Portfolios that Contain Risky Assets 12: Assessment of Independent, Identically-Distributed Models

C. David Levermore

University of Maryland, College Park, MD

Math 420: *Mathematical Modeling*
April 20, 2020 version

Portfolios that Contain Risky Assets
Part II: Stochastic Models

11. Independent, Identically-Distributed Models for Assets
12. Assessment of Independent, Identically-Distributed Models
13. Independent, Identically-Distributed Models for Portfolios
14. Kelly Objectives for Markowitz Portfolios
15. Cautious Objectives for Markowitz Portfolios
16. Optimization of Mean-Variance Objectives

Assessment of Independent, Identically-Distributed Models

1. Introduction
2. Comparing Distributions
3. Assessing Identical Distribution
4. Stationary Autoregression Models
5. Fitting Stationary Autoregression Models
6. Assessing Independence

Introduction

Independent, Identically Distributed (IID) models of returns make two simplifying assumptions.

1. **Independent.** That what happens on day d is independent of what has happened in the past.
2. **Identically Distributed.** What happens each day is statistically identical to what happens every other day.

In IID models the random numbers $\{R_d\}_{d=1}^{D}$ that mimic a return history are each drawn from $(-1, \infty)$ in accord with the *same* probability density.

The question arises as to how can we determine how well a given return history $\{r(d)\}_{d=1}^{D}$ is mimiced by such a model. Here we present ways by which the validity of each assumption can be assessed.

Introduction

First, we will examine how to assess the validity of the **identically distributed** assumption. This comes down to understanding how likely it is that two different return histories, say $\{r_1(d)\}_{d=1}^{D_1}$ and $\{r_2(d)\}_{d=1}^{D_2}$, might be drawn from the same probability density. We will take three approaches:

- graphical,
- comparing means and variances,
- comparing distributions.

Next, we will examine how to assess the validity of the **independent** assumption. This comes down to understanding how correlated each $r(d)$ is with earlier values, say with $r(d-1)$. We will take three approaches:

- graphical,
- comparing with an autoregression model,
- comparing autocovariance matrices.

Comparing Distributions

Comparing Distributions. In an IID model the random numbers $\{R_d\}_{d=1}^{D}$ are each drawn from $(-1,\infty)$ in accord with the *same* probability density $q(R)$. Therefore if we plot the points $\{(d,R_d)\}_{d=1}^{D}$ in the dr-plane they will usually be distributed in a way that looks uniform in d.

Therefore if the return history $\{r(d)\}_{d=1}^{D}$ is mimiced by such a model then the points $\{(d,r(d))\}_{d=1}^{D}$ plotted in the dr-plane should appear to be distributed in a way that is uniform in d.

This will be the case if every subsample of the return history $\{r(d)\}_{d=1}^{D}$ behaves as if it was drawn from the same probability density. Therefore the question that we must address is how to tell when two samples, $\{r_1(d)\}_{d=1}^{D_1}$ and $\{r_2(d)\}_{d=1}^{D_2}$, might be drawn from the same probability density.

Comparing Distributions

We start with a simpler question. How to compare two probability densities over $(-1, \infty)$, say $q_1(R)$ and $q_2(R)$ where $q_1(R) \geq 0$, $q_2(R) \geq 0$, and

$$\int_{-1}^{\infty} q_1(R)\,\mathrm{d}R = \int_{-1}^{\infty} q_2(R)\,\mathrm{d}R = 1\,.$$

One idea is to compare their distributions $Q_1(R)$ and $Q_2(R)$, which are

$$Q_1(R) = \int_{-1}^{R} q_1(R')\,\mathrm{d}R'\,, \qquad Q_2(R) = \int_{-1}^{R} q_2(R')\,\mathrm{d}R'\,.$$

These are nondecreasing functions of R over $(-1, \infty)$ such that

$$\lim_{R\to -1} Q_1(R) = \lim_{R\to -1} Q_2(R) = 0\,, \qquad \lim_{R\to\infty} Q_1(R) = \lim_{R\to\infty} Q_2(R) = 1\,.$$

Comparing Distributions

The *Kolmogorov-Smirnov* measure of the closeness of Q_1 and Q_2 is the sup norm of their difference:

$$\|Q_2 - Q_1\|_{\mathrm{KS}} = \sup\left\{|Q_2(R) - Q_1(R)| \,:\, R \in (-1,\infty)\right\}.$$

The *Kuiper* measure of the closeness of Q_1 and Q_2 is

$$\begin{aligned}\|Q_2 - Q_1\|_{\mathrm{Ku}} = \sup&\left\{Q_2(R) - Q_1(R) \,:\, R \in (-1,\infty)\right\}\\ &+ \sup\left\{Q_1(R) - Q_2(R) \,:\, R \in (-1,\infty)\right\}.\end{aligned}$$

It can be shown that

$$\|Q_2 - Q_1\|_{\mathrm{KS}} \le \|Q_2 - Q_1\|_{\mathrm{Ku}} \le 1\,.$$

Comparing Distributions

The *Cramer-von Mises* measure of the closeness of Q_1 and Q_2 is the L^2-norm of their difference:

$$\|Q_2 - Q_1\|_{\mathrm{CvM}} = \left(\int_{-1}^{\infty} \big(Q_2(R) - Q_1(R)\big)^2 \mathrm{d}R \right)^{\frac{1}{2}} .$$

This can clearly be generalized to any L^p-norm with respect to any positive measure over $(-1, \infty)$. Specifically, for every $p \in [1, \infty)$ we have

$$\|Q_2 - Q_1\|_{L^p} = \left(\int_{-1}^{\infty} \big(Q_2(R) - Q_1(R)\big)^p \mathrm{d}R \right)^{\frac{1}{p}} .$$

For simplicity we will stick to the Kolmogorov-Smirnov and Kuiper measures.

