A Practical Guide to GMM

Tom Arnold and Timothy Falcon Crack*

September 1999

*Assistant Professor, Ourso College of Business Administration, Finance Department, Louisiana State University, 2159 CEBA, Baton Rouge, LA 70803, E-mail: tmarnol@unix1.sncc.lsu.edu, Tel: (225) 388 6369, Fax: (225) 388 6266, and Assistant Professor, Finance Department, Kelley School of Business, Indiana University, 1309 East Tenth Street, Bloomington, IN 47405-1701. E-mail: tcrack@indiana.edu, Tel: (812) 855 2695; Fax: (812) 855 5875. We thank Louis Scott. A much-expanded version of the current paper "A Practical Guide to GMM (with Applications to Option Pricing)" is available upon request from the authors.

A Practical Guide to GMM

Abstract

We explain how and why Generalized Method of Moments (GMM) works. We identify problem areas in implementation and we give tactical GMM estimation advice, troubleshooting tips, and pseudo code. We pay particular attention to proper choice of moment conditions, exactly-identified versus over-identified estimation, estimation of Newey-West standard errors, and numerical optimization in the presence of multiple local extrema.

JEL Classification: A23, C13, C23, G13.

Keywords: Generalized Method of Moments, GMM, Newey-West.

I Introduction

Generalized Method of Moments (GMM) has been in existence since 1982, but a lack of adequate explanation in the literature has lead to underutilization in financial economics. We use a simple example to explain how and why GMM works. We then draw on our experience with GMM estimation of option pricing models to give tactical estimation advice, troubleshooting tips, and pseudo-code. Our intended audience includes empirical financial economics researchers and students of econometrics.

The paper proceeds as follows: in Section II we explain how and why GMM works via an extended yet simple example; in Section III we explain the attractive features of GMM; in Section IV we give GMM implementation advice; Section V concludes; and Appendix A presents pseudo code for GMM estimation.

II Understanding GMM – A Simple Example

Generalized method of moments is a generalization of the classical Method of Moments (MOM) estimation technique. The classical MOM technique equates sample and population moments to enable estimation of population

parameters. We think that MOM, GMM, and the relationship between them are best understood via a simple example (ours is a much-expanded version of one given by Hamilton (1994, pp409-412)). A much more sophisticated application of GMM to actual option pricing data appears in Section IV.

Suppose we have data $Y_1, \ldots, Y_T$ distributed Student-t with ν degrees of freedom, and we want to estimate ν. It is well known that the mean and variance of a Student-t are $E(Y_t) = 0$, and $E(Y_t^2) = \frac{\nu}{\nu-2}$, respectively, for $\nu > 2$ (Evans et al (1993)). By equating the sample second moment $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T} Y_t^2$ and the population second moment $\frac{\nu}{\nu-2}$, we may deduce a MOM estimator for ν as in Equation (1).

$$\hat{\nu}^{(1)} = \frac{2\hat{\sigma}^2}{\hat{\sigma}^2 - 1}. \tag{1}$$

For data distributed Student-t, it is also known that the population fourth moment satisfies $\mu_4 = E(Y_t^4) = \frac{3\nu^2}{(\nu-2)(\nu-4)}$ when $\nu > 4$ (Evans et al (1993)). Suppose we equate the sample fourth moment $\hat{\mu}_4 = \frac{1}{T}\sum_{t=1}^{T} Y_t^4$ and the population fourth moment $\frac{3\nu^2}{(\nu-2)(\nu-4)}$, and solve for ν. After manipulation, the quadratic nature of the problem yields two additional MOM solutions for ν.

Without loss of generality call both solutions "$\hat{\nu}^{(2)}$" as in Equation (2).

$$\hat{\nu}^{(2)} = \frac{3\hat{\mu}_4 \pm \sqrt{\hat{\mu}_4(\hat{\mu}_4 + 24)}}{(3\hat{\mu}_4 - 3)}. \tag{2}$$

In repeated simulations we find that only one of the two quadratic solutions in Equation (2) is close to the $\hat{\nu}^{(1)}$ of Equation (1) – the other is spurious and should be discarded.[1] Thus we obtain two estimators of the degrees of freedom for the Student-t distribution: $\hat{\nu}^{(1)}$, and $\hat{\nu}^{(2)}$.

