Strategy Discovery And Backtesting: How To Handle Multiple Testing ?

Study carried out by the Quantitative Practice
Special thanks to Pierre-Edouard THIERY.

canopee
empowering *ecosystem*

www.canopee-group.com

Summary

Introduction

Backtesting is a technique widely used in finance when it comes to strategy discovery: it simply consists in testing a potential strategy against the set of historical data. Morally, this amounts to observing how the strategy would have returned, had it been launched several years ago. This way of doing, albeit widespread across the industry, may prove to be trickier than it may seem at first sight. Tapping historical data exposes to a lot of statistical biases, and those biases must be properly taken into account if one does not want to face poor and disappointing results once the strategy is implemented in real life.

Inasmuch as there is inevitable data mining [1] [2] [3] when a researcher works with historical data, a common practice is to discount the Sharpe ratio of the trading strategy being assessed with a discount factor, generally 50%. Such questions are particularly important in the field of systematic strategies. However, as the reader may have guessed, a fixed discount factor, such as 50%, is not scientifically grounded; it is merely a rule of thumb.

More robust approaches are then necessary. The finding of a new strategy based on extensive mining of past data requires carrying out numerous tests. Indeed, in statistics, if a researcher thinks that a variable X explains a variable Y, he or she can carry out a statistical test: it gives him/her a way of assessing whether his/her assumption is false or not, i.e. whether X and Y are closely related. However, in practical situations, especially in finance, the researcher may try to find relationships between Y and various variables X_1, X_2, etc. If the researcher performs many tests, it is possible that, by chance, at least one of them will prove to be positive even though there is no relationship between the two considered variables.

The issue here is the one of multiple tests on a given set of data. In our paper, we provide a statistical framework to cope with the issue when backtesting financial strategies. To do so, in our first section, we remind the reader of the basics of statistical testing; then, in a second section, we extensively set forth how to deal with multiple tests when assessing financial strategies. In particular we present methods which are currently used, and which curtail the odds of finding a not-so-good strategy once put in production.

1 A Quick Reminder of Statistical Testing

1.1 The Philosophy Of Statistical Tests

When studying a statistical phenomenon, making assumptions on the observed data is a fairly common and intuitive practice. However, although the assumptions may sound rather obvious, it is important to statistically and mathematically assess them in order to evaluate whether or no they are validated. This is the point of statistical testing.

We assume we have at our disposal a set of observations, denoted $\mathscr{X}$. As an example, the reader may imagine we want to check if a coin is fair: to do so, he/she flips the coin n times before writing down the observations with zeros and ones (Bernoulli modeling), for instance:

$$\mathscr{X} = \{X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0, X_5 = 0, X_6 = 1, X_7 = 0, \mathit{etc.}\}$$

Based on the set of observations, the observer makes an assumption, called the null assumption and denoted H_0: for instance, the coin is fair. If we model the experiment with a Bernoulli random variable, where 1 stands for "head", it means that the average of the Bernoulli random variable $\mathscr{B}(\theta)$ representing our statistical phenomenon is worth $\theta = 0.5$.

To test our hypothesis, we must to design a function of the observations, denoted ϕ of the following form:

$$\phi = 1_{\mathscr{R}(\mathscr{X})}$$

where $\mathscr{R}(\mathscr{X})$ is called the rejection zone. If ϕ is worth 1, it means that we reject the null hypothesis.

Intuitively, with our simple example, we can compute the mean of our observations, $\hat{\theta}(\mathscr{X})$, and compare it to 0.5. So we can define the following rejection zone:

$$\mathscr{R}(\mathscr{X}) = \left\{ \left| \hat{\theta}(\mathscr{X}) - \frac{1}{2} \right| > c \right\}$$

with c a parameter. The quantity

$$\hat{\theta}(\mathscr{X}) - \frac{1}{2}$$

is sometimes referred to as the t-statistic of our test.

Here, it is pivotal to have in mind what the t-statistic is. It is a quantity which is computed based on our data. However, the t-statistic can be seen in two different ways: first, it is a mere **numerical value**, based on the observations; second, it also defines a **statistical distribution** based on the assumption we may have made on the observed statistical phenomenon.

