CrossMark

Sense, nonsense and the S&P500

L. C. G. Rogers[1] ⓘD

Abstract
The theory of financial markets is well developed, but before any of it can be applied
there are statistical questions to be answered: Are the hypotheses of proposed models
reasonably consistent with what data show? If so, how should we infer parameter values
from data? How do we quantify the error in our conclusions? This paper examines
these questions in the context of the two main areas of quantitative finance, portfolio
selection and derivative pricing. By looking at these two contexts, we get a very
clear understanding of the viability of the two main statistical paradigms, classical
(frequentist) statistics and Bayesian statistics.

Keywords Bayesian statistics · Frequentist statistics · Derivative pricing · Hedging

JEL Classification C11 · C18 · C58

1 Introduction

If S_t^i denotes the price of asset i ($i = 1, \ldots, d$) at the end of day t, then a very common
modelling assumption is that

$$X_t \sim N(\mu, V), \tag{1}$$

where

$$X_t \equiv \left(X_t^1, \ldots, X_t^d \right) \equiv \left(\log \left(S_t^1 / S_{t-1}^1 \right), \ldots, \log \left(S_t^d / S_{t-1}^d \right) \right)$$

is the vector of day-t returns. Here, the mean μ and the covariance V are unknown,
but it is assumed that they are constant over time and that returns on different days
are independent. If we ask the very natural question, 'How should we invest in this
market?', then there is no shortage of answers to this question; commonly, we suppose
that we know some objective that we wish to optimize, that we know the parameter $\theta \equiv$

✉ L. C. G. Rogers
  chris@statslab.cam.ac.uk

[1] Statistical Laboratory, University of Cambridge, Cambridge, UK

(μ, V) of the return distribution, and then we do some analysis to derive an investment strategy that optimizes our objective, and which will depend on the parameter θ. Assuming that we know the objective to be optimized is innocent, because we are free to choose it, but assuming that θ is known is not, and this is where statistics comes in. So in Sect. 2, we start with returns distributed as (1) and see in more detail how this would lead to an investment strategy—which of course must depend on θ— and how we would make use of statistics to represent our knowledge about θ. Even in this simplest setting, the inconsistencies and impossibilities of classical statistics are immediately visible. In contrast, Bayesian statistics buys us a path free from all of these problems for the price of making subjective inputs. In the struggle for the soul of statistics, the Bayesian approach is usually attacked because the statistician has to make a subjective choice of prior for the parameter θ. In truth, the weakness of subjectivity happens before that, when we choose the *family* of models to use— making an assumption about a prior distribution over that family is a much smaller leap of faith. But if we recognize that, then this subjectivity affects classical statistics in exactly the same way—subjectivity is not a weakness only of Bayesian statistics!

So Sect. 2 gives us some kind of framework for answering the question, 'How should we invest in this market?' In Sect. 3, we look at the question we should have asked first, namely 'Are the modelling assumptions reasonable?'—in other words, can we suppose that returns are IID multivariate Gaussian? Not surprisingly, the answer is 'No'. However, as we shall see, this is not as bad as it seems, because simple transformations change the data into something that is reasonably like IID Gaussian, and the theory developed in Sect. 2 may actually be fairly relevant. We then take a look in Sect. 4 at investing in the S&P500 and see what some of these ideas give us.

The next section of the paper looks at derivative pricing, and once again the inconsistencies of classical statistics surface in a big way almost immediately. Once again, Bayesian statistics offers an escape.

2 Portfolio selection

Recall that we are assuming (1) that returns are IID multivariate Gaussian. We let $\mathcal{F}_t$ denote the σ-field of information at the end of day t. The portfolio investment decision requires us to choose portfolio h_t at end of day $t-1$ to hold for day t; then,

$$w_t = (1+r)(w_{t-1} - h_t \cdot \mathbf{1}) + h_t \cdot (S_t/S_{t-1}). \tag{2}$$

Here, $\mathbf{1}$ is the vector with all entries equal to 1, and h_t^i denotes the dollar amount invested in the ith asset for day t, an $\mathcal{F}_{t-1}$-measurable random variable. Solving a problem with a multi-period objective involves dynamic programming of some form, which is rarely amenable to closed-form solution, so we just focus for now on a single-period objective, which already illustrates the issues: we aim to find

$$\sup_{h_t} E[U(w_t)|\mathcal{F}_{t-1}], \tag{3}$$

where U is C^2 strictly concave and increasing. In order to calculate this objective, we need to know the conditional law of X_t given $\mathcal{F}_{t-1}$, which is of course $N(\mu, V)$—but what are μ, V?