The estimators $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ are based on the sample statistics $\hat{\sigma}^2$, and $\hat{\mu}_4$ respectively. If these latter statistics are by chance equal to the true parameters (i.e. $\hat{\sigma}^2 = \sigma^2$, and $\hat{\mu}_4 = \mu_4$), then our two estimators $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ are identical and are equal to the true ν. However, both $\hat{\sigma}^2$, and $\hat{\mu}_4$ are necessarily estimated with error. Their sampling distributions are continuous probability densities, so it follows that although our two estimators of ν are similar, the chance that they are the same in practice is zero. Thus the parameter ν is over identified and no single $\hat{\nu}$ will solve both Equations (1) and (2).

To illustrate we simulate 1,000 independent drawings $Y_1, \ldots, Y_{1000}$ from a Student-t distribution with $\nu = 10$ degrees of freedom. For our simulated

data set we get $\hat{\nu}^{(1)} = 9.252$ and $\hat{\nu}^{(2)} = 10.408$ (the other quadratic root is $\hat{\nu}^{(2)} = 1.529$ and we discard it as spurious). We see that although both $\hat{\nu}^{(1)}$ and $\hat{\nu}^{(2)}$ are close to the true $\nu = 10$, the sampling error in the second and fourth moments respectively yields distinct estimators of the degrees of freedom ν.

Equations (1) and (2) are transformations of the original moment-matching conditions repeated here in Equations (3).

$$\hat{\sigma}^2 = \frac{\nu}{\nu - 2}, \quad \text{and} \quad \hat{\mu}_4 = \frac{3\nu^2}{(\nu - 2)(\nu - 4)}. \tag{3}$$

From our sampling error arguments it follows that no single MOM estimator $\hat{\nu}$ can solve both moment-matching conditions in Equations (3). With two estimators $\hat{\nu}^{(1)} = 9.252$ and $\hat{\nu}^{(2)} = 10.408$, how then are we to choose a single sensible MOM estimator for $\hat{\nu}$?

The solution is to generalize our MOM approach for estimating ν by introducing a 2×2 weighting matrix W that reflects our confidence in each moment-matching condition in Equations (3). We execute this by stacking

our previous moment-matching conditions into a vector g as in Equation (4).

$$g(\nu) \equiv \begin{bmatrix} \hat{\sigma}^2 - \frac{\nu}{\nu-2} \\ \\ \hat{\mu}_4 - \frac{3\nu^2}{(\nu-2)(\nu-4)} \end{bmatrix}. \tag{4}$$

We then minimize the scalar quadratic objective function $Q(\nu) = g(\nu)' W g(\nu)$ with respect to choice of ν. The contents of W describe the relative importance of each moment-matching condition in determining $\hat{\nu}$. If the matrix W is simply the 2×2 identity,[2] then the objective function reduces to

$$\begin{aligned} Q(\nu) &= g(\nu)' g(\nu) \\ &= \left[\hat{\sigma}^2 - \frac{\nu}{\nu-2}\right]^2 + \left[\hat{\mu}_4 - \frac{3\nu^2}{(\nu-2)(\nu-4)}\right]^2. \end{aligned}$$

If we think of the moment conditions as residuals (i.e. deviations from their ideal value of zero), then an identity weighting matrix reduces the quadratic objective function to a sum of squared residuals, and it reduces the optimization to a traditional least squares problem.

More generally, the weighting matrix W should place more weight on the moment-matching conditions in which we have more confidence. The obvious choice for a W that is directly related to our confidence in the moment

conditions in Equation (4) is the inverse of the variance-covariance matrix (VCV) of the moment conditions. In practice, rather than inverting the VCV of the vector g, we shall invert the asymptotic VCV defined as $\Omega \equiv \text{var}(\sqrt{T}g) = T\text{var}(g)$. This VCV is typically a function of the estimator itself, but we suppress the dependence of Ω on ν for ease of notation. Thus we shall minimize $Q(\nu) = g(\nu)'\Omega^{-1}g(\nu)$.

How are we to think of the minimization of the objective function $Q(\nu)$? A simple analogy is that when evaluated at the optimum $\hat{\nu}$, the objective $Q(\hat{\nu}) = g(\hat{\nu})'\hat{\Omega}^{-1}g(\hat{\nu})$ is similar to a squaring of a scaled version of the traditional t-statistic for testing whether a population mean is zero as reported in Equation (5).

$$\left\{ \underbrace{\left(\frac{1}{\sqrt{T}}\right)}_{\text{scale factor}} \cdot \underbrace{\left[\frac{\hat{\mu}-0}{\hat{\sigma}/\sqrt{T}}\right]}_{t\text{-statistic}} \right\}^2 = (\hat{\mu}-0)\left[T\left(\frac{\hat{\sigma}}{\sqrt{T}}\right)^2\right]^{-1}(\hat{\mu}-0). \tag{5}$$

