Abstract

This report is a brief discussion of the various functional forms and representations of a stochastic process as they appear in financial literature. We begin by considering a basic time series which is then appropriated as a series of returns and a portfolio of returns. The portfolio return maximization problem is then considered from the perspective of a decomposed stochastic process.

1 Two motivating examples

A random variable is a measurable function which maps from a possible set of outcomes to a measurable set. In our cases this set will be the real numbers. Typically one also takes the assumption that the sum of the probabilities are convex-linear. A *probability space* $(\Omega, \mathcal{F}, \mathbb{P})$, is a triple of any set of outcomes Ω combined with a particular σ-algebra of the subsets of this space $\mathcal{F}$, that is an algebra of sets over the power set, and a probability measure $\mathbb{P}$. A set $A \in \mathcal{F}$ is called an event and $\omega \in \Omega$ are defined as sample points. Consider the following two examples:

1. Let $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ be finite, and suppose that we have real numbers $0 \leq p_j \leq 1$ for index j such that $\sum p_j = 1$. Now, $\mathcal{F}$ is the set of all subsets of Ω which are closed under complement and countable unions. For a given set of outcomes $A = \{\omega_{ij}\}_{i=1}^{m}$, where $A \subset \mathcal{F}$, we define the probability of this set as:

$$\mathbb{P}(A) = \sum_{i=1}^{m} p_{ij} \leq 1$$

   This kind of a construction is fundamental to probability and statistics since we frequently experience the "chance" of an outcome as a non-unique mapping. That is, the chance of drawing a pair of Aces in a deck of cards is the same as drawing a pair of any other cards. And yet, the total chance of drawing all hands is unity. In this sense we may discuss the concept of the probability of a related set of events by their mutual chance or consider the algebraic relations of the set of possibilities simultaneously. We may also have recourse to the construct of functions of such random variables in an absolutely continuous sense.

2. The smallest σ-algebra containing all of the open subsets of $\mathbb{R}^n$ is called the Borel σ-algebra $\mathcal{B}$. Now, assume that there exists some f which is non-negative, integrable and has the property $\int_{\mathbb{R}^n} f \, \mathrm{d}x = 1$. We define:

   $$\mathbb{P}(B) = \int_B f(x) \, \mathrm{d}x \quad \text{for all} \quad B \subset \mathcal{B}$$

   Then we have the triple $(\mathbb{R}^n, \mathcal{B}, \mathbb{P})$ as a probability space and we say that f is the *density* of the measure $\mathbb{P}$. One last common construction follows. Suppose some singleton element is fixed, $\xi \in \mathbb{R}^n$. Then we may define:

   $$\mathbb{P}(B) := \begin{cases} 1 & \text{if} \quad \xi \in B \\ 0 & \text{if} \quad \xi \notin B \end{cases} \quad \text{for any } B \in \mathcal{B}$$

   Then $(\mathbb{R}^n, \mathcal{B}, \mathbb{P})$ is a probability space and we denote $\mathbb{P}$ as the **Dirac Mass** at ξ. With $\mathbb{P} = \delta_\xi$.

With these two examples as motivation we move onto the fundamental tools.

Wold Decomposition

By the Wold Decomposition we can write a covariance stationary stochastic process as a decomposition as a linear combination of lags of a white noise process and a process that can be predicted by a linear function of past observations. The classical decomposition returns a linear combination of a square summable sequence and fundamental innovations. Building on previous work, Ortu et al (2020) extends this using the Discrete Haar Transform so that X_t can be written in orthogonal frequency specific components.

$$x_t = \sum_{j=1}^{\infty} a_j Z_{t-j} + V_t$$

$$x_t = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} a_k^{(j)} z_{t-k2^j}^{(j)}$$

Where V_t is a deterministic component, z_t is a white noise process with positive second moments i.e. $z_t \sim \mathcal{N}(0, \sigma^2)$, and $\sum_{j=1}^{\infty} a_j^2 < \infty$. Finally, we also have that $\mathbb{E}\left[z_t \cdot V_t\right] = 0$.