2.1 What does classical statistics say?

The classical statistical paradigm is that the value of $\theta \equiv (\mu, V)$ is fixed but not known—so within that paradigm, we are *unable* to compute objective (3), because we do not know θ—and if we cannot compute the objective, we certainly cannot optimize it! The classical statistician would respond that observation of the data informs us about the possible values of θ, which would allow us to exclude values of θ which are poorly supported by the data. So after observing returns for some time, we would have some confidence set C for the values of θ, and we might then propose some minimax version of the original objective:

$$\sup_{h_t} \inf_{\theta \in C} E_\theta[U(w_t)]. \tag{4}$$

Now this might work for the simplest examples from Statistics 101, such as univariate Gaussian data with known variance, where we would be perverse to take the confidence set C to be anything other than some interval symmetric about the sample mean, but for multivariate Gaussian data, there is no obvious choice for C. We might try to exclude values of θ that are 'extreme' in some sense, but the definition of 'extreme' requires us to choose some statistic, and it is hard to see what we could pick here; how would we say that a covariance matrix V was too extreme given observations $X_1, \ldots, X_N$? Even if we could answer that, trying to calculate (4) is in general computationally infeasible, given that the set C (even if we could say what it was) is a subset of some high-dimensional Euclidean space. So in practice the classical approach to statistics is used by calculating some *estimator* $(\hat{\mu}, \hat{V})$ and pretending that these are the true values. This is not the case of course, but since it is so hard to quantify the error being made by this assumption, the preferred response in practice appears to be to ignore it.

So just by thinking briefly about classical statistics in the context of the portfolio selection problem (3), we see that *it just cannot work!*

2.2 What does Bayesian statistics say?

Bayesian statistics also treats the parameter θ as unknown, but proposes that it has a *known* distribution π_0. Then, after seeing $X_1, \ldots, X_t$, the distribution of θ has evolved to

$$\pi_t(d\theta) \propto \pi_0(d\theta) \prod_{s=1}^{t} f(X_s; \theta)$$

$$\propto \frac{\pi_0(d\theta)}{(\det V)^{t/2}} \exp\left\{-\tfrac{1}{2}\, t(\mu - \bar{X}_t) \cdot \tau(\mu - \bar{X}_t) - \tfrac{1}{2}\, t\ \mathrm{tr}\!\left[\tau \hat{V}_t\right]\right\}, \tag{5}$$

where $\tau \equiv V^{-1}$ is the precision matrix, and

$$\hat{\mu}_t = \bar{X}_t \equiv t^{-1} \sum_{s=1}^{t} X_s, \qquad (6)$$

$$\hat{V}_t = t^{-1} \sum_{s=1}^{t} (X_s - \bar{X}_t)(X_s - \bar{X}_t)^T. \qquad (7)$$

Now we *can* approach the optimization (3) of the objective because we really do know the law of X_t given $\mathcal{F}_{t-1}$—the law of $\theta = (\mu, V)$ is given by (5), and conditional on θ the law of X_t is $N(\mu, V)$. So all of the difficulties of the classical approach evaporate, provided we are willing to make the subjective choice of the prior π_0.

More has been said about the choice of the prior than we could ever summarize, but my view is that this is a relatively innocent subjective choice. In practice, one would run the analysis for a number of widely different priors as a diagnostic; if the answers are broadly similar, then the choice of prior was not particularly critical, and if the answers vary a lot, then we learn that there was not so much information in the data, again useful to know. A far more important subjective choice, already mentioned, is the choice of the family of models allowed.

As we shall see, taking a Bayesian view deals completely with all the theoretical aspects of statistical inference, but the price we end up paying is that the computational aspects become a lot more onerous.

3 Are S&P500 returns IID Gaussian?

We continue to illustrate the themes of this paper by simplifying the model (1) we began with to one asset, the S&P500 index. The model assumption is that the daily returns are IID Gaussian, but are they?

3.1 Are returns identically distributed?

Figure 1 plots the daily returns of the S&P500 from 1 July 1954 to 4 May 2017, and just looking at this plot, we would not believe that the returns are IID; there are obvious periods of higher and lower volatility, which would not happen if the returns were IID. Another plot which shows this quite clearly is to plot the cumulative sum of the squared returns (the realized quadratic variation), which we see in Fig. 2.

If the returns were IID, we should expect to see a plot that goes up roughly as a straight line, and this is obviously not the case. However, we can transform the returns into something much closer to IID by the simple trick of *vol rescaling*, which goes like this.

