WHY IS IT SO HARD TO ESTIMATE EXPECTED RETURNS?

ABSTRACT. A key part of experiment design is determining how much data to collect. When the data comes in the form of a timeseries, the sample size is expressed both by the count N of the observations and the duration T of the historical period over which observations were made. For forecasting the drift of an asset price process with continuous sample paths, it turns out that the duration is key. I demonstrate that the standard error of any unbiased estimator of the price of risk is bounded below by $1/\sqrt{T}$, which I believe this is higher than many practitioners realize.

1. Introduction

Robert Merton, who helped introduce the theory of stochastic processes into finance, was well aware of the statistical challenge of estimating the drift of financial asset prices. In Merton (1980), he presented this in the context of geometric brownian motion and argued for the need for better models for volatility. It turns out that this challenge is not met by better volatility models; rather it is fundamental to the nature of timeseries datasets.

The paper is organized as follows. Section 2 is background. Section 3 describes a framework for modeling heteroskedasticity. Section 4 introduces the change of measure argument. Section 5 introduces parameter estimation. Sections 6 and 7 discuss the result and conclude. Proofs are in the appendices.

2. Background

While a quantitative investment analyst is generally comfortable assuming the existence (if not necessary the uniqueness) of a martingale measure $\mathbb{Q}$ that can be used for hedging and derivatives valuation, he or she would never design an investment program around a martingal measure. Rather, he or she would build a proprietary *subjective* measure $\mathbb{P}'$ that encodes risk-reward assessments using techniques such as

Date: April 24, 2012.

2010 *Mathematics Subject Classification.* 60H30, 62F12, 62M09, 62M10, 62M20, 91B82, 91B84.

Key words and phrases. financial timeseries; parametric estimation; price of risk.

the framework in Black and Litterman (1992). See Meucci (2005) for an example of this program.

But even with the best and broadest fundamental research, an analyst will not have views on every possible investment that might be available to diversify the portfolio. So he or she face a need to construct what might be termed an *objective* measure $\mathbb{P}$ that can be used as a foundation upon which to construct subjective views[1].

It is natural that one would prefer to calibrate this measure to historical data. Backtesting provides some comfort that an objective $\mathbb{P}$ that generally models short-term market risk phenomena can indeed be estimated from data. But one is usually confounded by the corresponding forecasts for absolute or relative asset returns. The goal of this paper is to provide an intuitive explanation for why it is so difficult (if not futile) to forecast expected returns from historical data.

3. Martingale Measure

Let us start by considering a risky asset whose value (with reinvested dividends) is a process S_t with continuous sample paths, and an instantaneously risk-free investment with value process B_t ($dB_t = rB_t\,dt$ defines the risk-free interest rate r). Let $\mathcal{F}_t$ be the natural filtration generated jointly by S_t and B_t. If a risk-neutral measure $\mathbb{Q}$ exists, there exists a $\mathbb{Q}$-martingale Y_t with $Y_0 = 0$ and quadratic variation $\langle Y \rangle_t$ such that the discounted process is a Doléans-Dade exponential,

$$\frac{S_t}{B_t} = \frac{S_0}{B_0} e^{Y_t - \frac{1}{2}\langle Y \rangle_t}$$

We can write a stochastic differential equation for Y_t in terms of volatility $\sigma > 0$ as

$$dY_t \triangleq \sigma\, dW_t$$

where W_t is a brownian motion[2] under $\mathbb{Q}$ and σ is $\mathcal{F}_t$-measurable.

See for example Applebaum (2004), chapter 5, for background on this approach.

Now we can define our data generating process as

$$X_t \triangleq \log \frac{S_t}{B_t}$$

[1]Or perhaps for independent risk management purposes.

[2]We will focus here on (non-anticipating) Wiener processes rather than other more general Lévy processes or anticipating processes such as fractional brownian motion.

Then

$$X_t = X_0 + \int_0^t \sigma \, dW_s - \int_0^t \tfrac{1}{2}\sigma^2 \, ds \tag{1}$$

Note that this framework is not restricted to assets with normal returns. The form of the volatility is completely general at this juncture. The only restriction we make is to continuous sample paths for cumulative total return.

