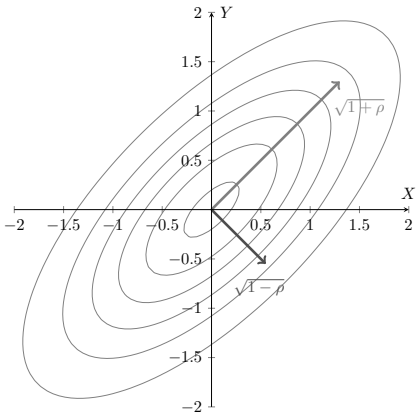


Figure 1.1: Density contours of a bivariate normal law, with $\rho = 0.7$

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:

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

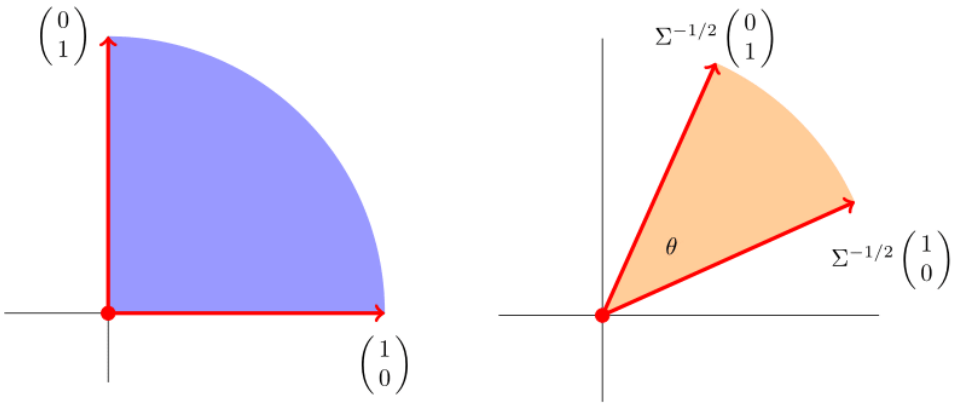


Figure 1.2: Area of the event $X > 0, Y > 0$ for $\rho = 0$ (left) and $\rho \neq 0$ (right). From here.

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 $\dfrac{1}{n}$ to get the same card we already have, thus $\dfrac{n-1}{n}$ to get a new card. This follows a geometric law: the expectation for such an event is $\dfrac{n}{n-1}$. 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:

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

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.

1.3 Change of variables

> Given X, Y, both following a standard normal distribution, compute the density of the random variable X/Y.

This random variable is well defined as the measure of the set $\{Y = 0\}$ is 0. The support is $\mathbb{R}$.

We will define a $\mathcal{C}^1-$diffeomorphism that handles this transformation of the vector (X, Y). Let $\phi\colon (x, y) \mapsto (x/y, y)$ such that $(u, v) = \phi(x, y)$. We will use the change of variable formula:

$$f_{(U,V)}(u, v) = f_{(X,Y)}(x(u, v), y(u, v))|\det J_{\phi^{-1}}(u, v)|.$$

We have to explicit several quantities in the above formula:

1. Expressing x and y as functions of (u, v) :

$$\begin{cases} u &= x/y \\ v &= y \end{cases} \Leftrightarrow \begin{cases} x &= uv \\ y &= v. \end{cases}$$

   Thus there exists $\phi^{-1}\colon (u, v) \mapsto (uv, v)$

2. Computing the Jacobian matrix of ϕ :

$$J_{\phi^{-1}}(u,v) = \begin{pmatrix} \dfrac{\partial \phi_1^{-1}}{\partial u} & \dfrac{\partial \phi_1^{-1}}{\partial v} \\ \dfrac{\partial \phi_2^{-1}}{\partial u} & \dfrac{\partial \phi_2^{-1}}{\partial v} \end{pmatrix} = \begin{pmatrix} v & u \\ 0 & 1 \end{pmatrix}$$

$|\det J_{\phi^{-1}}(u,v)| = v \neq 0$.
This indeed show that ϕ^{-1} is a $\mathcal{C}^1-$diffeomorphism.

Back to the formula :

$$f_{(U,V)}(u,v) = \frac{1}{2\pi} e^{-\frac{u^2v^2}{2}} e^{-\frac{v^2}{2}} |v|.$$

We need to marginalize this joint density by integrating with respect to v (on $\mathbb{R}^*$.

$$\begin{aligned} f_U(u) &= \int_{\mathbb{R}^*} \frac{1}{2\pi} e^{-\frac{u^2v^2}{2}} e^{-\frac{v^2}{2}} |v| dv \\ &= \frac{1}{\pi} \int_0^{+\infty} v e^{-\frac{v^2}{2}(u^2+1)} \\ &= \frac{1}{\pi} \Big[-\frac{1}{u^2+1} e^{-\frac{v^2}{2}(u^2+1)} \Big]_{v=0}^{v=+\infty} \\ &= \frac{1}{\pi(1+u^2)}. \end{aligned}$$

Thus, if $X, Y \sim \mathcal{N}(0,1)$, $X/Y \sim \text{Cauchy}(0,1)$.

1.4 Min and max covariance

> Let $U_i \sim \mathcal{U}[0,\theta]$ for $i \in \{1,\dots,n\}$. If $M = \max_{1\leq i\leq n} U_i$ and $m = \min_{1\leq i\leq n} U_i$, find $\text{Cov}(M,m)$.

There is no reason for these variables to be uncorrelated: the maximum depends on the value of the minimum; for instance it cannot be below.

We have to compute the densities for the max, the min and the joint distribution. To get them, we use the cumulative distribution functions. For $x \in [0,1]$:

$$\begin{aligned} \mathbb{P}(M \leq x) &= \mathbb{P}\left(\bigcap_{i=1}^n U_i \leq x \right) \\ &= \prod_{i=1}^n \mathbb{P}(U_i \leq x) \\ &= x^n. \end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(m \leq x) &= 1 - \mathbb{P}(m > x) \\ &= 1 - \mathbb{P}\left(\bigcap_{i=1}^{n} U_i > x\right) \\ &= 1 - \prod_{i=1}^{n}(1 - \mathbb{P}(U_i > x)) \\ &= 1 - (1-x)^n.\end{aligned}$$

Taking the derivatives of these continuous cdfs, we have :

$$f_M(x) = nx^{n-1}\mathbb{1}_{[0,1]}(x) \quad \text{and} \quad f_m(x) = n(1-x)^{n-1}\mathbb{1}_{[0,1]}(x).$$

For the joint density, the set we are looking for is a bit trickier : for $0 \leq y < x \leq 1$, we partition the event $\{M \leq x\}$ on the disjointed events $\{m \leq y\}$ and $\{m > y\}$:

$$\begin{aligned}\mathbb{P}(M \leq x, m \leq y) &= \mathbb{P}(M \leq x) - \mathbb{P}(m > y, M \leq x) \\ &= \mathbb{P}\left(\bigcap_{1 \leq i \leq n} (U_i \leq x)\right) - \mathbb{P}\left(\bigcap_{1 \leq i \leq n} (y < U_i \leq x)\right) \\ &= x^n - (x-y)^n.\end{aligned}$$

End by deriving with respect to x and y :

$$f_{(M,m)}(x,y) = n(n-1)(x-y)^{n-2}\mathbb{1}_{0 \leq y < x \leq 1}(x,y).$$

We check that this density corresponds to a probability measure by integrating it.

(We actually do not need to change this joint density with $\mathcal{C}^1$-diffeomorphism and the change of variable theorem.)