Make an initial estimate of the volatility of the returns series by choosing some integer N and calculating

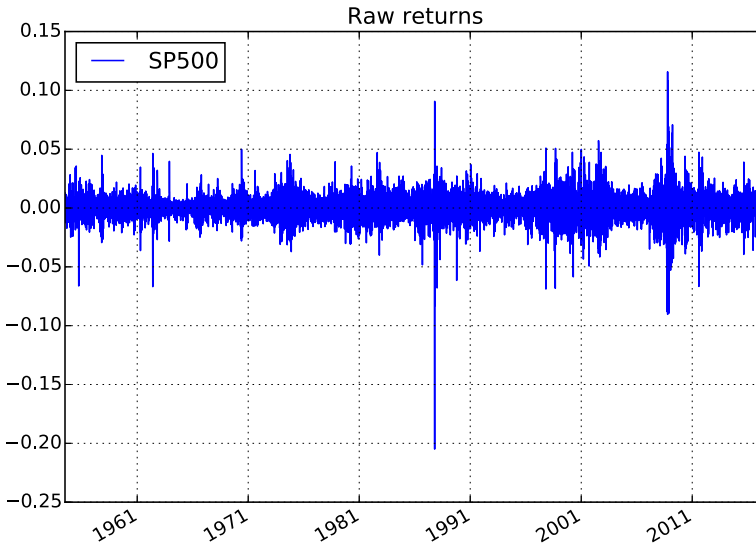$$\hat{\sigma}_0^2 = N^{-1} \sum_{t=1}^{N} X_t^2. \qquad (8)$$
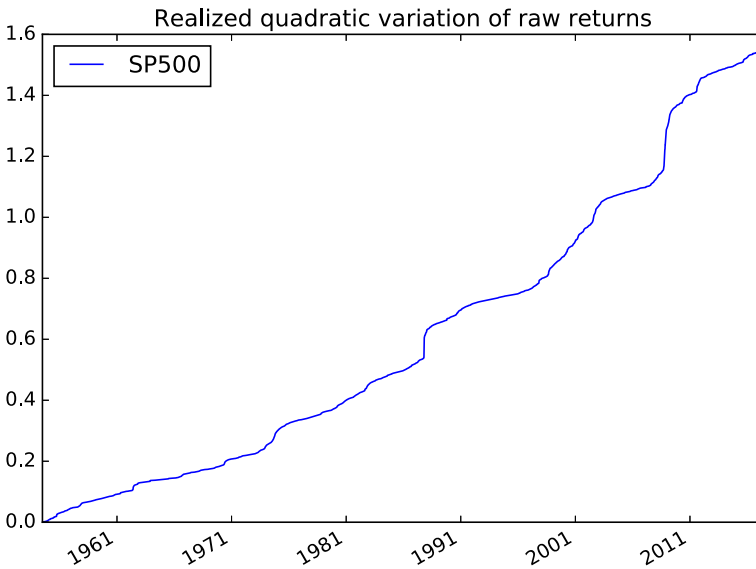
Fig. 1 Raw returns of the S&P500

Fig. 2 Cumulative sum of squared returns of the S&P500

Then, update recursively, starting at $t = 0$:

$$Y = \max\{-K\hat{\sigma}_t, \ \min\{K\hat{\sigma}_t, X_t\}\}$$
$$\hat{\sigma}_{t+1}^2 = \beta_W Y^2 + (1 - \beta_W)\hat{\sigma}_t^2$$
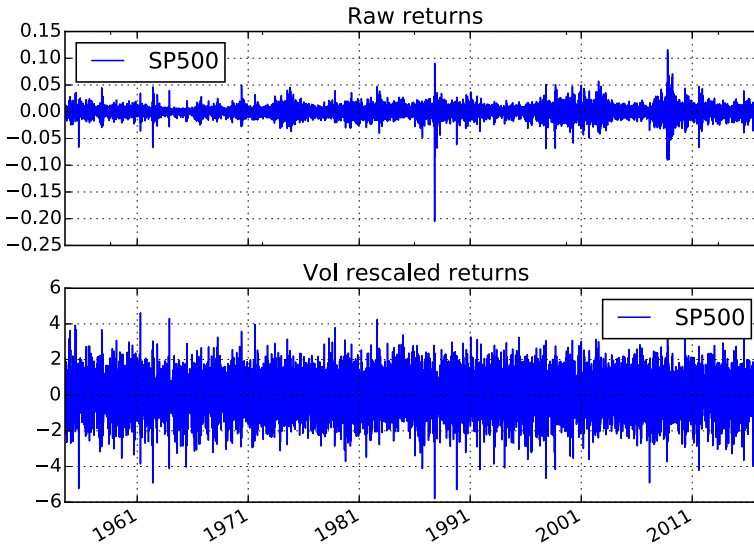$$\tilde{X}_{t+1} = X_{t+1}/\hat{\sigma}_{t+1}$$

Fig. 3 Raw and rescaled returns of the S&P500

Here, K is some cut-off constant ($K = 4$ would do) whose purpose is to prevent occasional very large returns from impacting the running vol estimate $\hat{\sigma}_t$ too much. The exponential weighting parameter $\beta_W \in (0, 1)$ smooths the vol estimates; in the calculations of this paper, it was taken to be 0.025. This corresponds to a mean lookback of 40 days, roughly 6 weeks—not too long, not too short. Other values could be used of course. Once we do this, the plots of the rescaled returns and the cumulative sum of rescaled returns are shown in Figs. 3 and 4.

