Financial Theory II Summary

Rebecca Sela

May 11, 2006

1 General Econometrics

Suppose we observe a sample, $\{y_t\}_{t=1}^T \in S$, where y_t may be a vector and S is the *sample space*. We have a model in the form of a density function, $f(y_t;\theta)$, with $\theta \in \Theta$, where the true data generating process is this density with the true parameter value, $\theta_0 \in \Theta$. An *estimator* is a function from S to Θ given by $\hat{\theta} = \theta(y_1, ..., y_T)$. (This function may also be a function of *priors* in Bayesian econometrics.)

Models come from both theory and looking at the data through diagnostic checking.

Definition A time series is *covariance stationary* if the mean is finite and constant and if the covariances are finite and don't depend on t.

Definition Given $f(y_1, ..., y_T;\theta)$, the *maximum likelihood estimator*, $\hat{\theta}_{MLE}$ is defined as:

$$\hat{\theta}_{MLE} = \arg\max_{\theta\in\Theta}(\log f(y_1, ..., y_T;\theta))$$

One could incorporate some uncertainty about the model by including more parameters.

Definition Suppose $g_t(y,\theta)$ is a $k \times 1$ vector such that $E(g_t(y_t,\theta_0)) = 0$. Let $g_T(\theta) = \frac{1}{T}\sum_{t=1}^T g_t(y_t,\theta)$. The *GMM estimator* is defined by:

$$\hat{\theta}_{GMM} = \arg\max_{\theta\in\Theta}\left(-g_T(\theta)'\hat{W}g_T(\theta)\right)$$

where $\hat{W}$ is a positive definite matrix (with probability one) which does not depend directly on θ.

The choice of moment conditions does not (necessarily) come directly from a density, but they may be implied by a model.

Proposition 1.1 *Maximum likelihood estimation is a special case of GMM, with moment conditions,* $g(y_t,\theta) = \frac{\partial}{\partial\theta}\log f(y_t;\theta)$.

In general, GMM is more robust to misspecification, but it is not necessarily optimal when the model is correct. If we use MLE but the density chosen is incorrect, this is called *quasi-maximum likelihood*.

Ways to check the robustness of the model include:

- Estimate the model for only a subsample of the data and compare the estimates.
- Compare out-of-sample forecasts.
- See if adding other variables changes the results.
- Try the same model with other data (for example, from other countries).
- Use other measures of the dependent or independent variables.
- Compare to other models.

Definition Let $w_t = (y_t, x_t)$. Suppose

$$f(w_t; \theta) = f(y_t|x_t; \theta_1) f(x_t; \theta_2)$$

where θ_1 and θ_2 are not related and only θ_1 is of interest. Then, we call $\{x_t\}$ *weakly exogenous*.

Proposition 1.2 *Ordinary least squares is MLE if the error term is normally distributed and the independent variables are weakly exogenous.*

Proof Suppose we observe data $\{y_t, x_t\}_{t=1}^T$. Then, we may write:

$$\begin{aligned} f(y_1, x_1, ..., y_T, x_T|\theta) &= \prod_{t=1}^{T} f(y_t, x_t|past, \theta) \\ &= \prod_{t=1}^{T} f(y_t|x_t, past, \theta_1) f(x_t|past, \theta_2) \end{aligned}$$

Alternatively, we may *concentrate* or *marginalize* the density, by computing $\int_X f_{xy}(x, y)dx = f_y(y)$. In either case, the resulting likelihood leads to an OLS estimator. ∎

Proposition 1.3 *OLS is GMM with certain moment conditions.*

Proof Set $g_t(\beta) = (y_t - \beta x_t)x_t$. These moment conditions are identical to the normal equations of the OLS and the first order conditions of maximum likelihood. ∎

If there is heteroskedasticity, then the moment conditions continue to hold, so that GMM (which is identical to quasi-maximum likelihood) is still consistent. For an optimal estimator, if σ_t^2 were known, one could use maximum likelihood, which would be identical to WLS. (Changing the weighting matrix in GMM is irrelevant in this case, since the parameters are exactly identified.)

Definition An *m-estimator* maximizes $Q_n(\{w\};\theta)$, where $Q_n = \sum_{t=1}^{n} m_t(w_t,\theta)$.

MLE is an m-estimator with $m_t(w_t,\theta) = \log f(w_t;\theta)$.

Asymptotic Theory

Definition A sequence of random variables, $\{X_n\}$, *converges in probability* to ξ if $\lim_{n\to\infty} P(|X_n - \xi| > \epsilon) = 0$ for all $\epsilon > 0$. This is also written as $plim(X_n) = \xi$ or $X_n \to_p \xi$.

Definition An estimator, $\hat{\theta}_n$, of θ_0 is *consistent* if $plim(\hat{\theta}_n) = \theta_0$.

Proposition 1.4 *Suppose $a(x)$ is a function which is continuous at ξ. Suppose $plim_{n\to\infty} X_n = \xi$. Then, $plim_{n\to\infty} a(X_n) = a(\xi)$. Note that X_n and ξ may be either scalars or vectors.*

Definition $f_n(X_n,\theta)$ converges to $f(\theta)$ *uniformly in probability* if $\sup_{\theta\in\Theta} |f_n(X_n,\theta) - f(\theta)| \to_p 0$.

Theorem 1.5 *Suppose $\hat{\theta}_n = \hat{\theta}_n(\{w_t\}) = \arg\max_{\theta\in\Theta} Q_n(\{w_t\};\theta)$. Assume that the following regularity conditions hold:*

- *Q_n is measurable with respect to $\{w_t\}$ (that is, it can actually be calculated),*
- *Θ is compact, and*
- *Q_n is continuous at all $\theta \in \Theta$ for all $\{w_t\}$.*

Suppose that the following conditions also hold:

- *there exists $\overline{Q}(\theta)$ which is uniquely maximized at $\theta_0 \in \Theta$, and*
- *$Q_n(\{w_t\},\theta)$ converges uniformly in probability to $\overline{Q}(\theta)$.*

Then, $\hat{\theta}_n \to_p \theta_0$.

Proof *(Sketch.)* Though Q_n is random, it eventually looks like $\overline{Q}$. Since $\overline{Q}$ is uniquely maximized at θ_0 and Q_n gets close to $\overline{Q}$, $\hat{\theta}$ must get close to θ_0. ∎

Proposition 1.6 *Suppose $\{y_t\}$ is stationary and ergodic. Assume the regularity conditions hold. Assume that:*

- $E(m(w_t;\theta))$ *is uniquely maximized at* θ_0*, and*
- $E(\sup_{\theta\in\Theta}|m(w_t;\theta)|) < \infty$*.*

Then, the m-estimator, $\hat{\theta}$*, is consistent for* θ_0*.*

Proof $E(\sup_{\theta\in\Theta}|m(w_t;\theta)|) < \infty$ is sufficient to ensure uniform convergence in probability.

Corollary 1.7 *Since the MLE is an m-estimator, it is consistent if:*

- *the regularity conditions hold,*
- $\{w_t\}$ *is stationary and ergodic,*
- $P(f(y_t|x_t;\theta) \neq f(y_t|x_t;\theta_0)) > 0$ *for all* $\theta \neq \theta_0$*, and*
- $E(\sup_{\theta\in\Theta}|\log f(y_t|x_t;\theta)|) < \infty$*.*

The second-to-last condition ensures that distinct parameter values lead to different probability distributions, so that the parameter is uniquely identified.

Proposition 1.8 *Suppose that the following regularity conditions hold:*

- $\hat{\theta}$ *maximizes* $-\hat{g}'\hat{W}\hat{g}$ *on* Θ*,*
- $\hat{W} \to_p W$*, with* W *positive definite,*
- $\{w_t\}$ *are stationary ergodic,*
- $g(w,\theta)$ *is measurable,*
- $g(w,\theta)$ *is continuous in* θ *for all* w*, and*
- $\theta_0 \in \Theta$ *and* Θ *is compact*

Further, suppose that:

- $E(g(w_t,\theta)) = 0$ *if and only if* $\theta = \theta_0$*, and*
- $E(\sup_{\theta\in\Theta}|g(w_t,\theta)|) < \infty$

Then, $\hat{\theta}$ *is consistent for* θ_0*.*

Definition The *non-linear least squares model* is given by:

$$y_t = \phi(x_t,\theta) + \epsilon_t$$

where ϕ is a known, non-linear function and $E(\epsilon_t) = 0$.

Non-linear least squares can be estimated consistently using GMM with moment conditions $g_n(w_t, \theta) = \frac{1}{n}\sum_{t=1}^{n}(y_t - \phi(x_t, \theta))x_t$, since $E(g(\theta)) = E_x(E(g(w_t;\theta)|x_t)) = E(\phi(x_t, \theta_0) - \phi(x_t, \theta)|x_t)$ If $\theta = \theta_0$, then the expectation will be zero. For most functions and distributions of x_t, the converse will hold as well. Θ may need to be bounded in some cases to ensure compactness.

Definition Let $F_n(\xi) = P(X_n \leq \xi)$ and $F(\xi) = P(X \leq \xi)$. We say that $\{X_n\}$ *converges in distribution* to X, and write $X_n \to_D X$ if $\lim_{n\to\infty} F_n(\xi) = F(\xi)$ for all ξ where F is continuous.

Theorem 1.9 Law of Large Numbers. *Under some conditions,* $\overline{X} \to_p E(X)$.

Theorem 1.10 Central Limit Theorem. *Under some conditions,*

$$\frac{\sum_{i=1}^{n}(X_t - E(X_t))}{\sqrt{\sum_{t=1}^{n} Var(X_t)}} = \sqrt{n}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(X_t - E(X_t))}{\sqrt{\frac{1}{n}\sum_{t=1}^{n} Var(X_t)}}\right) \to_D Normal(0,1)$$

Theorem 1.11 Mean Value Theorem/Taylor Series Expansion. *Let $h(x)$ be continuously differentiable. Then, we may write $h(x) = h(x_0) + \frac{\partial h}{\partial x}|_{\overline{x}\in(x,x_0)}(x - x_0)$. Equivalently, $\frac{h(x)-h(x_0)}{x-x_0} = \frac{\partial h}{\partial x}|_{\overline{x}\in(x,x_0)}$, for some $\overline{x}$ between x and x_0 .*

Lemma 1.12 Slutsky Condition. *If $X_n \to_D X$ and $Y_n \to_P \alpha$, then $X_nY_n \to_D X\alpha$. This applies to vectors and matrices as well, so that if $X_n \to_D X$ and $A_n \to_P A$ then $A_nX_N \to_D AX$ and $X_n'A_nX_n \to_D X'AX$.*

We may apply these theorems to find that:

$$\begin{aligned}
0 = \frac{\partial Q_n}{\partial \theta}(\{w\}, \hat{\theta}) &= \frac{\partial Q_n}{\partial \theta}(\{w\}, \theta_0) + \left.\frac{\partial^2 Q_n}{\partial\theta\partial\theta'}\right|_{\overline{\theta}}(\hat{\theta} - \theta_0) \\
\hat{\theta} - \theta_0 &= -\left(\left.\frac{\partial^2 Q_n}{\partial\theta\partial\theta'}\right|_{\overline{\theta}}\right)^{-1}\frac{\partial Q_n}{\partial\theta}(\{w\}, \theta_0) \\
\sqrt{n}(\hat{\theta} - \theta_0) &= -\left(\frac{1}{n}\cdot\left.\frac{\partial^2 Q_n}{\partial\theta\partial\theta'}\right|_{\overline{\theta}}\right)^{-1}\frac{1}{\sqrt{n}}\cdot\frac{\partial Q_n}{\partial\theta}(\{w\}, \theta_0)
\end{aligned}$$

The first term converges in probability to a constant, and the second term converges in distribution to a normal random variable, which shows that $\sqrt{n}(\hat{\theta} - \theta_0)$ converges to a normal distribution, with the mean and variance depending on the expressions on the right-hand side.

Proposition 1.13 *Assume that:*

- $\theta_0 \in Interior(\Theta)$,
- $m(w, \theta)$ *is twice-continuously differentiable in* θ *for all* w,
- $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial m}{\partial \theta}(w_t, \theta_0) \rightarrow_D Normal(0, \Sigma)$, *where* Σ *is positive definite,*
- *there is a neighborhood,* N, *of* θ_0, *with* $E(\sup_{\theta \in N} \| \frac{\partial^2 m}{\partial \theta \partial \theta'} \|) < \infty$, *and*
- $E(\frac{\partial^2 m}{\partial \theta \partial \theta'}|_{\theta_0}) = \overline{H}$, *where* $\overline{H}$ *is non-singular.*

Then, $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_D Normal(0, \overline{H}^{-1} \Sigma \overline{H}^{-1})$.

Proof Let $s = \frac{\partial m}{\partial \theta}$ (this is the *score*) and $H = \frac{\partial^2 m}{\partial \theta \partial \theta'}$ (this is the *Hessian*). Then, $s(w_t, \hat{\theta}) = s(w_t, \theta_0) + H(\overline{\theta})(\hat{\theta} - \theta_0)$. Summing over all n and recalling that the sum of the scores is zero at the maximum, we find that:

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{t=1}^{n} s(w_t, \theta_0) + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} H(\overline{\theta})(\hat{\theta} - \theta_0) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= \left(\frac{1}{n} \sum_{t=1}^{n} H(\overline{\theta}) \right)^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} s(w_t, \theta_0) \\ &\rightarrow_D Normal(0, \overline{H}^{-1} \Sigma \overline{H}^{-1}) \end{aligned}$$

since $\frac{1}{n} \sum_{t=1}^{n} H(\overline{\theta}) \rightarrow_p \overline{H}^{-1}$ by the law of large numbers. ∎

Corollary 1.14 *For maximum likelihood estimation,* $\overline{H} = \Sigma$ *by the information equality, so the MLE converges to a normal distribution with variance* Σ^{-1}.

Proposition 1.15 *Assume that:*

- $\theta_0 \in Interior(\Theta)$,
- $g(w_t, \theta)$ *is continuously differentiable,*
- $\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g(w_t, \theta_0) \rightarrow_D Normal(0, S)$,
- $E(\sup_{\theta \in N} \| \frac{\partial g}{\partial \theta} \|) < \infty$ *for some neighborhood,* N, *about* θ_0, *and*
- $E(\frac{\partial g}{\partial \theta}) = G$ *is of full column rank.*

Then, $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_D Normal(0, (G'WG)^{-1} G'WSWG(G'WG)^{-1})$.