The multivariate form of X_t is a natural extension where the a_j sequence instead becomes a sequence of absolutely summable matrices instead.

$$A_n = (a_{i,j,n}) : \sum_{n=1}^{\infty} |a_{i,j,n}| < \infty$$

We note that this is equivalent to demanding that the spectrum of A is inside the unit circle for all n, $\{\lambda \in \sigma(A) : |\lambda| < 1\}$. This is a common requirement for convergent numerical schemes and is satisfied by a rich class of matrices, including positive semi-definite matrices like covariances. One could proceed as in Bandi et al (2021) to write the returns from a portfolio of assets using the extended Wold Decomposition. Factor models, which estimate returns as a function of market conditions, are consistent with the extended Wold Decomposition. Where the risk components are decomposed in a frequency domain to form factors related to economic cycles of various duration.

Portfolio Theory

We begin by constructing a portfolio of assets and writing the returns from the time interval $[t-1, t)$ as the vector $\mathbf{R} = (R_i, \ldots, R_N)'$. The classical optimization solution for maximizing the expected return, $\mathbb{E}\left[\mathbf{R}\right] = \boldsymbol{\mu}$, is due to Harry Markowitz and involves the special purpose Critical Line Algorithm. The weights for portfolio allocation are a convex combination of assets that may be defined using a weighted average of valuations or other methods.

$$\text{Portfolio weights: } \sum_{i=1}^{N} w_i = \mathbf{1}'\mathbf{w} = 1; \qquad \text{Portfolio Return: } R_p = \mathbf{w}'\mathbf{R}$$

$$\mathbf{w}^* = \arg\min_{w} \mathbf{w}'\Sigma\mathbf{w} \qquad \text{subject to} \qquad \boldsymbol{\mu}'\mathbf{w} = \boldsymbol{\mu}_*$$

Where $\mathbb{E}\left[R_p\right] = \boldsymbol{\mu}'\mathbf{w}$, and $\boldsymbol{\mu}_*$ is a fixed target portfolio return. This method requires the covariance, $\mathbf{w}'\Sigma\mathbf{w} = Cov(\mathbf{R}_p, \mathbf{R}_p')$, and expected returns as inputs. For which it is often referred to as mean-variance portfolio construction.

Risk Factor

A common model for theoretically determining the rate of return for a portfolio of assets is the factor model. Where an asset return is a linear combination of sensitivity to systemic risk, β, with market factors and stochastic shocks.

Assume we have M factors, f_m, with $1 \leq m \leq M$. Write the return for any given asset as:

$$R_i = \alpha_i + \sum_{m=1}^{M} \beta_{i,m} f_m + \varepsilon_i$$

Collecting all of the assets held in a portfolio then gives the vectorized form for a returns series.

$$\mathbf{R} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \vdots \\ \alpha_M \end{bmatrix} + \begin{bmatrix} \beta_{11} & \beta_{12} & \beta_{13} & \dots & \beta_{1M} \\ \beta_{21} & \beta_{22} & \beta_{23} & \dots & \beta_{2M} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta_{N1} & \beta_{N2} & \beta_{N3} & \dots & \beta_{NM} \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ \vdots \\ f_M \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \vdots \\ \varepsilon_M \end{bmatrix}$$

$$= \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{f} + \boldsymbol{\varepsilon}$$

Asserting that asset shocks are uncorrelated, i.e. $Cov(\varepsilon_i, \varepsilon_j) = 0$, allows one to write $\mathbb{E}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\right] = D$. Where D is a diagonal matrix. Writing the covariance of the factors as $C = Cov(\mathbf{f}, \mathbf{f}')$ and taking $\boldsymbol{\beta} = Cov(\mathbf{R}\mathbf{f}')C^{-1}$ as the least squares projection gives the following convenient form for the covariance of returns.

$$\Sigma = Cov(\mathbf{R}, \mathbf{R}') = \boldsymbol{\beta} C \boldsymbol{\beta}' + D$$

Which is a typical matrix form for optimization of a quadratic programming problem under constraints, e.g. $\boldsymbol{\lambda}'\mathbf{1} = D$. Under these assumptions we can also decompose the factors into orthogonal frequency components using the extended Wold Decomposition. Empirical estimation of J covariance matrices is then a possible improvement on classical factor models.