Comparing Distributions

Now we return to our original question. Given two samples, $\{r_1(d)\}_{d=1}^{D_1}$ and $\{r_2(d)\}_{d=1}^{D_2}$, we construct their so-called *emperical distributions*

$$\widehat{Q}_1(R) = \frac{\#\{d \,:\, r_1(d) \leq R\}}{D_1}\,, \qquad \widehat{Q}_2(R) = \frac{\#\{d \,:\, r_2(d) \leq R\}}{D_2}\,.$$

Here $\#S$ denotes the number of elements in a set S. These approximate the unknown true distributions Q_1 and Q_2 because

$$Q_1(R) = \mathrm{Pr}\{r_1(d) \leq R\}\,, \qquad Q_2(R) = \mathrm{Pr}\{r_2(d) \leq R\}\,.$$

Then the Kolmogorov-Smirnov and Kuiper measures of the difference $\widehat{Q}_2 - \widehat{Q}_1$ give us a way to quantify the likelihood that samples are drawn from similar distributions.

Comparing Distributions

Because $\widehat{Q}_1$ and $\widehat{Q}_2$ are step functions, we see that

$$\begin{aligned}
\|\widehat{Q}_2 - \widehat{Q}_1\|_{\mathrm{KS}} &= \max\{|\widehat{Q}_2(R) - \widehat{Q}_1(R)| \,:\, R \in (-1,\infty)\}\,.\\
\|\widehat{Q}_2 - \widehat{Q}_1\|_{\mathrm{Ku}} &= \max\{\widehat{Q}_2(R) - \widehat{Q}_1(R) \,:\, R \in (-1,\infty)\}\\
&\quad + \max\{\widehat{Q}_1(R) - \widehat{Q}_2(R) \,:\, R \in (-1,\infty)\}\,.
\end{aligned}$$

Fortunately statisticians have provided software that efficiently computes these values given any two samples $\{r_1(d)\}_{d=1}^{D_1}$ and $\{r_2(d)\}_{d=1}^{D_2}$. These are called respectively the *two-sample KS test* and the *two-sample Kuiper test*.

Assessing Identical Distribution

Assessing Identical Distribution. We will now present three ways to assess how much a given return history $\{r(d)\}_{d=1}^{D}$ is consistent with the *identical distribution assumption*. More specifically, we will present:

- a graphical assessment,
- a mean and a variance assessment,
- two distribution assessments.

The first is purely visual, but can be used to build understanding of the data. The other two are analytical. They will yield measures ω^{m}, ω^{v}, ω^{KS}, and ω^{Ku} of how consistent the given data is with the identical distribution assumption. As before, these measures will take values in the interval $[0,1]$ with higher values indicating greater consistency with the identical distribution assumption.

Assessing Identical Distribution

Graphical Assessment. In an IID model the random numbers $\{R_d\}_{d=1}^D$ are each drawn from $(-1, \infty)$ in accord with the *same* probability density $q(R)$. Therefore, if we plot the points $\{(d, R_d)\}_{d=1}^D$ in the dr-plane they will usually be distributed in a way that looks uniform in d.

Therefore if the return history $\{r(d)\}_{d=1}^D$ is mimiced by such a model then the points $\{(d, r(d))\}_{d=1}^D$ scatter plotted in the dr-plane should appear to be distributed in a way that is uniform in d.

Remark. Of course, determining whether such a scatter plot is distributed in a way that is uniform in d simply by looking at it is subjective. However, sometimes this graphical approach can make it quite clear that the identical distribution assumption is flawed! Henceforth, we will present quantitative approaches.

Assessing Identical Distribution

Mean and Variance Assessments. Given any two samples $\{r_1(d)\}_{d=1}^{D_1}$ and $\{r_2(d)\}_{d=1}^{D_2}$, we can compute their sample means and variances as

$$
\begin{aligned}
m_1 &= \frac{1}{D_1}\sum_{d=1}^{D_1} r_1(d)\,, & m_2 &= \frac{1}{D_2}\sum_{d=1}^{D_2} r_2(d)\,,\\
v_1 &= \frac{1}{D_1}\sum_{d=1}^{D_1} \big(r_1(d) - m_1\big)^2\,, & v_2 &= \frac{1}{D_2}\sum_{d=1}^{D_2} \big(r_2(d) - m_2\big)^2\,.
\end{aligned}
$$

Our goal is to develop measures of how close m_1 is to m_2 and v_1 is to v_2.

Assessing Identical Distribution

We begin by assessing the closeness of v_1 and v_2 because it is easier. We will assume that D_1 and D_2 are sufficiently large to insure that v_1 and v_2 are positive. Then the relative difference of v_1 and v_2 is

$$\frac{v_1 - v_2}{v_1 + v_2} .$$

This ratio takes values in the interval $(-1, 1)$. When its absolute value is small then v_1 and v_2 are relatively close.