With

$$\phi\colon (x,y) \mapsto xy.$$

$$\mathbb{E}[\phi(M,m)] = \int_{\mathbb{R}}\int_{\mathbb{R}} \phi(x,y) f_{(M,m)}(x,y)dydx.$$

Using integration by parts for the inner integral :

$$\begin{aligned}\mathbb{E}[\phi(M,m)] &= \int_0^1 \int_0^x xyn(n-1)(x-y)^{n-2}dydx \\ &= n(n-1)\int_0^1 \frac{1}{n(n-1)}x^n dx \\ &= \frac{1}{n+2}.\end{aligned}$$

Finally,

$$\mathrm{Cov}(M,n) = \mathbb{E}[Mm] - \mathbb{E}[M]\mathbb{E}[m] = \frac{1}{n+2} - \frac{n}{(n+1)^2}.$$

We have a covariance that decreases to zero as n goes to infinity.

1.5 Queueing theory: M/M/1/K

> We are considering a store with capacity K where there is 1 employee that fulfills the needs of each client in a random memoryless time (that is each customer stay follows an $\mathcal{E}(\mu)$) and customers arrive through the front door as a $\mathcal{P}(\lambda)$ process. Find the expected number of people in the store.

With Kendall's notation, this is a M/M/1/K problem.

We are interested in the stationary distribution. Let $p_{t,k}$ denote the probability to be in state k at time t. For $1 \leq k \leq K$, we write the differential of this probability as "the entrance flow minus the exit flow":

$$dp_{t,k} = \lambda p_{t-1,k} + \mu p_{t+1,k} - (\lambda+\mu)p_{t,k}.$$

The equilibrium is such that these probabilities do not move. With the initialization and $\rho = \frac{\lambda}{\mu}$, $\pi_1 = \rho\pi_0$, then for each $1 \leq i \leq K$, $\pi_i = \rho^i\pi_0$.
We have $\sum_{i=0}^{K} \pi_i = 1 = \frac{1-\rho^{K+1}}{1-\rho}\pi_0$.

Thus, for each $0 \leq i \leq K$,

$$\pi_i = \frac{\rho^i(1-\rho)}{1-\rho^{K+1}}.$$

We may now calculate the expectancy of the number of people in the store:

$$
\begin{aligned}
\mathbb{E}[L] &= \sum_{i=0}^{K} i\pi_i \\
&= \sum_{i=1}^{K} i\frac{\rho^i(1-\rho)}{1-\rho^{K+1}} \\
&= \frac{(1-\rho)}{1-\rho^{K+1}}\frac{1}{\rho}\left(\frac{(1-\rho^{K+1})}{1-\rho}\right)' \\
&= \frac{(1-\rho)}{1-\rho^{K+1}}\frac{1}{\rho}\frac{(1-\rho)(-(K+1)\rho^K)+(1-\rho^{K+1})}{(1-\rho)^2} \\
&= \frac{\rho(1-(K+1)\rho^K+K\rho^{K+1})}{(1-\rho)(1-\rho^{K+1})}
\end{aligned}
$$

Which gives

$$
\mathbb{E}[L] = \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}.
$$

That is the expectancy where the store has an infinite capacity, minus a term depending on the capacity.

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 :

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

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

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 : $B(\hat{\theta}^{\text{MLE}}) = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta \neq 0$. We can introduce a corrected estimator that we will consider too : $\hat{\theta}^{\text{MLE}}_{\text{corr}} = \frac{n+1}{n}\hat{\theta}^{\text{MLE}}$, 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 :

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

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 Estimating parameters from normal observations

Suppose that you have $x_1, \dots, x_n$ observations from a normal law $X \sim \mathcal{N}(\mu, \sigma^2)$, where both parameters are unknown. Give several estimators for this set of parameters and compare them.

We will compute several estimators from classic methods (MM, MLE, MAP) and analyze their behaviour.

- **Method of moments :**
  By definition, the first and second order moments are equal to the parameters of the normal distribution. Thus, a natural idea is to take the empirical counterpart of these quantities :

$$\begin{cases} \hat{\mu} & = \dfrac{1}{n}\sum_{i=1}^{n} x_i \\ \sigma^2\widehat{\sigma^2} & = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{cases}$$

  We remember that the observations are identically distributed and independent. Thus, $\mathbb{E}[\hat{\mu}] = \mathbb{E}\Big[\dfrac{1}{n}\sum_{i=1}^{n} x_i\Big] = \frac{1}{n} n\mathbb{E}[x_1] = \mu$. This estimator is thus unbiased.

  Let's write the variance estimator a bit differently :

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \end{aligned}$$

  Thus,

$$\begin{aligned} \mathbb{E}[\sigma^2\widehat{\sigma^2}] &= \frac{1}{n}\mathbb{E}\Big[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\Big] \\ &= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(\mathrm{Var}[x_i] + \mathbb{E}[x_i]^2) - n(\mathrm{Var}[\bar{x}] + \mathbb{E}[\bar{x}]^2)\right] \\ &= \frac{1}{n}\Big(n\sigma^2 + n\mu^2 - n\frac{1}{n^2}n\sigma^2 - n\mu^2\Big) \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

There is bias in the estimator of the variance. We can correct this, using *Bessel's correction* : $\sigma^2\widehat{\sigma^2}_{corr} = \dfrac{n}{n-1}\sigma^2\widehat{\sigma^2}$.

- **Maximum likelihood estimator :**

Let's write the likelihood of this model :

$$\begin{aligned} L((x_1,\dots,x_n),\mu,\sigma^2) &= \prod_{i=1}^{n} f_X(x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\dfrac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(\sum_{i=1}^{n} -\frac{(x_i-\mu)^2}{2\sigma^2}\Big). \end{aligned}$$

Here, it seems clearer to work with the log-likelihood :

$$\ell_n((x_1,\dots,x_n),\mu,\sigma^2) = -n/2\ln(\sqrt{2\pi}) - n/2\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

And we want to solve :

$$(\hat{\mu}, \sigma^2\widehat{\sigma^2}) \in \operatorname{argmax}_{(\mu,\sigma^2)\in\mathbb{R}\times\mathbb{R}^+} \ell_n((x_1,\dots,x_n),\mu,\sigma^2).$$

$$\left\{ \begin{array}{ll} \dfrac{\partial \ell_n}{\partial \mu} & = 0 \\ \dfrac{\partial \ell_n}{\partial \sigma^2} & = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} \frac{1}{\sigma^2}\sum_i (x_i-\mu) & = 0 \\ -\dfrac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_i (x_i-\mu)^2 & = 0 \end{array} \right. \Leftrightarrow \left\{ \begin{array}{ll} \hat{\mu} & = \frac{1}{n}\sum_i x_i \\ \sigma^2\widehat{\sigma^2} & = \frac{1}{n}\sum_i (x_i-\mu)^2 \end{array} \right.$$

We end up on the same estimators as given by the method of moments.

2.3 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.

2.4 Central limit theorem

What is the central limit theorem? Highlight the proof steps. Give a counter example for when the sequence of random variables lacks independence; and when it is not identically distributed.

Theorem statement: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $(X_n)_{n\in\mathbb{N}^*}$ a sequence of independent and identically distributed (i.i.d.) real random variables with finite variance (they belong to $L^2(\Omega, \mathcal{F}, \mathbb{P})$). Assume this variance is non null. Then,

$$\sqrt{n}(\bar{X}_n - \mathbb{E}[X_1]) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \mathrm{Var}(X_1)).$$

Proof. Let's consider without loss of generality the case where $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}(X_1) = 1$ (otherwise switch to considering $Y_n = \frac{X_n - \mathbb{E}[X_1]}{\mathrm{Var}(X_1)}$).

Our goal is to use *Lévy's continuity theorem*: the pointwise convergence of the sequence of characteristic function is equivalent to the convergence in distribution.