Proof In this case, $Q = -\frac{1}{2} g_n' \hat{W} g_n$, and $\frac{\partial Q}{\partial \theta}|_{\theta=\hat{\theta}} = -G(\hat{\theta})' \hat{W} g_n(\hat{\theta})$. By the mean value theorem,

$$\begin{aligned} g_n(\hat{\theta}) &= g_n(\theta_0) + G(\overline{\theta})(\hat{\theta} - \theta_0) \\ 0 &= G(\hat{\theta})' \hat{W} g_n(\hat{\theta}) \\ &= G(\hat{\theta})' \hat{W} g_n(\theta_0) + G(\hat{\theta})' \hat{W} G(\overline{\theta})(\hat{\theta} - \theta_0) \\ \sqrt{n}(\hat{\theta} - \theta_0) &= -\left(\frac{1}{n} G(\hat{\theta}) \hat{W} G(\overline{\theta}) \right)^{-1} \frac{1}{\sqrt{n}} G(\hat{\theta})' \hat{W} g_n(\theta_0) \end{aligned}$$

The first term converges in probability to $(G'WG)^{-1}$ and the second term converges in distribution to $Normal(0, G'WSWG)$. Thus, $\sqrt{n}(\hat{\theta} - \theta_0) \rightarrow_D Normal(0, (G'WG)^{-1} G'WSWG (G'WG)^{-1})$. ∎

To estimate the standard errors, we note that W is chosen before estimation and G can be computed based on the derivatives of the moment conditions (it is consistent to evaluate it at $\hat{\theta}$ instead of θ_0 as long as the derivatives are continuous). To estimate S, we note that $S = Var\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g(w_t, \theta_0) \right)$, which is the long-run variance. The best method to estimate S depends on the assumptions we can make:

- Suppose $g(w_t, \theta)$ is serially uncorrelated. Then, $Var(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g(w_t, \theta_0)) = Var(g(w_t, \theta_0))$ since there are no covariances. We may estimate $\hat{S} = \frac{1}{n} \sum_{t=1}^{n} g(w_t, \hat{\theta}) g(w_t, \hat{\theta})'$. (In the context of maximum likelihood estimators, this is called the *outer product of the gradient*.) This is the *heteroskedasticity-consistent estimator*.

- If there is serial correlation,

$$\begin{aligned} Var\left(\frac{1}{\sqrt{n}} \sum_{t=1}^{n} g(w_t, \theta_0) \right) &= \frac{1}{n} E\left(\sum g(w_t, \theta_0) g(w_t, \theta_0)' \right) \\ &\quad + \frac{1}{n} E\left(\sum (g(w_t, \theta_0) g(w_{t+1}, \theta_0)' + g(w_{t+1}, \theta_0) g(w_t, \theta_0)') \right) \\ &\quad + ... \end{aligned}$$

  (Note that each term of the sums is symmetric, since one term is the transpose of the other.) As the number of lags increases there are fewer elements in each sum, which may lead to noisy standard errors or a matrix that is not positive definite. To solve this, the sums at different lags may be weighted differently. *Newey-West* uses Barlett weights to weight the sums, based on the user's choice of the number of lags. Non-parametric methods may be used to smooth out the terms, based on automatic choices of bandwidths. (In this case, a kernel might be helpful in estimation.)

- One may also *pre-whiten* residuals by fitting a time series model to $\{g(w_t, \theta)\}$ and assuming that any remaining residuals are white noise. The VAR will then imply the long-run variance.

For GMM, the most efficient estimator sets $W = S^{-1}$, in which case the covariance matrix becomes $(GS^{-1}G)^{-1}$. This choice of weighting matrix gives more weight to the less variable moment conditions. Since S is estimated, it might not be positive definite or might be nearly singular, which can cause problems. In such a case, using $W = I$ is safer (and more robust). If GMM is exactly identified, G is invertible and the weighting matrix does not matter; in this case, the covariance matrix is $G^{-1}SG^{-1}$.

For an m-estimator, the covariance matrix is $H^{-1}SH^{-1}$. In the case of the MLE, $H = S$ by the information matrix inequality, so that the covariance matrix is H^{-1}, which makes MLE the most efficient m-estimator (it also achieves the Cramer-Rao lower bound). This means we may calculate H based on the Hessian, or we may use the information matrix, $E(H)$ instead. The three estimates agree asymptotically; if they disagree radically for a particular sample, there could be a problem with the model.

To test GMM for adequacy, we use the *J-test* for overidentifying assumptions. We set $J = \min_{\theta \in \Theta} g'Wg$, where $W = S^{-1}$. Then, under the null hypothesis that all the moment conditions have expected value 0, $\frac{1}{\sqrt{n}} \sum g(w_t; \theta) \rightarrow_D Normal(0, S)$, and $(n-p)J \sim \chi^2_{k-p}$, where k is the number of moment conditions and p is the number of parameters estimated. We reject for large values of the J statistic.

The J-test will fail to reject either if the model is right or if the instruments are bad (that is, they are only weakly correlated with the endogenous variables). In such a case, the moment conditions are true, but not useful in estimation. Weak instruments can lead to different asymptotics, misleading estimates, different confidence intervals, and a nearly singular long-run variance matrix. To check for weak instruments, one should check the correlation between the instrument and the (estimated) moment condition.

Wald tests can be performed, based on the estimated covariance matrix. Likelihood ratio tests are generally helpful only when the likelihood is known. LM tests also have their place.

2 Asset Pricing

Suppose an asset pays off X_{t+1} in period $t+1$. Then, its price today is $p_t = E_t(X_{t+1}M_{t+1})$, where M_{t+1} is called the *pricing kernel* or *stochastic discount factor*, and adjusts for the risk and the waiting time, and is assumed to be unknown at time t. The *return* is $R_{t+1} = \frac{X_{t+1}}{p_t}$. Then,

$$\begin{aligned} 1 &= E_t\left(\frac{X_{t+1}}{p_t} M_{t+1}\right) \\ &= E_t(R_{t+1}M_{t+1}) \\ &= Cov_t(R_{t+1}, M_{t+1}) + E_t(R_{t+1})E_t(M_{t+1}) \end{aligned}$$

The expected return is:

$$E_t(R_{t+1}) = \frac{1}{E_t(M_{t+1})}(1 - Cov_t(R_{t+1}, M_{t+1}))$$

Definition Suppose $X_{t+1} = 1$ in all states of nature. Then, the asset is called a *riskless asset*.

For a riskless asset,

$$R^f = \frac{1}{p_t} = \frac{1}{E_t(M_{t+1})}$$

Then, for a general asset:

$$\begin{aligned} E_t(R_{t+1}) &= R^f - \frac{Cov(R_t, M_{t+1})}{Var(M_{t+1})} \cdot \frac{Var(M_{t+1})}{E(M_{t+1})} \\ &= R^f - \beta_t \lambda_t \end{aligned}$$

where $\beta_t = \frac{Cov(R_{t+1}, M_{t+1})}{Var(M_{t+1})}$ is the population regression coefficient and $\lambda_t = \frac{Var(M_{t+1})}{E(M_{t+1})}$ is the coefficient of variation or the *price of risk*. λ_t does not depend on the individual asset.

In a cross-section, if β were known, one could check if λ were constant across all stocks in that period, by regressing the mean returns on the β's of the stocks to estimate λ.

One could assume that each β is constant over time and that λ_t varies each period. This yields the moment condition:

$$g_t(y_t, \theta) = r_t^e + \beta \lambda_t$$

where r_t^e is the return, adjusted for the risk-free rate. A simpler method notes that the returns on any two assets are always related by λ, which allows for another test of the model.

Definition The CAPM model for the pricing kernel is:

$$r_{it} - r_t^f = \alpha + \beta(r_t^m - r_t^f) + \epsilon_{it}$$

where r_t^f is risk-free rate, r_t^m is the market return, and r_{it} is the individual stock's return.

We may test whether $\alpha = 0$, to ensure that once risk-adjustment factors (such as excess return) are included, there is no additional return. Also, the estimate of the risk-free rate should be constant across all assets. Alternatively, one may write:

$$r_{it} = \alpha + (1 - \beta) r_t^f + \beta r_t^m + \epsilon_{it}$$

If the risk-free rate is constant over time and $\beta \approx 1$, then the constant term is approximately α. (If dividends are not adjusted for, then α will include dividend-yield as well.)

In this model, r_t^m is endogenous. However, it is still true that $1 = E(MR) = Cov(M, R) + E(M)E(R)$, so we may compute the population regression coefficient, $\frac{Cov(r_{it}, r_t^m)}{Var(r_t^m)}$. That is, the regression will still estimate $E(r_{it}|r_t^m)$, though we may no longer be estimating the structural parameter.

Definition In *arbitrage pricing theory*, we assume that markets are competitive and frictionless, and that returns are described by:

$$R_{it} = \alpha + \beta_i f_t + \epsilon_{it}$$

where f_t is a $K \times 1$ vector of factors common across all assets and ϵ_{it} are sufficiently uncorrelated across assets.

In arbitrage pricing theory, λ_t is approximated by a vector of factors. Often, one of the factors is a market index, designed to be a proxy for the market portfolio. Other factors may be industry specific or factors that affect only cyclical companies.

Definition For any utility function, *risk aversion* is defined by $-\frac{u''(c)c}{u'(c)}$. Constant relative risk aversion preferences are defined by $u(c) = \beta c^{1-\gamma}$; in this case, the coefficient of risk aversion is γ. If $\gamma = 1$, consumers are risk-neutral. If $\gamma = 0$, utility is linear, and asset pricing is independent of the consumption level. In most contexts, $\gamma \geq 2$ is considered reasonable.

Since M_t depends on the marginal utility of consumption, Hansen and Singleton used the utility function, $u(c) = c^\gamma$, to compute $M_t = \frac{u'(c_t)}{u'(c_{t-1})} = (\frac{c_t}{c_{t-1}})^{\gamma-1}$. We expect that $0 < \gamma < 1$. If this utility function is correct and if we have a measure of consumption, we may then apply GMM to the moment condition:

$$g_t = \beta R_t \left(\frac{c_t}{c_{t-1}}\right)^{\gamma-1} - 1$$

Since there are two parameters (β and γ), we need at least two moment conditions (otherwise, for any γ, we may set $\beta = \frac{1}{E((\frac{c_{t+1}}{c_t})^{\gamma-1} R_{t+1})}$, and the moment condition will be 0). For more moment conditions, we may use the returns on multiple assets (since neither parameter depends on the asset chosen). Also, since the expectation is 0 with respect to all time t information, we must have $E(x_t(\beta(\frac{c_{t+1}}{c_t})^{\gamma-1} R_{t+1} - 1)) = 0$ for any x_t known at time t, which gives additional moment conditions.

Using maximum likelihood instead would require a joint distribution for (R_t, C_t), which could be complicated. Log-normal distributions are sometimes used, but the tests often have low power.

3 Volatility

Volatility measures the amplitude of returns, regardless of sign. Volatility can be predictable, while returns themselves are unpredictable, using tests like the Q-test for autocorrelation. Volatility tends to stay high or low for a long period; this is called *volatility clustering*.

Definition In the *stochastic volatility model*, we model $r_t = \sigma_t \epsilon_t$, where ϵ_t is a zero-mean, variance one random variable and $\log \sigma_t = \alpha + \beta \log \sigma_{t-1} + \kappa \eta_t$. Often, we assume that ϵ_t, η_t are serially uncorrelated and independent of each other. In this model, volatility is a latent variable. (There are other possible forms for the volatility, such as long memory or other functional forms in the equation.)

In this model, there are two error terms affecting a single series. This causes problems for both prediction and estimation. However, this model can be applied in continuous time.

Definition A *conditional volatility model* is one that satisfies $r_t = \sqrt{h_t}\epsilon_t$ where $h_t = E_{t-1}(r_t^2)$. In this model, h_t is called the *conditional variance*.

These models are covariance stationary for r_t if the unconditional variance is finite and does not depend on t. One should check the coefficients to ensure that the conditional variance cannot be negative.

Since h_t is a conditional expectation, it must be measurable at time $t-1$, so there is no error term at time t. This model specifically builds forecasting into the model. Since $E_{t-1}\left(\left(\frac{r_t}{\sqrt{h_t}}\right)^2\right) = 1$, we must have $Var(\epsilon_t) = 1$. If returns are not forecastable, $E\left(\frac{r_t}{\sqrt{h_t}}\right) = E(\epsilon_t) = 0$.

Every stochastic volatility model implies a conditional variance model, but the derivation is generally not simple.

Definition In the *autoregressive conditional heteroskedasticity* model, or *ARCH(p)* model, we have $h_t = \omega + \alpha_1 r_{t-1}^2 + ... + \alpha_p r_{t-p}^2$. If $\alpha_1 = ... = \alpha_p = 0$, this is a *constant variance model*. If $\alpha_j = \frac{1}{p}$ and $\omega = 0$, this is a *moving average (historical volatility) model*, $h_t = \frac{1}{p}\sum_{j=1}^{p} r_{t-j}^2$.

Definition In the *generalized autoregressive conditional heteroskedasticity model*, or *GARCH* model, $h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1}$. In general, the *GARCH(p,q)* model is $h_t = \omega + \sum_{j=1}^{p} \alpha_j r_{t-j}^2 + \sum_{j=1}^{q} \beta_j h_{t-j}$. This is an *exponential smoother* if $\omega = 0$ and $\alpha + \beta = 1$, so that $h_t = \alpha r_{t-1}^2 + (1-\alpha) h_{t-1}$.

Both the ARCH and GARCH models can be estimated from a single time series of data. Empirically, one lag of the GARCH term, h_t, is generally enough in a GARCH model.

Using repeated substitution for h_{t-1}, a GARCH(1,1) model can be written as an $ARCH(\infty)$ model:

$$\begin{aligned} h_t &= \omega + \alpha r_{t-1}^2 + \beta(\omega + \alpha r_{t-2}^2 + \beta h_{t-2} \\ &= \omega(1+\beta) + \alpha(r_{t-1}^2 + \beta r_{t-2}^2) + \beta^2 h_{t-2} \\ &= ... \\ &= \frac{\omega}{1-\beta} + \sum_{j=1}^{\infty} \alpha\beta^{j-1} r_{t-j}^2 \end{aligned}$$

For the GARCH(1,1) model, we find the unconditional variance using the law of iterated expectations:

$$\begin{aligned} \theta_t &= E(r_t^2) = E(h_t \epsilon_t^2) \\ &= E(h_t E_{t-1}(\epsilon_t^2)) = E(h_t) \\ &= E(\omega + \alpha r_{t-1}^2 + \beta h_{t-1}) \\ &= \omega + \alpha\theta_{t-1} + \beta\theta_{t-1} \\ &= \omega + (\alpha+\beta)\theta_{t-1} \\ \theta_t &= \frac{\omega}{1-\alpha-\beta} + (\alpha+\beta)^t \theta_0 \end{aligned}$$

where θ_0 may be any positive initial value. If $|\alpha+\beta| > 1$, the variance will explode over time. If $|\alpha+\beta| < 1$, the variance process will be mean-reverting. If $\alpha+\beta = 1$, as in the exponential smoother, $\theta_t = \frac{0}{0} + \theta_0 1^t$, which is undefined. If the unconditional variance is constant, then $\theta_t = \theta_{t-1} = \frac{\omega}{1-\alpha-\beta}$.