When this ratio is squared and subtracted from 1 we get

$$1 - \frac{(v_1 - v_2)^2}{(v_1 + v_2)^2} = \frac{4 v_1 v_2}{(v_1 + v_2)^2} . \tag{3.1}$$

This quantity takes values in the interval $(0, 1]$. Its value is closer to 1 when v_1 and v_2 are relatively closer.

Assessing Identical Distribution

We now assess the closeness of m_1 and m_2. Using relative difference does not work because m_1 and m_2 might have opposite signs and $m_1 + m_2$ might be zero or nearly zero. Rather, because the variances associated with m_1 and m_2 are estimated by $\frac{1}{D_1} v_1$ and $\frac{1}{D_2} v_2$, we use the ratio

$$\frac{(m_1 - m_2)^2}{\frac{1}{D_1} v_1 + \frac{1}{D_2} v_2} .$$

This ratio takes values in the interval $[0, \infty)$. It is close to 0 when $|m_1 - m_2|$ is small compared to either standard deviation.

When this ratio is added to 1 and the reciprocal taken we get

$$\left(1 + \frac{(m_1 - m_2)^2}{\frac{1}{D_1} v_1 + \frac{1}{D_2} v_2} \right)^{-1} . \tag{3.2}$$

This quantity takes values in the interval $(0, 1]$. Its value is closer to 1 when m_1 and m_2 are relatively closer.

Assessing Identical Distribution

Finally, given return histories over a year $\{r(d)\}_{d=1}^{D}$, we can split the year into quarters and compare the mean and variance of each quarter with that of another quarter or with that of the other three quarters combined. The maximum of all such comparisons made is the score for the year. For example, motivated by (3.2) and (3.1), for each year we might define

$$\begin{aligned}\omega^{\mathrm{m}} &= \min\left\{ \left(1 + \frac{(m_1 - m_2)^2}{\frac{1}{D_1} v_1 + \frac{1}{D_2} v_2} \right)^{-1} \ : \ \text{all comparisons made} \right\} , \\ \omega^{\mathrm{v}} &= \min\left\{ \frac{4 v_1 v_2}{(v_1 + v_2)^2} \ : \ \text{all comparisons made} \right\} .\end{aligned} \tag{3.3}$$

If we compare quarters with each other then six comparisons are made. If we compare each quarter with the other three quarters combined then four comparisons are made. Notice that the means are closer when ω^{m} is nearer 1, and that the variances are closer when ω^{v} is nearer 1.

Assessing Identical Distribution

Distribution Assessments. Similarly, given return histories over a year $\{r(d)\}_{d=1}^{D}$, we can split the year into quarters and compare the emperical distribution of each quarter with that of another quarter or with that of the other three quarters combined. The maximum of all such comparisons made is the score for the year. For example, for each year we might define

$$\omega^{\mathrm{KS}} = 1 - \max\left\{ \|\widehat{Q}_2 - \widehat{Q}_1\|_{\mathrm{KS}} \,:\, \text{all comparisons made} \right\},$$
$$\omega^{\mathrm{Ku}} = 1 - \max\left\{ \|\widehat{Q}_2 - \widehat{Q}_1\|_{\mathrm{Ku}} \,:\, \text{all comparisons made} \right\}.$$

If we choose to compare quarters with each other then six comparisons are made. If we choose to compare each quarter with the other three quarters combined then four comparisons are made. Notice that $\omega^{\mathrm{Ku}} \leq \omega^{\mathrm{KS}} \leq 1$, and that the distributions are closer when ω^{Ku} is nearer 1.

Stationary Autoregression Models

Stationary Autoregression Models. One way to quantify how well a return history $\{r(d)\}_{d=1}^{D}$ is mimicked by an IID model is to fit it to a more complicated model and then measure how far that fit is from an IID model. We illustrate this approach using the family of *stationary autoregression models*. These models have the form

$$R_d = a + b\,R_{d-1} + Z_d \quad \text{for } d = 1, \cdots, D\,, \tag{4.4}$$

where a and b are real numbers, R_0 is a random variable and $\{Z_d\}_{d=1}^{\infty}$ is a sequence of IID random variable with mean zero.

Definition. An autoregression model in the form (4.4) is called *stationary* when for every $d \in \{1, \cdots, \infty\}$ the random variable R_d has the same statistical behavior as R_0.

Remark. We will see that stationarity implies that $|b| < 1$.