These plots show that the rescaled returns look quite time homogeneous.

3.2 Are returns Gaussian?

The classical diagnostic for a Gaussian distribution is to take the sample and make a q–q plot of it. When we do this, we see Figs. 5 and 6. The first is quite close to a straight line in $[-2, 2]$, which includes more than 90% of the range of the standard Gaussian, but the second is quite close to a straight line in $[-3, 3]$ (which includes 99.9% of the standard Gaussian), so this looks closer to Gaussian. We cannot expect a perfect straight line, because there are going to be days when something big happens and the return on those days will be out of line with the usual behaviour, but if we have a story that is good for 999 out of 1000 days, this is saying that an unusual day happens roughly once every four years, which seems plausible.

3.3 Are returns independent?

If returns are independent, then when we plot the autocorrelation function (ACF) of the time series of returns, we should see something that is essentially zero for all
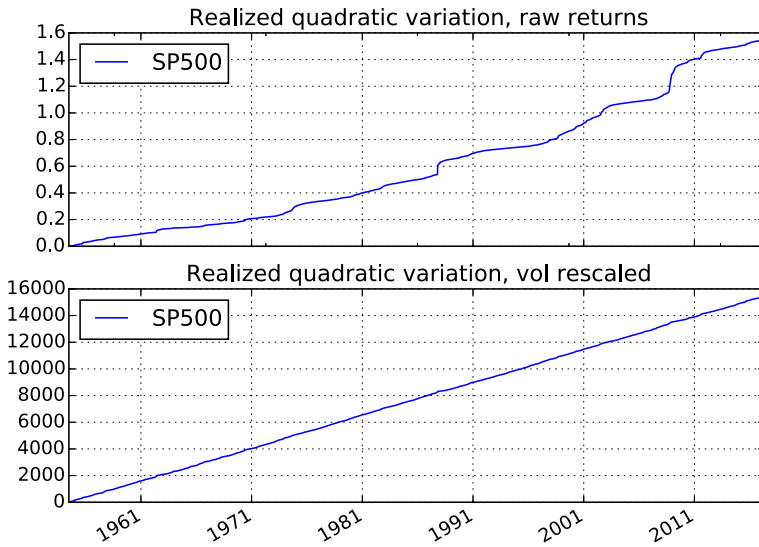
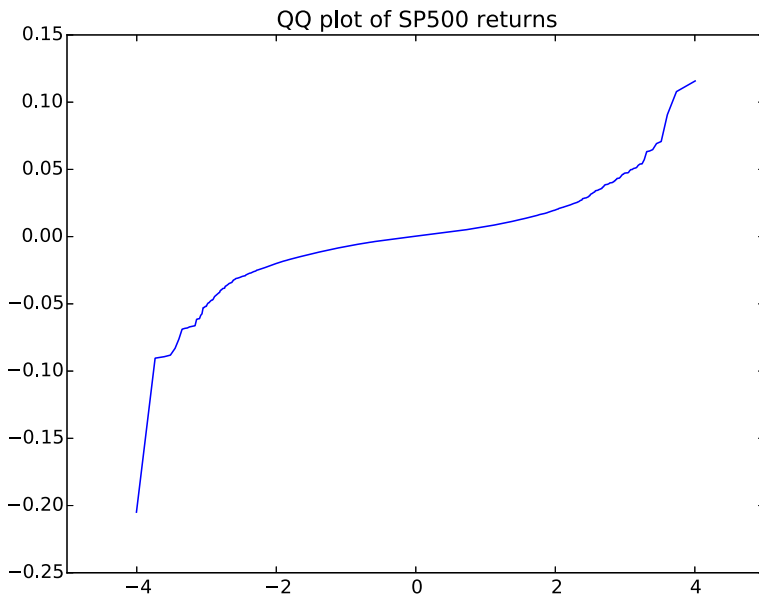**Fig. 4** Cumulative sum of squared returns and squared rescaled returns of the S&P500