We also calculate the higher moments:

$$E(r_t^3) = E(h_t^{3/2} E_{t-1}(\epsilon_t^3))$$

If ϵ_t is symmetric, then this will be 0.

If the *conditional kurtosis*, $E_{t-1}(\epsilon_t^4)$ is constant and equal to k^c, then,

$$\begin{aligned} E(r_t^4) &= E(h_t^2 E_{t-1}(\epsilon_t^4)) \\ &= k^c E(h_t^2) \end{aligned}$$

By Jensen's inequality, we find that the unconditional kurtosis satisfies:

$$\begin{aligned} k &= \frac{E(r_t^4)}{E(r_t^2)^2} \\ &= \frac{k^c E(h_t^2)}{E(h_t)^2} \geq k^c \end{aligned}$$

Thus, ARCH and GARCH data will have more kurtosis unconditionally than conditionally; this allows a normal distribution for ϵ_t to lead to a leptokurtotic (fat-tailed) distribution for r_t.

In practice, $\alpha + \beta$ tends to be close to one, and ω is close to 0.

While the linearity of the GARCH model makes taking expectations and forecasting either, it might not be correct empirically.

Definition In an *asymmetric volatility model*, also called a *threshold ARCH (TARCH) model* or a *GJR-GARCH* model, we model $h_t = \omega + \alpha r_{t-1}^2 + \beta h_{t-1} + \gamma r_{t-1}^2 d_{t-1}$, where $d_t = 1$ if $r_t < 0$ and 0 otherwise.

Definition In the *exponential GARCH (EGARCH) model*, $\log h_t = \omega + \beta \log h_{t-1} + \alpha \frac{|r_{t-1}|}{\sqrt{h_{t-1}}} + \gamma \frac{r_{t-1}}{\sqrt{h_{t-1}}}$.

In the TARCH model, squared positive returns increase volatility by α while squared negative returns increase volatility by $\alpha + \gamma$. In a TARCH model, $\alpha + \beta + \gamma$ may be greater than 1. In EGARCH, positive and negative returns have different effects ($\alpha + \gamma$ and $\alpha - \gamma$ respectively); if $\gamma < 0$, then negative returns increase volatility. Since we are now modeling the logarithm of volatility, we no longer need to constrain anything to be non-negative.

For forecasting with TARCH, we note that h_{t+1} is known at time t, assume that r_t is symmetric about zero (so that $E_t(d_{t+1}) = \frac{1}{2}$), and compute:

$$\begin{aligned} h_{t+2} &= \omega + \alpha r_{t+1}^2 + \beta h_{t+1} + \gamma r_{t+1}^2 d_{t+1} \\ E_t(r_{t+2}^2) &= E_t(h_{t+2}) \\ &= \omega + \alpha E_t(r_{t+1}^2) + \beta E_t(h_{t+1}) + \gamma E_t(r_{t+1}^2 d_{t+1}) \\ &= \omega + \alpha h_{t+1} + \beta h_{t+1} + \gamma \left(\frac{1}{2} h_{t+1}\right) \\ &= \omega + \left(\alpha + \beta + \frac{\gamma}{2}\right) h_{t+1} \end{aligned}$$

If we assume stationarity, then we can compute the unconditional variance as $E(r_t^2) = \frac{\omega}{1-(\alpha+\beta+\gamma/2)}$. For this to be finite, we must have $\alpha + \beta + \frac{\gamma}{2} < 1$.

All of these models assume that $E(r_t) = 0$, which is not exactly correct (since returns are positive over the very long run). However, the mean is close enough to 0 that it can be approximated by 0 for daily data or higher frequency data.

TARCH can also be modeled as having a different threshold. However, this makes the model more complicated because it no longer takes advantage of symmetry.

In TARCH, ARCH or GARCH, the conditional forecast of the variance at long horizons reverts exponentially fast to the unconditional variance:

$$\begin{aligned} E_t(h_{t+2}) &= E(h_t)\left(1 - \left(\alpha + \beta + \frac{\gamma}{2}\right)\right) + \left(\alpha + \beta + \frac{\gamma}{2}\right) h_{t+1} \\ E_t(h_{t+k}) - E(h_t) &= \left(\alpha + \beta + \frac{\gamma}{2}\right)^{k-1} (h_{t+1} - E(h_t)) \end{aligned}$$

If $\alpha + \beta + \frac{\gamma}{2}$ is close to 1, then the mean-reversion occurs more slowly. This (and many of the other results) assumes that $E(h_t)$ is constant over time.

Definition Another volatility model is *PARCH* or *Power ARCH*, where $h_t = \alpha(r_t^2)^\beta h_t^{1-\beta}$.

To select a model for volatility, we may first use *diagnostic checks*. One should first test the returns or the residuals from a model for the returns (using a correlogram) to ensure that there is no autocorrelation. Then, one can use LM tests that regress squared standardized residuals (that is, corrected for h_t) on past squared residuals to check for autocorrelation in volatility. One can also use a correlogram on the squared, standardized residuals after fitted to ensure that no correlation remains. A histogram of the standardized residuals should have mean 0 and variance 1, as well as skewness or kurtosis that matches the distribution for ϵ_t (one can also use the Jacques-Berra test to ensure that normality is not rejected). Second, for model selection, we use an information criterion. Though the likelihood always increases if there are more parameters, we may use a likelihood ratio test to compare nested models. In general, the AIC or BIC can be used to choose between models (we wish to minimize the criterion of choice).

For estimation, we note that any of these models can be written as $r_t = \sqrt{h_t}\epsilon_t + m_t$, where m_t is the conditional mean, h_t is the conditional variance, and $\epsilon_t \sim Normal(0,1)$ independent of m_t, h_t. Then, the conditional likelihood is:

$$\begin{aligned} r_t|\mathcal{I}_{t-1} &\sim Normal(m_t, h_t) \\ \log f_t(r_t|\mathcal{I}_{t-1}) &= 1 - \frac{1}{2}\log(2\pi h_t) - \frac{1}{2h_t}(r_t - m_t)^2 \end{aligned}$$

Summing up leads to the joint log likelihood:

$$l = \sum_{t=1}^{T} \log f_t(r_t) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(h_t) - \frac{1}{2}\sum_{t=1}^{T}\frac{(r_t - m_t)^2}{h_t}$$

We may then insert any functional forms for h_t, m_t. Taking the derivatives with respect to the parameters, θ, that affect h_t (and assuming that m_t does not depend on the same parameters!) shows that:

$$\begin{aligned} \frac{\partial l}{\partial \theta} &= -\frac{1}{2}\sum_{t=1}^{T}\frac{1}{h_t}\frac{\partial h_t}{\partial \theta} - \frac{1}{2}\sum_{t=1}^{T}\left(-\frac{(r_t - m_t)^2}{h_t^2}\right)\frac{\partial h_t}{\partial \theta} \\ &= -\frac{1}{2}\sum_{t=1}^{T}\frac{h_t - (r_t - m_t)^2}{h_t^2}\frac{\partial h_t}{\partial \theta} \end{aligned}$$

Since h_t depends only on the past, the expectation of $\frac{\partial l}{\partial \theta}$ is 0 (as it always should be if the likelihood is correct), since $E\left((r_t - m_t)^2\right) = h_t$ and $\frac{\partial h_t}{\partial \theta}$ depends only on the past.

If we treat this as an m-estimator (quasi-MLE) instead, then the estimates will be consistent even if normality does not hold, as long as $E\left((r_t - m_t)^2\right) = h_t$. In this case, Bollerslev-Wooldridge standard errors should be used; these assume that the scores are serially uncorrelated (which is reasonable since the returns are serially uncorrelated), and may lead to much bigger standard errors.

We could also use the likelihood function for a different distribution (for example, the Student's t distribution with the degrees of freedom estimated; smaller degrees of freedom lead to larger kurtosis). However, if the distribution is wrong, the estimator is no longer even a quasi-MLE.

One can also estimate the density from the first-stage residuals. This is called an *adaptive* or *semiparametric* estimator, and requires more data to really get the tail behavior right.

3.1 Economic Reasons

For most equity time series, $\delta > 0$. Economic reasons for this may include:

- *Leverage Effects:* A negative return implies that the debt/equity ratio increases. This means that the volatility of equity must increase.
- *Risk aversion:* News about volatility (since price decreases predict higher future volatility) may cause investors to change holdings.

The leverage theory has two problems. First, the effects on volatility seem disproportionately large. Second, the effect seems to be stronger for indices than for individual stocks.

Volatility "news" may come in two forms. First, one may know that something (like a macroeconomic announcement) is going to happen, but not know whether it will lead to a positive or a negative movement. (This can be seen in options price data.) Second, knowing today's volatility forecasts tomorrow's volatility. If people are rational, then volatility must come from news; any volatility clustering would come from the clustering of news. However, if there is also price discovery, commentary on news, or reaction to the trades based on the news, then the volatility associated with a single news event may last longer, leading to volatility clustering.

3.2 Skewness

Skewness can be used as a measure of downside risk. Asymmetry in volatility leads to negative skewness, since price increases lead to low volatility and therefore little movement the next day, while price declines lead to high volatility, and potentially bigger declines. As data is aggregated over days, skewness increases at first (up to 30-40 days) and then declines. Returns have been more skewed since the 1987 crash.

Skewness should affect the risk-neutral distribution, through risk aversion and through the empirical distribution of returns. Therefore, skewness affects options prices and any other prices.

Empirically, individual stocks are positively skewed while indexes are negatively skewed.

4 Options

Definition *Options* are contracts that give the holder the right but not the obligation to do something. A *call* option gives the right to buy a certain stock at a pre-specified *strike price*. A *put* option gives the right to sell a certain stock at a pre-specified strike price. A *European* option must be exercised at a particular *expiration date*, while an *American* option can be exercised at any time up to the expiration date. A *Bermuda* option can be exercised a a fixed number of dates in an interval.

Definition An option is *in-the-money* if it will lead to a positive payoff at the current price of the underlying stock, *out-of-the-money* if it won't, and *at-the-money* if the underlying stock price is equal to the (discounted) strike price. An option is *forward at-the-money* if the present value of the strike price equals the underlying price ($K = S_0 e^{rT}$). (This distinction matters most for larger risk-free rates and larger expirations dates.)

A call or put option is described by the strike price, K, the current underlying price of the asset, S_t (or S_0 if we start the clock today), and the date it expires, T. The price today is C_t or P_t. At time T, the payoff (and therefore the price, if it were traded) is:

$$\begin{aligned} C_T &= \max(0, S_T - K) \\ P_T &= \max(0, K - S_T) \end{aligned}$$

Note that the possible gain on a call option is unbounded, and that both options have bounded losses. This means that short and long positions in options are not symmetrical. Since the payoff is non-linear, the price of the option must be sensitive to the whole distribution. In particular, the payoff depends on the volatility of the underlying stock price, σ.

For any $t < T$, even if the option is out-of-the-money, the price has time to move so that there might still be a positive payoff. This means that the price lies above the payoff curve. Thus, an option is more valuable than an equivalent position in the stock.

The *Greeks* for call options are:

- *Delta*: $\frac{\partial C_t}{\partial S_t}$. This measures how the option price is affected by the price of the underlying. $0 < \frac{\partial C_t}{\partial S_t} < 1$ in all cases, since the option price must increase with the underlying price (since the expected payoff is higher), but it doesn't go up as fast (because it is asymptotically equal to the underlying?). Delta goes to zero for longer maturities.

- *Gamma*: $\frac{\partial^2 C_t}{\partial S_t^2}$. The gamma must always be positive, since the price curve as a function of the underlying price is convex; if the curve were not convex, one could use *butterfly spreads* which one buys two options for the middle strike price and sells an option on either side of that, which would lead to an arbitrage profit or $C(K+\epsilon) - 2C(K) + C(K-\epsilon)$. Gamma is larger when the option is close to at-the-money and at short maturities at-the-money (as the curve goes the final payoff which has infinite curvature at the money).

- *Vega*: $\frac{\partial C_t}{\partial \sigma}$. This is positive, since the likelihood of the underlying price going up more or down more increases, but downside risks are bounded for options. Note that $\frac{\partial C_t}{\partial \sigma}$ goes to 0 as the option is deeper into and out of the money.

- *Theta*: $\frac{\partial C_t}{\partial T}$. This is usually positive, since having more time to maturity generally leads to a larger variance in the stock price. On the other hand, if the risk-free interest rate is high, then the theta might be negative for a European option, since one has to wait longer to exercise the option; this will only happen in the much longer run, and is more likely with puts (where the payoff is bounded above by the strike price). Mean-reverting volatility can also lead to this sort of effect. (Theta can also be defined as $\frac{\partial C_t}{\partial (T-t)}$, which reverses the sign and gives the interpretation as the time rate of decay on options prices.)

For stocks, the delta is always 1, while the gamma and vega are 0. For puts, the delta is always negative, while the gamma, vega, and theta are always positive. For two options on the same underlying security, the Greeks for the portfolio is just the sum of the individual Greeks, since derivatives are linear and the value of a portfolio is the sum of the individual values.

The most active options in trading are usually the ones that are at-the-money. Also, the options that are due to expire soon (but not too soon) tend to be more active. Analysis should be based on these, if possible. Since many options are not frequently traded, analysis should often be done using the midpoint of the current bid and ask prices, since that is always kept current. Using the last trade will lead to a less smooth curve (where it is even defined), since the bid-ask spread no longer needs to contain the last price. The curves of bids and asks are based on market makers' formulas; it is not necessarily true that people would want to trade at these prices.

4.1 Options Pricing Models

Suppose we have a pricing kernel, M^*, or, equivalently, a risk-neutral density, f^*. Then, the prices must be:

$$\begin{aligned} P_t &= E_t(\max(K - S_T, 0)M^*) \\ &= e^{-r(T-t)} E_t^*(\max(K - S_T, 0)) \\ C_t &= e^{-r(T-t)} E_t^*(\max(0, S_T - K)) \end{aligned}$$

where E_t^* is the expectation at time t using the risk-neutral density. Note that the risk premium is:

$$e^{r(T-t)}(E_t(x) - E_t^*(x))$$

Since the risk-neutral density is constant across all types of derivatives based on the same underlying asset, $S_t = e^{-r(T-t)}E_t^*(S_T)$.

Definition *Put-call parity* is defined by:

$$C_0 - P_0 = e^{-rT}E_t^*(S_T - K) = S_0 - Ke^{-rT}$$

which must hold as long as the risk-neutral density is the same for call and put options with the same strike price and expirations.