With $X \sim \mathcal{N}(0,1)$, the characteristic function $\phi_x(\cdot)$ is:

$$\begin{aligned}
\phi_X(t) &= \mathbb{E}[e^{itX}] \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx - x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} e^{(it)^2/2} \int_{-\infty}^{+\infty} e^{-1/2(x-it)^2} dx \\
&= e^{-t/2}.
\end{aligned}$$

Now, let $S_n = \sum_{i=1}^n X_i$. Let's study the convergence of the sequence of characteristic function induced by X_i: $\left(\phi_{S_n/\sqrt{n}}\right)_{n\in\mathbb{N}^*}$.

First, let's explicit this characteristic function:

$$\begin{aligned}
\phi_{S_n/\sqrt{n}}(t) &= \mathbb{E}[\exp(itS_n/\sqrt{n})] \\
&= \mathbb{E}\left[\exp(\frac{it}{\sqrt{n}} \sum_{i=1}^n X_i)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^n \exp(\frac{it}{\sqrt{n}} X_i)\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[\exp(\frac{it}{\sqrt{n}} X_i)\right] \quad \text{by independence of the } X_i\text{'s} \\
&= \prod_{i=1}^n \phi_{X_i}(t/\sqrt{n}) \\
&= \left(\phi_{X_1}(t/\sqrt{n})\right)^n \quad \text{as the } X_i\text{'s are identically distributed.}
\end{aligned}$$

X_1 belongs to $L^2(\Omega, \mathcal{F}, \mathbb{P})$ thus the characteristic function is twice differentiable, and $\phi'_{X_1}(0) = i\mathbb{E}[X_1] = 0$ while $\phi''_{X_1}(0) = -\mathbb{E}[X_1^2] = -\operatorname{Var}(X_1) = -1$. Thus in the neighbourhood of 0, there exists a development of the characteristic function of the form:

$$\phi_{X_1}(h) = 1 - h^2/2 + \underset{h\to 0}{o}(h^2).$$

Let's remark that $\left(1 - \frac{t^2}{2n}\right)^n \underset{n\to\infty}{\longrightarrow} e^{-t^2/2}$, thus we consider $\left|\phi_{S_n/\sqrt{n}}(t) - (1-\frac{t^2}{2n})^n\right| = \left| (\phi_{X_1}(t/\sqrt{n}))^n - (1 - \frac{t^2}{2n})^n\right|$.

We need an upper bound for quantities of the type $|a^n - b^n|$, with $|a| < 1$ and $|b| < 1$.

Fortunately,

$$\begin{aligned}
|a^n - b^n| &= \left|(a-b)\sum_{k=0}^{n-1} a^{n-1-k}b^k\right| \\
&\leq |a-b|\sum_{k=0}^{n-1}\left|a^{n-1-k}b^k\right| \\
&\leq n|a-b|.
\end{aligned}$$

Thus, plugging it all together:

$$\begin{aligned}
\left|\phi_{S_n/\sqrt{n}}(t) - (1-t^2/2n)^n\right| &= \left|\left(\phi_{X_1}(t/\sqrt{n})\right)^n - (1-t^2/2)^n\right| \\
&\leq n\left|\phi_{X_1}(t/\sqrt{n})(1-\frac{t^2}{2n})\right| \\
&= n\left|1 - t^2/2n + t^2/n\varepsilon(t/\sqrt{n}) - 1 + \frac{t^2}{2n}\right| \\
&= t^2|\varepsilon(t/\sqrt{n})| \\
&\underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Applying Lévy's continuity theorem, we get the convergence (in law) towards the standard normal distribution.

□

What if we have non-identical distributions? Let's consider a sequence of independent but non-identical random variables, with mean μ_i and variance σ_i^2. There are conditions like Lyapunov's (moments of order $2+\delta$ with some boundary of the rate of growth of the moments) or Lindeberg's for which ther CLT still holds: with $s_n = \sum_{i=1}^n \sigma_i^2$, $\frac{1}{s_n}\sum_{i=1}^n (X_i - \mu_i) \xrightarrow{(d)} \mathcal{N}(0,1)$.

We can build a counterexample, by trying to break the above conditions for instance. An other natural goal could be choosing a sequence of random variables whose sum has a finite support, as it will obviously not converge towards a normal distribution.
Let's start with i.i.d. $Z_k \overset{i.i.d.}{\sim} \mathcal{U}[-1,1]$, and $X_k = a^k Z_k$ where $0 < a < 1$. The X_k's

are independent uniform on $\left[-a^k, a^k\right]$, with mean $\mathbb{E}\left[X_k\right] = 0$ and variance $\sigma_k^2 = \frac{1}{12}(a^k - (-a^k))^2 = a^{2k}/3 < \infty \quad \forall k \in \mathbb{N}^*$. However, $\sum X_k \in \left[-\frac{1}{1-a}, \frac{1}{1-a}\right]$ thus cannot converge to a Gaussian distribution.

What if the random variables are non-independent? A trivial counterexample would be taking a sequence where each random variable coincides with the first one.

A more interesting counterexample involving timeseries would be the $MA(1)$ model:

$$X_t = \theta Z_{t-1} + Z_t, \quad Z_t \sim \mathcal{N}(0, \sigma_Z^2).$$

The autocorrelation function $\gamma(\cdot)$ would be:

$$\begin{aligned}
\gamma(0) &= \sigma_X^2 = (\theta^2 + 1)\sigma_Z^2, \\
\gamma(1) &= \theta\sigma_Z^2, \\
\gamma(k) &= 0 \quad \forall k > 1.
\end{aligned}$$

With such non-zero autocorrelation, the limit variance of $\sqrt{n}(\bar{X}_n - \mu)$ no longer is $(\theta^2 + 1)\sigma_Z^2$ but rather $\gamma(0) + 2\sum_{k=1}^{\infty}\gamma(k) = (\theta + 1)^2\sigma_Z^2$.

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,

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

we can just minimize each $f_i: x \mapsto -2\hat{\beta}_i^{\text{OLS}}\beta_i + \beta_i^2 + \lambda|\beta_i|$. 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 $(2\eta(X) - 1)\left(\mathbb{1}\{g^*(X) = 1\} - \mathbb{1}\{h(X) = 1\}\right) \geq 0$.

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** if 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 $\mathbb{P}_y(\cdot)$ means that we are starting at 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] $S(x) := \int_{x_0}^{x} \exp\left\{-2\int_{x_0}^{y} \frac{b(z)}{\sigma^2(z)}\,dz\right\} dy$ is a solution of $\mathcal{A}f = 0$,[2] thus provides a recurrence criterion: the process X is recurrent if

[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.

and only if $S(+\infty) = +\infty$ and $S(-\infty) = -\infty$.

Similarly we introduce the speed measure $M(x) \coloneqq \int_{x_0}^{x} \frac{2}{\sigma^2(y)} \exp\left\{2\int_{x_0}^{y} \frac{b(z)}{\sigma^2(z)}\,dz\right\} dy$ whose density writes $m(x) = \frac{2}{\sigma^2(x)S'(x)}$. Then the diffusion is recurrent positive if and only if this measure is σ-finite.

For the Brownian motion, $S(x) = \int_{x_0}^{x} dy$ thus is recurrent; however the speed measure $m(\,dx) = dx$ is not σ-finite. Thus the Brownian motion is only null recurrent.

When it comes to the Ornstein-Uhlenbeck process, $S(x) = \int_{x_0}^{x} \exp\left\{2\int_{x_0}^{y} \frac{\mu(z-m)}{\sigma^2}\,dz\right\} dy = \int_{x_0}^{x} \exp\left\{\frac{\mu}{\sigma^2}(y-m)^2\right\} dy$ is a space transform if and only if $\mu \geq 0$.
The speed measure $m(dx) = \frac{dx}{\exp(\mu(x-m)^2)}$ is σ-finite if and only if $\mu > 0$.

4.2 Heston model

> Let's consider the Heston model. Tell me about its dynamics, pricing properties, simulation schemes, calibration.