3.1. **Parameterization.** In practice, we need to make some assumptions about the form of the volatility σ in order to model it. For example, let us say it is a function of some finite collection of past levels of the data generating process and possibly time. To make this more familiar, let us discretize time according to the partition $0 = t_0 < t_1 < \cdots < t_N = T$ and let

$$\int_0^{t_n} \sigma \, dW_s = \sum_{i=1}^{n} \sqrt{\frac{h_i}{t_i - t_{i-1}}} \left(W_{t_i} - W_{t_{i-1}} \right) \tag{2}$$

where each h_i is $\mathcal{F}_{t_{i-1}}$-measurable. Then we have

$$\begin{aligned} \epsilon_i &\triangleq X_{t_i} - \mathrm{E}^{\mathbb{Q}} \left[X_{t_i} \middle| \mathcal{F}_{t_{i-1}} \right] \\ &= \sqrt{\frac{h_i}{t_i - t_{i-1}}} \left(W_{t_i} - W_{t_{i-1}} \right) \end{aligned}$$

For example, the popular GARCH(1,1) model (see Engle (1982)) can be expressed in these terms as

$$h_i \triangleq \omega + \alpha \epsilon_{i-1}^2 + \beta h_{i-1}$$

with parameters ω, α, β, and h_0 (one usually assumes $\epsilon_0 = 0$).

Let us refer collectively to the parameters for the volatility model[3] as θ.

Since under (2),

$$X_{t_i} | \mathcal{F}_{t_{i-1}} \sim \mathrm{N} \left(X_{t_{i-1}} - \tfrac{1}{2} h_i, h_i \right)$$

one could define the joint density $f_X^{\mathbb{Q}}(x;\theta)$ associated with this model for the data generating process and use it as a likelihood or quasi-likelihood function to estimate the volatility parameters[4].

---

[3]The implied volatility model parameters calibrated from derivatives prices characterize a potentially different risk-neutral measure $\mathbb{Q}'$. One observes this, for example, in the Engle (2012) NYU Stern V-lab studies that compare the CBOE® VIX® implied volatility index to various GARCH estimates for S&P 500® stock index volatility.

[4]Some authors omit the $-\frac{1}{2}h_i$ term in the mean, which is usually inconsequential.

The approach above is purposefully concrete. Aït-Sahalia (2002) presents an alternate approach to the model parameterization in terms of a series approximation. Bandi and Phillips (2003) present an attractive non-parametric estimation approach for time-homogeneous processes. There are corresponding versions of θ for each.

4. Change of Measure

Of course the discounted price process for the risky asset will not be a martingale under the objective probability measure $\mathbb{P}$. We handle this in the natural way by introducing the Radon–Nikodým derivative,

$$\frac{d\mathbb{P}}{d\mathbb{Q}} \triangleq e^{Z_t - \frac{1}{2}\langle Z\rangle_t}$$

where

$$Z_t \triangleq \int_0^t \lambda \, dW_s$$

with the $\mathcal{F}_t$-measurable value λ termed the "price of risk".

We will explore the problem of estimating the price of risk using historical data. Let us restrict ourselves to the simplest model where λ is a constant. In this case,

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = e^{\lambda W_t - \frac{1}{2}\lambda^2 t} \quad \text{for constant } \lambda \tag{3}$$

Introducing the concept of asset price drift, μ, our model can be expressed in more conventional terms as

$$dS_t = \mu S_t \, dt + \sigma S_t \, d\tilde{W}_t$$

where $\tilde{W}_t \triangleq W_t - \lambda t$ is a $\mathbb{P}$-martingale and $\mu \triangleq r + \lambda\sigma$. In our construction μ, unlike λ, is not a parameter that can be estimated *per se* because r and σ are dynamic; although one could propagate the forecast errors for r and σ in combination with the standard error for λ to estimate the error of a forecast for μ.

Let us represent the concatenation of the volatility parameters θ with λ as $\theta \mathbin{+\!\!+} \lambda$.

5. Parameter Estimation

Let us say we have a time series of $N+1$ observations collected over a period of T years, $x = \left(x_0, x_{t_1}, \ldots, x_{t_{N-1}}, x_T\right)'$, which we can think of as a realization of the random vector $X = (X_0, \ldots, X_T)'$ whose likelihood under the objective measure is

$$f_X^{\mathbb{P}}(x; \theta, \lambda) = f_{X_{t_1}|\mathcal{F}_0}^{\mathbb{P}}(x_{t_1}) \cdots f_{X_T|\mathcal{F}_{t_{N-1}}}^{\mathbb{P}}(x_T)$$

The key observation comes from the relationship between the $\mathbb{P}$ and $\mathbb{Q}$ versions of this characterization from (3):

$$\log f_X^{\mathbb{P}}(X;\theta,\lambda) = \log f_X^{\mathbb{Q}}(X;\theta) + \lambda W_T - \tfrac{1}{2}\lambda^2 T \tag{4}$$

where

$$f_X^{\mathbb{Q}}(X;\theta) \triangleq f_X^{\mathbb{P}}(X;\theta,0)$$

and

$$W_T = \int_0^T \frac{dX_t}{\sigma} + \int_0^T \frac{\sigma}{2}\,dt$$

from (1).