For instance, with our coin example, X_i is both a numerical value and a random variable with law $\mathscr{B}(\theta)$. When defining the t-statistic, it is important to find mathematical construction which can be easily interpreted as distribution: for instance here, the sum of n independent Bernoulli variables is famously known as a binomial distribution.

Two kinds of errors are derived from a statistical test. The first one is the so-called type I error, when we reject the hypothesis, although it is correct. The second one is the so-called type II error, when we accept the assumption, although it is not correct.

if we assume the null hypothesis can be parametrized in the following manner: $\theta \in \Theta_0$, then the type I error can be written:

$$\sup_{\theta \in \Theta_0} \mathbb{P}\left[\phi = 1 \,|\theta\right]$$

and the type II error:

$$\sup_{\theta \notin \Theta_0} \mathbb{P}\left[\phi = 0 \,|\theta\right]$$

It is also important to assess the level of our test. The test ϕ of the assumption H_0 is said to be of level α when:

$$\sup_{\theta \in \Theta_0} \mathbb{P}\left[\phi = 1 \,|\theta\right] \leq \alpha$$

Setting a level α is a way of finding the corresponding value c in the rejection zone. With our simple example, the inequality may be rewritten:

$$\mathbb{P}\left[\frac{n}{2} - nc < \sum_{i=1}^{n} X_i < \frac{n}{2} + nc \middle| \theta = \frac{1}{2}\right] \leq \alpha$$

where the sum of the independent variables X_i is binomial distribution with parameters 0.5 and n. Since this law is widely known, it is possible to find the value of c to make sure that our test reaches the desired level α.

Another interesting concept is the p-value. It consists in comparing the distribution of the t-statistic with its value: the former is denoted T and and the latter t. The p-value can be defined in various ways depending on what we want to study. For instance:

$$p = \mathbb{P}\left[T \geq t \,|H_0\right]$$

if we consider a p-value defined on the right part of the distribution, or

$$p = \mathbb{P}\left[|T| \geq |t| \,|H_0\right]$$

for a symmetric p-value.

As the reader may have guessed so far, statistical tests is more of a flexible framework than a rigid process. Computing the p-value amounts to checking if the t-statistic value is extreme compared to the theoretical law of the t-statistic under the null hypothesis. If the p-value is small (meaning that t is very far in one of the distribution tails), it is either that the null hypothesis is false, or the something highly unlikely has happened. Formally, we reject the null hypothesis. Of course, this is done by comparing the p-value with pre-defined threshold α.

Often, we use $\alpha = 0.05$. If the p-value is above this level, it means the t-statistic value is the outcome of a situation which my happen in more than 95 cases out of 100.

1.2 P-Value In Practice: The Coin Example

In this subsection, we explicitly compute the p-value in the case of our simple example. The most natural t-statistic consists in computing:

$$\frac{\sum_{i=1}^{n} X_i}{n} - \frac{1}{2}$$

However, as mentioned above, flexibility is key in making a statistical test. Here, we notice that the sum of n independent Bernoulli variables is well-known: it is a binomial distribution. The t-statistic law is denoted T. The t-statistic value counts the number of heads.

We decide to compute the p-value on the right of the distribution, to ensure that the coin is not biased towards falling head:

$$\begin{aligned}\mathbb{P}\left[T \geq t \middle| \theta = \frac{1}{2}\right] &= \sum_{i=t}^{n} \mathbb{P}\left[T = i \middle| \theta = \frac{1}{2}\right] \\ &= \frac{1}{2^n} \sum_{i=t}^{n} \frac{n!}{i!(n-i)!}\end{aligned}$$

For a numerical example, imagine we flip a coin 20 times, and we get 14 heads. In this case, the p-value is close to 0.058. If we choose an α level of 0.05, the p-value exceeds α, meaning that we are in a situation which would happen 95% of the time if the coin is fair. Therefore our test does not lead to the rejection of the null hypothesis.

1.3 Gaussian Tests On The Mean

In this subsection we extensively set forth one of the main categories of statistical tests: those which are done on the mean of a Gaussian distribution. This class of tests indeed proves to be pivotal in finance when studying the distribution of returns of an investment, as we will see in our next section.