Heston model introduced in [Hes93] is a stochastic volatility model that assumes the instantaneous variance dynamics follows a CIR (or square-root) process. It has dynamics:

$$\begin{cases} dS_t = \mu S_t\,dt + \sqrt{V_t} S_t\,dW_t \\ dV_t = \kappa(\theta - V_t)\,dt + \sigma\sqrt{V_t}\,dB_t, \end{cases}$$

with the two Brownians having correlation ρ: $\langle dW, dB\rangle_t = \rho dt$. In the variance process, parameters are the long term variance level θ, the mean-reversion speed κ and the vol-of-vol σ. The set of parameter is $\Theta = \{\mu, \kappa, \theta, \sigma\}$.

Stochastic volatility has been introduced to allow calibration to the smiled and skewed shapes empirically when looking at quoted implied volatilities across different strikes on the market. Although for very short maturities with exploding wings stochastic volatility models may underestimate the options prices (hence the introduction of jumps with Bates model), they remain a very interesting and relevant class of models.

Thorough mathematical analysis

Existence of a solution non lipschitz coeffs, still there exists a unique solution vol vol of vol corr assumed... feller condition

The Heston model belongs to the larger family of *affine models*: the joint process $(\log S, V)$ has an explicit characteristic function that allows for fast and accurate pricing and hedging using Fourier inversion techniques. In particular Heston model is an affine Markovian

model; the popular Stein-Stein model belongs to this class as well, while the Bergomi and Hull-White models are non-affine Markovian models.

Characteristic function The Fourier-Laplace transform for the joint process $(\log S, V)$ is:

$$\mathbb{E}\left[e^{u \log S_T} \mid \mathcal{F}_t\right] = \exp\{u \log S_t + \phi(T-t) + \psi(T-t)V_t\}, \tag{4.1}$$

where ϕ, ψ are the solutions of Riccati equations – this allows for tractability and fast calibration of the model[3].

Proof. Let's start by finding the pricing partial differential equation associated with the model. Feynman-Kac links SDEs to PDEs: for a call price,

$$\frac{\partial C}{\partial t} + \mathcal{A}C - rC = 0,$$

with $\mathcal{A}$ the infinitesimal operator of the diffusion,

$$\mathcal{A} = rS\frac{\partial}{\partial S} + \kappa(\theta - V)\frac{\partial}{\partial V} + \rho\sigma SV\frac{\partial^2}{\partial S \partial V} + \frac{1}{2}VS^2\frac{\partial^2}{\partial S^2} + \frac{1}{2}\sigma^2 V\frac{\partial^2}{\partial V^2}.$$

with $r = \mu - \frac{1}{2}\sigma^2$ for switching from the statistical measure to the risk-neutral measure. A financial construction, applying Itô to an hedged portfolio leads to the same pricing PDE.

Then we make an *ansatz* for the solution: we assume it looks like

$$C(S_t, V_t, t, T) = S_t P_1 - K e^{-r(T-t)} P_2[4].$$

To simplify this, we transform the first argument to the log-moneyness $x = \log(F/K)$, F the $T-$forward price of the stock seen from t, and the time to maturity $\tau = T - t$. The ansatz becomes $C(x, y, \tau) = K\left[e^x P_1(x, y, \tau) - P_2(x, y, \tau)\right]$.

Plugging this ansatz into the pricing PDE,

□

FFT expression: then we can derive a model-free formula for vanilla option prices.

Under a risk-neutral measure $\mathbb{Q}$, the price of a the European call option with maturity T and strike K is:

$$C_t = e^{-r(T-t)}\mathbb{E}^{\mathbb{Q}}\left[(S_T - K)^+\right] = e^{-r(T-t)}(\mathbb{E}^{\mathbb{Q}}\left[K\mathbb{1}_{S_T>K}\right] - \mathbb{E}^{\mathbb{Q}}\left[S_T\mathbb{1}_{S_T>K}\right]).$$

The second expectation can be computed as $\mathbb{E}^{\mathbb{Q}}\left[S_T\mathbb{1}_{S_T>K}\right] = K\mathbb{Q}(S_T > K)$ while for the first one we can introduce a change of numéraire [EGR95]. Introducing the change of measure $\left.\frac{d\mathbb{Q}_S}{d\mathbb{Q}}\right|_{\mathcal{F}_t}$, setting the underlying price as numéraire,

$$\mathbb{E}^{\mathbb{Q}}\left[S_T\mathbb{1}_{S_T>K}\right] = \mathbb{E}^{\mathbb{Q}}[S_T]\mathbb{E}^{\mathbb{Q}}\left[\frac{S_T}{\mathbb{E}^{\mathbb{Q}}[S_T]}\mathbb{1}_{S_T>K}\right] = e^{r(T-t)}S_t\mathbb{Q}_S(S_T > K)$$

[3]The log-price satisfies 4.1, however there exists similar equations for other quantities of interest, for instance spot variance and integrated spot variance.

[4]This is not a surprise when we know Black-Scholes formula for vanillas. Financial practitioners can read these quantities as in-the-money probabilities: $P_1 = \mathbb{Q}^S(S_T > K), P_2 = \mathbb{Q}(S_T > K)$, with the probability measures under different numeraires.

.

Eventually the option price at time t is

$$C_t = S_t \mathbb{Q}_S(S_T > K) - e^{-r(T-t)} K \mathbb{Q}(S_T > K).$$

Then, for a random variable, the Gil-Pelaez inversion theorem [Gil51] relates the cumulative distribution function F and its characteristic function ϕ:

$$F(x) = \frac{1}{2} + \frac{1}{2\pi} \int_0^\infty \frac{e^{itx}\phi(-t) - e^{-itx}\phi(t)}{it} dt.$$

Further computations using real and imaginary parts of an complex quantity z (recall $\mathfrak{Re}(z) = \frac{z+z^*}{2}$ and $\mathfrak{Im}(z) = \frac{z-z^*}{2}$):

$$\mathbb{P}(X > x) = \frac{1}{2} + \frac{1}{\pi} \int_0^\infty \mathfrak{Re}\left[\frac{e^{-itx}\phi(t)}{it}\right] dt.$$

Simulating

Euler scheme For simulating such a process, we have to be careful about the values of the instantaneous variance. Indeed, nothing assures the positivity of the variance, but we have to take its square-root in the log price dynamics. Hence a naive Euler scheme is not feasible.

Exact χ^2 sampling

Calibration Estimating the parameters

Other quantities of interest

Forward variance The forward variance at time t for maturity T is defined as:

$$\xi_t^T = \mathbb{E}\left[V_T \mid \mathcal{F}_t\right].$$

Taking the expectation in the Heston dynamics yields $d\xi_t^T = \kappa(\theta - \xi_t^T)dT$, satisfied by $\xi_t^T = \theta + e^{-\kappa(T-t)}(\theta - V_t)$.
Differentiating, we get

$$d\xi_t^T = e^{-\kappa(T-t)}\sigma\sqrt{V_t}dB_t.$$

In particular this allows to write Heston under an affine forward variance form.

Such a quantity can be used when computing VIX-related metrics, like pricing VIX vanilla options.

ATMF skew a v cool object haha

Link between volatility and vol-of-vol Let's consider the log-volatility process $\ln(\sigma_t)$, $\sigma_t = \sqrt{V_t}$, not to be confused with the vol-of-vol parameter σ.

let's consider the log-volatility dynamics as a transformation of the process V: $\ln(\sigma_t) = f(V_t, t)$ with $f\colon (x,t) \mapsto \ln(\sqrt{x})$.

