

Estimation Methods

Wai Kong Yuen
Wayne Pushka

Dec 1999

Introduction

In this article, we discuss how to estimate the input variables from historical data and why the inputs are not particularly reliable. We shall explain three approaches of estimation and discuss their strengths and weaknesses.

The major inputs to estimate in our model are:

- the annual expected values,
- the standard deviations and the correlation coefficients of the rate of returns of the assets and factors in the future time period.

Once these values are determined, the expected return of each asset/factor follows an approximate geometric Brownian motion. If we know the values *exactly*, we can make some very good predictions on the future expected returns and risk probabilities and consequently some excellent investment decisions.

Unfortunately, the future is not known. We have to make some predictions of the inputs today. The common way is to estimate the inputs from the past. This can be done by calculating the sample mean, sample variance and sample correlation matrix from the data from each time period in the past. So, the behavior of each asset/factor is determined by the estimates from the historical data. Therefore, we are assuming that the asset/factor should perform exactly like the past in the future.

There are two problems in this assumption. First, it is hard to decide how many years of data we should use in the past to estimate the future. For example, to estimate the monthly expected return rates of an asset, should we use the historical data from the last 10 years or 30 years? They can be very different in general and mathematically, we don't know which one is better. Whichever we use, it is just a reference to what will happen in the future. So, we have to make a subjective decision, which is not always reliable, to choose the right period to model the future.

Second, there is an estimation error from the historical data. Our optimizer maximizes the expected return and minimizes the risk of the portfolio (like most optimizers do). It significantly over-weights (under-weights) those assets that have large (small) estimated returns, negative (positive) correlations and small (large) variances. These assets are those performed very well with relatively low risk in a certain period in the past. Unfortunately, they are also the ones most likely to have large estimation errors. It often happens in finance that an asset that performs well above average in one time period does far less well in a subsequent one (see Michaud and Carty 1999). So, our portfolio may end up weighting too much on some high return hard-to-predict assets, which in turn

increasing the risk (see Hensel and Turner). Therefore, it can be very risky to use historical values to predict the future without considering estimation errors.

Before looking into how to solve the problems, we describe briefly what the three estimation approaches are. They are: Classical, Bayesian and Empirical Bayesian.

In the **Classical** approach, we assume no knowledge of the parameters in the model. Suppose a set of data points from the past are outcomes of an experiment designed for the model. Consider the event that the outcomes of the experiment are exactly the above data set. Even though we do not know the parameters yet, we can still find the joint probability of such an event in terms of the unknown parameters. An intuitive choice of the parameters is the set that maximizes the probability of the event. Such estimators are called *maximum likelihood estimators* (MLE) (see Section II for detail). In our model, for example, it can be shown that the MLE for the expected return rate is the sample mean, which is the easiest and most widely used estimator. MLE for the standard deviation and the correlation coefficients are the sample standard deviation and the sample correlation coefficients.

In the **Bayesian** approach, we have some *prior belief* in the unknown parameters. We assume the parameters follow a *prior* distribution with parameters known as *hyper-parameters*. In this case, we assume the hyper-parameters are constant. Then we can calculate the distribution of the parameter conditioned on the historical outcomes (for detail, see Section III). This distribution is known as the *posterior* distribution. Now, we can estimate the parameters with the posterior distribution. For example, we can take the mean of the posterior distribution to estimate the parameters.

In the **Empirical Bayesian** approach, we also assume the parameters follow a prior distribution with hyper-parameters. However, some of them may be unknown now. So, before finding the posterior distribution, we have to estimate the hyper-parameters from the historical data (e.g. by classical methods). Then, as in the Bayesian case, we can estimate the parameters from the posterior distribution.

The classical approach is very different from both the Bayesian and Empirical Bayesian approaches because we do not assume any knowledge of the parameters. It is more objective but on the other hand, we lose some information from what we know from experience. The Bayesian and Empirical Bayesian approaches are very similar in the way that we have a prior belief in the distribution of the parameters. The only difference is that we have to estimate the hyper-parameters in the Empirical Bayesian approach. With the additional information available, our estimates will likely bias towards the values with higher prior density values.