Stationary Autoregression Models

Let μ and ξ be the mean and variance of the random variable R_0. Then stationarity implies that

$$\mathrm{Ex}(R_d) = \mu\,, \quad \mathrm{Var}(R_d) = \xi\,, \quad \text{for every } d \in \{0, \cdots, \infty\}\,. \tag{4.5a}$$

Let ξ_d denote the covariance of R_d with R_0, so that

$$\begin{gathered}\xi_d = \mathrm{Cov}(R_0, R_d) = \mathrm{Ex}\big((R_0 - \mu)\,(R_d - \mu)\big) \\ \text{for every } d \in \{0, \cdots, \infty\}\,.\end{gathered} \tag{4.5b}$$

(Notice that $\xi_0 = \xi$.) Then stationarity implies that

$$\begin{gathered}\mathrm{Cov}(R_d, R_{d'}) = \mathrm{Ex}\big((R_d - \mu)\,(R_{d'} - \mu)\big) = \xi_{|d-d'|}\,, \\ \text{for every } d, d' \in \{0, \cdots, \infty\}\,.\end{gathered} \tag{4.5c}$$

Stationary Autoregression Models

Let η be the variance of the IID mean-zero variables Z_d. Then

$$\mathrm{Ex}(Z_d) = 0\,, \quad \mathrm{Var}(Z_d) = \eta\,, \quad \text{for every } d \in \{1, \cdots, \infty\}\,. \tag{4.6a}$$

Because the random variables $\{Z_d\}_{d=1}^{\infty}$ are IID, we have

$$\begin{gathered}\mathrm{Cov}(Z_d, Z_{d'}) = \mathrm{Ex}(Z_d\, Z_{d'}) = 0\,, \\ \text{for every } d, d' \in \{1, \cdots, \infty\} \text{ with } d \neq d'\,.\end{gathered} \tag{4.6b}$$

Because the random variable R_0 is independent of each Z_d, we have

$$\begin{gathered}\mathrm{Cov}(R_0, Z_d) = \mathrm{Ex}((R_0 - \mu)\, Z_d) = 0\,, \\ \text{for every } d \in \{1, \cdots, \infty\}\,.\end{gathered} \tag{4.6c}$$

Stationary Autoregression Models

Given the five parameters a, b, μ, ξ, and η we will now derive two relationships between these parameters as well as formulas in terms of these parameters for the covariances

$$\begin{aligned} &\mathrm{Cov}(R_d, R_{d'}) \qquad \text{for every } d, d' \in \{1, \cdots, \infty\}\,, \\ &\mathrm{Cov}(R_d, Z_{d'}) \qquad \text{for every } d, d' \in \{1, \cdots, \infty\}\,. \end{aligned}$$

We will thereby show that the mean-variance statistics of stationary autoregression models in the form (4.4) are specified by just three parameters.

Stationary Autoregression Models

Because each Z_d has mean zero, by taking expected values in (4.4) while using (4.6) we see that

$$\mu = \mathrm{Ex}(R_d) = a + b\,\mathrm{Ex}(R_{d-1}) + \mathrm{Ex}(Z_d) = a + b\mu\,.$$

Therefore a, b, and μ are related by

$$\mu = a + b\mu\,. \tag{4.7}$$

By using this relation to eliminate a from the form (4.4), we obtain

$$R_d = \mu + b\,(R_{d-1} - \mu) + Z_d \quad \text{for } d = 1, \cdots, \infty\,,$$

which can be recast as

$$R_d - \mu = b\,(R_{d-1} - \mu) + Z_d \quad \text{for } d = 1, \cdots, \infty\,. \tag{4.8}$$

Stationary Autoregression Models

Multiplying (4.8) by $Z_{d'}$ and taking expected values we obtain

$$\begin{gathered}\mathrm{Ex}\big((R_d-\mu)\,Z_{d'}\big)=b\,\mathrm{Ex}\big((R_{d-1}-\mu)\,Z_{d'}\big)+\mathrm{Ex}\big(Z_d\,Z_{d'}\big)\;,\\ \text{for every } d,d'\in\{1,\cdots,\infty\}\,.\end{gathered}\tag{4.9}$$

By using (4.6b) we see from (4.9) that

$$\begin{gathered}\mathrm{Ex}\big((R_d-\mu)\,Z_{d'}\big)=b\,\mathrm{Ex}\big((R_{d-1}-\mu)\,Z_{d'}\big)\;,\\ \text{for every } d,d'\in\{1,\cdots,\infty\} \text{ with } d<d'\,.\end{gathered}$$

Then by using (4.6c) we can prove by induction that

$$\begin{gathered}\mathrm{Cov}(R_d,Z_{d'})=\mathrm{Ex}\big((R_d-\mu)\,Z_{d'}\big)=0\,,\\ \text{for every } d,d'\in\{0,\cdots,\infty\} \text{ with } d<d'\,.\end{gathered}\tag{4.10}$$

Stationary Autoregression Models

By squaring (4.8) and taking expected values while using (4.5), (4.6), and (4.10), we see that

$$\begin{aligned}\xi = \mathrm{Var}(R_d) &= \mathrm{Ex}\Big((R_d - \mu)^2\Big) = \mathrm{Ex}\Big(\big(b\,(R_{d-1} - \mu) + Z_d\big)^2\Big)\\ &= b^2\mathrm{Ex}\Big((R_{d-1} - \mu)^2\Big) + 2b\,\mathrm{Ex}((R_{d-1} - \mu)\,Z_d) + \mathrm{Ex}\Big(Z_d^2\Big)\\ &= b^2\mathrm{Var}(R_{d-1}) + \mathrm{Var}(Z_d) = b^2\xi + \eta\,.\end{aligned}$$

Therefore b, ξ, and η are related by

$$(1 - b^2)\xi = \eta\,. \tag{4.11}$$

Because the variances ξ and η are positive, we see that

$$b^2 < 1\,, \qquad \eta \leq \xi\,.$$

Notice that if $b = 0$ then $\xi = \eta$ and the stationary autoregression model (4.8) reduces to an IID model.