Applying Itô, we get :

$$\begin{aligned} d\ln(\sigma_t) &= \frac{\partial f}{\partial x} dV_t + \frac{1}{2}\frac{\partial^2 f}{\partial x^2} d\langle V\rangle_t \\ &= \frac{1}{2V_t} dV_t + \frac{1}{2}\left(\frac{1}{-V_t^2}\right)\sigma^2 V_t dt \\ &= \frac{1}{\sigma_t^2}\left(\kappa(\theta - \sigma_t^2) - \frac{\sigma^2}{2}\right)dt + \frac{\sigma}{2\sigma_t} dW_t. \end{aligned}$$

We get a vol-of-vol of $\frac{\sigma}{2\sigma_t}$ in the Heston model, which implies that the volatility and the volatility of volatility will move in opposite directions. This is unconsistent with the empirical market observations.

Chapter 5

Finance

5.1 Around variance swaps

A desk is interested in expressing a view on volatility. To do so they suggest entering a variance swap contract with payoff:

$$\frac{1}{T}\int_0^T \sigma_t^2 dt - \sigma_K^2$$

On the sell side, how would you price and hedge such a contract?
On the buy side, if your goal is to tail-hedge your portfolio, could you suggest other approaches than buying a varswap?

Variance swaps are heavily important instruments: while vanilla options reveals the implied density of the underlying at maturity, varswaps give some information about the path the asset will take. These products are also used to construct the forward variance curve which is taken as an input in Bergomi models.

Hedging varswaps

Could we find a smooth transformation of the asset that would help in the replication of the payoff? Applying Itô's formula to $f(S_t)$, we get:

$$f(S_t) = f(S_0) + \int_0^T f'(S_t)\, dS_t + \frac{1}{2}\int_0^T f''(S_t)\, d\langle S_t\rangle = f(S_0) + \int_0^T f'(S_t)\, dS_t + \frac{1}{2}\int_0^T f''(S_t)\sigma_t^2 S_t^2\, dt$$

To make the variance swap variable leg appear, remark that taking $f = \log$ indeed yields $f''(x) = x^{-2}$. This yields:

$$\int_0^T \sigma_t^2 dt = 2\int_0^T \frac{dS_t}{S_t} - 2\log\frac{S_T}{S_0}.$$

The difficulty here is in the replciation of the log contract. As we are in the case of a twice differentiable payoff ϕ, we can apply the *Carr-Madan formula*[1] that gives the replication

[1] First derived in Peter Carr and Dilip Madan. "Towards a theory of volatility trading". In: *Volatility: New estimation techniques for pricing derivatives* 29 (1998), pp. 417–427.

in terms of calls and puts:

$$\phi(S) = \phi(x) + \phi'(x)(S-x) + \int_0^x \phi''(K)(K-S)^+ dK + \int_0^x \phi''(K)(S-K)^+ dK$$

Around gamma swaps

While varswaps give constant dollar Gamma ($\$\Gamma$) exposure, one might want it to be linear – it lowers skew exposure. It is possible to build such a product through a weighted variance swap:

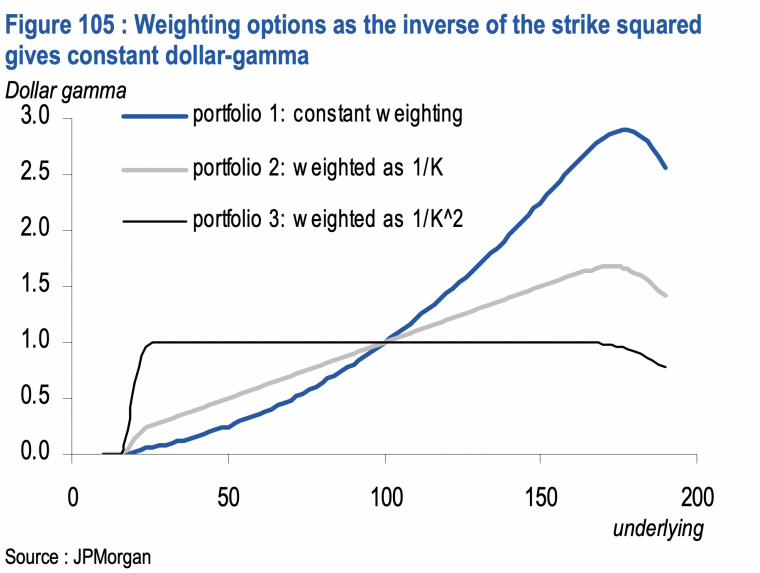


Figure 5.1: Dollar-gamma exposure for different swap flavours. Constant for variance swap, liner for gamma swap.

here the correct choice would be $f : x \mapsto x \log x - x$, which indeed yields the variable leg $\int_0^T \frac{S_t}{S_0}\sigma_t^2 dt$.

There exists some other flavours of swaps: corridor swaps, entropy swaps[2], correlation swaps, *etc.* Volswaps are not really a thing as they cannot be statically replicated by a portfolio of options.

Extracting the forward variance curve

Forward variance curve $\xi_0(\cdot)$ should be extracted from variance swap replication as a piece-wise constant càdlàg function, where each next level is built thanks to the SVI interpolation of the strip of calls and puts prices. The quantity $\xi_t(u)$ is the instantaneous variance at time $u > t$ seen from t.
Having extracted forward variance curve values ξ_i at time T_i for $0 \leq i \leq n$, we can

- built a piece-wise constant curve between $[T_i, T_{i+1})$ such that $\xi_0(t) = \sum_{i=1}^n \xi_i \mathbb{1}_{t\in[T_i,T_{i+1})}$,
- or decide to run a cubic spline interpolation with knots (t_i, x_i) with $t_i = (T_i + T_{i+1})/2$ and $x_i = \sqrt{\int_{T_i}^{T_{i+1}} \xi_0(s)ds}$ and then square the interpolation to ensure positivity of the forward variance curve [JI+22],

[2]Check out Hans Buehler's thesis.

- other parametrization with mean-reversion dynamics.

5.2 Interest rates models

> Whar are the most populat interest rates models? We will denote $B(t,T)$ the zero-coupon bond price at time t that gives one unit at time T.

and Musiela papers!

5.3 Options P&L attribution

> How would you break down the profits and losses of an options portfolio?

For short time horizon, the canonical method is the so-called Greeks decomposition, which splits P&L across the various derivatives of the Black-Scholes formula. This decomposition highlights contributions that come from the passage of time versus that which comes from changes in market variables.

Denoting the portfolio Π as a function of time t, spot level S and volatility σ, an application of Itô's formula (read stochastic Taylor expansion) yields, assuming the portfolio is composed of a single call option C:

$$d\Pi(t,S,\sigma) = \overbrace{\frac{\partial C}{\partial t}}^{\text{theta }\Theta}\, dt + \underbrace{\frac{\partial C}{\partial S}}_{\text{delta }\Delta}\, dS + \frac{1}{2}\overbrace{\frac{\partial^2 C}{\partial S^2}}^{\text{gamma }\Gamma}\, d\langle S\rangle + \frac{1}{2}\underbrace{\frac{\partial^2 C}{\partial \sigma^2}}_{\text{volga}} d\langle\sigma\rangle + \overbrace{\frac{\partial^2 C}{\partial S \partial \sigma}}^{\text{vanna}} d\langle S,\sigma\rangle$$

$$= \frac{\partial C}{\partial t}dt + \frac{\partial C}{\partial S}dS + \overbrace{\frac{1}{2}S^2\sigma^2\frac{\partial^2 C}{\partial S^2}}^{\text{gamma P\&L}} dS^2 + \frac{1}{2}\frac{\partial^2 C}{\partial \sigma^2}d\sigma^2 + \frac{\partial^2 C}{\partial S \partial \sigma}dSd\sigma.$$

This framework gives daily P&L breakdown that can be summed to explain longer term P&L.

5.4 Pricing a cliquet option with change of numéraire

> What are cliquet options? Which exposure do they give? How to price them?