Different versions of Stein estimator can be derived from both the Bayesian and Empirical Bayesian approaches. We assume the expected return rate of each asset follows a prior distribution with mean equal to the global mean. The resulting estimators shrink the expected return rates of the assets towards the global mean. This is done

because assets with extreme returns often revert back to less extreme return in the following period. This effect is called the *mean reversion*.

There are two major types of estimation errors. The first type is due to the insufficient data. In order to estimate the parameters, large samples from the model are required. Otherwise, the statistical errors of the estimators can be very large. The second is due to the mean reversion. The samples we have are only the samples for the model in the past, not for the future. So, the mean reversion effect described in the previous paragraph can be very significant.

To get around the insufficient data, we can add an uncertainty to the estimated mean of an asset. One possibility is to derive an approximate confident distribution statistically from the given data using classical method. Instead of saying that the mean is a fixed value, we can describe it by a distribution. It is similar to using a prior distribution in the Bayesian approach, except that we use it not just as a mean to estimate the parameters, but as a real distribution in the model. Therefore, the effective standard deviation of each asset is higher. This will reduce the risk of having insufficient data and drawing bad conclusion. Another possibility is to introduce some *prediction error*. Note that the distribution in the future may change but we are using the distribution in the past. So, we can model the difference between prediction and real return rate by introducing an extra random variable. Its distribution can be estimated by using the historical predictions from economists and the real return rate in the same period. Again, this will increase the standard deviation of each asset. The advantage of this method is it models the reality directly.

Finally, if we want to deal with errors due to the mean reversion effect, we can use the Stein estimator. The estimator reverts the extreme mean towards the global mean (which we believe is a good reference point). This reduces the chance of over-weighting assets with high estimated returns. Besides, the estimator is a 'better' estimation than the MLE (see Section V). Therefore, it is strongly suggested (see Michaud 1999) that the Stein estimator should be used instead of the historical mean. The same errors can also affect the estimation of the standard deviation and correlation coefficients. Ledoit (1999) introduced a similar Stein method to estimate them. However, since we are considering a longer horizon, changing the mean will change the optimal portfolio a lot. So, it is more important to pay more attention to the mean. We should also be careful on grouping the assets into different classes with different global means.

The above suggestions should be examined (especially on estimating the parameters in the new distribution and choosing the global mean) before being applied.

We begin discussion of the mathematical theory of the estimations. Let X_i be n independent identically distributed random variables with density $f(x_i; \theta)$ where θ is a parameter (can be a vector) we want to estimate in terms of the observed value of X_i . So, it shall be helpful to think of X_i as known values. There are 3 different approaches

of estimating θ : classical, Bayesian, and Empirical Bayesian. We shall discuss the theory behind each approach in Section I, II and III. In Section IV, we shall extend the technique to a more general setting and discuss the Stein estimator. In Section V, we shall relate the methods to our model.

Section I

The Classical Approach

In the classical approach, we assume that we have no knowledge on the parameter θ . Then the joint density of X_1, X_2, \dots, X_n is given by

$$f(x_1, x_2, \dots, x_n; \theta) = f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta) \quad (0.1)$$

This function is called the *likelihood function*. Intuitively, this function specifies how likely (in terms of the density value) is the occurrence of the event

$X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ for a particular value of θ . So, given observed values of X_1, X_2, \dots, X_n , we want to choose θ such that the above density value is maximized. An estimator obtained in this way is called a *maximum likelihood estimator* (MLE).