Fig. 5 q–q plot of the returns of the S&P500

positive lags. The same should be true when we plot the ACF of *absolute* returns. The corresponding plots for the raw and rescaled returns of the S&P500 are shown in Figs. 7 and 8 and are entirely typical of what these plots show. The ACF of raw returns and of rescaled returns is essentially zero at all positive lags, but the ACF of *absolute* raw returns remains positive for many lags, which is explained by the fact that the
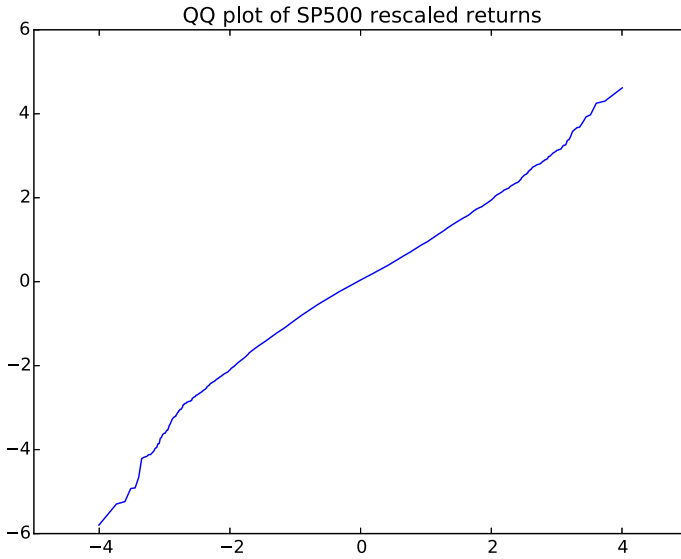
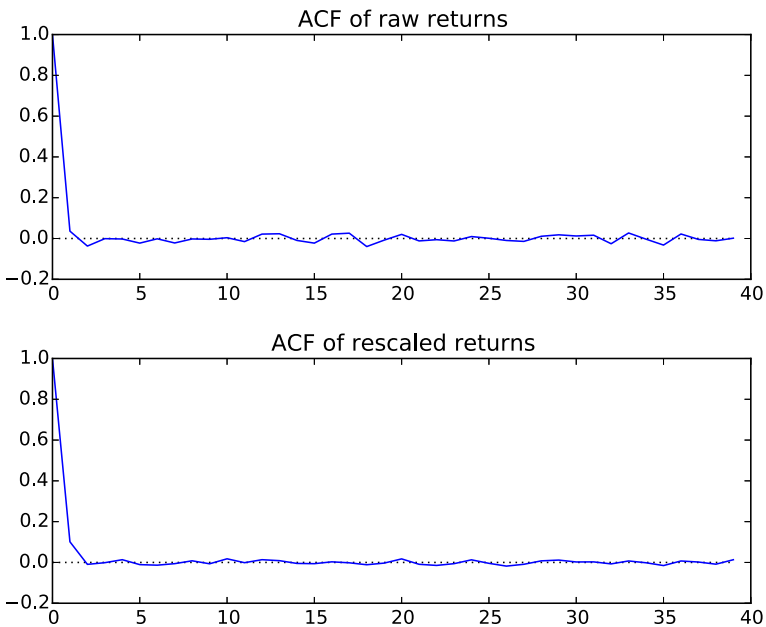**Fig. 6** *q–q* plot of the rescaled returns of the S&P500

Fig. 7 ACF of the raw and scaled returns of the S&P500

raw returns exhibit volatility clusters, with big returns coming together. Interestingly though, the ACF of the rescaled absolute returns is close to zero at all positive lags. This is a necessary (but not sufficient) condition for the returns to be independent.
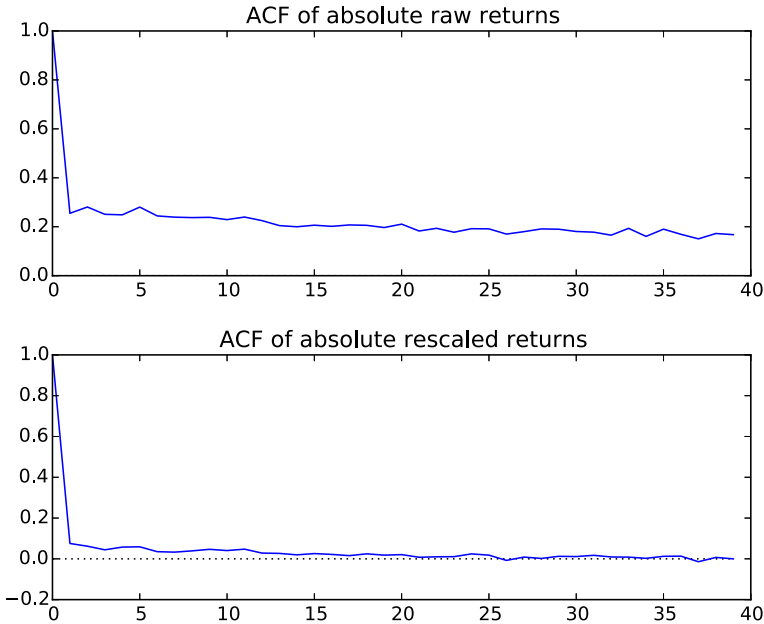
Fig. 8 ACF of the raw and rescaled absolute returns of the S&P500

So to summarize, on the basis of these simple exploratory analyses, we may make the working hypothesis that *the vol-rescaled returns are IID Gaussians.* For multivariate return data, we may need to be more circumspect, but for this univariate return series, we can suppose that the rescaled return series are IID Gaussian, with variance equal to 1. Our interest then focuses on understanding the mean μ, which we expect to be quite small relative to the variance (otherwise, it would be a simple matter to generate huge profits from investment). In practical terms, we can take the original return data and rescale them, treating the rescaled return data as if they were the actual returns; because our portfolio analysis will tell us each day how many units of rescaled asset we should hold through tomorrow, and then we can immediately work out how many units of the original asset we need to hold. For portfolio selection problems then, broadly speaking

non-constant vol does not matter, non-constant mean returns do.