Cliquets – we restrict ourselves to simple ratchets – are a kind of exotic options consisting in a series of forward starting options. It can be seen as a sequence of pre-purchased ATM options which become active in turn consecutively. The strikes of the options are not known at inception of the product and are resetted when the new option becomes active.
Such a structure is motivated by the insurance sector and the selling to retail investors

of annuities (think fixed index annuities for retirement schemes). These contracts offer downside protection while maintaining potential upside thus have become popular post-crisis.
It is highly sensitive to future implied volatility and ATM forward skew, thus the need for a model that catches those dynamics correctly[3].

Below we highlight pricing schemes for cliquet-style options (that includes accumulators, reverse cliquets and Napoleons).

The change of numéraire method

Change of numéraire methods are very powerful when introducing stochastic rates in modelling – typically when pricing and hedging interest rates derivatives.

Let X_T be the payoff of a contingent claim maturing at time T (*e.g.* $(S_T - K)^+$ for a vanilla call). We suppose we are at time t. According to the risk-neutral valuation principle, the t-price of this contract (under the risk-neutral measure $\mathbb{Q}$) is:

$$X_t = \mathbb{E}_t \left[e^{-\int_t^T r_s ds} X_T \right].$$

This can be resolved when the rate is deterministic, taking the actualisation term out of the expectancy, but requires more work when the rate process is stochastic.
Let's define the *forward measure* $\mathbb{Q}^T$ associated with the *numéraire* $B(t,T)$ through the Radon-Nikodym derivative:

$$\left. \frac{d\mathbb{Q}^T}{d\mathbb{Q}} \right|_{\mathcal{F}_t} = \frac{B(T,T)}{B(t,T)} = Z_t.$$

Then, using Bayes rule for conditional expectation,

$$\begin{aligned}
\mathbb{E}_t \left[e^{-\int_t^T r_s ds} X_T \right] &= \mathbb{E}_t \left[\frac{B(t,T)}{B(T,T)} X_T \right] \\
&= \frac{1}{Z_t} \times Z_t \mathbb{E}_t^{\mathbb{Q}} \left[\frac{1}{Z_t} X_T \right] \\
&= \frac{1}{Z_t} \mathbb{E}_t^{\mathbb{Q}^T} \left[X_T \right] \\
&= B(t,T) \mathbb{E}_t^{\mathbb{Q}^T} \left[X_T \right].
\end{aligned}$$

Forward start option

Let $t < T$ and $\theta > 0$, with t the current time at which we want to price the payoff

$$(S_{T+\theta} - K S_T)^+$$

paid in $T + \theta$.

[3] Stochastic volatility models that specify dynamics for the entire forward variance curve rather than only the instantaneous variance are thus preferred, *e.g.* Bergomi one and two-factor models [Ber04; Ber05], where the skew can even be part of the error function in the calibration to ensure its dynamics are well caught.

In practise: which is the best model to price cliquets?

For payoffs with path-dependent features, capturing the volality surface evolution is crucial. Typically, stochastic volatility models are to be preferred to local volatility models to capture dynamics, as the latter assume deterministic changes in the volatility (with regards to the underlying price and time).
LV is a static model calibrated on today's vol surface: it will produce much flatter forward smiles. The whole dynamic comes from the volatility surface at time t, which is too strong of an assumption to replicate accurate surface behaviour. Thus LV will tend to underprice the option as it underestimates the persistence of the forward skew.
As a rule-of-thumb, forward skew decays as $\tau^{-1/2}$ in this model [Lee05] while we generally observe a powerlaw decay with an exponent lesser than 1/2 for ATM skew[4].

One may be looking at LSV models to compensate drawbacks of simple LV.

5.5 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**:

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

[4]We observe a more persistent ATM skew decay: $\mathcal{S}_T \propto T^{-1/2+H}$, $H \in (0, 1/2)$, see the interest of rough models [Fuk17] and considerations about explosions in the very short term [GE22].

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 $C = Se^{-qT}(d_1) - Ke^{-rT}(d_2)$ such that $\displaystyle\int_K^{+\infty} Sp(S,t)dS = \frac{1}{D(t,T)}\left(C - K\frac{\partial C}{\partial K}\right)$.

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.

Numerical implementation

While this formula for $\sigma(S_t, t)$ (or $\sigma(K, T)$) is straighforward, computation requires careful treatment to ensure the resulting surface is absent of static arbitrage.

5.6 Leveraged ETFs and volatility drag

> How would an asset manager build and manage a leveraged ETF? How to cope with volatility drag?

Let's say we want to offer a "2× AAPL" product to our clients. This exchange-traded product should replicate the daily performance of two times the one of the underlying common Apple stock.

Product design: Two instruments can be used on top of the udnerlying to deliver a leveraged performance: total return swaps and futures. Following [CM09], let S_t be the price of one Apple share at time t, and r_i be the return of the underlying between t_{i-1} and t_i: $r_i = \frac{S_{t_i}}{S_{t_{i-1}}} - 1$. $r_{i:j}$ will denote the return between times t_i and t_j. We will denote x the leveraged multiple, which usually takes values in $\{-2, -1, 2, 3\}$.
There is a path-dependency in the delivered returns, as the compounded move has no reason to be the same as the multipled return:

$$\prod_{i=1}^{N}(1 + xr_i) \neq (1 + xr_{0:N}).$$

Let's consider the ETP has a NAV of A_n at time t_n. The notional of the total return swaps required is $L_n = xA_n$. With a daily rebalancing, at the end of the day the TRS exposure is $xA_n(1 + r_{n+1})$ while the fund's NAV grew to $xA_n(1 + xr_{n+1})$. To account for this disparity, the exposure to the TRS has to be adjsuted by $\Delta_n = A_n(x^2 - x)r_{n+1}$.

Volatility drag is the difference between summed and compounded returns: it is harder to recover from a much lower starting point. Indeed, while we need a 20% drop to go from \$100 to \$80, we need 25% to make it back to \$100, and $-20\% + 25\% \neq 0$.
This is referred to as vol drag, vol tax, or even variance drain.

If we consider a series of fair coin flips where, starting at \$100 you bet one percent of your bankroll, double your bet on heads and lose it on tails, with $v_n = 100 \times 0.99^n \times 1.01^n$, you'd have $\forall n \in \mathbb{N}^* v_n < 100$ and $\lim_{n\to+\infty} v_n = 0$.

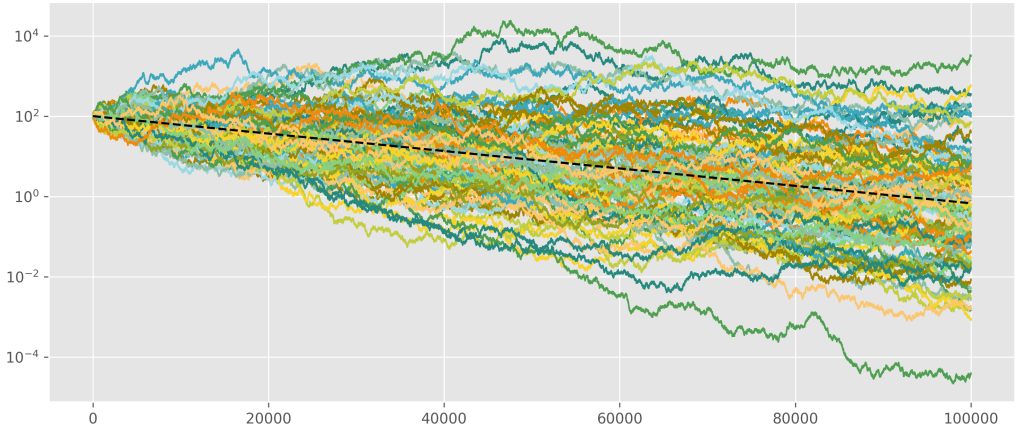


Figure 5.2: Wealth after n coin flips, with the drag trend. See code snippet.

Are all leveraged ETFs doomed to ruin though? No, but they appear not suited for long term investors as only a positively skewed return distribution could compensate the effect of the drag.