Section II

The Bayesian Approach

In the Bayesian approach, we assume that we have some prior beliefs in the distribution of θ . Suppose the density of such distribution is $p(\theta)$ and θ is independent of

X_1, X_2, \dots, X_n . By Bayes' theorem, the density of θ given X_1, X_2, \dots, X_n is given by:

$$f(\theta | x_1, x_2, \dots, x_n) \propto f(x_1; \theta) \cdot f(x_2; \theta) \cdots f(x_n; \theta) \cdot p(\theta) \quad (0.2)$$

for fixed values of x_1, x_2, \dots, x_n . The proportional sign in the formula means the integral of the right hand side may not be 1. Therefore, if we have observed the values of x_1, x_2, \dots, x_n , we have a distribution of θ in terms of some known parameters.

$f(\theta | x_1, x_2, \dots, x_n)$ is called the *posterior distribution* of θ and $p(\theta)$ is called the *prior distribution* of θ . So, we know how θ is distributed from the experiment.

The problem of this function f is that we still don't know how to estimate the value of θ . Therefore, we introduce a decision theory to choose the optimal θ from the observed values of x_1, x_2, \dots, x_n . A naive choice is then

$$\hat{\theta}_n = E(\theta | x_1, x_2, \dots, x_n) = \int_{\theta} f(\theta | x_1, x_2, \dots, x_n) d\theta, \text{ which is a function of } x_1, x_2, \dots, x_n \text{ as}$$

required. However, mathematically, it is not clear why we choose the mean of the posterior distribution instead of the mode or the median. In fact, we have to specify the decision rule by using the concept of a *loss function* $L(\hat{\theta}, \theta)$ and different rules can give different estimators. A loss function is one that measures the distance between the real

parameter θ and an estimated parameter $\hat{\theta}$. An example of a loss function is the *quadratic loss* $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$. Notice that θ is a distribution and so is $L(\hat{\theta}, \theta)$. This is the major difference between the Bayesian approach and the classical approach, in which θ is a constant parameter. The idea is then to choose $\hat{\theta}$ to ‘minimize’ the distribution $L(\hat{\theta}, \theta)$ in a certain sense. A *Bayes’ estimator* is the $\hat{\theta}$ (a function of X_1, X_2, \dots, X_n) that minimizes $E(L(\hat{\theta}, \theta) | x_1, x_2, \dots, x_n)$. In case of the quadratic loss function, it turns out that $E(\theta | x_1, x_2, \dots, x_n)$ is the Bayes’ estimator. See LEE p.205-210.

Section III

The Empirical Bayesian Approach

The Empirical Bayesian approach is very similar to the Bayesian approach except on the prior distribution. In the Bayesian approach, we have a prior belief in the distribution of the parameter θ we want to estimate. In other words, the density $p(\theta)$ is a known function. In the Empirical Bayesian approach, however, we assume the distribution of θ depends further on some other parameter μ which is unknown. Therefore, the distribution of θ is not completely known. So, we have to find some ways to estimate μ before estimating θ from X_1, X_2, \dots, X_n . It is not hard to estimate μ though. Notice that the distributions of X_1, X_2, \dots, X_n can be expressed in terms of μ only (since θ depends on μ). We can then treat μ as the only unknown parameter and we often use the classical approach to estimate μ and therefore the density $p(\theta)$ is known now. We can carry out the Bayesian procedure described in Section III to estimate θ . Unfortunately, this approach is not mathematically rigorous as it violates the Bayes’ theorem. To see this, μ is estimated from the observed value of X_1, X_2, \dots, X_n and so the distribution of θ depends on X_1, X_2, \dots, X_n and violates the independence assumption in the Bayes’ theorem. Even though it is not coherent to use such estimators, Empirical Bayesian estimator can be viewed as approximate Bayesian estimator. See PRESS p.42-43.

Section IV

The Stein Estimator

The Classical approach is the most popular approach because it makes no subjective assumptions on the parameter θ . However, under some reasonable decision rules, it can be shown that the MLE, for example, may not be the best estimator. We will consider the Stein estimator.