4 How well does Bayesian model averaging work?

If we invested a constant \$1 in the S&P500 over the 63 years of data used above, then the Sharpe ratio is 23.39%. If we invest a constant \$1 in the vol-rescaled returns, we get a slight improvement to a Sharpe ratio of 26.14%. These are rather primitive strategies however. In a Bayesian model averaging, we take some finite family of J models, each of which makes an assumption about the conditional distribution of X_t given $\mathcal{F}_{t-1}$, and we let Bayes' theorem update the posterior distribution over the models. In more
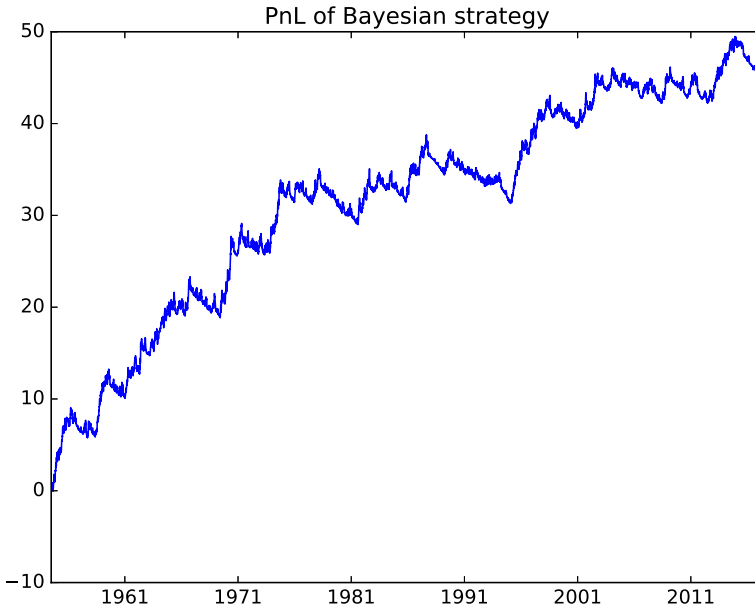
Fig. 9 P&L from simple Bayesian model averaging

detail, if model j says that[1]

$$\tilde{X}_t | \mathcal{F}_{t-1} \sim N(\mu_t(j), 1) \tag{9}$$

then the updating of the posterior probabilities is

$$\pi_t(k) \propto \sum_j \pi_{t-1}(j) \, p_{jk} \, \gamma(\tilde{X}_t - \mu_t(k)), \tag{10}$$

where γ is the standard Gaussian density. The trading strategy comes from a simple myopic rule, where we choose h_t to maximize (3), where $U(x) = -\exp(-x)$. To keep the story realistic, we assume there are proportional transaction costs $\varepsilon |h_t - h_{t-1}|$ when we switch positions, so there will be situations where the cost of switching exceeds the gain in utility, and we therefore choose not to switch portfolio.

In Fig. 9, we see the P&L generated when we take just two models, the first of which thinks that $\mu_t(1) = 0.15$ for all t and the second of which thinks that $\mu_t(2) = -0.15$ for all t—in other words, the index is either growing at 15% per annum, or shrinking at 15% per annum. Taking the transactions costs to be 3 bp, we find that the Sharpe ratio of the strategy is 41.00%, substantially higher than the two constant-dollar strategies. The P&L shown in Fig. 9 displays relatively little drawdown. The positions shown in Fig. 10 fluctuate between -0.15 and 0.15, the extremes we would expect if the posterior probabilities were at their extreme values. It is interesting to see that periods

---

[1] We suppose that the variance is 1, since we have done volatility rescaling.
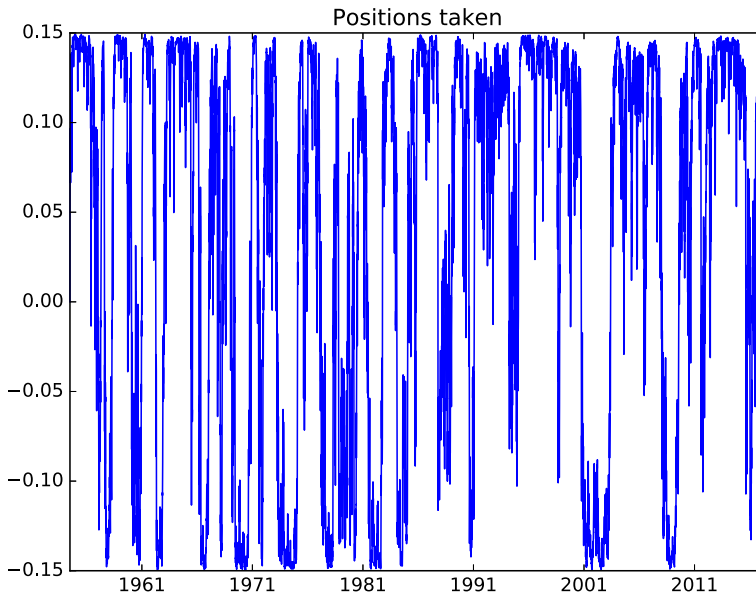
Fig. 10 Positions in simple Bayesian model averaging