In the *Black-Scholes* pricing method, the risk-neutral density satisfies:

$$dS = rSdt + \sigma Sdz$$

where dz is the random Gaussian measure, and both σ and r are assumed to be constant. Using this risk-neutral density, we calculate the expected payoff:

$$\begin{aligned}
E(S_T) &= S_0 e^{rT} \\
\log\left(\frac{S_T}{S_0}\right) &\sim Normal\left(\left(r - \frac{\sigma^2}{2}\right)T, \sigma^2 T\right) \\
E(\max(V-K,0)) &= E(V)\Phi(d_1) - K\Phi(d_2) \\
C_0 &= S_0\Phi(d_1) - Ke^{-rT}\Phi(d_2) \\
P_0 &= Ke^{-rT}\Phi(-d_2) - S_0\Phi(-d_1) \\
d_1 &= \frac{1}{\sigma\sqrt{T}}\left(\ln\left(\frac{S_0}{Ke^{-rT}}\right) - \frac{T\sigma^2}{2}\right) \\
d_2 &= d_1 - \sigma\sqrt{T}
\end{aligned}$$

For an at-the-money forward call option, this can be approximated by $c_0 \approx S_0\sigma\sqrt{\frac{T}{2\pi}}$.

Definition The *implied volatility* is the value of σ in the Black-Scholes formula that would make the actual price of the option agree with the formula.

Implied volatilities are generally better predictors of future volatilities than GARCH predictions, since implied volatilities built in the expectations of the people trading the options.

If Black-Scholes were correct, implied volatility would be the same for options at any maturity or expiration.

Because of put-call parity, put and call options at the same strike price imply the same volatility (approximately). Empirically, implied volatility is highest for puts or calls that are far out of the money, and lower at-the-money. This shape is called the *volatility smile*. For indexes, there tends to be a *volatility smirk*, where the implied volatility is lower at higher strike prices, turning up less for out-of-the-money puts. The smirk tends to be steeper over longer horizons, perhaps because the long-run risks are greater, with either volatility being expected to rise or more uncertainty in financial markets in the long run. For stocks, there tends to be more of a smile; this may come from higher implied volatility if a stock could be overvalued (then, puts are a hedge against a correction).

In the *Hull and White* pricing method, they compute:

$$\begin{aligned} P_t &= E_t^* \left(E_t(\max(K - S_t, 0)|\overline{\sigma}) \right) \\ C_t &= E_t^* \left(E_t(\max(S_t - K, 0)|\overline{\sigma}) \right) \end{aligned}$$

where $\overline{\sigma}$ is the realized volatility for the period. This will simply be $E_t^*(BS(\overline{\sigma}, \frac{K}{S_0}))$ if the following assumptions hold:

- We are modeling volatility, σ, not variance, σ^2,
- the density for the volatility is the risk-neutral density,
- S_t is conditionally log normal (given the volatility), and
- $Cov(S_t, \overline{\sigma}) = 0$.

Since the Black-Scholes formula is not linear, $E_t^*(BS(\overline{\sigma}, \frac{K}{S})) \neq BS(E(\overline{\sigma}, \frac{K}{S}))$. That would give the *plug-in Hull and White* formula, $P_t = BS(E_t^*(\overline{\sigma}, \frac{K}{S}))$. In general, the implied volatility from the Hull and White prices will not agree with the average volatility, but it will be closest for at-the-money options. The Hull and White formula can lead to a relatively flat smile, but never a smirk.

Suppose we can assume that a certain risk-neutral density, f^*, of S_T is known. Then, we can use *simulation-based methods* to price options. To do this:

1. Simulate $\{S_{T,i}\}_{i=1}^N$ from the risk-neutral distribution.
2. Compute $E_t^*(\max(K - S_T, 0)) \approx \frac{1}{N} \sum_{i=1}^N \max(K - S_{T,i}, 0)$.
3. Discount back to find $\hat{P}_t = e^{-r(T-t)} \frac{1}{N} \sum_{i=1}^N \max(K - S_{T,i}, 0)$.

One might want to ensure that the asset itself is priced correctly; that is, that $S_t = e^{-r(T-t)} \frac{1}{N} \sum_{i=1}^N S_{T,i}$ exactly. To do this, one can adjust each $S_{T,i}$ slightly, though this introduces some small dependence into the observations. Note that this method only prices European options and is only valid for options that expire at time T. Because we are modeling stock prices directly, we must be sure that we are using the risk-neutral distribution, not the empirical distribution.

To choose a risk-neutral distribution, $p_0(K,T)$, one might specify a parametric form that depends on θ and then choose θ to minimize $\sum_K \sum_T (p_0(K,T) - \hat{p}_0(K,T))^2$. This allows θ to determined across expiration dates and strike prices. However, this would require many simulations, since a new simulation is needed for each T.

Finding $\hat{p}_0(K,T)$ for successive days can be used to compute how implied volatility changes over time. This uses the approximation:

$$\frac{\partial c}{\partial S} = \frac{c(S,K) - c(S+\epsilon,K)}{\epsilon}$$

However, this method requires additional simulations. Alternatively, if a pricing formula, $c(S_0,K)$, is homogenous of degree one:

$$\begin{aligned} c(S_0,K) &= \frac{1}{t}c(tS_0,tK) \\ c(S,K) &= c_K K + c_S S \\ c_S &= -c_K \frac{K}{S} + \frac{1}{S}c(S,K) \end{aligned}$$

Simulation is also homogenous of degree one if $S_{T,i} = S_0 g_i$ where g_i is the random variable being simulated, since

$$e^{rT}\frac{1}{N}\sum \max(S_0 g_i - K, 0) = \frac{1}{t}e^{rT}\frac{1}{N}\sum \max\left(\frac{S_0}{t}g_i - \frac{K}{t}, 0\right)$$

and we may use the formula as before. Since $c_K, \frac{K}{S}, \frac{1}{S}, c(S,K)$ can all be computed from a single simulation, this is an efficient way of finding the delta.

For market-makers, it is more important to have an *arbitrage free model*, even if the parameter values change over time.

In *least squares Monte Carlo* (Longstaff and Schwarz), one estimates the payoff on an American option from the regression on a constant and a function of last period's price, that is,

$$Payoff_i = c + \epsilon_i + f(s_{T-1})$$

if it is exercised today. f may be estimated using a kernel estimator; it would be expected that f would be close to 0 when options are out of the money. f may depend on lagged options prices as well. With such a structure, we can determine which values of $f(s_{t-1})$ should lead to an early exercise of the option.

Other models for options pricing include:

- *Binomial Trees*: This method assumes that the underlying asset can take on a finite number of values in each period and uses the expectation to price the option.
- *Heston Square Root Model*

- *GARCH Trees*: This uses GARCH to price options (since it gives a distribution for future stock prices?). To do this, we fix the returns a GARCH model to make them risk-neutral:

$$\begin{aligned}\ln\left(\frac{S_{t+1}}{S_t}\right) &= r - \frac{h_{t+1}}{2} + \sqrt{h_{t+1}}z_t \\ h_{t+1} &= \omega + \alpha h_t(z_t - c - \lambda_t)^2 + \beta h_t\end{aligned}$$

- One might need to adjust the risk-neutral distribution to add additional fear of crashes.

- *Component Models* (Cristoferson, Jacobs and Wang): Suppose the risk-neutral density is

$$\begin{aligned}\ln\left(\frac{S_{t+1}}{S_t}\right) &= r + \lambda h_{t+1} + \sqrt{h_{t+1}}z_{t+1} \\ h_{t+1} &= \omega_t + b_1 h_t + a_1(z - c_1\sqrt{h_t})^2\end{aligned}$$

  where w_t is slowly time-varying. Then, the long-run variance changes as $\sigma_t^2 = \frac{\omega_t}{1-b_1-a_1c_1^2}$. This builds in the short-run dynamics of a GARCH model and the long-run dynamics from ω_t. This can help in pricing different maturities.

If options markets are incomplete, then the volatility across the two markets might differ.

Suppose $P_t = BS(\overline{\sigma}, \frac{K}{S})$. By definition, $\frac{\partial P_t}{\partial \sigma} = \Lambda$ (vega, as before), and

$$\frac{\partial P_t}{\partial S_t^2} = \Gamma + \frac{\partial BS}{\partial \sigma}\frac{\partial^2 \sigma}{\partial S_t^2} = \Gamma + \Lambda \cdot VM$$

where $VM = \frac{\partial \sigma^2}{\partial S_t^2}$ is the *variance multiplier*. The variance multiplier is generally smaller for longer maturities (because σ is stationary in the long run), while Γ is larger for long-run maturities. A method called *GARCH GAMMA* corrects for this.

Theorem 4.1 Breedan and Litzenberger. *Suppose we options prices, $c(K)$ for every strike price K. Then, we may find the risk-neutral density, f^*, using $\frac{\partial^2 c}{\partial K^2} = e^{rT}f^*(K)$.*

Proof

$$\begin{aligned}c(K) &= e^{-rT}\int_0^\infty \max(S_T - K, 0)f^*(S_T)dS_T \\ &= e^{-rT}\int_K^\infty (S_T - K)f^*(S_T)dS_T \\ \frac{\partial c}{\partial K} &= e^{-rT}\int_K^\infty (-1)f^*(S_T)dS_T - e^{-rT}\max(K-K,0)f^*(K) \\ \frac{\partial^2 c}{\partial K^2} &= e^{-rT}f^*(K)\end{aligned}$$

Unfortunately, approximations of $\frac{\partial^2 c}{\partial K^2}$ based on a small number of strike prices are noisy. One could interpolate between strike prices to try to smooth out the estimates. Also, using this result, one could get estimates of the pricing kernel M^*_{t+1}, based on the ratio of the empirical density to the risk-neutral density. To smooth this out, M^*_{t+1} could be projected onto variables known at time t, using $M^*(r_{t+1}) = \theta_t r_{t+1}^{-\theta_t}$. With this specification, the pricing kernel will be monotonically decreasing in returns; more flexible specifications allow people to seem to be risk-loving around 0 returns.

4.2 Trading Volatility

To buy volatility (as a hedge against volatility increases), one could use various portfolios:

- *Straddle*: Buy a call and a put at the same strike (usually, at-the-money) and maturity.
- *Strangle*: Buy a call and a put at slightly different strikes.
- Buy an option and a delta hedge (?).

Such methods require constant adjustment as the underlying price changes.

The Volatility Index (VIX) averages out-of-the-money calls and puts to approximate the future realized variance of the S & P 500. This is like a portfolio of strangles. This is based on the exact Taylor series expansion:

$$g(S) = g(\overline{S}) + g'(\overline{S})(S-\overline{S}) + \int_K^\infty g''(K)(K-S)^+ dK + \int_0^K g''(K)(S-K)^+ dK$$

In particular, if $g(S)$ is the payoff and S_t is the present value (known today), then

$$\begin{aligned} E(e^{-r(T-t)}g(S)) &= (g(S_t) - g'(S_t)S_t)e^{-r(T-t)} + g'(S_t)S_t \\ &\quad + \int_K^\infty g''(K)C(K,t,T)dK + \int_0^K g''(K)P(K,t,T)dK \end{aligned}$$

where $C(K,t,T)$ and $P(K,t,T)$ are the holdings of calls and puts at the given strikes and expirations. Using $g(S_T) = \log(\frac{S_T}{S_t})$, the weights on the calls and puts should be $\frac{2(1-\log(K/S_t))}{K^2}$. There must be interpolation since there is not an option at each point. This is how the VIX is created, using no assumptions about the risk-neutral distribution.

In general, the VIX is higher than the GARCH estimates of volatility; this may come from the risk premium being built into the VIX and not GARCH. Furthermore, VIX usually gives a better forecast than GARCH (which in turn is better than just using historical volatility).

This can be extended to trading more complicated moments:

- *Correlations*: One could buy the variance of an index and sell the variance of its components. Suppose $\rho_{ij} = \rho$ for all $i \neq j$. Then:

$$Var\left(\sum w_i r_{it}\right) = \sum w_i^s Var(r_{it}) + 2\rho \sum_{i \neq j} w_i w_j \sqrt{Var(r_i) Var(r_j)}$$

- Skewness (which could be used to hedge downside risk) and kurtosis could also conceivably be traded, but they are not yet.

5 Extreme Value Theory

Extreme value theory is used to measure the frequency with which extreme movements (particularly crashes) occur in the market. This may be more challenging for stocks (than for natural events) because of the human component built into stock trading. In this section, we often talk about the maximum of n events; when we are interested in minima instead, we simply talk about the maximum of their negatives. Throughout this section, we only consider the unconditional distribution of returns, which might not be the best option in reality.

Proposition 5.1 *Let $M_n = \max(X_1, ..., X_n)$. If $X_1, ..., X_n$ are independent and identically distributed with distribution F, then:*

$$\begin{aligned} F_n(z) &= P(M_n < z) \\ &= P(X_1 < z)P(X_2 < z)...P(X_n < z) \\ &= F(z)^n \end{aligned}$$

Theorem 5.2 Fisher-Tippett. *Suppose X_i is independent and identically distributed, and there exist a sequence of constants c_n, d_n, such that $\frac{M_n - d_n}{C_n} \rightarrow_D H$, where H is a non-degenerate distribution. Then, there are three possible cumulative distribution function for H:*

- *Frechet Distribution (also known as the Pareto distribution): For some $\alpha > 0$,*

$$\Phi_\alpha(x) = \begin{cases} \exp(-x^{-\alpha}) & x > 0 \\ 0 & x \leq 0 \end{cases}$$

- *Weibull Distribution: For some $\alpha > 0$,*

$$\Psi_\alpha(x) = \begin{cases} \exp(-(-x)^{\alpha}) & x \leq 0 \\ 1 & x > 0 \end{cases}$$

- *Gumbel Distribution:* $\Lambda(x) = \exp(-e^{-x})$

The three distributions are related. In particular, if $x \sim \Phi_\alpha$, then $(\log x)^\alpha \sim \Lambda$ and $-\frac{1}{x} \sim \Psi_\alpha$ (SAME α?).

Definition Let L be a function. If $\lim_{x\to\infty} \frac{L(tx)}{L(x)} = 1$ for all t, then L is a *slowly-varying function.*

Definition Suppose $\lim_{x\to\infty} \frac{h(tx)}{h(x)} = t^\rho$ for all t. Then, h is a *regularly-varying function.*

The logarithm is a slowly-varying function. Polynomials are regularly varying functions with ρ equal to the highest order of the polynomial. e^x is neither. The product of a slowly-varying function and a regularly varying function is another regularly-varying function.

Definition If F_n converges to a certain one of the distributions, Θ, in the Fisher-Tippett Theorem, then we say that F is in the *maximal domain of attraction* of that distribution, or $F \in MDA(\Theta)$.