Suppose θ is a vector of parameters and $\theta = (\theta_1, \theta_2, \dots, \theta_k)$. For each one of them we observe a random variable

$$X_i \square N(\theta_i, 1) \quad (0.3)$$

Suppose we pick an estimator $\delta = (\delta_1, \delta_2, \dots, \delta_k)$ where each δ_i is a function of X_1, X_2, \dots, X_k . In order to compare the performance of different estimators, we have to define the loss of using δ to estimate θ , which is the concept of the loss function in Section III. The easiest loss function to work with is the *mean squared error loss*:

$$L(\delta, \theta) = \frac{1}{k} \sum_{i=1}^k (\delta_i - \theta_i)^2 \quad (0.4)$$

It measures the distance of an estimator δ from the actual parameter θ and is a distribution with parameter θ since δ is a function of X_1, X_2, \dots, X_k . So, a good estimator is one which the expectation of the distance $E(L(\delta, \theta))$ is small for all θ (think of L as a distribution of a parameter θ). Since we only have one observation X_i for each parameter θ_i , a natural choice of the parameter is the MLE

$\theta^{\text{MLE}} = (X_1, X_2, \dots, X_k) = \mathbf{X}$. In fact, $E(L(\theta^{\text{MLE}}, \theta)) = 1$. However, Stein showed that if $k \geq 3$, $S = \sum_{i=1}^k X_i^2$ and

$$\theta^{\text{STEIN}} = \left(1 - \frac{k-2}{S}\right) \mathbf{X}, \quad (0.5)$$

we have $E(L(\theta^{\text{STEIN}}, \theta)) < 1 = E(L(\theta^{\text{MLE}}, \theta))$ for all θ . Therefore, θ^{STEIN} is a better estimator than θ^{MLE} under such decision rule.

It can also be proved that θ^{STEIN} is the same as the estimator obtained by using the Empirical Bayes' approach, assuming that the prior distributions of $\theta_1, \theta_2, \dots, \theta_k$ are independent normal with mean 0. More generally, suppose $X_i \square N(\theta_i, \alpha)$ and we have a prior belief that for each i , $\theta_i \square N(\mu, \phi)$ for some fixed μ and α . By using the Empirical Bayesian approach, we have a general form of Stein estimator:

$$\theta^{\text{STEIN}} = \mu + \left(1 - \frac{k-2}{S_\mu}\right) (\mathbf{X} - \mu) \quad (0.6)$$

where $\mu = (\mu, \mu, \dots, \mu)$ and $S = \sum_{i=1}^k (X_i - \mu)^2$. Such an estimator is also better than the MLE. The result is somewhat surprising because the stein estimator 'shrinks' the obvious estimator MLE towards an arbitrary origin μ .

Section V

Application to our model

a) The Stein estimator

We shall only explain briefly the concept of including global mean into the estimation. Suppose we have k assets and the expected return of asset k is distributed as $N(\theta_i, \alpha)$

and we have a prior belief that for each i , $\theta_i \sim N(\mu, \phi)$ for some fixed α . Here, μ can be interpreted as the global mean of the assets. In other words, we believe that the mean of each asset should be related to the global mean so that on average, if we pick an asset randomly, the mean return of the asset should be closed to the global mean return. Under this assumption, we can use the Stein estimator and choose μ to be the global mean.

Suppose the sample means of the assets are $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ and the global sample mean is \bar{X} . Intuitively, $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_k$ are estimates of $(\theta_1, \theta_2, \dots, \theta_k)$ and \bar{X} is an estimate of μ . Then, by using a similar approach as in the previous section (except μ is unknown now), we have the following form of Stein estimator:

$$\theta_i^{\text{STEIN}} = \bar{X} + c_i(\bar{X}_i - \bar{X}) \quad (0.7)$$

for some constant $c_i = \max\left\{0, 1 - (k-3)\sigma_i^2 / \sum (\bar{X}_i - \bar{X})^2\right\}$. This estimator is better than the MLE (with respect to the mean squared error loss) and the origin is chosen under the Bayesian prior belief.