Our set of observations $\mathscr{X}$ is made of n observations of a variable, and we assume these observations Y_i are independent drawing of a random variable Y with normal distribution with mean μ and variance σ^2.

$$\mathscr{X} = \{Y_i, 1 \leq i \leq n\}$$

Our null assumption H_0 is that μ is equal to a given level μ_0: we want to statistically assess the validity of such hypothesis. An intuitive solution consists in computing the mean $\bar{Y}_n$ of our observations, insofar as we know this mean should converge towards the mean of the distribution (law of large numbers). We then set a rejection zone of kind:

$$\mathscr{R}(\mathscr{X}) = \{|\bar{Y}_n - \mu_0| \geq c(\alpha)\}$$

where we determine $c(\alpha)$ in accordance with the desired level of α. However, even though this way of doing seems rather simple, it is also possible to present a second approach, which is more widely used when it comes to practice.

Indeed, we can define the following t-statistic for our problem:

$$\frac{\sqrt{n}(\bar{Y}_n - \mu_0)}{s_n}$$

where s_n is the unbiased standard deviation estimator, i.e.

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_n\right)^2$$

Theoretical results gives us that such t-statistics, under the null hypothesis, follows a Student law with n-1 parameters, which is a well-known law. So it is possible to define the following rejection zone:

$$R(\mathscr{X}) = \left\{\left|\frac{\sqrt{n}(\bar{Y}_n - \mu_0)}{s_n}\right| > c(\alpha)\right\}$$

where the value of $c(\alpha)$, as already mentioned, is determined thanks to the quantiles of the t-statistic distribution.

Similarly, we can compute the p-value of our t-statistic. For instance, if we consider it is important to assess what may happen at both sides of the distribution, we could compute, where T is a Student distribution with $n-1$ parameters, and t is the t-statistic value:

$$\mathbb{P}\left[|T| > |t|\right]$$

2 Multiple Testing In The Context Of Backtesting and Strategy Discovery

2.1 Sharpe Ratio And Simple Tests

Let us assume we devise an investment strategy, which generates a set of realized returns over a period of length T_h. We denote r_t the return between times $t-1$ and t, for $1 \leq t \leq T_h$. Simply having a look at the set of past returns is the first idea that may cross someone's mind: if the path of returns looks good, one may be inclined to believe that the strategy is also good. However, we would like to formalize the assessment thanks to a statistical test.

So, the purpose of our test is to check whether the strategy is able to generate true profits in the future. We want our statistical test to tell us if the returns of our strategy are different from 0. Our null hypothesis (H_0) is that the historical returns follow a normal distribution with mean μ and variance σ^2, that those returns are independent and identically distributed, and that $\mu = 0$. If the null assumption is rejected, we admit we have found a strategy able to generate non-zero returns; otherwise we disregard the strategy.

A mentioned in the final subsection of our first part, a natural t-statistic is the following:

$$\frac{\sqrt{T_h}\bar{\mu}}{\bar{\sigma}}$$

where $\bar{\mu}$ is the estimation of the average returns based on the set of observed historical returns, i.e.:

$$\bar{\mu} = \frac{\sum_{i=1}^{T_h} r_t}{T_h}$$

and $\bar{\sigma}$ is the unbiased estimation of the standard deviation of the set of observed historical returns, i.e.:

$$\bar{\sigma} = \sqrt{\frac{1}{T_h - 1}\sum_{i=1}^{T_h}(r_t - \bar{\mu})^2}$$

Under the assumption H_0, the t-statistic, seen as a distribution, follows a Student law with $T_h - 1$ parameters. Therefore, it is straightforward to assess whether the null assumption must be rejected or not, for example by computing a p-value. For instance, if the p-value is above 0.05, we do not reject the null assumption: if the average of the returns is close to 0, this scenario is one which falls in the category of the 95% outcomes.

It also worth noticing that the Sharpe ratio is closely related to our t-statistic. Indeed, in our case, the Sharpe ratio is simply:

$$SR = \frac{\bar{\mu}}{\bar{\sigma}}$$

So the Sharpe ratio is the product of the t-statistic value with $\sqrt{T_h}$. Thus, for a given period length T_h, a higher t-statistic value is equivalent to a higher Sharpe ratio, then to a lower p-value. The lower the p-value, the higher the chance of rejecting the null assumption, the higher the significance of the investment strategy.

However we have only presented the single test scenario so far. In the next subsection, we introduce the multiple testing issue.