when the position is strongly negative, such as 1973–1974, 2001–2003, 2008–2010, correspond to periods when the global economy was under significant stress.

This looks like (and is) an impressive demonstration of the power of Bayesian modelling techniques. But it is worth underlining that some cherry-picking has been going on here; if we include a third model into the comparison which says $\tilde{X}_t | \mathcal{F}_{t-1} \sim N(0, 1)$, then the same analysis leads to a Sharpe ratio of 30.18%. Changing the various parameters of the model can make a big difference to the conclusion, and we need to be aware of this; searching around for a 'sweet spot' is a form of data snooping. We would be outraged if someone proposed a trading strategy that needed to know all future returns, but if we search for 'good' parameter values in some parametric model, we are in effect making use of information about the entire future evolution of returns, even if the individual model selected at the end does not. I have seen this done in practice; a model gets adopted and then fails to deliver the returns that historical analysis gave. It is good practice to leave several years of data locked up until the model has been chosen, and then see what happens once those data are unlocked— out-of-sample testing. Even so, the future may not cooperate.

5 Derivative pricing

When it comes to derivative pricing, we work in the pricing measure in which the growth rate of the asset is replaced by the riskless rate. The great industry of the implied volatility surface shows that the non-constancy of the volatility is a very important matter in derivative pricing, so in contrast to the situation with portfolio selection,

non-constant mean returns do not matter, non-constant vol does.

Some derivatives are very liquid, so their prices are taken to be the market prices—any model should match those prices very closely, if not perfectly. More exotic derivatives on the other hand are made to order, and there is no market price, so the price has to come from some (parametric) model, as a function of observable state variables X_t and unobserved parameters θ. The parameter θ of the model will not be known, so we have to carry out some statistical procedure to identify it, and as with portfolio selection, there are the two main paradigms to consider.

5.1 What does classical statistics say?

The conventional model calibration procedure of the industry takes the prices Y_t^a, $a = 1, \ldots, A$ of some liquid derivatives and compares those to the model prices $\varphi^a(X_t, \theta)$. Then, some 'best-fitting' choice θ_t^* of the parameter is found by solving

$$\inf_\theta \sum_a |Y_t^a - \varphi^a(X_t, \theta)|^2. \tag{11}$$

Then, the price of some exotic is calculated by assuming that $\theta = \theta_t^*$. There are various issues with this approach, some more important than others.

1. The model prices $\varphi^a(X_t, \theta)$ may not exactly match market prices Y_t^a.
2. Tomorrow we recalibrate and arrive at a value θ_{t+1}^*—so how do we mark-to-market and hedge a derivative that we sold on day t? Using $\theta = \theta_t^*$? Using θ_{t+1}^*? Using some other θ value?
3. Would some other model be 'better'?
4. θ_t^* is an estimate—what account do we take of estimation error?

The first point need not cause insuperable problems, because market prices are not always taken simultaneously, and the market price is in any case a bid-ask spread, so exactly fitting some ideal value is not essential. But the second issue is very real—the derivative priced and sold on day t was calculated on the assumption that θ_t^* was the true parameter value, unchanging for all time, and yet on the very next day, we abandon that assumption by saying that $\theta = \theta_{t+1}^*$! This is a fundamental inconsistency of the calibration approach. The third point cannot be answered in this framework, because no models outside the chosen parametric model are admitted. The fourth point is again unanswerable in most situations, because of the difficulty of specifying a confidence set and searching over it for extremes; so the estimation error is either ignored completely or treated in a very crude manner.