Theorem 5.3 *F is in the maximal domain of attraction of Φ if and only if $\overline{F}(x) = 1 - F(x) = L(x)x^{-1/\xi}$, with $\xi > 0$ and $L(x)$ a slowly varying function.*

The Frechet distribution is in its own MDA. If $F \in MDA(\Phi)$, F has an infinite right endpoint. If $F \in MDA(\Psi)$, it has a finite right endpoint.

Definition We call $\alpha = \frac{1}{\xi}$ in the previous theorem the *tail index.* If α is small, the distribution will have fat tails.

The tail index determines which moments of the distribution are finite. For the Frechet distribution, there are α finite moments. The Student's t distribution with ν degrees of freedom has $\alpha = \nu$.

Theorem 5.4 Von Mises Condition. *$F \in MDA(\Phi)$ if and only if $\lim_{x\to\infty} \frac{xf(x)}{F(x)} = \alpha$ (or $\lim_{x\to\infty} \frac{f(x)}{1-F(x)} = c$). Suppose x_F is the maximum value that F can take; $F \in MDA(\Psi)$ if and only if $\lim_{x\uparrow x_F} \frac{(x_F - x)f(x)}{1-F(x)} = \alpha > 0$. In the latter case, α is the tail index.*

The maximal domain of attraction of the Frechet includes: Frechet distribution, Cauchy distribution, Burr distribution, Stable distribution with $\alpha < 2$, and the t distribution. The maximal domain of attraction of the Weibull includes: Uniform, Beta, some Power Law distributions. The maximal domain of attraction of the Gumbel includes: Exponential, Normal, Weibull, Erlang, Log Normal, Gamma.

Definition The *generalized extreme value distribution* is given by

$$H_\xi(x) = \begin{cases} \exp(-(1+\xi x)^{-1/\xi}) & \xi \neq 0, 1+\xi x > 0 \\ \exp(-e^{-x}) & \xi = 0 \end{cases}$$

In the tails, $\xi > 0$ agree with the Frechet distribution, $\xi < 0$ agrees with the Weibull distribution, and $\xi = 0$ agrees with the Gumbel distribution. Note that the change in the neighborhood of $\xi = 0$ is continuous.

5.1 Estimation

To find the probabilities of extreme events, we must estimate the tails of the distribution by estimating α.

One can use *block maxima*, where the maximum is found over fixed consecutive blocks of data, and the Frechet distribution is based on only those maxima. This throws out a lot of data, which is inefficient.

With *threshold exceedances*, we pick a threshold, u, and concentrate on the data that exceed it, $\{(x_i - u)^+\}$. For the Pareto distribution, the distribution of $(x_i - u)^+$ is $f(x) = \frac{\alpha u^\alpha}{x^{\alpha+1}}, x > u$. Corresponding to the generalized extreme value distribution in this context is the *Generalized Pareto distribution*:

$$G_{\xi,\beta}(x) = \begin{cases} 1 - (1 + \frac{\xi x}{\beta})^{-1/\xi} & \xi \neq 0, 1 - \frac{\xi}{\beta} \geq 0 \\ 1 - \exp(-\frac{x}{\beta}) & \xi = 0 \end{cases}$$

where ξ matches the parameter from the generalized extreme value distribution and β depends on the threshold. The distribution of threshold exceedances converges to the GPD as u increases.

Theorem 5.5 Pickens-Balkma-DeHaan. *Let X be a random variable with distribution function $P(X < x) = F(x)$. We define:*

$$\begin{aligned} F_u(x) &= P(X - u < x | X > u) \\ &= \frac{P(u < X < x + u)}{P(u < X)} \\ &= \frac{F(x + u) - F(u)}{1 - F(u)} \end{aligned}$$

Let x_F be X, which may be infinite. $\lim_{x \to x_F} \sup_{0 \leq x \leq x_F - u} |F_u(x) - G_{\xi,\beta}(x)| = 0$ if and only if $F \in MDA(H_\xi)$.

For maximum likelihood estimation, the density and log-likelihood are:

$$\begin{aligned} g(x; \xi, \beta) &= \frac{\partial G}{\partial x} = \frac{1}{\beta} \left(1 + \frac{\xi x}{\beta}\right)^{-1-1/\xi} \\ l &= \sum \log g(y_i) = -n \log \beta - \sum \left(1 + \frac{1}{\xi}\right) \log \left(1 + \frac{\xi y_i}{\beta}\right) \end{aligned}$$

where $y_i = x_i - u$ for those $x_i > u$.

The choice of the threshold, u, can be a challenge. If u is too big, there will be very few observations, which will lead to larger standard errors. If u is too small, the observations chosen might not be far enough in the tail, which will lead to bias. One can use the mean

squared error or just "window carpentry" (eyeballing where bias seems to be a problem) to choose u.

Another option is the *Hill estimator*, which assumes that the distribution of extreme values is $f(x) = cx^{-\alpha}$. This is the Pareto distribution, which is a stronger assumption but still matches the extreme tails. This estimator has a closed form solution.

5.2 Quantile Regression

Suppose $y_t = \beta + u_t$. We may estimate β in a variety of ways. If we choose $\hat{\beta}$ to minimize $\sum u_t^2$, then $\hat{\beta}$ will be the mean of y_t. If we choose it to minimize $\sum |u_t|$ instead, then $\hat{\beta}$ will be the median of y_t. If we minimize $\left(\sum_{u_t<0} \alpha|u_t| + \sum_{u_t>0}(1-\alpha)|u_t|\right)$, then $\hat{\beta}$ will be the α-quantile of y_t. Note that this last method weights positive and negative residuals differently.

More generally, *quantile regression* is based on the model $y_t = f(x_t, \beta) + u_t$ and choosing $\hat{\beta}$ to minimize $\left(\sum_{u_t<0} \alpha|u_t| + \sum_{u_t>0}(1-\alpha)|u_t|\right)$. *Least squares* minimizes $\sum u_t^2$ and *least absolute deviations* sets $\alpha = 0.5$. It is possible to find estimators for more than one quantile, but in finite samples the quantiles might not overlap or might intersect; asymptotically, they should all work out.

Suppose that X_t has a hypothesized cumulative distribution function F. To test if the distribution is correctly specified, we have a few options:

- We may compare the moments of the empirical distribution to the moments of the hypothesized distribution.
- $U_t = F(X_t)$ is a random variable with a uniform distribution. If X_t actually has fatter tails than is hypothesized, then the distribution of U_t will be too high in the middle and too low in the tails (since we divide by the variance first).
- We may also compute $Y_t = F^{-1}(U_t)$ and check if the resulting distribution matches the properties of F.

If there is possible time dependence, we would want to test that the CDF is uniform at each time, not just on average.

In general, any monotonic transformation of a random variable will keep the observations in the same order; their spacing is the only part that is changed.

5.3 Copulas

Definition A *copula* is any joint distribution of uniform random variables. That is, $U = (U_1, ., , , U_k) \sim C(u_1, ..., u_k)$.

To check if C is a copula, it is sufficient to check that:

- $C(u) = 0$ if any argument is 0.

- $C(u)$ is monotonically increasing in each argument.
- $C(u) = 1$ if and only if $u = 1$.

The copula for independence is $C(u_1, ..., u_k) = u_1...u_k$.

Theorem 5.6 Sklar's Theorem. *If F is a k-dimensional cumulative distribution function with univariate continuous marginals, $F_1, ..., F_k$, then there exists a unique k-dimensional copula, C, such that, for all x,*

$$F(x_1, ..., x_k) = C(F_1(x_1), ..., F_k(x_k))$$

Conversely, if C is a copula and $F_1, ..., F_k$ are univariate cumulative distribution functions, then F as defined above is a joint cumulative distribution function with the given marginals.

We may also write: $C(u) = F(F_1^{-1}(u_1), ..., F_k^{-1}(u_k))$

Definition If C is a copula, then the *copula density* is $c(u) = \frac{\partial^k C(u)}{\partial u_1 ... \partial u_k}$.

The joint density is the product of the marginal densities and the copula: $f(y) = c(u)f_1(y_1)...f_k(y_k)$.

The correlation of variables cannot be associated with a particular copula, because it is sensitive to non-linear monotonic transformations of the variables (that is, different marginals). However, Kendall's τ (where $\tau = P((Y_1 - Y_1')(Y_2 - Y_2') > 0) - P((Y_1 - Y_1')(Y_2 - Y_2') < 0)$) and Spearman's rank correlation (which is equivalent to finding the correlation of U_1, U_2) are invariant to the marginals.

Definition *Upper tail dependence* is defined by:

$$\begin{aligned} \lambda_U &= \lim_{u \to 1} P(U_1 > u | U_2 > u) = \lim_{u \to 1} P(U_2 > u | U_1 > u) \\ &= \lim_{u \to 1} \frac{1 - 2u + C(u, u)}{1 - u} \end{aligned}$$

Lower tail dependence is defined by:

$$\lambda_L = \lim_{u \to 0} \frac{C(u, u)}{u}$$

Since tail dependence is a probability, it lies between 0 and 1. For joint normal distributions, $\lambda_U = \lambda_L = 0$, even if the correlation is high. Other copulas may have one or both be non-zero; in financial applications, we would like to have more lower tail dependence (for example, in measuring correlations between defaults).

Tail dependence may help describe *contagions*, where extreme declines occur; in these situations, correlations and volatility both increase. The multivariate normal distribution

therefore underprices extreme events, such as defaults, and the hedges against them (like *CDO tranches*, which protect against a combination of defaults).

Consider three measure of volatility:

$$\begin{aligned} |r_t|^2 &= \Psi_t \epsilon_t \\ (\log(range))^2 &= \phi_t \xi_t \\ (RealizedVolatility)^2 &= \theta_t \nu_t \\ \epsilon_t &\sim (1, \sigma_1^2) \\ \xi_t &\sim (1, \sigma_2^2) \\ \nu_t &\sim (1, \sigma_3^2) \end{aligned}$$

In some applications, we assume that ϵ_t, ξ_t, ν_t all have Gamma distributions. We may also be interested in the joint distribution of the three terms; we want a copula with Gamma marginals. One possibility is a normal copula.

Copulas can also be used in modeling the implied volatility of various stocks. To do this we must find the mean and errors of each, the correlations, and then the joint distribution of the errors.

Suppose $r_{it} = \beta_i r_t^m + \epsilon_{it}$. If (r_t^m, ϵ_{it}) is jointly normal, then each r_{it} will also be normal and there will be no tail dependence in returns across assets. However, if r_t^m has a asymmetric distribution, even though ϵ_{it} is still normal, the joint distribution will have lower tail dependence. In particular, if $\beta = 1$ and $Var(\epsilon_{it}) = 1$, then

$$P(r_i < d_i) = E\left(P(r_i < d_i | r^m)\right) = E(\Phi(d_i - r^m))$$

and skewness satisfies $s = s^m \rho^3$; that is, individual return skewness is attenuated relative to the skewness of the market return. If default occurs when the return is below d, then the joint probability of default is:

$$P(r_i < d \cap r_j < d) = E(\Phi(d - r^m)^2) = Var(\Phi) + P(r_i < d)^2$$

The heavier the tails of r^m, the greater the tail dependence in returns.

5.4 Value at Risk

Definition The *value at risk* of a portfolio, given a confidence level, α, and a horizon, T, is a value, $VaR(t, T, \alpha)$, such that the return over horizon T will exceed it α of the time (with α confidence). That is, $P_t(r_{t,T} < -VaR(t, T, \alpha)) = \alpha$.

The value at risk must be measurable in the current period. In general, $VaR(t) = f(VaR(t-1), r(t-1), parameters)$. In the *Adaptive Model*,

$$VaR(t) = VaR(t-1) + \beta(I(r(t-1) < -VaR(t-1)) - \theta)$$

There are many other models as well.

In the *CAViaR Strategy*, we have the following steps:

- Define a quantile model with unknown parameters.
- Construct the quantile criterion function.
- Optimize the criterion over the history.
- Use diagnostic checks to ensure that the model is adequate.

One quantile criterion function (Koenker and Bassett, 1978) is:

$$Q(\beta) = \sum I(r(t) < VaR(t))(r(t) - VaR(t))$$

which treats positive and negative errors differently (it is harsher on errors when the return is below the VaR.

6 Multivariate Models

6.1 Factor Models and Principal Components

Definition A *K-factor model* or *multiple indicator model* for y_t is defined by $y_t = \sum_{k=1}^{K} \beta_k f_{kt} + \epsilon_t$, where y_t is an observed vector, f_{kt} are factors, and β_k are vectors of coefficients. The factors can be either *unobservable (latent)* or *observable*.

We use $E(y_t), Var(y_t)$ to understand the behavior of the unobservable factors. Specifying a certain factor model is equivalent to a restriction of the moment condition.

Suppose we have an unobservable factor model, $y_t = BF_t + \epsilon_t$, where F_t is a vector of k factors and B is an $N \times K$ matrix of coefficients. Since we do not observe the factors, we require some additional identifying (orthogonality) assumptions:

- $E(F_t) = 0$
- $E(F_t F_t') = I$
- $E(F_t \epsilon_t') = 0$
- $E(\epsilon_t \epsilon_t') = D$, where D is a diagonal matrix

Note that $Var(Y_t) = B(E(F_t F_t'))B' + E(\epsilon_t \epsilon_t') + 0 = BB' + D$. Then, if we assumed that $E(F_t F_t') = PP'$ more generally, the coefficient, B would simply become $B^* = BP$. The assumptions above are still not sufficient to identify the factors, since for any rotation matrix C, $F_t^* = CF_t$ would lead to the same variance and would still satisfy $E(F_t^* F_t^*) = CIC' = I$. To fix this, we may add zero restrictions on the B matrix. However, the choice of which rotation is used matters only in the "naming" of factors to correspond to economic theory; predictions would be the same.

In this model, any dependence in the y_t is captured by the factors, since ϵ_t is diagonal. This gives a parsimonious way to describe correlations.

In general, the data gives $\frac{N(N+1)}{2}$ moments (from the lower triangle of the covariance matrix), while there are $NK + N$ parameters (NK elements of B and N non-zero elements of D). Thus, adding more assets identifies or over-identifies the model (and over-identification can lead to tests of adequacy), while adding more factors may require more assets or more restrictions to identify the model.

One could estimate this model by maximum likelihood, under the assumption that ϵ_t is multivariate normal:

$$L = -\frac{T}{2}\log|NN' + D| - \frac{1}{2}\sum y_t'(BB' + D)^{-1}y_t$$

However, this is generally hard.