2.2 Sharpe Ratio and Multiple Tests

We now set forth a statistical framework to handle the multiple tests question when devising investment strategies. The use of the Sharpe ratio may be misleading due to the extensive data mining made by practitioners from a limited set of historical data. It is then possible to discover profitable strategies which are not as good as expected in real life. In this subsection, our purpose is to show how to adjust the Sharpe ratio to take into account data mining.

To begin with, we resume from where we stood at the end of our previous subsection: given a set of T_h historical returns, it is possible to test the null assumption of i.i.d, normal returns with mean 0 by computing the p-value:

$$p^S = P\left[|T| > |t|\right]$$

The capital S is here to dwell on the fact that this is the p-value from a single test. T is a student law with $T_h - 1$ parameters, and t is our t-statistic value.

As of now, here is the situation: we assume that the researcher has tried, not only one strategy, but N different strategies, and he or she wants to choose the most profitable one. To do so, when backtesting his/her N strategies, he/she computes the Sharpe ratio for each one of them, and he/she then keeps only the strategy with the highest ratio.

The question is: are we sure that the chosen strategy is really good? Indeed, the finding of this strategy is only the conclusion of a process made of multiple tests. It is possible that, by chance, we finally end up with a great strategy, but on paper!

We formalize this question in the following manner. t is now the t-statistic value of the chosen strategy. The null hypothesis H_0 is that none of the N strategies can generate non-zero returns, and that the N t-statistics distribution, denoted T_i, for each strategy are independent.

We can now define the p-value p^M of our multiple tests situation: it is the probability

$$p^M = \mathbb{P}\left[\max_{1\leq i\leq N} |T_i| \geq |t|\right]$$

where T_i are independent Student distribution with $T_h - 1$ parameters. Since the T_i are independent, p^M can be rewritten:

$$p^M = 1 - \prod_{i=1}^{N} \mathbb{P}\left[|T_i| < |t|\right]$$

$$= 1 - (1 - p^S)^N$$

Mathematically, we directly see that, when N grows bigger, then the p-value p^M tends towards 1.

A numerical example may help the reader visualize the impact of multiple testing: let us imagine that the single p-value p^S is worth 0.049. If we consider the chosen strategy and see it as a unique strategy, with no attention to the previous work of "testing" N strategies, the p-value is under the α level of 0.05. We confidently reject the null hypothesis: our strategy is deemed profitable.

However, if we now consider the multiple testing p-value p^M, it is worth 0.3949, which is far above the α level. The null hypothesis is not rejected, and the strategy is deemed to have zero returns.

This approach is interesting since it provides us with a mean of determining a correct haircut for the Sharpe ratio, instead of an exogenous value such as 50%. To do so, we equate p^M, which is known, with what the p-value of a single test would be:

$$p^M = \mathbb{P}\left[|T| > |t|\right]$$

The crucial element here is to bear in mind that our t-value is equal to the Sharpe ratio multiplied by the square root of T_h, the number of observations. Since we want to find the haircut coefficient for the Sharpe ratio, we replace the Sharpe ratio by a haircut Sharpe ratio: $HSR = \gamma \times SR$. So:

$$p^M = p^M = \mathbb{P}\left[|T| > \left|HSR\sqrt{T_h}\right|\right]$$

It is then possible to numerically determine the coefficient γ. Let us assume we have 20 years of monthly returns, meaning that $T_h = 240$, , and that an annual Sharpe ratio of 0.75 leads to a p-value of 0.0008 for a single test. For $N = 200$, $p^M = 0.15$. The adjusted Sharpe ratio must be 0.32 according to the above equation. So carrying out 200 tests leads to a reduction of the original Sharpe ratio by approximately 60%.

As we have seen in this subsection, the single p-value is no longer helpful when it comes to assessing the statistical significance of the strategy. The multiple testing p-value p^M is a more appropriate measure.

However, we have so far worked in the simplest case, when all the statistical tests are independent. In real life, this is not the case; in the next subsection we set forth a framework to handle properly the p-value adjustment in the case of non-independent tests.

2.3 P-Value Adjustment: Bonferroni And Holm Methods

In this part, we now assume we want to test M assumptions, denoted H_i for $1 \leq i \leq M$; each one leads to a p-value. So we have a set of M p-values, denoted p_i for $1 \leq i \leq M$. Among the M assumptions which are tested, R are rejected: they can be rejected either for a good reason (the assumption H_i is false), or for a bad reason (false positive). In finance, M would be the number of tested strategies: for each one of them we want to test an assumption on its ability to generate non-zero returns.