$$\mathbf{R} = \boldsymbol{\alpha} + \sum_{j=1}^{J} \left(\boldsymbol{\beta}^{(j)} \mathbf{f}^{(j)} + \boldsymbol{\varepsilon}^{(j)} \right)$$

$$\Sigma^{(j)} = \boldsymbol{\beta}^{(j)} C^{(j)} \boldsymbol{\beta'}^{(j)} + D^{(j)}$$

Other improvements to factor models could include shrinkage of the empirical covariance matrix or improved return estimation using predictive models like LSTM or GARCH models on the volatility of the returns series. The following section is a primer on GARCH models and their use for volatility estimation.

Conditionally Heteroskedastic ARMA: ARCH

For financial time series the variance of the stochastic process under study is itself stochastic. The variance of returns of a price series fall into such a class and their study is referred to as stochastic volatility modelling. Original work by Robert Engle introduced ARCH models and a basic framework follows as adapted from Tsay & Chen(2019). Assume z is a stochastic process and σ is a time dependent standard deviation.

$$\begin{aligned} \text{returns at time t} &= r_t = \mu_t + \epsilon_t \\ &= \mathbb{E}\left[r_t \,|\, \mathcal{F}_{t-1}\right] + \sigma_t \cdot z_{t-1} \\ &= \mathbb{E}\left[r_t \,|\, \mathcal{F}_{t-1}\right] + \sqrt{\mathbb{E}\left[(r_t - \mu_t)^2 \,|\, \mathcal{F}_{t-1}\right]} \cdot z_{t-1} \end{aligned}$$

Just as with ARMA models, we can begin with a moving average term and then add an auto-regressive term. Here we are hoping to produce a formal series for the volatility or standard deviation of returns. Note that our calculated expected volatility from a time series may already be the product of a windowing method since sampling a variety of frequencies returns the daily, monthly, or annual volatility as scaled by $\sqrt{T}$.

$$\text{ARCH}(p) : \sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2$$

$$\text{GARCH}(p,q) : \sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \epsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

The most common GARCH model is typically a $(1, 1)$ process. That is we represent the volatility as a constant plus a weighted square of a residual and a weighted expected volatility from the previous time index.

$$\text{GARCH}(1,1): \quad \sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\omega_{t-1}^2$$

The necessary conditions for a GARCH model to hold are as follows:

- Non-negativity: α, β, $\omega \geq 0$.
- Mean-Reversion: $\alpha + \beta < 1$.
- Long-Run Variance: $\sigma_T^2 \xrightarrow{T\to\infty} \frac{\omega}{1-(\alpha+\beta)}$

Just as with ARMA models, there is an intuitive way to understand the formula. Large α and large β correspond to the effect of ”shocks” in the moving average and the rate of decay for ”shocks” respectively.

Information Entropy

The following is an introduction to simple betting schemes using material adapted from Cover & Thomas. A stochastic process is said to be *strongly stationary* if the joint distribution of any subset of the sequence of random variables is invariant with respect to shifts in the time index.

$$\begin{aligned}&Pr\{X_1 = x_1, \dots, X_n = x_n\}\\&= Pr\{X_{1+k} = x_{1+k}, \dots, X_{n+k} = x_{n+k}\}\\&\text{for any shift } k \text{ and for all } x_1, x_2, \dots, x_n \in \mathbb{R}\end{aligned}$$

Such a stochastic process satisfies the following relation:

$$\lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \dots X_n) = \lim_{n\to\infty} H(X_n |\, X_{n-1}, \dots, X_1)$$

The first is the per symbol entropy of the n random variables, and the second is the conditional entropy of the last random variable given the past. For stationary processes both the limits exist and are equal. The equivalence of these two definitions of entropy occurs due to the Asymptotic Equipartition Property. Which for i.i.d. variables only requires the weak law of large numbers.