Stationary Autoregression Models

By multiplying (4.8) by $(R_0 - \mu)$ and taking expected values while using (4.5b) and (4.6c) we see that

$$\begin{aligned} \xi_d &= \mathrm{Ex}\big((R_0 - \mu)\,(R_d - \mu)\big) \\ &= b\,\mathrm{Ex}\big((R_0 - \mu)\,(R_{d-1} - \mu)\big) + \mathrm{Ex}\big((R_0 - \mu)\,Z_d\big) \\ &= b\,\xi_{d-1}\,. \end{aligned}$$

Because $\xi_0 = \xi$, by induction we can show that

$$\xi_d = \xi\, b^d \quad \text{for every } d \in \{1, \cdots, \infty\}\,. \tag{4.12}$$

Because $|b| < 1$, we see that ξ_d decays as d increases.

Stationary Autoregression Models

By setting $d' = d$ in (4.9) while using (4.5a) and (4.10) we obtain

$$\mathrm{Cov}(R_d, Z_d) = \mathrm{Var}(Z_d) = \eta \,, \quad \text{for every } d \in \{1, \cdots, \infty\} \,. \tag{4.13}$$

By using (4.6b) we see from (4.9) that

$$\mathrm{Ex}\big((R_d - \mu)\, Z_{d'}\big) = b\, \mathrm{Ex}\big((R_{d-1} - \mu)\, Z_{d'}\big) \,, \\ \text{for every } d, d' \in \{1, \cdots, \infty\} \text{ with } d' < d \,.$$

Then by using (4.13) we can prove by induction that

$$\mathrm{Cov}(R_d, Z_{d'}) = \mathrm{Ex}\big((R_d - \mu)\, Z_{d'}\big) = \eta\, b^{d-d'} \,, \\ \text{for every } d, d' \in \{1, \cdots, \infty\} \text{ with } d' \leq d \,. \tag{4.14}$$

Because $|b| < 1$, we see that $\mathrm{Cov}(R_d, Z_{d'})$ decays as d increases.

Stationary Autoregression Models

The *autocorrelation time* t_{ar} of the stationary autoregression model (4.4) is defined by

$$\frac{1}{t_{\mathrm{ar}}} = \log\left(\frac{1}{|b|}\right) , \tag{4.15}$$

so that by (4.12) we have

$$|\xi_d| = \xi \exp\left(-\frac{d}{t_{\mathrm{ar}}}\right) , \qquad \text{for every } d \in \{0, \cdots, \infty\} ,$$

and by (4.14) we have

$$|\mathrm{Cov}(R_d, Z_{d'})| = \eta \exp\left(-\frac{d-d'}{t_{\mathrm{ar}}}\right) ,$$
$$\text{for every } d, d' \in \{1, \cdots, \infty\} \text{ with } d' \leq d .$$

The smaller t_{ar} the closer the stationary autoregression model is to an IID model.

Stationary Autoregression Models

In summary: from (4.5a) we have for every $d \in \{0, \cdots, \infty\}$ that

$$\mathrm{Ex}(R_r) = \mu\,, \qquad \mathrm{Var}(R_r) = \xi\,; \tag{4.16a}$$

from (4.12) we have for every $d, d' \in \{0, \cdots, \infty\}$ that

$$\mathrm{Cov}(R_d, R_{d'}) = \xi\, b^{|d-d'|}\,; \tag{4.16b}$$

from (4.10) and (4.14) we have for every $d \in \{0, \cdots, \infty\}$ and $d' \in \{1, \cdots, \infty\}$ that

$$\mathrm{Cov}(R_d, Z_{d'}) = \begin{cases} 0 & \text{if } d < d'\,, \\ \eta\, b^{d-d'} & \text{if } d' \leq d\,. \end{cases} \tag{4.16c}$$

Stationary Autoregression Models

We have seen that a stationary autoregression model in the form (4.4) is specified by three parameters. These can be $a \in \mathbb{R}$, $b \in (-1, 1)$, and $\eta > 0$, in which case μ, ξ, and ξ_1 are given by

$$\mu = \frac{a}{1-b}\,, \qquad \xi = \frac{\eta}{1-b^2}\,, \qquad \xi_1 = \frac{\eta\, b}{1-b^2}\,.$$

Alternatively, they can be $\mu \in \mathbb{R}$, $\xi > 0$, and $\xi_1 \in (-\xi, \xi)$, in which case a, b, and η are given by

$$a = \left(1 - \frac{\xi_1}{\xi}\right) \mu\,, \qquad b = \frac{\xi_1}{\xi}\,, \qquad \eta = \xi - \frac{{\xi_1}^2}{\xi}\,.$$

In the next section we will show how to pick the parameters to best fit a given data set.