Note that W_T in (4) is the *terminal value* of the latent driving process for our risky asset. Intermediate values cancel out.

5.1. **Fisher information.** The Fisher information for our experiment can be expressed as the negative curvature of the log-likelihood averaged over all possible realizations.

$$\mathcal{I}(\theta,\lambda) \triangleq \mathrm{E}^{\mathbb{P}}\left[-\frac{\partial^2 \log f_X^{\mathbb{P}}(X;\theta,\lambda)}{\partial(\theta + \lambda)'\,\partial(\theta + \lambda)}\right]$$

An experiment with a strongly-peaked log-likelihood gives more precise joint estimates for the unknown parameters than an experiment with a flatter log-likelihood.

In this case, we can use the change of measure to isolate and focus on the market price of risk parameter.

$$\mathcal{I}(\theta,\lambda) = \begin{pmatrix} \mathcal{I}(\theta) - \lambda\,\mathrm{E}^{\mathbb{P}}\frac{\partial^2 W_T}{\partial\theta'\partial\theta} & -\mathrm{E}^{\mathbb{P}}\frac{\partial W_T}{\partial\theta'} \\ -\mathrm{E}^{\mathbb{P}}\frac{\partial W_T}{\partial\theta} & T \end{pmatrix} \tag{5}$$

where

$$\mathcal{I}(\theta) \triangleq \mathrm{E}^{\mathbb{P}}\left[-\frac{\partial^2 \log f_X^{\mathbb{Q}}(X;\theta)}{\partial\theta'\,\partial\theta}\right]$$

5.2. **Cramér–Rao bound.** The Cramér–Rao bound is a lower bound on the covariance of any *unbiased* estimator. See appendix B for a proof. In this case,

$$\mathrm{cov}^{\mathbb{P}}\left[\hat\theta + \hat\lambda\right] \geq \mathcal{I}(\theta,\lambda)^{-1}$$

The diagonal entries of the inverse Fisher information are the squared standard errors. Since the Fisher information (5) is positive semidefinite, the result (7) from appendix A, on a bound for the bottom right entry of $\mathcal{I}(\theta,\lambda)^{-1}$, leads directly to the main result,

$$\boxed{\hat\lambda \text{ unbiased} \implies \mathrm{s.e.}\,\hat\lambda \geq \frac{1}{\sqrt{T}}} \tag{6}$$

6. Discussion

While there may be unbiased estimators for the parameters of the volatility whose standard errors can be reduced by increasing the *size* of the sample N (see Florens-Zmirou (1993) for example), all unbiased estimators of the excess return will have standard errors fundamentally bound by the *length* of the historical period T.

Estimating the price of risk for an asset on an *ex ante* basis is closely related to measuring the Sharpe ratio of an asset on an *ex post* basis. See for example Lo (2002). To a performance analyst, a value of of $+0.5\ \mathrm{yr}^{-1/2}$ for the Sharpe ratio would be considered acceptable, while a value of $0.0\ \mathrm{yr}^{-1/2}$ would not. In order to reliably discriminate between $\hat{\lambda} = 0.5\ \mathrm{yr}^{-1/2}$ and $\hat{\lambda} = 0.0\ \mathrm{yr}^{-1/2}$, a statistician would probably insist on $\mathrm{SE}\,\hat{\lambda} \lessapprox 0.25\ \mathrm{yr}^{-1/2}$ if not less.

Our result implies that an analyst in this situation would conclude he or she needs $T \gtrapprox 16$ years if he or she wanted to use an unbiased estimator to make statistically significant claims about the relative prospects of assets. It is difficult to argue that the economic basis for an asset's relative return remains stationary over such a long period. With the presence of corporate actions, technology and marketplace innovations, and management changes, many professional equity analysts are reluctant to use data more than two years old for forecasting.

Assuming that an analyst is aware of this result—or at least sensitive to statistical significance—how should he or she proceed? Below is a brief survey.

6.1. **Long history.**

Setting aside doubts about structural stationarity, perhaps one can follow Merton's lead and develop a sufficiently sophisticated volatility model to capture long-term historical dynamics. Recent history puts the scale of this challenge into perspective: the last sixteen years have included several regimes punctuated by a number of significant events such as: the 1997 Asian financial crisis, the collapse of the "dot-com" bubble, the accounting scandals leading to Sarbanes–Oxley[5], the 1998 Russian default and the failure of LTCM, the August 2008 "quant crisis" and the collapse of the "subprime" mortgage bubble, the financial crisis leading to Dodd–Frank[6],

[5]Entitled, "An act to protect investors by improving the accuracy and reliability of corporate disclosures made pursuant to the securities laws, and for other purposes." Public Law 107–204 approved July 30, 2002.