Instead, we use a similar (but not identical) model for estimation. In *principal components analysis*, we assume that $Var(y_t) = \Omega$, and we want to choose a vector a to maximize $a'\Omega a$ such that $a'a = 1$. This is the unit vector that maximizes $Var(a'y_t)$. Using the Lagrangian, we find that:

$$\begin{aligned} L &= a'\Omega a - \lambda(a'a - 1) \\ \Omega a - \lambda a &= 0 \end{aligned}$$

which implies that a is the eigenvalue of Ω corresponding to the largest eigenvector, λ. If $a_1, ..., a_N$ are the eigenvectors corresponding to eigenvalues $\lambda_1 > ... > \lambda_N$, then $Var(y_t) = \sum_{i=1}^{N} \lambda_i a_i a_i'$, and a_i is the linear combination of y_t which has the largest variance, subject to the restriction that a_i is orthogonal to $a_1, ..., a_{i-1}$. If $A = [a_1, ..., a_N]$, then $Z = A'y$ is the vector of *principal components*. Note that

$$E(z_t z_t') = E(A'y_t y_t' A) = A'\Omega A = \Lambda$$

where Λ is the diagonal matrix consisting of the eigenvalues.

We might be able to approximate the series using a smaller number of principal components. In particular, we note that $trace(Var(y_t)) = trace(A\Lambda A') = \sum_{i=1}^{N} \lambda_i$, so that $\frac{\lambda_i}{\sum \lambda_j}$ is the importance of each component.

We can also compute the principal components using the eigenvectors of the correlation matrix. In this case, the trace is always N. The results will be similar, particularly if the series have similar variances, but they need not be identical.

By looking at the eigenvectors, we may "name" the components. For example, if all the returns have roughly equal weights (or weights approximately equal to the market capitalization), then the component is like the market return. Other factors might be cyclical factors or industry-based factors. Graphing the factors over time may also be helpful in naming.

If one or more eigenvalues were 0, then one could form a portfolio with 0 variance (if this had a positive return, this would be *statistical arbitrage*). In this case, the covariance matrix must be singular.

To relate this to a K-factor model, we may write

$$\Omega = \sum_{i=1}^{K} \lambda_i a_i a_i' + \sum_{i=K+1}^{N} \lambda_i a_i a_i'$$

If $\sum_{i=K+1}^{N} \lambda_i a_i a_i'$ is approximately diagonal, then we may write $\sum_{i=1}^{K} \lambda_i a_i a_i' = BB'$ and this is the factor structure. If we assume that $y_t = BF_t + \epsilon_t$ with $D = \sigma^2 I$ (equal variances in the errors), then the eigenvectors of $Var(y_t) = BB' + \sigma^2 I$ must be the eigenvectors of BB'; the corresponding eigenvalues will be σ^2 greater. This means that the first K factors will have eigenvalues greater than σ^2, while the remaining $N - K$ factors will all have eigenvalues of σ^2, corresponding to the $N - K$ zero eigenvalues of BB'.

In an *observable factor model*, we have $y_t = Bx_t + \epsilon_t$, where x_t is observable and we no longer restrict $Var(\epsilon_t) = \Omega$. This may also be the reduced form of simultaneous equation system, a vector autoregression (if x_t contains lagged y_t), a model of expected returns, or CAPM or APT where the x_t are the risk factors. This should be estimated as a *seemingly unrelated regressions (SUR) model*. If $\epsilon_t \sim Normal(0, \Omega)$ and use maximum likelihood:

$$\begin{aligned} l &= -\frac{T}{2}\log|\Omega| - \frac{1}{2}\sum_t (y_t - B'x_t)'\Omega^{-1}(y_t - Bx_t') \\ &= -\frac{T}{2}\log|\Omega| - \frac{1}{2} trace\left(\sum_t \Omega^{-1}(y_t - B'x_t)(y_t - B'x_t)'\right) \\ \frac{\partial l}{\partial \beta} &= \frac{2}{2} trace\left(\sum_t \Omega^{-1}(y_t - \beta' x_t)x_t'\right) \end{aligned}$$

or if we use the method of moments, we also find the usual equation-by-equation OLS estimator for the coefficients:

$$\hat{\beta} = \sum y_t x_t' (\sum x_t x_t')^{-1}$$

This works as long as the same x_t are in each equation.

As an intermediate case between observable and unobservable models, we might have the model:

$$\begin{aligned} y_t &= \sum_{k=1}^{K} \beta_k f_{kt} + \epsilon_t \\ f_{kt} &= \gamma_k' x_{tk} + \eta_{kt} \end{aligned}$$

We may then substitute the equation for the factors back into the equation for y_t to show that this is equivalent to a regression with a more complicated error structure:

$$\begin{aligned} y_t &= \sum_{k=1}^{K} \beta_k(\gamma_k' x_{tk} + \eta_{kt}) + \epsilon_t \\ &= \sum_{k=1}^{K} \beta_k \gamma_k' x_{tk} + \left(\sum_{k=1}^{K} \beta_k \eta_{kt} + \epsilon_t \right) \end{aligned}$$

This is called a *multiple indicators, multiple causes (MIMC) model*.

If the model specifies that there are fewer factors than variables in x_t (as might be implied by APT or MIMC), then B no longer should have full rank. We test this using the *canonical correlations* test. This test chooses vectors a, b to maximize $Corr(a'Y, b'X)$ subject to the restriction that $Var(a'Y) = Var(b'X) = 1$. If some of the correlations are equal to 0, then B has reduced rank. The subsequent correlations are the *canonical correlations*.

6.2 Multivariate GARCH Models

In a multivariate GARCH model, we want to allow both time-varying correlations and time-varying volatilities.

Definition A matrix, M, is *positive definite* if $x'Mx > 0$ for all $x \neq 0$. M is *positive semi-definite* if $x'Mx \geq 0$ for all x.

Proposition 6.1 *The sum of a positive definite matrix and a positive semi-definite matrix is positive definite.*

Definition The *Hadamard product* of two matrices, $A \circ B$, is the element-by-element product, that is, $[A \circ B]_{ij} = [a_{ij} b_{ij}]$.

Proposition 6.2 *$A \circ (rr')$ is positive (semi-)definite if A is positive (semi-)definite and $r \neq 0$.*

Proof

$$\begin{aligned} A \circ (rr') &= \begin{bmatrix} A_{11} r_1^2 & \ldots & A_{n1} r_1 r_n \\ \ldots & \ldots & \ldots \\ A_{1n} r_1 r_n & \ldots & A_{nn} r_n^2 \end{bmatrix} \\ &= diag(r) A diag(r) \end{aligned}$$

where $diag(r)$ is the diagonal matrix with r along the diagonal. If A is positive definite and $r \neq 0$, then $diag(r)Adiag(r)$ is positive definite. ∎

Proposition 6.3 *$A \circ B$ is positive definite if both matrices are positive definite. $A \circ B$ is positive semi-definite if both matrices are positive semi-definite.*

Proof Suppose $B = \sum_{i=1}^{n} \lambda_i b_i b_i'$ is the eigenvector decomposition of B. Then,

$$\begin{aligned} A \circ B &= A \circ (\sum_{i=1}^{n} \lambda_i b_i b_i') \\ &= \sum_{i=1}^{n} \lambda_i A \circ (b_i c_i') \end{aligned}$$

AND? ∎

Definition The operator, $vec(A)$, converts an $n \times k$ matrix to a vector of length nk by stacking the columns.

Proposition 6.4 $vec(ABC) = (C^T \otimes A)vec(B)$.

6.2.1 Models for Covariance Matrices

Definition The *conditional covariance matrix* is given by $H_t = E_{t-1}(r_t r_t')$.

For a portfolio with weights w, this leads to a conditional variance of returns of $Var_{t-1}(w'r_t) = w'H_t w$. For two portfolios with weights w_1, w_2, the correlation of their returns is $Corr(w_1'r_t, w_2'r_t) = \frac{w_1'H_t w_2}{\sqrt{w_1'H_t w_1 w_2'H_t w_2}}$.

Definition Two models for the conditional covariances are the *moving average* model with N lags (where N must be larger than the number of assets to ensure that H_t is positive definite), with $H_t = \frac{1}{N}\sum_{k=1}^{N} r_{t-k}r_{t-k}'$, and the *exponential smoothing* model with parameter λ, where $H_t = \lambda r_{t-1}r_{t-1}' + (1-\lambda)H_{t-1}$.

Definition The *diagonal multivariate GARCH model* is given by:

$$\begin{aligned} \sigma_{ijt} &= \omega_{ij} + \beta_{ij}\sigma_{ij,t-1} + \alpha_{ij}r_{i,t-1}r_{j,t-1} \\ H_t &= \Omega + A \circ (r_{t-1}r_{t-1}') + B \circ H_{t-1} \end{aligned}$$

This yields $\frac{n(n-1)}{2}$ equations. Note that σ_{ijt} depends only on lags of $r_{i,t-1}r_{j,t-1}$, not of any other returns or either asset's own covariance. (This is why the model is called diagonal.)

For the predicted covariance matrix to be positive definite, we require that Ω is positive definite and that A, B are positive semi-definite. This requires restrictions in estimation (or perhaps a parameterization of the matrices that will ensure that they are positive semi-definite).

Definition The *BEKK model* is given by

$$H_t = \Omega + A' r_{t-1} r_{t-1}' A + B' H_{t-1} B$$

If A, B are diagonal in the BEKK model, then the model reduces to a diagonal multivariate GARCH model. If they are not diagonal, this allows one covariance to depend on the returns of other assets.

Definition The *VEC model* is given by:

$$vec(H_t) = vec(\Omega) + Avec(r_{t-1} r_{t-1}') + Bvec(H_{t-1})$$

This model allows all squares and cross-products to affect each other. We may write the BEKK model as a special case of the VEC model:

$$vec(H_t) = vec(\Omega) + (A' \otimes A) vec(r_{t-1} r_{t-1}') + (B' \otimes B) vec(H_{t-1})$$

This shows that the BEKK model is a special case of the VEC model. To reduce the VEC model to the diagonal model, A, B must be diagonal. The VEC model need not produce positive definite covariance matrices. Also, it requires the estimation of $2n^4 + 1$ parameters (since A, B are $n^2 \times n^2$).

For forecasting with the VEC model,

$$\begin{aligned}
vec(H_t) &= vec(\Omega) + Avec(r_{t-1} r_{t-1}') + Bvec(H_{t-1}) \\
E_{t-1}(vec(r_t r_t')) &= vec(H_t) \\
E_{t-1}(vec(r_{t+k} r_{t+k}')) &= E_{t-1}(vec(H_{t+k})) \\
&= vec(\Omega) + AE_{t-1}(vec(r_{t+k-1} r_{t+k-1}')) + BE_{t-1}(vec(H_{t+k-1})) \\
&= vec(\Omega) + AE_{t-1}(vec(H_{t+k-1})) + BE_{t-1}(vec(H_{t+k-1})) \\
&= vec(\Omega) + (A+B)E_{t-1}(vec(H_{t+k-1}))
\end{aligned}$$

where we use the law of iterated expectations and the linearity of the *vec* operator. If we iterate on k, we find that the unconditional variance is $E(vec(H_t)) = (I - A - B)^{-1} vec(\Omega)$, assuming that the model is stationary. This also shows that the model is linear is the squares and cross-products, which makes forecasting easier.

Definition In the *K-factor ARCH model*, we have:

$$\begin{aligned}
H_t &= \Omega + BF_t B' \\
\sigma_{ijt} &= \omega_{ij} + \sum_{f=1}^{K} \sum_{f'=1}^{K} \beta_{if} \beta_{jf'} \sigma_{f,f',t}^{FACTOR}
\end{aligned}$$

where B is an $N \times K$ matrix and F is a $K \times K$ matrix with the K factors. Then, H_t varies with only the factors, and the factors change the relative importance of B and Ω in affecting H_t.

In this model, it is a restriction to assume that F is diagonal.

The arbitrage pricing theory model leads to a similar model. In this model:

$$r_{jt} - r^0 = \sum_{k=1}^{K} \beta_{jk}(f_{kt} - r^0) + \epsilon_{jt}$$

which leads to an expected return of $E_{t-1}(r_{jt} - r^0) = \sum_{k=1}^{K} \beta_{jk}\mu_{kt} = B\mu_t$. If we assume that the ϵ_t has constant covariance, Ω, then $H_t = \Omega + BF_tB'$. However, constant variance portfolios are in the null space of B, which is hard to find and makes the assumption of a constant Ω unlikely. Even if the returns don't have constant variance, they might have an ARCH model, which could be extended to a model for the variance. This yields a relationship between returns and volatility.

For estimation, we assume that the errors have a normal distribution, which gives the model and likelihood:

$$\begin{aligned} r_t &= \mu + \epsilon_t \\ \epsilon_t &\sim Normal(0, H_t) \\ L &= -\frac{1}{2}\sum_{t=1}^{T}(\log|H_t| + \epsilon_t' H_t^{-1}\epsilon_t) \end{aligned}$$

For diagnostic checking, we use the standardized residuals, $\epsilon_t = H_t^{-1/2}(r_t - E_{t-1}(r_t))$. If the model is correct, the following should hold:

- $Cov(\epsilon_t) = I$.
- ϵ_t^2 has no autocorrelation.
- ϵ_{it}^2 and ϵ_{jt}^2 have no cross-asset autocorrelation.
- $\epsilon_{it}\epsilon_{jt}$ has no autocorrelation.
- No other asymmetries in the residuals.

With this many tests, most models are rejected.

Estimating the intercept can be hard since Ω is often close to 0 but must be estimated to be positive definite. This leads to the GMM moment condition:

$$(I - A - B)vec\left(\frac{1}{T}\sum_{t=1}^{T} y_t y_t'\right) - vec(\Omega) = 0$$

If A, B are known, then Ω is exactly identified. Let $\hat{\Sigma} = \frac{1}{T}\sum_{t=1}^{T} y_t y_t'$; this is an estimator of $(I - A - B)vec(\Omega)$. Then, we may estimate A, B using:

$$vec(H_t) = \hat{\Sigma} + A(vec(r_{t-1}r_{t-1}') - vec(\hat{\Sigma})) + B(vec(H_{t-1}) - vec(\hat{\Sigma}))$$

Since this is a two-step estimator, it is no longer a maximum likelihood estimator, even if A, B are estimated by MLE. If y_t is very persistent, then $\hat{\Sigma}$ can be quite variable if the sample period is too short.

6.2.2 Conditional Correlations Models

Definition The *conditional correlation* is given by:

$$\rho_{12,t} = \frac{E_{t-1}(r_{1t}r_{2t})}{\sqrt{E_{t-1}(r_{1t}^2)E_{t-1}(r_{2t}^2)}}$$

If $x_t = \sqrt{h_{xt}}\epsilon_{xt}$ and $y_t = \sqrt{h_{yt}}\epsilon_{yt}$, where $Var(\epsilon_{xt}) = Var(\epsilon_{yt}) = 1$, then $\rho_{xyt} = E_{t-1}(\epsilon_{xt}\epsilon_{yt})$.