Fitting Stationary Autoregression Models

Fitting Stationary Autoregression Models. Given a return history $\{r(d)\}_{d=0}^{D}$ and a choice of positive weights $\{w_d\}_{d=1}^{D}$ that sum to 1 we can use least squares to fit a stationary autoregression model of the form (4.4). Specifically, this approach constructs estmators $\hat{a}$ and $\hat{b}$ such

$$(\hat{a}, \hat{b}) = \arg\min\left\{ \sum_{d=1}^{D} w_d \big| r(d) - a - b\, r(d-1) \big|^2 \right\} , \qquad (5.17)$$

and then construct the estmator $\hat{\eta}$ by

$$\begin{aligned} \hat{\eta} &= \min\left\{ \sum_{d=1}^{D} w_d \big| r(d) - a - b\, r(d-1) \big|^2 \right\} \\ &= \sum_{d=1}^{D} w_d \big| r(d) - \hat{a} - \hat{b}\, r(d-1) \big|^2 . \end{aligned} \qquad (5.18)$$

Fitting Stationary Autoregression Models

It is helpful to define the return sample means

$$m_0 = \sum_{d=1}^{D} w_d \, r(d) \,, \qquad m_1 = \sum_{d=1}^{D} w_d \, r(d-1) \,, \tag{5.19a}$$

the return sample variances

$$v_{00} = \sum_{d=1}^{D} w_d \big(r(d) - m_0 \big)^2 \,, \qquad v_{11} = \sum_{d=1}^{D} w_d \big(r(d-1) - m_1 \big)^2 \,, \tag{5.19b}$$

and the return sample autocovariance

$$v_{10} = \sum_{d=1}^{D} w_d \big(r(d-1) - m_0 \big) \big(r(d) - m_1 \big) \,. \tag{5.19c}$$

It is also helpful to replace a with $\tilde{a}$ that is defined by

$$a = m_0 - b \, m_1 + \tilde{a} \,. \tag{5.20}$$

Fitting Stationary Autoregression Models

Then

$$\begin{aligned} z(d) &= r(d) - a - b\, r(d-1) \\ &= \big(r(d) - m_0\big) - b\,\big(r(d-1) - m_1\big) + \tilde{a} \\ &= \tilde{r}_0(d) - b\tilde{r}_1(d) + \tilde{a}\,, \end{aligned}$$

where we define

$$\tilde{r}_0(d) = r(d) - m_0\,, \qquad \tilde{r}_1(d) = r(d-1) - m_1\,. \tag{5.21}$$

Therefore

$$\begin{aligned} |z(d)|^2 = {} & |\tilde{r}_0(d)|^2 + b^2|\tilde{r}_1(d)|^2 + \tilde{a}^2 \\ & - 2b\,\tilde{r}_1(d)\,\tilde{r}_0(d) + 2\tilde{a}\,\tilde{r}_0(d) - 2\tilde{a}b\,\tilde{r}_1(d)\,. \end{aligned} \tag{5.22}$$

Fitting Stationary Autoregression Models

It is evident from (5.19) and (5.21) that $\{\tilde{r}_0(d)\}_{d=1}^D$ and $\{\tilde{r}_1(d)\}_{d=1}^D$ satisfy

$$\sum_{d=1}^{D} w_d \tilde{r}_0(d) = 0\,, \qquad \sum_{d=1}^{D} w_d \tilde{r}_1(d) = 0\,,$$

$$\sum_{d=1}^{D} w_d |\tilde{r}_0(d)|^2 = v_{00}\,, \quad \sum_{d=1}^{D} w_d |\tilde{r}_1(d)|^2 = v_{11}\,,$$

$$\sum_{d=1}^{D} w_d \tilde{r}_1(d)\, \tilde{r}_0(d) = v_{10}\,.$$

By using these facts we see from (5.22) that

$$\sum_{d=1}^{D} w_d |z(d)|^2 = v_{00} + b^2 v_{11} + \tilde{a}^2 - 2b\, v_{10}\,.$$

Fitting Stationary Autoregression Models

Because $v_{11} > 0$, the foregoing quantity is clearly minimized when

$$\tilde{a} = 0\,, \qquad b = \frac{v_{10}}{v_{11}}\,,$$

and that

$$\min\left\{ \sum_{d=1}^{D} w_d |z(d)|^2 \right\} = v_{00} - \frac{v_{10}^2}{v_{11}}\,.$$

Recalling (5.17), (5.18), and (5.20), this suggests using the estimators

$$\hat{a} = m_0 - \frac{v_{10}}{v_{11}}\, m_1\,, \qquad \hat{b} = \frac{v_{10}}{v_{11}}\,, \qquad \hat{\eta} = v_{00} - \frac{v_{10}^2}{v_{11}}\,. \tag{5.23}$$

Fitting Stationary Autoregression Models

However, the estimators (5.23) given by the least squares fit have a problem. Specifically, the formula for $\hat{b}$ can give values that lie outside of the interval $(-1,1)$. So rather than use the estimators (5.23), we will use the estimators

$$\hat{a} = m_0 - \frac{v_{10}}{v_{11}}\, m_1\,, \qquad \hat{b} = \frac{v_{10}}{\sqrt{v_{00}\, v_{11}}}\,, \qquad \hat{\eta} = v_{00} - \frac{v_{10}^2}{v_{11}}\,. \tag{5.24}$$

These estimators will satisfy $\hat{b} \in (-1,1)$ and $\hat{\eta} > 0$ if and only if the *autocovariance matrix* V is positive definite, where

$$V = \begin{pmatrix} v_{00} & v_{10} \\ v_{10} & v_{11} \end{pmatrix}\,. \tag{5.25}$$

This condition is always met in practice.