Definition 1.1 (Entropy of a Random Variable)**.** The *differential entropy* of a continuous random variable X with density function $f(X)$:

$$H(X) = -\int_S f(X)\, \log(f(X))\, \mathrm{d}x$$

Where S is the support set of the random variable. Similarly, we may apply the same definition for a sequence of i.i.d. random variables $\{X_n\}$ with a joint distribution $f(X_1, \dots, X_n)$. Even further two related but often more useful quantities may be defined.

Definition 1.2 (Conditional Entropy)**.** By conditioning a joint distribution we can produce:

$$H(X \,|\, Y) = -\int_S f(X, Y) \log(f(X \,|\, Y))\, \mathrm{d}x\, \mathrm{d}y$$

$$\text{Clearly then } H(X \,|\, Y) = H(X, Y) - H(Y)$$

Definition 1.3 (Relative Entropy)**.** The Relative Entropy or Kullbach Leibler Divergence is given as:

$$D(P||Q) = \int_S P \log\left(\frac{P}{Q}\right)$$

Remark. The relative entropy is finite only if the support of P is contained in the support of Q. In the context of the Radon-Nikodym Derivative, if the two probability distributions are both absolutely continuous with respect to some measure μ over Ω this may be rewritten using the natural partitions $\mathcal{Q}$ and $\mathcal{P}$:

$$D(\mathcal{Q}||\mathcal{P}) = \int_\Omega q \log\frac{q}{p}\,\mathrm{d}\mu = \int_\Omega \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P$$

$$\text{Where } \; p = \frac{\mathrm{d}P}{\mathrm{d}\mu} \text{ and } \; q = \frac{\mathrm{d}Q}{\mathrm{d}\mu}$$

Definition 1.4 (Mutual Information)**.** The mutual information between two random variables with a joint density is defined as:

$$I(X;Y) = \int_S f(X,Y)\log\left(\frac{f(X,Y)}{f(X)f(Y)}\right)\mathrm{d}x\,\mathrm{d}y$$

$$I(X;Y) = D(f(X,Y)||f(X)f(Y)) = H(Y) - H(Y\,|\,X)$$

Remark. Note that for a given quantization of continuous random variables X, Y by the partitions $\mathcal{P}$ and $\mathcal{Q}$:

$$I(X;Y) = \sup_{\mathcal{P},\,\mathcal{Q}}(I([X]_\mathcal{P};[Y]_\mathcal{Q}))$$

Horse Betting

Applications of variational calculus using entropy as a minimizer are quite useful for betting schemes and portfolio optimization. Now if we consider a gambler's wealth over time as a function of random variables: $S(X) = o(X)b(X)$, where $o(X)$ is the odds given by the track for payout on horse i and $b(X)$ is the the fraction of wealth invested in each race. The number of horses for each race is m and the number of consecutive races is n.

$$\text{Then } \; b_i \geq 0 \;\text{ and }\; \sum_{i=1}^{m} b_i = 1$$

$$\text{Payout over n races: } S_n(X) = \prod_{i=1}^{n} S_i(X)$$

$$\text{And the } \textbf{Doubling Rate} \text{ is } W(b,p) = \mathbb{E}\left[\log S(X)\right] = \sum_{i=1}^{m} p_i \log b_i o_i$$

And now by the Weak Law of Large Numbers for $\{X_n\}_{n\geq 1}$ which are i.i.d. $X_i \sim p(x)$:

$$\frac{1}{n}\sum_{i=1}^{n} \log S(X_i) \to \mathbb{E}\left[\log S(X)\right]$$

$$\therefore \qquad S_n \doteq 2^{nW(b,p)}$$

The optimum doubling rate is the maximum doubling rate over all possible choices of a portfolio: **b**. The optimum doubling rate for fractional betting is then given by the best-case portfolio weights or bet sizes for each race:

$$W^*(p) = \sup_b W(b,p) = \sum_{i=1}^{m} p_i \log o_i - H(p)$$

The equality can be seen by since:

$$W(b,p) = \sum_{i=1}^{m} p_i \log\left(\frac{b_i}{p_i} p_i o_i\right) = \sum_{i=1}^{m} p_i \log o_i - H(p) - D(p||b)$$

One can also verify this by checking the extremal given by:

$$J(\mathbf{b}) = \mathbb{E}\left[\log(S(X))\right] + \lambda \sum_{i=1}^{m} b_i$$

Where the usefulness of defining the relative entropy is now quite clear. Furthermore, we have equality if and only if $b \equiv p$. That is, the gambler places bets in proportion to the odds of winning.