So overall *the conventional calibration approach is inconsistent and cannot account for estimation error.*

5.2 What does Bayesian statistics say?

In the Bayesian approach, we choose and fix a finite set of J models: under model j, the underlying state process X is Markovian, with transition density

$$p_j(x, x') = P_j(X_h \in dx'|X_0 = x)/dx' \quad (j = 1, \dots, J), \tag{12}$$

where $h > 0$ is the time step. As before, we give ourselves some prior distribution $\pi_j(0)$ over the possible models. Model j has pricing function $\varphi_j^a(\cdot)$ for derivative a. We select some loss function $Q(\varphi_j(X), Y_t)$, which for the sake of the discussion we might take to be

$$Q(y, y') = \alpha\|y - y'\|^2 \tag{13}$$

for some $\alpha > 0$. The log-likelihood $\ell_j(t)$ of model j at time t then updates as

$$\ell_j(t) = \ell_j(t - h) + \log p_j(X_{t-h}, X_t) - Q(\varphi_j(X_t), Y_t). \tag{14}$$

In practice, it is a good idea to allow the data-generating model to change with a small probability each period, according to some Markov chain with transition matrix P. This prevents the Bayesian inference from getting stuck at some long-term average values as the number of time steps increases, and reflects a natural requirement that data from the distant past should have less influence on our inference than more recent data. The posterior distribution then updates as

$$\pi_j(t) \propto \sum_k \pi_k(t - h)\, p_{kj}\, \exp(\ell_j(t)). \tag{15}$$

Now everything is easy:

- The law of X_{t+h} conditional on $\mathcal{F}_t$ has density $\sum_j \pi_j(t)\, p_j(X_t, \cdot)$;
- If model j gives the price of an exotic to be ξ_j, then take the overall price to be

$$\bar{\xi} \equiv \sum_j \pi_j(t)\, \xi_j, \tag{16}$$

  the posterior mean;
- What is the error in $\bar{\xi}$? It is the mean of a discrete distribution over the values ξ_j with weights $p_j(t)$, so we know the variance and all other moments;
- If model j gives delta hedge[2] H_j, then to first order we have a delta hedge given by $\sum_j \pi_j(t) H_j$.

If we revisit the issues that were problematic for the classical approach, we have answers:

1. The model prices $\varphi^a(X_t, \theta)$ may not exactly match market prices Y_t^a. *The Bayesian approach does not say that the prices* must *be any particular value—it says that any price is a random variable whose distribution we know completely.*

---

[2] That is, a hedge which to first order cancels out the effect of moves of the underlying.

2. Tomorrow we recalibrate and arrive at a value θ^*_{t+1}—so how do we mark-to-market and hedge a derivative that we sold on day t? Using $\theta = \theta^*_t$? Using θ^*_{t+1}? Using some other θ value? *At all times, the price from the Bayesian approach* is *the posterior mean of the price—there is no inconsistency;*

3. Would some other model be 'better'? *Other models can be compared simply by adding them to the universe of models in the Bayesian comparison;*

4. θ^*_t is an estimate—what account do we take of estimation error? *Nothing is estimated in the Bayesian approach.*

At this point, it might appear that the Bayesian approach to inference deals triumphantly with all the conceptual difficulties and inconsistencies of the classical approach, which it does. However, this is not to say that all problems have been eliminated, and in fact there remain very considerable difficulties in applying the Bayesian methodology effectively, to do with computation. To apply the Bayesian approach in the way we have just described requires us in the first place to make a choice of the finite family of models considered, and this is the major issue. If we were only going to consider a one-parameter family of models, we could select a finite set of parameter values (perhaps just a few thousand) which effectively cover the parameter space, and the computational analysis will run ahead with no issues. But if we were looking at a family of models indexed by some parameter $\theta \in \mathbb{R}^8$, then it will be hard to distribute even one million points in the parameter space in such a way as to cover reasonably effectively, and at this point the computational Bayesian method starts to struggle. We are talking here about particle filtering (also known as sequential Monte Carlo), and although much effort has in the last 30 years been directed towards doing this well, it remains far from a finished technology. All manner of variants of the basic approach have been proposed—more than we could possibly begin to summarize here—which just goes to show that obvious general implementations must often fail.

6 Summary

This survey has taken a look at how statistical methodology helps in the analysis of financial asset returns, whether for portfolio selection or for derivative pricing. The conclusion is that statistics helps up to a point, but falls far short of what we would like to be able to do. The classical paradigm is an unworkable conceptual framework for studying data; its shortcomings may be hidden when we look at experimental data from the physical sciences, where the signal-to-noise ratio is much smaller than in financial data, but once we try to use it in finance and economics, it simply fails. Nevertheless, the methods of classical statistics provide very useful exploratory tools; if we were given ten years of daily returns on 2000 assets, we would almost certainly begin by calculating sample mean returns, and the sample covariance matrix, then we might try to pull out some principal components. Such calculations would very quickly tell us stylized facts of the data and direct our attention to questions of interest.

Hopefully, this article has well made the point that if we want a statistical methodology that is consistent, then it has to be Bayesian. Sadly, when it comes to trying to use Bayesian statistics in practice, the computational challenges quickly become overwhelming. Nevertheless, with patience and computational resources, we can make

progress. As always, choosing very simple models pays off, and it is here that some judicious use of classical methodology to discover stylized facts and then using a more thorough Bayesian analysis of a simple model expressing those facts can be successful. Though the tools of statistics have changed little over time, there is no uniform recipe for using them; in the end, applying experience and an open-minded approach to a new data context is the best we can do.