Fitting Stationary Autoregression Models

Remark. Given a return history $\{r(d)\}_{d=0}^{D}$ of any market index, we can use the autoregression estimator $\hat{b}$ given by (5.24) to estimate a autocorrelation time for that index. Motivated by formula (4.15), we define $\hat{t}_{\mathrm{ar}}$ by

$$\frac{1}{\hat{t}_{\mathrm{ar}}} = \log\left(\frac{1}{|\hat{b}|}\right) . \tag{5.26}$$

Because the history has length D, we would like $\hat{t}_{\mathrm{ar}} \ll D$ in order to have some confidence in our estimators of the return mean μ and the return variance ξ.

Assessing Independence

Assessing Independence. We will now present three ways to assess how much a given return history $\{r(d)\}_{d=1}^{D}$ that is consistent with the *identical distribution assumption* of an IID model is also consistent with the *independence assumption* of an IID model. More specifically, we will present:

- a graphical assessment,
- an autoregression assessment,
- an autocovariance assessment.

The first is purely visual, but can be used to build understanding of the data. The other two are analytical. They will yield measures ω^{ar} and ω^{ac} of how consistent the given data is with the independence assumption. As before, these measures will take values in the interval $[0, 1]$ with higher values indicating greater consistency with the independence assumption.

Assessing Independence

Graphical Assessment. In an IID model the random numbers $\{R_d\}_{d=1}^{D}$ are drawn from $(-1,\infty)$ in accord with the probability density $q(R)$ *independent* of each other. This means that there is no correlation between R_d and $R_{d'}$ when $d \neq d'$. Because of this, if we *scatter plot* the points $\{(R_d, R_{d+c})\}_{d=1}^{D-c}$ in the rr'-plane for any $c > 0$ then they will be distributed in accord with the probability density $q(R)q(R')$.

Therefore if the return history $\{r(d)\}_{d=1}^{D}$ is mimiced by such a model then when the points $\{(r(d), r(d+c))\}_{d=1}^{D-c}$ are scatter plotted in the rr'-plane they should appear to be distributed in a way consistant with the probability density $q(r)q(r')$.

We expect that the strongest correlation should be seen when $c = 1$ because the behavior of an asset price on any given trading day seems to correlate with its behavior on the previous trading day.

Assessing Independence

Autoregression Assessment. Given a return history $\{r(d)\}_{d=0}^{D}$ and a choice of positive weights $\{w_d\}_{d=1}^{D}$ that sum to 1, we define the return sample means

$$m_0 = \sum_{d=1}^{D} w_d \, r(d) \,, \qquad m_1 = \sum_{d=1}^{D} w_d \, r(d-1) \,,$$

the return sample variances

$$v_{00} = \sum_{d=1}^{D} w_d \big(r(d) - m_0\big)^2 \,, \qquad v_{11} = \sum_{d=1}^{D} w_d \big(r(d-1) - m_1\big)^2 \,,$$

and the return sample autocovariance

$$v_{10} = \sum_{d=1}^{D} w_d \big(r(d-1) - m_0\big)\big(r(d) - m_1\big) \,.$$

This is often done with uniform weights $w_d = 1/D$.

Assessing Independence

The estimators (5.24) for the autoregression model of the return history $\{r(d)\}_{d=0}^{D}$ are then given by

$$\hat{a} = m_0 - \frac{v_{10}}{v_{11}}\, m_1 \,, \qquad \hat{b} = \frac{v_{10}}{\sqrt{v_{00}\, v_{11}}} \,, \qquad \hat{\eta} = v_{00} - \frac{v_{10}^2}{v_{11}} \,. \qquad (6.27)$$

Notice that the last two estimators satisfy

$$\hat{\eta} = v_{00}\left(1 - \hat{b}^2\right) \,.$$

Because v_{00} is the sample variance of $\{r(d)\}_{d=1}^{D}$ while $\hat{\eta}$ is the sample variance of $\{z(d)\}_{d=1}^{D}$, we see that:

- *$\hat{b}^2$ is the fraction of the sample variance of $\{r(d)\}_{d=1}^{D}$ that is contributed by the autoregression term;*
- *$1 - \hat{b}^2$ is the fraction of the sample variance of $\{r(d)\}_{d=1}^{D}$ that is contributed by the the nugget term.*

Assessing Independence

This suggests that a natural measure of how well the history $\{r(d)\}_{d=1}^{D}$ can be mimicked by an IID model is

$$\omega^{\mathrm{ar}} = 1 - \hat{b}^2 = 1 - \frac{v_{10}^2}{v_{00}\, v_{11}} \,. \tag{6.28}$$

The closer ω^{ar} is to 1, the better the IID model.