An Important Example

Consider a race track where there is no track take and the given odds are fair. That is, $\sum_{i=1}^{m}(o_i)^{-1} = 1$. Introduce a pdf over the horses as $r_i := (o_i)^{-1}$.

In this case we have a particularly elegant closed form.

$$W(b,p) = \sum_{i=1}^{m} p_i \log\left(\frac{b_i}{p_i}\frac{p_i}{r_i}\right) = D(p||r) - D(p||b)$$

Where we can see that the doubling rate is the difference between the distance of the bookie's estimate from the true distribution and the distance of the gambler's estimate from the true distribution. In other words, a gambler makes money only if their estimate, contained in b, is better than the bookie's.

Connections to Maximum Likelihood and Kernels

A kernel, or more specifically a reproducing kernel hilbert space, may be defined using the Kullback-Leibler Divergence. Consider a model where we are interested in estimating a parameter given some data, $\hat{\theta}$. For example, the Fisher Kernel where $\mathcal{I}$ is defined as the fisher information matrix.

$$k(x,x') = \nabla_\theta \log\Big(p(x,\hat{\theta})\Big)\mathcal{I}^{-1}\nabla_\theta \log\Big(p(x',\hat{\theta})\Big)$$

Note that if we are drawing our distributions, p, from an exponential family we may write a suitable form with a sufficient statistic of interest. Although Tsuda et al (2004) have shown its effectiveness in other cases.

$$\begin{aligned} p(x|\theta) &= \exp\{\theta^T s(X) + \varphi(\theta)\} \\ \implies \nabla_\theta \log(p(x,\theta)) &= s(X) + \nabla_\theta\varphi(\theta) \end{aligned}$$

Where it is patently clear why we may be interested in this formulation since a potentially non-linear relationship between a parameter of interest and our data points may be captured and calculated.

$$D_{KL}(\underbrace{p(x|\hat{\theta})}_{data}, \underbrace{p(x|\theta))}_{model} = \mathbb{E}\left[\log p(x|\hat{\theta})\right] - \mathbb{E}\left[\log p(x|\theta)\right]$$

$$\therefore \qquad H(p(x|\theta)) = D_{KL}(\underbrace{p(x|\hat{\theta})}_{data}, \underbrace{p(x|\theta))}_{model} + H(p(x|\hat{\theta}))$$

$$\implies \min_\theta \; H(p(x|\theta)) \qquad \propto \qquad \min_\theta \; D_{KL}(\underbrace{p(x|\hat{\theta})}_{data}, \underbrace{p(x|\theta))}_{model}$$

$$\text{But, } \min_{\theta} \;\; H(p(x|\theta)) \qquad \propto \qquad \max_{\theta} \;\; \mathcal{L}\{\theta, X\}$$

So we see that MLEs will also satisfy an optimization criteria for minimizing the Kullback-Leibler Divergence.

References

- Tsay, R. S., &; Chen, R. (2019). Nonlinear time series analysis. John Wiley &; Sons.

- Cover, T. M., &; Thomas, J. A. (2010). Elements of information theory. Wiley.

- Ortu, F., Severino, F., Tamoni, A., &; Tebaldi, C. (2020). A persistence-based wold-type decomposition for stationary time series. Quantitative Economics, 11(1), 203–230. https://doi.org/10.3982/qe994

- Tsuda, K., Akaho, S., Kawanabe, M., &; Müller, K.-R. (2004). Asymptotic properties of the fisher kernel. Neural Computation, 16(1), 115–137. https://doi.org/10.1162/08997660460734029

- Bandi, F. M., Chaudhuri, S. E., Lo, A. W., &; Tamoni, A. (2021). Spectral factor models. Journal of Financial Economics, 142(1), 214–238. https://doi.org/10.1016/j.jfineco.2021.04.024