This shows that we may model the conditional correlations based on the standardized residuals. That is, $r_t \ F(0, H_t)$ and $H_t = D_t^{1/2}R_tD_t^{1/2}$. Then, the method is:

1. Collect the standardized residuals for each asset, by estimating volatility models for each separately. (The volatility models can be different for each asset.)

2. Estimate the correlations based on the standardized residuals using one of the models below.

Because any estimate of the correlation is a ratio of random variables, it need not be unbiased. Since $-1 \leq \rho_{12,t} \leq 1$, estimates are likely to be non-linear, particularly near the endpoints.

We focus on contemporaneous correlations, because those tend to be the only significant ones (otherwise, there would be a way to make a profit on one stock using information from the previous return of the other stock).

Definition The *constant conditional correlation model* assumes that $\sigma_{ijt} = \rho_{ij}\sigma_{it}\sigma_{jt}$, so that the correlation between two returns is constant over time. Then, $H_t = D_t^{1/2}RD_t^{1/2}$, where D_t is the matrix of conditional variances.

This model is easy to estimate, but the assumption that the correlation is constant is too strong. Just because r_t has constant conditional correlations does not imply that linear combinations of the returns have constant conditional correlations.

Definition The *exponential smoother model* for conditional correlations is given by:

$$\rho_t = \frac{\sum_{s=1}^{\infty}\lambda^s\epsilon_{1,t-s}\epsilon_{2,t-s}}{\sqrt{\sum_{s=1}^{\infty}\lambda^s\epsilon_{1,t-s}^2\sum_{s=1}^{\infty}\lambda^s\epsilon_{2,t-s}^2}}$$

Equivalently, this may be written as $\rho = \frac{q_{12t}}{\sqrt{q_{11t}q_{22t}}}$ where $q_{ijt} = (1-\lambda)\epsilon_{i,t-1}\epsilon_{j,t-1} + \lambda q_{ij,t-1}$. In matrix terms, this can be written as $Q_t = (1-\lambda)\epsilon_{t-1}\epsilon_{t-1}' + \lambda Q_{t-1}$ with $R_t = (Q_t^*)^{-1/2}Q_t(Q_t^*)^{-1/2}$, where Q_t^* has the diagonal of Q_t and zeroes elsewhere.

Definition The *mean reverting DCC model* is given by $\rho_{ijt} = \frac{q_{12t}}{\sqrt{q_{11t}q_{22t}}}$ where

$$q_{ijt} = \overline{\rho}_{ij}(1-\alpha-\beta) + \alpha\epsilon_{i,t-1}\epsilon_{j,t-1} + \beta q_{ij,t-1}$$

In matrix form, this can be written as $Q_t = \overline{R}(1-\alpha-\beta) + \alpha\epsilon_{t-1}\epsilon'_{t-1} + \beta Q_{t-1}$, with $R_t = (Q_t^*)^{-1/2}Q_t(Q_t^*)^{-1/2}$.

The form above uses the variance targeting assumption for R. Because of the non-linear transformation, the assumption is not exactly correct (it should be based on Q), but it should be close.

Note that the Q_t in both models are only covariance matrices (for the ϵ_t). The diagonal elements need not be one and the other elements need not be bounded between -1 and 1; this is why we need the final, non-linear step.

Other models include:

- *Tse and Tsui*: In this model, R is modeled directly as $R_t = \overline{R}(1-\theta_1-\theta_2) + \theta_1\hat{r}^k_{t-1} + \theta_2 R_{t-1}$, where

  $$r^k_{t-1} = diag\left(\frac{1}{k}\sum_{s=t-k}^{t-1}\epsilon_s\epsilon'_s\right)^{-1/2}\left(\frac{1}{k}\sum_{s=t-k}^{t-1}\epsilon_s\epsilon'_s\right)diag\left(\frac{1}{k}\sum_{s=t-k}^{t-1}\epsilon_s\epsilon'_s\right)^{-1/2}$$

  is the sample correlation matrix based on the last k observations. Since all three terms are correlation matrices, the sum will always be a correlation matrix. However, because r^k_{t-1} depends on the last k observations, this measure is slower to react to changes.

- *DCC(p,q)*: One can add more lags to the DCC model, so that $Q_t = \overline{R} + \sum_{i=1}^p \alpha_i(\epsilon_{t-i}\epsilon'_{t-i} - \overline{R}) + \sum_{i=1}^q \beta_j(Q_{t-j} - \overline{R})$.

- *Generalized DCC*: We may add parameters for each asset. Assume that $A = \alpha\alpha'$ and $B = \beta\beta'$. Let $Q_t = \overline{R} + A \circ (\epsilon_{t-1}\epsilon'_{t-1} - \overline{R}) + B \circ (Q_{t-1} - \overline{R})$. With the restrictions on A, B above, Q_t will be positive definite.

- *Asymmetric DCC*: We may also include a term of the form $\gamma\eta_{it}\eta_{jt}$, with $\eta_t = \min(\epsilon_t, 0)$, in the usual DCC equation. If $\gamma > 0$, then the correlation will be increased when both assets decline. This picks up negative tail dependence and contagions. This model may be generalized with additional lags or more general parameters as well.

For estimation, the likelihood is:

$$\begin{aligned} L &= -\frac{1}{2}\sum_t (\log(2\pi) + \log|D_t R_t D_t| + r_t' D_t^{-1} R_t^{-1} D_t^1 r_t) \\ &= -\frac{1}{2}\sum_t (\log(2\pi) + 2\log|D_t| + \log|R_t| + \epsilon_t' R_t^{-1}\epsilon_t) \\ &= -\frac{1}{2}\sum_t (\log(2\pi) + 2\log|D_t| + r_t' D_t^{-2} r_t - r_t' D_t^{-2} r_t + \log|R_t| + \epsilon_t' R_t^{-1}\epsilon_t) \end{aligned}$$

The first few terms correspond to running univariate GARCH estimation while the last few correspond to running a DCC, conditional on the GARCH estimation. The term $-r_t' D_t^{-2} r_t$ is not included in either, but goes to a constant. This leads to the a Two-Step QMLE with the method outlined at the beginning of the section. To compute standard errors, note that we may write the overall quasi-likelihood as:

$$QL(\phi, \theta) = QL_1(\phi) + QL_2(\phi, \theta)$$

where the first term is the GARCH likelihood and the second term is the correlation part of the likelihood. Any multivariate GARCH that is correct for the first two moments and satisfies the usual regularity conditions is a QMLE; this does not apply to the two-step estimator. If we take the derivatives of the likelihood, this leads to the equivalent 2-step GMM estimator with k_1 moment conditions in $g_1(\phi)$ and k_2 moment conditions in $g_2(\phi, \theta)$. The solution to these exactly identified moment conditions is equivalent to the QMLE. Checking for consistency, we find that:

$$\begin{aligned} g_1 = \frac{\partial}{\partial \phi} QL_1 &= -\frac{1}{2T}\nabla_\phi \sum_t (\log(2\pi) + 2\log|D_t| + r_t' D_t^2 r_t) \\ &= -\frac{1}{2T}\nabla_\phi \sum_k \sum_t \log(h_{kt}) + \frac{r_{kt}^2}{h_{kt}} \end{aligned}$$

where D_t is the diagonal matrix with $(h_{1t}, ..., h_{Kt})$ along the diagonal. This shows that the GARCH estimation part is consistent. Then, the standard errors are:

$$\begin{aligned} G &= \begin{bmatrix} \nabla_\phi g_1 & 0 \\ \nabla_\phi g_2 & \nabla_\theta g_2 \end{bmatrix} \\ \sqrt{T}\begin{bmatrix} \hat{\phi} - \phi_0 \\ \hat{\theta} - \theta_0 \end{bmatrix} &\to_D Normal(0, G^{-1}\Omega G^{-1}) \end{aligned}$$

We may use variance targeting to simplify our GARCH estimation. In this case, we note that $h_t = \sigma^2(1 - \alpha - \beta) + \alpha y_{t-1}^2 + \beta h_{t-1}$, and we replace σ^2 by $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^T y_t^2$. This

ensures that $\alpha + \beta < 1$ and $\sigma^2 > 0$. This is not an MLE, so it is not efficient. In this case, $g = \frac{1}{T}$???. Then, the Two-Step DCC is based on:

$$\begin{aligned} g &= \frac{1}{T}\sum_{t=1}^{T}\begin{bmatrix} \nabla_\phi(\log(h_{it}) + \frac{y_{it}^2}{h_{it}}) \\ \nabla_\theta(\log|R_t| + \epsilon_t' R_t^{-1}\epsilon_t) \end{bmatrix} \\ h_{it} &= \phi_{i0} + \phi_{i1} y_{i,t-1}^2 + \phi_{i2} h_{i,t-1} \end{aligned}$$

where $i = 1, ..., K$. We may also use Three-Step DCC in which we also estimate the unconditional correlations, so that we add the additional moment condition:

$$\frac{1}{T}\sum_{t=1}^{T}\epsilon_{it}\epsilon_{jt} - \overline{\rho}_{ij} = 0$$

Definition The *orthogonal GARCH model* assumes that each principal component of the returns has a GARCH model and fits the model accordingly.

To fit this model:

1. Use the returns to compute the principal components.
2. Fit a GARCH model to each principal components (each model may be different).
3. For forecasting, compute the predictions for the principal components individually and then combine them back to the original returns.

Empirically, this does not work as well as DCC.

Empirically, the correlations from DCC are estimated to be smooth (β large) and close to integrated ($\alpha + \beta$ close to 1). Other methods may lead to noisier estimates for correlations, especially when the same parameters are used to fit both correlations and volatility (and end up being influenced mostly by volatility).

If the normal assumption on the residuals is not palatable, then one one could rank the GARCH residuals and build a DCC model on the ranks instead; this model is more robust. Alternatively, one could convert the ranked residuals back to a Gaussian process (that is, keeping their order but giving them a Gaussian shape) and fit a DCC model on the new residuals. Since correlations are better for normal residuals and influenced by outliers, this might improve results.

Because the correlation matrix differs each period, there would be a different principal components model each period as well.

One could also assume that certain pairs of correlations are all equal, or use a factor structure in the DCC.

These methods can also be used on cross-country returns.

To test the usefulness of these models, we test whether we could make money or reduce variance using it. Given $\{H_t\}$, the *portfolio problem* is to choose weights, $\{w_t\}$, to minimize

the variance, $w_t' H_t w_t$ subject to the constraint that $w_t' \mu_t > r_0$, where μ_t is the vector of expected returns. If μ is fixed, the weights are:

$$w_y = (\mu' H_t^{-1} \mu)^{-1} H_t^{-1} \mu r_0$$

If H is not the true covariance matrix, Ω, then we may compare the variance from the estimated covariance matrix and the true covariance matrix:

$$\begin{aligned} \sigma^H &= \frac{r_0^H \sqrt{\mu' H^{-1} \Omega H^{-1} \mu}}{\mu' H^{-1} \mu} \\ \sigma^\Omega &= \frac{r_0^\Omega}{\sqrt{\mu' \Omega^{-1} \mu}} \end{aligned}$$

To compare two models or weighting schemes, the Diebold-Mariano test suggests that we estimate:

$$(w_{it}' r_t)^2 - (w_{2t}' r_t)^2 = \xi + u_t$$

and test whether $\xi = 0$. We could also use a weighted test:

$$\frac{(w_{it}' r_t)^2 - (w_{2t}' r_t)^2}{\sqrt{(w_{1t}' H_{1t} w_{1t})(w_{2t}' H_{2t} w_{2t})}} = \xi + u_t$$

BEKK and the many forms of DCC tends to lead to models with similar portfolio variances, and one cannot be rejected in favor of another.

7 Market Microstructure Econometrics

Market microstructure econometrics looks at the minute details of financial markets in order to figure out how information is incorporated into prices (not just where they eventually end up).

In general, we assume that there are buyers and sellers with various quantities of a single good. We require a market structure to match the two groups. One agent posts a price and waits until another agent is willing to take the other side of the transaction at that price. Sometimes, the agents do this directly (retail stores or real estate sellers on the selling side, employers on the buying side); in other cases, there are intermediaries (wholesalers) who buy from one and sell to another, and are compensated for their waiting time and risk by the spread in their buying and selling prices.

In a financial market, the intermediary is called a *market maker*. Market makers hold inventory and charge a spread between buying and selling prices. Their risks include bankruptcy, price changes (when they buy the stock at a higher price than they can sell it again), and trading with informed traders (asymmetric information that can lead to arbitrage for the trader). Some exchanges (like NASDAQ) have competing market makers

for each asset; all market makers also have to compete with market makers on other regional and global exchanges selling the same asset. Their spread is also limited by *limit orders* (orders that are posted at prespecified prices and which will be filled before the market maker can buy or sell at the same price) and by regulation in some exchanges (like the NYSE). Much of microstructure econometrics tries to model how market makers behave.

Reasons prices might move:

- Inventory Models: Suppose a buy order comes in. That depletes inventory, so that the market maker wants to encourage sellers. This increases the bid price (and possibly the ask, so that the spread stays the same; this model doesn't necessarily say). After the inventory has been replenished, the price might return to the original level. Some particular models of this include:
  - Garman (1976): If buy and sell orders follow a Poisson process, the market maker must change a spread so that neither his number of shares nor his amount of cash reaches 0.
  - Amihud and Mendelsohn (1980): Prices should be a function of a market maker's inventory.
  - Stoll (1978): If the market maker is risk averse, then he must be compensated for possible deviations from his optimal inventory by the spread.
- Asymmetric Information Models: If some of the traders have more information (about the true stock value, fundamentals, future trading plans, or who is informed) than the market maker, then the market maker must infer future prices from their trading strategies. Over time, the market maker will learn the information, and prices will adjust to the efficient price (this may be a slow process if not all the traders are informed traders). During the adjustment, information traders can only make limited profits, since they may not be able to trade immediately, and because their trades have a price impact.
  - Sequential Trading Model: Suppose that people must trade one at a time, and not necessarily at the instant they want to. The market maker forms expectations:

$$\begin{aligned} E(value|history, buy) &= P_{ask} \\ E(value|history, sell) &= P_{bid} \end{aligned}$$

    That is, the current bid and ask are the best guesses of the price if the next order order is a sell or a buy.
  - Easley and O'Hara (1992): Suppose that there are three possible types of event (good news, bad news, and no news), that this information is known ahead

of time to some traders, and that the market makers want to figure out this information. There are three possible trading moves, buy, sell, and nothing. The amount the price moves depends on the market maker's beliefs about the probabilities of types of news and of traders' being informed (more informed traders leads to faster price changes; this also tends to be reflected in a higher spread). One can estimate these probabilities empirically in a discrete choice model. When there are more buyer initiated trades than seller initiated trades, there is good news. When there is low volume, there is no news. The distribution of the three types of actions depends on the news. (Empirically, the spread tends to be greatest in the morning, which is when there is the biggest probability of news.)