We can quantity the difference between arithmetic and geometric returns: $r_g \approx r_a - \sigma^2/2$. This applies to all assets, however leveraged products can only amplify this phenomenon and the 2× AAPL product would come with quadruple the volatility drag of owning shares spot, with manager fees.

Chapter 6

Computer science

6.1 Generating random variables

> When querying samples from a known distribution, what really happens under the hood? How do computer generate randomness? Explain how you would build `np.random.normal` or `np.random.exponential` from scratch.

PRNGs

Before jumping to random variables, we first need to generate random numbers. Pseudo-random number generators (PRNGs) define a deterministic recurrence relation between x_{n+1} and x_n such that a sequence looks random (that's why setting the first input through `np.random.seed(x_0)` fixes the randomness of the generator).

The most popular PRNGs are:

- Mersenne twister: through matrix multiplication and deterministic operations, it generates sequences by batches of 32 bits (for MT19937, others choices are available), however it is fully predictable after 624 outputs.
  At the hardware level, this involves using a linear-feedback shift register (LFSR) with a XOR logic gate and a well-chosen feedback function (insights from group theory are useful to make a good choice here).
  It is used as the default PRNG in Python.

- PCG family: developped by Melissa O'Neill in 2014, it combines nicer properties. It stands for Permuted Congruential Generator. It applies a congruential operation to update the random state (the "CG" – process from the historical first PRNGs but statistatically weak) but instead of returning it direcly as a random integer, applies a permutation function – the "P".
  `numpy`'s default switched to PCG64.

- There exists cryptographically secure PRNGs where non-predictability is important.

There are batteries of statistical tests for measuring the quality of a random number generator: George Marsaglia's diehard or L'Ecuyer's TestU01.

From uniform distribution to other densities

Knowning how to get "random" integers between, say $[0 \,..\, 2^{64} - 1]$, we can shrink the output to effectively sample from $\mathcal{U}([0, 1])$.

A classic way to sample from any known density is to invert its cumulative distribution function – this is the *inverse transform method*: for X a random variable and F_X its cdf, U a standard uniform. Then,

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(U \leq F_X(x)) = \mathbb{P}(F_X^{-1}(U) \leq x),$$

thus we can sample from X by sampling from $F_X^{-1}(U)$.

Sampling from the exponential distribution by computing its cdf: if $X \sim \mathcal{E}(\lambda)$, $F_X(x) = 1 - e^{-\lambda x}$. Inverting it, we just need to compute $-\frac{1}{\lambda}\ln(1-u)$ where u is a uniform sample. Moreover $U \overset{(d)}{=} 1 - U$ so it is even more efficient to compute $-\frac{1}{\lambda}\ln(u)$.

We have an efficient method when the cdf inverse has a closed form, and something that works numerically if we can query values of the cdf (is most of the cases, although it induces bias if the support is not finite and a choice of partition). But there are better ways to sample from popular distributions.

6.2 Using `git`

> `git` is the go-to version control system to track changes in files and collaborate with other programmers. Below is a list of the basic survival commands, as well as some more in-depth routines – always refer to a good documentation though, where all the possible arguments are detailed.

There are 145 git commands. We'd like to highlight the most important ones to have an effective workflow.

Basics

Let's say it is your first day at an investment bank as a quant. Soon enough you're going to get developper access to a bunch of libraries (used for pricing, monitoring, or whatever) that you are going to contribute to.
`git clone`
is going to be your first command to get your local version of the remote repository. If you were to create a new repository, you'd use `git init`.

Then you're going to add features to the current library, for instance implement this fancy stochastic volatility model you've heard about. There is the `main` or `master` branch that everyone takes as reference, and you'll want to create your changes on a parallel version to not interfere. Instead of using `git branch` to create your own branch, you can directly go with:
`git checkout -b $BRANCH_NAME`
Now you can start adding / modifying files. Switch between branches with `git switch` and toggle between the current and the last seen with `git checkout -`. Check the state

of the modified files with `git status` (and add the `-s` flag for a shorter summary). And use `git diff` to see the differences, although it is better to have a nice GUI for this.

A nice routine to add, commit and push modifications is:
```
git add .
git commit -m "$COMMIT_MESSAGE"
git push
```

A good commit meassage should look like `feat: fast pricing for heston model`, `fix(params): update default buffer for edge case` or `docs: added equations in the docstring`, specifying the type of the modification and even (sometimes) an additional scope in parentheses.

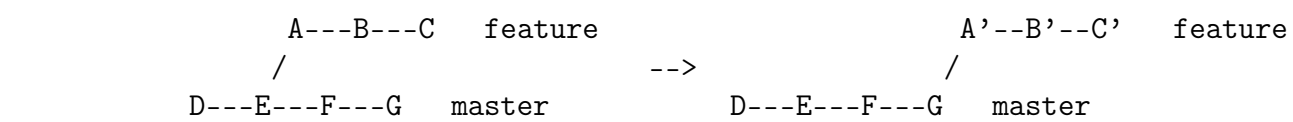
To stay on top of the modifications of your colleagues and avoid conflicts, keep regularly pulling the latest changes:
`git pull origin master`
which is a shortcut for `git fetch origin` and `git merge`. Careful with pulling though, as each time it creates a merge commit and can make it hard to navigate the commit history; try rebasing (more on that later, with `git pull --rebase`).

`git log` is useful to show an history of the commits, with messages and commit references. A better log can be seen with:
`git log --pretty=oneline --abbrev-commit --graph --decorate --all`

Intermediate

You are now well established in the `git` world. Several features are being developped simultaneously on different branches. There is an emergency and you need to switch to the branch `feature1` while you are developping on `feature2`. However you cannot add your changes to the staging area as they are not quite ready yet. You can `git stash` the current changes in a dirty working directory to keep the current state. The stash works like a stack, you can visualize it with `git stash list`, `pop` it to apply the oldest change and delete it from the stash, or `apply` it while keeping it in the ditry working directory. Now you can switch to the urgent feature and come back to the current and reapply the modifications.

`git rebase` allows to play with branches in such a way:



This is what we were talking about, advocating for `git pull --rebase` instead of `git pull` that adds a merge commit. In case of a merge conflict, use `git rebase --abort`, pull normally and solve it.

`git cherry-pick <commit>` applies the changes of a commit.

Also, you're now caring more about the quality of your commits, with a defined goal and scope for each one. If you have changes that should have been included in the latest commit and don't want to create a new commit, use `git commit --amend --no-edit`.

Advanced

My aliases

Some of the above commands can be long to type, here is a list of aliases I use for a smoother workflow:

```
git config --global alias.amend "commit --amend --no-edit"
git config --global alias.pr "pull --rebase"
```

6.3 Automatic differentiation

> Explain the principle of automatic differentiation, and highlight example use in finance and machine learning.

When given a function, there are severl ways of computing its derivative:

- **Numerical differentiation** makes use finite differences to get an approximation at a given point x according to $f'(x) \approx \dfrac{f(x+h) - f(x-h)}{2h}$,
- **Symbolic differentiation** manipulates the mathematical expression of the function to get a plug-in formula for the derivative, relying on chain rule (and simplifying the resulting expression that can get quite lengthy) – this is the process we are taught in school,
- **Automatic differentiation**

The key data structure in automatic differentiation (AD or AAD for adjoint algorithmic differentiation) is a directed acyclic graph (DAG). Given a mathematical expression, we can break it down into elementary operations.

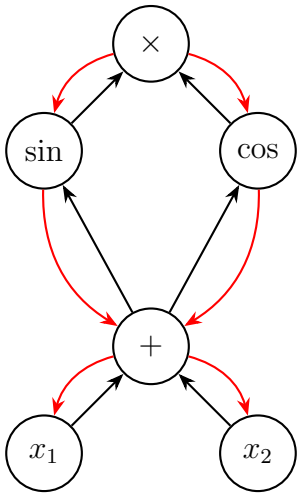


Figure 6.1: Expression DAG for $f : (x_1, x_2) \mapsto \cos(x_1 + x_2)\sin(x_1 + x_2)$.