[6]Entitled, "An act to promote the financial stability of the United States by improving accountability and transparency in the financial system, to end "too big to fail", to protect the American taxpayer by ending bailouts, to protect consumers

the May 2010 "flash crash", and the August 2011 volatility shock and European sovereign debt crisis.

6.2. **Equilibrium.** One approach is to appeal to the Capital Asset Pricing Model and suppose that the price of risk is shared across several assets as a common "market price of risk" or even a "world price of risk" as in Harvey (1991). The advantage of this is that under a joint volatility model, one has a cross section of time series of latent sample paths all exhibiting the same drift. To the extent that these processes are not mutually dependent, the benefits of combining these data into a panel is not entirely diluted.

6.3. **Representative agents.** Another more fundamental approach is to reverse-engineer some benchmark portfolio or some set of heuristic constraints on a practicable portfolio implementation along with an assumption about the preferences of a representative agent. See Sharpe (1974). This leads to a solution or restricted domain for the allocation-implied asset returns.

6.4. **Bias.** Finally, one could reject unbiased estimators altogether. This is becoming more common in practice, with explicitly biased James–Stein shrinkage estimators or more subtly biased Bayesian estimators. For example, Merton (1980) recommends using a prior whose support excludes negative values.

This approach is less problematic for statisticians working in finance, who often have the luxury of being their own clients[7].

7. Conclusion

For an asset that can be financed at the risk-free rate, and whose value process is continuous, the price of risk is the only meaningful measure of its relative worth. I have shown that one would need to work with almost two decades of historical data in order to be able to reliably estimate this quantity using an unbiased estimator. Such data presents unusual challenges which may be insurmountable. Alternative approaches involve additional assumptions which may be challenging to defend.

from abusive financial services practices, and for other purposes." Public Law 111–203 approved July 21, 2010.

[7] In any case one should be open about one's use of bias in analyses and studies that a colleague or reader might assume to be objective.

8. Acknowledgments

I am grateful to Simon Babbs, Long Kang, and Nancy Verma of The Options Clearing Corporation and Jingnan Zhang of the University of Minnesota for their helpful comments and suggestions.

Appendix A. Block Positive Semi-Definite Matrices

Let A be a symmetric matrix, a a vector, and α a scalar and define

$$M = \begin{pmatrix} A & a \\ a' & \alpha \end{pmatrix}$$

Assuming it exists,

$$M^{-1} = \begin{pmatrix} A^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{\alpha - a'A^{-1}a} \begin{pmatrix} A^{-1}aa'A^{-1} & -A^{-1}a \\ -a'A^{-1} & 1 \end{pmatrix}$$

where we recognize in the second term the inverse of the Schur complement of A in M. Note that this may be relevant in extending the result of this paper to more general descriptions of the price of risk.

Let us further assume that both A and M are positive semi-definite. Therefore, there is a lower-diagonal Cholesky matrix B such that $A = BB'$, a vector b, and a scalar β such that

$$M = \begin{pmatrix} B & 0 \\ b' & \beta \end{pmatrix} \begin{pmatrix} B' & b \\ 0' & \beta \end{pmatrix} = \begin{pmatrix} A & Bb \\ b'B' & b'b + \beta^2 \end{pmatrix}$$

so $a = Bb$ thus $b = B^{-1}a$, and $\alpha = a'B^{-1\prime}B^{-1}a + \beta^2$. So

$$0 \le a'A^{-1}a \le \alpha$$

must hold, or

$$\frac{1}{\alpha - a'A^{-1}a} \ge \frac{1}{\alpha} \tag{7}$$

That is, the bottom right entry of M^{-1} must be equal to or greater than the reciprocal of the bottom right entry of M.

Appendix B. Cramér–Rao lower bound

The Cramér–Rao lower bound is a classic result in statistics. I provide below an outline of a proof in the multi-parameter setting. See for example Keener (2010).