These models contradict the *efficient markets hypothesis* that even people with private information can never make excess profits. Instead, they suggest that private information leads to short-term profits, but not long-term profits; the speed of the transition back to an efficient market depends on the market characteristics, the market maker's knowledge about informed traders, and transparency.

Informed traders are more likely to be a hurry (so they are trading during short duration times) and prefer large volumes. If the market maker expects more informed traders, then the spread is likely to be larger.

Definition The *depth* measures the number of shares that one can trade at or near the current price. The *quoted depth* is the number of shares that the market maker is willing to buy (sell) at the quoted bid (ask). One can also measure the depth as the number of shares that can be traded for up to a fixed amount of price deterioration (by looking at the limit order book). There are different depths for selling and for buying.

The *market reaction curve* plots sell/buy volume on the x-axis and the price on the y-axis. The line is broken at 0 (by the bid-ask spread), flat for the length of the quoted depth on either side of 0, and then upward-sloping, according to the limit order book (and the way the market maker would react).

Quote changes might happen because the market maker is responding to the information conveyed by an order or because a big trade starts clearing the limit order book.

For econometric models we may be interested in modeling:

- Timing: How quickly the market is moving
  - Trade duration: The time between trades
  - Quote duration: The time between quotes
  - Order duration: The time between orders
  - Execution time: The time between the submission of the order and the transaction

  - Trade-quote duration: The time from a transaction to a new quote
  - Price duration: The time needed for a price to move a certain amount

- Spread: The difference between the prices at which one can buy and one can sell. The spreads may differ for retail trades and institutional trades, since institutional trades tend to be larger.
  - Spread: $q^A - q^B$
  - Effective spread: $2|p_{t_t} - \frac{1}{2}(q^A_{t_0} + q^B_{t_0})|$, where p_{t_t} is the transaction price. This is the spread if you trade at the bid or ask. *Price improvement* occurs when the bid and ask move so that the effective spread is less than the spread because of the time delay.
  - Realized spread: $2|p_{t_t} - \frac{1}{2}(q^A_{t_t+5} + q^B_{t_t+5})|$. This is the difference between the transaction price and a future quote; this measures the price impact.
  - Information component: $q^A_{t+5} + q^B_{t+5} - (q^A_t + q^B_t)$. This measures the difference between the realized and effective spread.

When plotting data, consider whether to do it in transaction time (with one tick for each transaction) or calendar time (so that transactions are further apart on the graph if they happened further apart in time).

7.1 Conditional Durations Model

Suppose we model the time to the next price change as a random duration; this time is related to the inverse of volatility (especially if the size of the price change is constant). Then, price changes are a *point process*. The *conditional intensity process (hazard function)* of price changes can be written generally as:

$$\lambda(t, N(t); t_1, ..., t_{N(t)}) = \lim_{\Delta t \to 0} \frac{P(N(t + \Delta t) > N(t) | N(t), t_1, ..., t_{N(t)})}{\Delta t}$$

where $N(t)$ is the number of events up to time t, which occurred at times $t_1, ..., t_{N(t)}$. This is the limit of the conditional probability of the next event occurring between t and $t + \Delta t$, scaled for the length of the interval.

Definition In the *conditional durations model*, $x_i = t_i - t_{i-1}$ is described by:

$$\begin{aligned} x_i &= \psi_i \epsilon_i \\ \psi_i(t_{i-1}, ..., t_1; \theta) &= E(x_i | t_{i-1}, ..., t_1) \end{aligned}$$

where ϵ_i is independent and identically distributed with mean 1, non-negative support, and some distribution. In the *autoregressive conditional durations model*,

$$\psi_i = \omega + \sum \alpha_j x_{t-j} + \beta \psi_{i-1}$$

Other models have ψ_i as a function of exogenous variables, $y_i, ..., y_1, z_i$.

One possible distribution is the exponential, which has no memory and therefore a constant intensity process whose level shifts with ψ_i. Others include Weibull, generalized gamma, and non-parametric (all of which will have non-constant intensity processes and therefore duration dependence). Empirically, we expect a decreasing intensity function, since the longer it has been since a trade, the more likely it is that we have entered a period of low volatility and low trading, which will lead to longer durations.

For the exponential distribution, the likelihood is:

$$L = -\sum_i \log(\psi_i) + \frac{x_i}{\psi_i}$$

This is analogous to a GARCH model with x_i in place of r_t^2. Then, by using $\sqrt{x_i}$, one can use standard GARCH software to fit the model.

One could add additional exogenous variables, such as the number of transactions in the previous period (in a model of price durations, not trade durations), the current spread, or the volume per transaction. Empirically, these models show that more active markets lead to more volatility and less liquidity in the future.

Definition Suppose that arrival times, x_i, are associated with other characteristics, y_i, called *marks*. This is called a *marked point process*, and we may model (x_i, y_i) jointly.

In this context, the marks may include price, volume, and other trade characteristics.

The joint density of (x_i, y_i) is given by:

$$\begin{aligned} (x_i, y_i)|\mathcal{F}_{i-1} &\sim f(x_i, y_i|x_{i-1}, y_{i-1}, ..., x_1, y_1; \theta_i) \\ f(x_i, y_i|x_{i-1}, y_{i-1}, ..., x_1, y_1; \theta_i) &= g(x_i|x_{i-1}, y_{i-1}, ..., x_1, y_1; \theta_{1i}) q(y_i|x_i, x_{i-1}, y_{i-1}, ..., x_1, y_1; \theta_{2i}) \end{aligned}$$

This decomposes the likelihood of a marked point process into an ACD process and a process for the marks conditional on the arrival times.

Empirically, short durations (that is, higher speeds) leads to more volatility. At the tick-by-tick level, the bid-ask bounce appears (this occurs because the next trade (after a sell) will either be at the same ask or at a lower bid); this leads to negative autocorrelation in transactions prices.

7.2 Price Impact Models

Definition Each order and transaction affects the trading environment and moves prices a little. The *price impact* of a trade is how much the price moves in reaction to it.

Let P be the midquote price, R be the log change in price, T be the time between transactions, V be the volume, and x_t be -1 if the transaction price is less than the midquote (this is considered a buy) and $+1$ if the transaction price is above the midquote. Some stylized facts include:

- $Corr(|R|, T_{-1}) < 0$: A short previous time between trades implies more volatility in future prices. (This is generally explained by an information flow argument.)
- $Corr(spread, T_{-1}) < 0$: If trades are closer together, the spread tends to increase.
- $Corr(|R|, V_{-1}) > 0$: A high volume in the previous trade tends to lead to more volatility in the future.

The *Hasbrouck model* is a VAR of returns and trade directions (x_t above), so that the errors measure the new information provided by the latest trade, which may affect future prices. This can be generalized for duration effects in the price impact:

$$\begin{aligned} r_t &= \lambda^r_{open} D_t x_t + \sum_{i=0}^{5} a_i r_{t-i} + \sum_{i=0}^{5} \left(\gamma^r_i + \delta^r_i \ln(T_{t-i})\right) x_{t-i} + \epsilon_{1t} \\ x_t &= \lambda^x_{open} D_{t-1} x_{t-1} + \sum_{i=0}^{5} c_i r_{t-i} + \sum_{i=0}^{5} \left(\gamma^x_i + \delta^x_i \ln(T_{t-i})\right) x_{t-i} + \epsilon_{2t} \end{aligned}$$

This model shows that as the time between trades increases, the marginal effect of the other variables is smaller. Also, the price change associated with an unexpected buy is permanent.

Definition The *liquidity* of a market is how frictionless the market is. Liquidity is associated with low transactions costs, including the execution price and spread, the uncertainty in a transaction, and the speed of transactions.

In general, having more buyers and sellers involved in a market increases liquidity, but no market is frictionless. Three common measures of liquidity are:

- Bid-Ask Spread: This is the execution cost for a small trade.
- Depth
- Price impact of each trade

Volume is also used to measure liquidity. Though volume across assets is positively correlated with the liquidity in those asset markets, volume over time for a given asset is negatively correlated with liquidity.

In general, new information leads to an active market with a wider bid-ask spread and more price impact. At the same time, it leads to more volume (as people rebalance because of the news) and more volatility. Thus, there is a negative correlation between volume and liquidity for an individual asset because of the impact of news on both.

Definition The process through which new information is transmitted to the stock price through trading is called *price discovery*.

Definition A *liquidity trader* is person who trades an asset for exogenous reasons (such as needing cash or having excess cash) instead of for reasons related to the asset.

Engle and Patton fit an error correction model to bid and ask prices (since the two series must be cointegrated because the spread seems to be mean-reverting), including many additional independent variables. This model shows that:

- A larger spread leads to bid and ask adjustment to return the spread to its usual level.
- After a buy transaction, both the bid and the ask increase, but the ask increases more, so that the spread widens. This may occur because of asymmetric information (where the ask is raised in case there is good news, but the bid is raised less because there might not be news after all). It may also occur because of the effect on the limit order book (if this were the only effect, then the bid would not increase at all).
- The effect of a buy is smaller for stocks with more volume.
- The spread increases after any transaction but declines during periods with no trades.
- Larger trades have more impact on prices, especially for low-volume stocks.
- Short duration trades have more price impact, which means that active markets are less liquid. They also increase the spread.
- If the ask depth is greater than the bid depth, both prices will tend to decline.
- In the long run, buy orders increase the price, with the biggest impact being from medium-sized short-duration trades on infrequently traded stocks. However, the effect on the spread mostly disappears in the long run.

Liquidity can be affected by many things. Tick size may affect liquidity. Different exchanges may have different liquidity for the same asset. The trading system (electronic versus broker) also matters.

7.3 Execution Risk and Trading Strategies

Definition *Transaction costs* are the price of trading. *Implicit costs* are measured as the ratio of the execution price to the cost at the initial price minus 1 (this shows how the price changed as the trade was being executed, because of price impact, for example). *Explicit costs* measure commissions and fees as a percentage of price.

Trades that are larger or happen in markets with lower liquidity are considered more difficult, and therefore tend to have higher execution costs.

Definition *Execution risk* is the risk that the price will move during a trade during the delay between choosing to trade and when the trade is finished. This is also called the *implementation shortfall*.

In general, it is costly to decrease the risk on the trade (by trading faster), so there is a risk-return tradeoff in the trading strategy (as well as in choosing the stock itself).

Let (x_t, p_t) be the sequence of portfolios from $t = 0$ to $t = T$, with $x_T = 0$ (x_0 can be negative if the purpose is to buy a portfolio). The (implicit?) transaction cost of moving from x_0 to x_T is $TC = \sum \Delta x_t'(\tilde{p}_t - p_0)$, where $\tilde{p}_t$ is the transaction price and p_0 is the initial (midquote) stock price. In general, $\tilde{p}_t > p_0$ for buys and $\tilde{p}_t < p_0$ for sells, so that $E(TC) > 0$. We may also write:

$$TC = \sum \Delta x_t'(\tilde{p}_t - p_t) + \sum (x_T - x_{t-1})' \Delta p_t$$

The first term is the loss due to the spread and the second term is the cost of the most recent price impact (on the shares that remain to be bought or sold).

To optimize the trading strategy, one should use the same risk tolerance, λ, that is used for the usual portfolio optimization. This yields the trading optimization problem:

$$\max E\left(\sum_{t=1}^{T}\left(x_{t-1}'\Delta p_t - \Delta x_t'(\tilde{p}_t - p_t)\right)\right) + Var\left(\sum_{t=1}^{T}\left(x_{t-1}'\Delta p_t - \Delta x_t'(\tilde{p}_t - p_t)\right)\right)$$

Then, for choosing a portfolio and a trading strategy, assuming that there are no covariances between the transactions costs and the returns, there is a simple two-step method:

1. Use the standard portfolio problem to choose the target portfolio, x_T, to hold during the interval, $[T_1, T_2]$.

2. Holding x_T fixed, choose the optimal trading strategy to buy it during the interval, $[0, T_1]$.

Definition The *Sharpe ratio* is defined as $\frac{E(x_T'(p_T - p_0) - TC - RF)}{\sqrt{Var(x_T'(p_T - p_0) - TC)}}$, where RF is the risk-free rate.

If $Var(\tilde{p}_t - p_t) = 0$ (that is, the cost of immediate execution is known) and $Var(\Delta p_t) = 0$, then:

$$Var(x_T'(p_T - p_0) - TC) = \sum_{t=1}^{T} x_{t-1}' \Omega x_{t-1}$$

Note that transactions costs must be included in the estimates of both the risk and the expected return in order to correctly assess the tradeoff.

In order to choose the optimal trading strategy, we require a model for price impacts. Suppose we assume that losses due to spreads are constant and transitory and that price impacts have a permanent effect and some variance. That is,

$$\begin{aligned} Var(\tilde{p}_t - p_t|\{x_t\}) &= 0 \\ E(\tilde{p}_t - p_t|\{x_t\}) &= \tau_t \\ Var(\Delta p_t|\{x_t\}) &= \Omega \\ E(\Delta p_t|\{x_t\}) &= \pi_t \end{aligned}$$

Then, the problem is to maximize the return, with a penalty for the risks:

$$\max \sum_{t=1}^{T} \left(x_{t-1}' \pi_t - \Delta x_t' \tau_t + \lambda x_T' \Omega x_T - \lambda (x_T - x_{t-1})' \Omega (x_T - x_{t-1}) \right)$$

If π_t and τ_t are linear in x_t, this is a linear-quadratic optimization problem. This shows that the remaining risk at any t depends only on trades that have not yet been made. This shows the tradeoff between getting to x_T quickly and the transaction costs of doing this. A risk-neutral individual would trade equal amounts each period, while a risk-averse individual would trade more at the beginning.

One could also trade in a second asset to hedge the trading risks, with the eventual position in that asset being 0 again. This would require a high covariance between the two asset types.

Econometrically, one could model:

$$\%TC_i = \exp(x_i'\beta) + \exp\left(\frac{x_i'\gamma}{2}\right)\epsilon_i$$

where $\epsilon_i \sim Normal(0, 1)$. This model ensures that the average cost is always positive, but allows the transactions costs to be negative (for the case in which the price moves in an advantageous direction during trading). β and γ imply a risk-return frontier. Empirically, increases in spread, volatility, and the ratio of value (of the trade?) to the total volume in the stock lead to increases in the risk of the trade. As the trading strategy used is more "urgent" (that is, it happens faster), the variance of the costs goes down (and, therefore, the risk goes down).