Each real number in this flow graph is actually a dual: the value and its gradient. To get the latter, we compute the Jacobian matrix of the primitive operation at the value point (in a vectorized manner). This is the *Jacobian-vector product* step (jvp). And that's it:

the accumulation of these primitive operations allow for the computation of the gradient at a given point.
Or rather that's it for *forward-mode* AD. Most of the times, we are trying to differentiate a function $f : \mathbb{R}^n \to \mathbb{R}$, where n can be large (think of a neural network with a lot of features as inputs and a single loss value at the end). We can switch to *reverse-mode* AD where the dual (value, gradient) gets propagated backwards – red arrows on the example DAG. We are now doing a *vector-Jacobian product* (vjp).

Let's break down f into elementary functions f_i and store the intermediate results x_i. Then, for $i = 1, \dots, n, x_i = f_i(x_1, \dots, x_{i-1})$ and we can compute the adjoint[1] with the chain rule:

$$\overline{x_i} = \frac{\partial x_n}{\partial x_i} = \sum_{j=i+1}^{n} \frac{\partial x_n}{\partial x_j} \frac{\partial f_j}{\partial x_i} = \sum_{j=i+1}^{n} \overline{x_j} \frac{\partial f_j}{\partial x_i},$$

where the previous adjoints have already been computed.

Financial applications: AAD is used in the financial industry to compute sensitivity of a portfolio exposure[2], or calibrate a model to market data (and actually whenever numerical minimization is involved).
While finite differences work up to a certain scale, AAD alleviates the need to write bespoke code for each new traded product and can save on computational cost. [Gee+17] clearly highlights the underlying motivation behing such a framework to keep up with regulatory demand. XVA computations involving nested Monte Carlo simulations can also be considerably sped up switching to AD.

When trying to find the optimal parameter for a model (*e.g.* the SABR parameters to fit the SPX smile) the objective function often is a variation of mean square error. Brent or Levenberg-Marquardt are popular methods there.

In finance, second-order sensitivities can be harder, as vanilla payoff are not twice differentiable[3]. Thus, for higher order sensitivities, AD can be refined to outperform (complex) finite differences and Malliavin calculus.

Machine learning applications: The main reason behind the popularity of ML frameworks – beside the convenience of writing an entire neural network in a snippet of a few lines of code – resides in their handling of gradient computation through automatic differentiation to efficiently minimize loss functions.

To zoom into these areas of the libraries, one can have a look at projects like `autograd` that automatically differentiate native `numpy` code, or Karpathy's `micrograd`.

[1] AAD uses adjoints rather tha tangents.

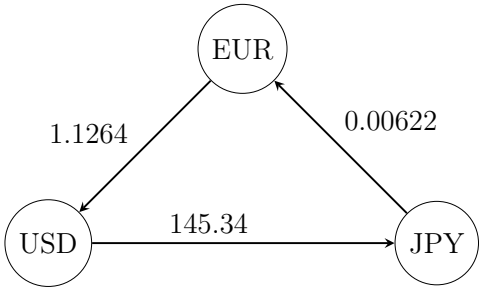
[2] See the standardised approach highlighted in BIS's Minimum capital requirements for Market Risk for the Basel framework.

[3] Sometimes practitioners rely on a smooth approximation of the Heaviside step function to bypass the difficulties of a discrete function. Tons of approximations here.

6.4 Currency arbitrage

> Consider the FX spot market, where we can see live exchange rates among several currencies (say USD, EUR, GBP, JPY, CHF, PLN, BRL, CNH). Design an algorithm that finds pure arbitrage opportunities on this market.

Let's represent this market as a network where currencies are nodes and available exchange rates are edges weight.



In the above example, you can roundrip from one dollar to $145.34 \times 0.00622 \times 1.1264 = \$1.0183$. If we scale the network, potential arbitrages won't be as obvious and we need an efficient algorithmic way to exploit them.

To scan such a graph for arbitrage opportunities, we won't go through the $\approx n!$ paths possible, but rather focus on finding interesting cycles.

Bellman-Ford algorithm finds negative-weight cycles in time complexity $O(|V| \cdot |E|)$ by repeatedly relaxing the edges. We just need to transform the edges weights noticing that

$$\text{USDJPY} \times \text{JPYEUR} \times \text{EURUSD} > 1 \Leftrightarrow -\log(\text{USDJPY}) - \log(\text{JPYEUR}) - \log(\text{EURUSD}) < 0.$$

One can also use *Floyd-Warshall* which relies on dynamic programming[4].

Algorithm	**Time Complexity**	**Space Complexity**
Bellman-Ford	$O(\|V\| \cdot \|E\|)$	$O(\|V\|)$
Floyd-Warshall	$O(\|V\|^3)$	$O(\|V\|^2)$

Table 6.1: Algorithms to find negative-weight cycles

Going further we could add bid and ask data, as well as factoring in transaction fees and liquidity constraints.

[4]Shortest Path Faster Algorithm that blends Bellman-Ford and breadth-first search ideas (see this SPFA implementation), while some other solutions use a divide-and-conquer approch [YK02]; however we do not use Dijkstra as we can have edges with negative weights. Techniques from Kruskal's algorithm using union-find would work on an undirected graph, which is not the case here.

References

[Gil51] J Gil-Pelaez. "Note on the inversion theorem". In: *Biometrika* 38.3-4 (1951), pp. 481–482.

[Hes93] Steven L Heston. "A closed-form solution for options with stochastic volatility with applications to bond and currency options". In: *The review of financial studies* 6.2 (1993), pp. 327–343.

[EGR95] Nicole El Karoui, Helyette Geman, and Jean-Charles Rochet. "Changes of numeraire, changes of probability measure and option pricing". In: *Journal of Applied Probability* 32.2 (1995), pp. 443–458.

[CM98] Peter Carr and Dilip Madan. "Towards a theory of volatility trading". In: *Volatility: New estimation techniques for pricing derivatives* 29 (1998), pp. 417–427.

[YK02] Takeo Yamada and Harunobu Kinoshita. "Finding all the negative cycles in a directed graph". In: *Discrete Applied Mathematics* 118.3 (2002), pp. 279–291.

[Ber04] Lorenzo Bergomi. "Smile dynamics I". In: *Available at SSRN 1493294* (2004).

[Ber05] L Bergomi. *Smile Dynamics II. Risk Magazine*. 2005.

[Lee05] Roger W Lee. "Implied volatility: Statics, dynamics, and probabilistic interpretation". In: *Recent advances in applied probability* (2005), pp. 241–268.

[CM09] Minder Cheng and Ananth Madhavan. "The dynamics of leveraged and inverse exchange-traded funds". In: *Journal of investment management* 16.4 (2009), p. 43.

[Fuk17] Masaaki Fukasawa. "Short-time at-the-money skew and rough fractional volatility". In: *Quantitative Finance* 17.2 (2017), pp. 189–198.

[Gee+17] Sébastien Geeraert et al. "Mini-symposium on automatic differentiation and its applications in the financial industry". In: *ESAIM: Proceedings and Surveys* 59 (2017), pp. 56–75.

[GE22] Julien Guyon and Mehdi El Amrani. "Does the Term-Structure of Equity At-the-Money Skew Really Follow a Power Law?" In: *Available at SSRN 4174538* (2022).

[JI+22] Eduardo Abi Jaber, Camille Illand, et al. "The quintic Ornstein-Uhlenbeck volatility model that jointly calibrates SPX & VIX smiles". In: *arXiv preprint arXiv:2212.10917* (2022).