We denote N_r the total number of false positives. In our case of financial strategies, it would correspond to the number of strategies which are deemed profitable when they are not, since the null assumption H_i is, as above, that strategy i displays zero returns. What is called the family-wise error rate, denoted FWER, is the probability of finding at least one false positive:

$$FWER = \mathbb{P}\left[N_r \geq 1\right]$$

The FWER is a generalization of the type I error, set forth in the first section of this paper.

The Bonferroni [4] and Holm [5] [6] methods are a way of adjusting the p-values in the case of FWER. To do so, we sort the M p-values by increasing order:

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$$

For the sake of clarity, p-value $p_{(i)}$ is associated with the test of hypothesis $H_{(i)}$.

Bonferroni method consists in adjusting each p-value equally, by multiplying all of them by the number of tests:

$$p^B_{(i)} = \min\left(M \times p_{(i)}; 1\right)$$

So, if we observe 10 strategies and one of them displays a p-value of 0.05, the adjusted p-value is equal to 0.5, meaning that the strategy is not significant at 50%.

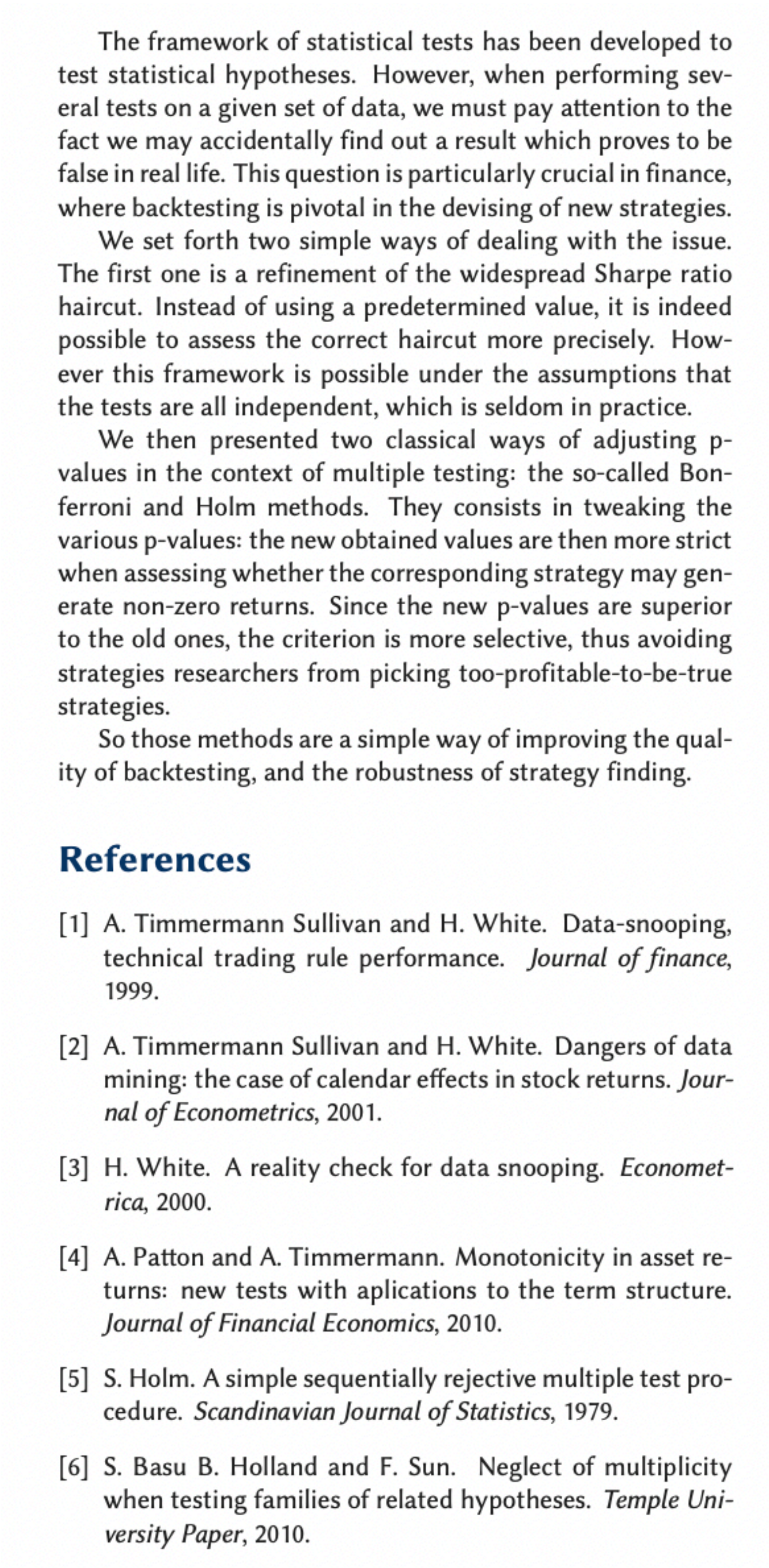
Holm's method consists in adjusting the p-values sequentially. To do so, we compute, for i from 1 to M:

$$p^H_{(i)} = \min\left(\max_{j\leq i}\left((M - j + 1)p_{(j)}\right); 1\right)$$

In both cases, we see that for each strategy, we increase the p-value, meaning that we are more selective when checking whether one of them deserves to be considered as a profitable strategy.

Conclusion

In this paper, we provide the reader with some insight into the issue of backtesting and multiple testing. Carrying out many tests on the same set of data is plagued with many biases, which must be taken into account when devising financial strategies.



The framework of statistical tests has been developed to test statistical hypotheses. However, when performing several tests on a given set of data, we must pay attention to the fact we may accidentally find out a result which proves to be false in real life. This question is particularly crucial in finance, where backtesting is pivotal in the devising of new strategies.

We set forth two simple ways of dealing with the issue. The first one is a refinement of the widespread Sharpe ratio haircut. Instead of using a predetermined value, it is indeed possible to assess the correct haircut more precisely. However this framework is possible under the assumptions that the tests are all independent, which is seldom in practice.

We then presented two classical ways of adjusting p-values in the context of multiple testing: the so-called Bonferroni and Holm methods. They consists in tweaking the various p-values: the new obtained values are then more strict when assessing whether the corresponding strategy may generate non-zero returns. Since the new p-values are superior to the old ones, the criterion is more selective, thus avoiding strategies researchers from picking too-profitable-to-be-true strategies.

So those methods are a simple way of improving the quality of backtesting, and the robustness of strategy finding.

References

[1] A. Timmermann Sullivan and H. White. Data-snooping, technical trading rule performance. *Journal of finance*, 1999.

[2] A. Timmermann Sullivan and H. White. Dangers of data mining: the case of calendar effects in stock returns. *Journal of Econometrics*, 2001.

[3] H. White. A reality check for data snooping. *Econometrica*, 2000.

[4] A. Patton and A. Timmermann. Monotonicity in asset returns: new tests with aplications to the term structure. *Journal of Financial Economics*, 2010.

[5] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 1979.

[6] S. Basu B. Holland and F. Sun. Neglect of multiplicity when testing families of related hypotheses. *Temple University Paper*, 2010.

Nous sommes Canopee, un cabinet de conseil indépendant, spécialiste en Finance, DATA et transformation digitale.

Nous sommes l'*empowering ecosystem* ! Un écosystème en perpétuel mouvement.
Un écosystème qui donne le pouvoir à nos collaborateurs, nos projets et nos clients de se nourrir de l'émulation collective pour lui donner de la force.

Depuis 2009, nous intervenons dans les secteurs de la BFI de l'Asset Management, la banque privée ou de détail, les services financiers et l'assurance et cela auprès de clients tels que SG, HSBC,BNP Paribas ou encore ALLIANZ GI et AXA IM. Des écosystèmes réglementaires et spécifiques qui ont été nos premiers terrains de jeu.

C'est ici que nous avons su développer notre agilité, nos compétences, notre sens de l'engagement. Aujourd'hui, nous grandissons et continuons à nous investir et nous engager sur des écosystèmes spécifiques, techniques, mais diversifiés tels que l'industrie, le retail ou encore la pharma.
Toujours autour de nos 3 expertises que sont la finance, la data et la transformation digitale.

canopee
empowering *ecosystem*

www.canopee-group.com