Assessing Independence

Autocovariance Assessment. Consider the 2 × 2 *autocovariance matrix*

$$V = \begin{pmatrix} v_{00} & v_{10} \\ v_{10} & v_{11} \end{pmatrix} . \qquad (6.29)$$

This matrix is symmetric and is usually positive definite. If the data was drawn from an IID process with mean μ and variance ξ then it can be shown that

$$\mathrm{Ex}(V) = \xi W \,, \quad \text{where} \quad W = \begin{pmatrix} 1 - \bar{w} & -\bar{w}_1 \\ -\bar{w}_1 & 1 - \bar{w} \end{pmatrix} , \qquad (6.30)$$

with

$$\bar{w} = \sum_{d=1}^{D} w_d^2 \,, \qquad \bar{w}_1 = \sum_{d=2}^{D} w_d \, w_{d-1} \,.$$

Assessing Independence

The matrix W is known. For uniform weights $w_d = 1/D$ we have

$$\bar{w} = \frac{1}{D}\,, \qquad \bar{w}_1 = \frac{D-1}{D^2}\,,$$

whereby

$$W = \begin{pmatrix} 1 - \frac{1}{D} & -\frac{D-1}{D^2} \\ -\frac{D-1}{D^2} & 1 - \frac{1}{D} \end{pmatrix}.$$

It can be shown for $D > 1$ that in general we have

$$0 < \bar{w}_1 < \bar{w}\,, \qquad \bar{w} + \bar{w}_1 < 1\,, \tag{6.31}$$

which implies that the symmetric matrix W given by (6.30) is always *diagonally dominant* and thereby is always *positive definite*.

Assessing Independence

The deviation of V given by (6.29) from the form (6.30) measures of how well an IID model mimics the data. For example, its size can be measured with the Frobenius norm, which for any real matrix A is determined by

$$\|A\|_{\mathrm{F}}^2 = \mathrm{tr}\big(A^{\mathrm{T}} A\big) \,.$$

We first estimate ξ in the form (6.30) to give the best least squares fit with respect to this norm. In other words, we set

$$\hat{\xi} = \arg\min \Big\{ \mathrm{tr}\big((V - \xi W)^2\big) \Big\}$$

Because

$$\mathrm{tr}\big((V - \xi W)^2\big) = \mathrm{tr}\big(V^2\big) - 2\xi \; \mathrm{tr}\big(W\, V\big) + \xi^2 \, \mathrm{tr}\big(W^2\big) \,,$$

we see that

$$\hat{\xi} = \frac{\mathrm{tr}\big(W\, V\big)}{\mathrm{tr}\big(W^2\big)} \,. \tag{6.32}$$

Assessing Independence

When the estimator $\hat{\xi}$ is expressed in terms of the entries of the matrices V and W given by (6.29) and (6.30) we have

$$\hat{\xi} = \frac{(1-\bar{w})\big(v_{00}+v_{11}\big) - 2\bar{w}_1 v_{10}}{2\big((1-\bar{w})^2 + \bar{w}_1{}^2\big)} \,.$$

The fact that $\hat{\xi} > 0$ whenever $V \neq 0$ is can be seen directly from (6.32) and the following general fact, the proof of which is left as an exercise.

Fact. If A and B are symmetric matrices of the same size such that A is positive definite, B is nonnegative definite, and $B \neq 0$ then $\mathrm{tr}(AB) > 0$.

Moreover, it is evident from (6.30) and (6.32) that

$$\mathrm{Ex}(\hat{\xi}) = \frac{\mathrm{tr}\big(W\,\mathrm{Ex}(V)\big)}{\mathrm{tr}\big(W^2\big)} = \frac{\mathrm{tr}\big(\xi\, W^2\big)}{\mathrm{tr}\big(W^2\big)} = \xi \,.$$

Therefore $\hat{\xi}$ is an unbiased estimator of ξ.

Assessing Independence

The size of the deviation of V given by (6.29) from the form (6.30) is thereby quantified by

$$\frac{\| V - \hat{\xi} W \|_{\mathrm{F}}^2}{\| V \|_{\mathrm{F}}^2} = 1 - \frac{\mathrm{tr}(W\,V)^2}{\mathrm{tr}\big(V^2\big)\ \mathrm{tr}\big(W^2\big)}\,.$$

Therfore we defined the measure

$$\omega^{\mathrm{ac}} = \frac{\mathrm{tr}(W\,V)^2}{\mathrm{tr}\big(V^2\big)\ \mathrm{tr}\big(W^2\big)}\,. \tag{6.33}$$

This is the square of the cosine of the angle between V and W as determined by the Frobenius scalar product. The closer ω^{ac} is to 1, the better an IID model mimics the data.

Assessing Independence

Remark. From (6.33) we can show by using (6.29) and (6.30) that

$$1 - \omega^{\mathrm{ac}} = \delta^2 + \left(1 - \delta^2\right) \cos(\phi)^2 \,,$$

where

$$\delta^2 = \frac{(v_{00} - v_{11})^2}{(v_{00} - v_{11})^2 + (v_{00} + v_{11})^2 + 4v_{10}^2} \,,$$

$$\cos(\phi)^2 = \frac{\left(2(1 - \bar{w})v_{10} + \hat{w}_1(v_{00} + v_{11})\right)^2}{\left((1 - \bar{w})^2 + \bar{w}_1^2\right)\left((v_{00} + v_{11})^2 + 4v_{10}^2\right)} \,.$$

This shows that ω^{ac} is near 1 if and only if both δ and $\cos(\phi)$ are small. The first condition holds if and only if v_{00} and v_{11} are relatively close. The second holds if and only if the vectors $(1 - \bar{w}, \bar{w}_1)$ and $(2v_{10}, v_{00} + v_{11})$ are nearly orthogonal.