The analogy follows because: $g(\hat{\nu})$ and $(\hat{\mu}-0)$ should be zero under their respective null hypotheses; $g(\hat{\nu})$ and $(\hat{\mu}-0)$ appear fore and aft in their respective expressions; $\hat{\Omega}$ and $[T(\hat{\sigma}/\sqrt{T})^2]$ are the estimated asymptotic variances of $g(\hat{\nu})$ and $\hat{\mu}$ respectively (i.e. you divide them by T to get actual vari-

ances of $g(\hat{\nu})$ and $\hat{\mu}$ respectively in large sample); and finally, the asymptotic variances are inverted in the kernels of both expressions.

When minimizing the objective function $Q(\nu)$, we are building through choice of $\hat{\nu}$ a test statistic least likely to reject the hypothesis that the moments are zero. If we multiply the optimized objective function by T, we get an asymptotically chi-squared test statistic for whether the moment conditions are zero.[3] Thus by construction the GMM estimator of ν is that value of ν statistically least likely to reject the null hypothesis that the moments $g(\nu)$ are zero. This optimal ν is selected via weighted least squares minimization of a quadratic function of moment conditions.

In the case of our simulation of 1,000 observations from a Student-t, the objective function $Q(\hat{\nu}) = g(\hat{\nu})'\hat{\Omega}^{-1}g(\hat{\nu})$ is shown in Figure 1. A simple numerical minimization of the objective function in Figure 1 locates the GMM estimator $\hat{\nu} = 11.337$. Our two MOM estimators $\hat{\nu}^{(1)} = 9.252$ and $\hat{\nu}^{(2)} = 10.408$ bracket the true ν, but the GMM estimator is outside of this range.[4] The numerical ordering of $\hat{\nu}^{(1)}$, $\hat{\nu}^{(2)}$, and the GMM estimator vary with the random seed used in the simulation.

The next question is how to get the standard error of our GMM estimator. We need to give more details of the formal GMM setup and the

VCV estimations to answer this question. Hansen's (1982) GMM is a formalization of our technique. Assume the underlying data X_t are stationary and ergodic.[5] We use economic theory and intuition (see examples in our Section IV) to obtain q unconditional moment restrictions on f (a vector of functions) for true parameter vector β_0:

$$f(X_t, \beta_0) = \begin{bmatrix} f_1(X_t, \beta_0) \\ f_2(X_t, \beta_0) \\ \vdots \\ f_q(X_t, \beta_0) \end{bmatrix}, \quad \text{where} \quad E[f(X_t, \beta_0)] = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

For admissible β, we let $g_T(\beta) \equiv \frac{1}{T}\sum_{t=1}^{T} f(X_t, \beta)$. In our Student-$t$ example we are implicitly using

$$f(Y_t, \nu) = \begin{bmatrix} f_1(Y_t, \nu) \\ f_2(Y_t, \nu) \end{bmatrix} = \begin{bmatrix} Y_t^2 - \frac{\nu}{\nu-2} \\ Y_t^4 - \frac{3\nu^2}{(\nu-2)(\nu-4)} \end{bmatrix} \text{ for } t = 1, \ldots, T$$

$$\text{to get} \quad g(\nu) = \frac{1}{T}\sum_{t=1}^{T} f(Y_t, \nu) = \begin{bmatrix} \hat{\sigma}^2 - \frac{\nu}{\nu-2} \\ \hat{\mu}_4 - \frac{3\nu^2}{(\nu-2)(\nu-4)} \end{bmatrix}.$$

Let W_T be positive definite such that $\lim_{T\to\infty} W_T = W$, where W is posi-

tive definite, then the GMM estimator $\hat{\beta}_{GMM}$ is the choice of β that minimizes the scalar quadratic objective function $Q_T(\beta) = g_T(\beta)' W_T g_T(\beta)$ (we argue shortly that $W_T = \Omega^{-1}$). If we assume that a weak law of large numbers applies to the average g, so that $g_T(\beta) \xrightarrow{T} E[g_T(\beta)]$, and in particular $g_T(\beta_0) \xrightarrow{p} 0$, then $\hat{\beta}_{GMM}$ is consistent for the true β_0. Assume a central limit theorem applies to $f(X_t, \beta_0)$, so that the (appropriately scaled) sample mean g of the f_t's satisfies $\sqrt{T} g_T(\beta_0) \overset{\text{"a"}}{\sim} \mathcal{N}(0, \Omega)$ (where Ω is the asymptotic VCV of g).