Consider an unbiased estimator $\hat{\theta}(x)$ for an unknown parameter vector θ with likelihood $f_X(x)$ at sample x.

$$0 = \mathrm{E}\left[\hat{\theta}(X) - \theta\right] = \int \left(\hat{\theta}(x) - \theta\right) f_X(x)\, dx$$

If we take the (vector) derivative with respect to the parameters, and we are allowed to distribute it, we get

$$0 = \int \left(\hat{\theta}(x) - \theta\right) \frac{\partial f_X(x)}{\partial \theta} \, dx - I \int f_X(x) \, dx$$

or, with some manipulation,

$$\int \left(\left(\hat{\theta}(x) - \theta\right) \sqrt{f_X(x)}\right) \left(\frac{\partial \log f_X(x)}{\partial \theta} \sqrt{f_X(x)}\right) dx = I$$

Consider any vectors a and b in parameter space. The previous result means

$$\int \left(a' \left(\hat{\theta}(x) - \theta\right) \sqrt{f_X(x)}\right) \left(\frac{\partial \log f_X(x)}{\partial \theta} \sqrt{f_X(x)}\, b\right) dx = a'b$$

This can be thought of as an inner product in the Hilbert space L^2, which means we can apply Cauchy-Schwarz to get

$$a' \left(\int \left(\hat{\theta}(x) - \theta\right) \left(\hat{\theta}(x) - \theta\right)' f_X(x) \, dx\right) a$$
$$b' \left(\int \frac{\partial \log f_X(x)}{\partial \theta'} \frac{\partial \log f_X(x)}{\partial \theta} f_X(x) \, dx\right) b \geq (a'b)^2$$

Define the Fisher information to be

$$\begin{aligned}
\mathcal{I}(\theta) &\triangleq \mathrm{E}\left[\frac{\partial \log f_X(X)}{\partial \theta'} \frac{\partial \log f_X(X)}{\partial \theta}\right] \\
&= \mathrm{cov}\left[\frac{\partial \log f_X(X)}{\partial \theta'}\right] \\
&= E\left[-\frac{\partial^2}{\partial \theta' \, \partial \theta} \log f_X(X)\right]
\end{aligned}$$

if the log-likelihood is twice differentiable on its support in the last instance.

With $b \triangleq \mathcal{I}^{-1}(\theta)\, a$, the previous result translates to

$$\left(a' \,\mathrm{cov}\left[\hat{\theta}(X)\right] a\right) \left(a' \mathcal{I}^{-1}(\theta)\, a\right) \geq \left(a' \mathcal{I}^{-1}(\theta)\, a\right)^2$$

So we can conclude that

$$a' \left(\mathrm{cov}\left[\hat{\theta}(X)\right] - \mathcal{I}^{-1}(\theta)\right) a \geq 0$$

for all vectors a.

This conforms with the definition of a positive semi-definite matrix, and can be written as

$$\mathrm{cov}\left[\hat{\theta}(X)\right] \geq \mathcal{I}^{-1}(\theta) \tag{8}$$

Note that the Cramér–Rao lower bound is a special case of the Kullback inequality about the relative entropy of one measure with respect to another, which may be useful in extending the result of this paper to a wider class of processes.

References

Y. Aït-Sahalia. Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. *Econometrica*, 70(1):223–262, 2002.

D. Applebaum. *Lévy Processes and Stochastic Calculus*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, first edition, 2004.

F. M. Bandi and P. C. B. Phillips. Fully nonparametric estimation of scalar diffusion models. *Econometrica*, 71(1):241–283, 2003.

F. S. Black and R. B. Litterman. Global portfolio optimization. *Financial Analysts Journal*, 48(5):28–43, 1992.

R. F. Engle. Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation. *Econometrica*, 50:987–1008, 1982.

R. F. Engle. V-lab: real-time financial volatility and correlation measurement, modeling, and forecasting, January 2012. `http://vlab.stern.nyu.edu/`.

D. Florens-Zmirou. On estimating the diffusion coefficient from discrete observations. *Journal of Applied Probability*, 30(4):790–804, 1993.

C. R. Harvey. The world price of covariance risk. *Journal of Finance*, 46(1):111–157, 1991.

R. W. Keener. *Theoretical Statistics: Topics for a Core Course*. Springer Texts in Statistics. Springer Science+Business Media, New York, first edition, 2010.

A. W. Lo. The statistics of Sharpe ratios. *Financial Analysts Journal*, 58(4):36–52, 2002.

R. C. Merton. On estimating the expected return on the market. *Journal of Financial Economics*, 8(4):323–361, 1980.

A. Meucci. *Risk and Asset Allocation*. Springer Finance. Springer Science+Business Media, New York, first edition, 2005.

W. F. Sharpe. Imputing expected security returns from portfolio composition. *Journal of Financial and Quantitative Analysis*, 9(3):463–472, 1974.