If $q = p$ (the number of restrictions in f equals the number of parameters in β_0), then β_0 is exactly identified, and $\hat{\beta}_{GMM}$ is independent of choice of the weighting matrix. However, if $q > p$, β_0 is over identified. In this case different weighting matrices lead to different $\hat{\beta}_{GMM}$. In either case, a numerical optimization routine (e.g. Newton-Raphson, or Berndt et al (1974)) is typically but not always needed to find $\hat{\beta}_{GMM}$. Hansen (1982) shows that in the over-identified case, $W_0 = \Omega^{-1}$ gives the asymptotically efficient GMM estimator (consistent with our earlier intuition that the weights should be inversely related to our uncertainty regarding the moments). The GMM

estimator $\hat{\beta}_{GMM}$ is asymptotically Normal, with

$$\begin{aligned}
\sqrt{T}(\hat{\beta}_{GMM} - \beta_0) &\overset{\text{"a"}}{\sim} \mathcal{N}[0, V_{GMM}], \quad \text{where} \\
V_{GMM} &= (\Gamma'\Omega^{-1}\Gamma)^{-1}, \\
\Gamma &= E\left(\frac{\partial g_T(\beta_0)}{\partial \beta'}\right) = E\left(\frac{1}{T}\sum_{t=1}^{T}\frac{\partial f(X_t, \beta_0)}{\partial \beta'}\right), \quad \text{and} \\
\Omega &= E[\sqrt{T}g_T(\beta_0) \cdot \sqrt{T}g_T(\beta_0)'] \\
&= E\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}\underbrace{f(X_t, \beta_0)}_{q\times 1}\underbrace{f(X_s, \beta_0)'}_{1\times q}\right].
\end{aligned}$$

The matrix Ω gives the asymptotic VCV of the moment conditions g.[6] To get from Ω to the asymptotic VCV of the parameter vector $\hat{\beta}_{GMM}$ we need the matrix Γ to capture the relationship between the moments and the parameters. This is why Γ pre- and post-multiplies Ω^{-1} in the calculation of the VCV matrix of the parameters. Γ can sometimes be calculated explicitly. Otherwise it is estimated using $\hat{\Gamma} = \frac{1}{T}\sum_{t=1}^{T}\frac{\partial f(X_t, \hat{\beta}_{GMM})}{\partial \beta'}$, which is typically a function of $\hat{\beta}_{GMM}$. In our Student-t example, the data are IID, so Γ reduces to

$$\Gamma = E\left[\frac{\partial f(Y_t, \nu)}{\partial \nu}\right] = \begin{bmatrix} \frac{2}{(\nu-2)^2} \\ \frac{6\nu(3\nu-8)}{(\nu-2)^2(\nu-4)^2} \end{bmatrix}, \tag{6}$$

and we estimate Γ during the optimization by substituting in $\hat{\nu}$.

Assume the components of the moment vector $f(X_t, \beta_0)$ are not auto- or cross-correlated at any non-zero lag (i.e. $E[f_i(X_t, \beta_0) f_j(X_s, \beta_0)] = 0$ for all $t \neq s$ and for any i and j). If f is potentially heteroskedastic (i.e. $\text{var}(f_i) \neq \text{var}(f_j)$ for some i and j), then the matrix Ω may be estimated using the White (1980) estimator in Equation (7).

$$\hat{\Omega}_{WHITE} = \frac{1}{T} \sum_{t=1}^{T} f(X_t, \hat{\beta}_{GMM}) f(X_t, \hat{\beta}_{GMM})'. \tag{7}$$

It may be seen that the ij^{th} element of $\hat{\Omega}_{WHITE}$ estimates the ij^{th} element of the asymptotic VCV of the vector g as follows.

$$\begin{aligned}
\text{var}\left(\sqrt{T}g\right)_{ij} &= \text{cov}\left(\sqrt{T}g_i, \sqrt{T}g_j\right) \\
&= E\left(\sqrt{T}g_i \sqrt{T}g_j\right) \\
&= TE(g_i g_j) \\
&= TE\left[\frac{1}{T}\sum_{t=1}^{T} f_i(X_t, \beta_0) \cdot \frac{1}{T}\sum_{s=1}^{T} f_j(X_s, \beta_0)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} E[f_i(X_t, \beta_0) f_j(X_s, \beta_0)] \\
&= \frac{1}{T}\sum_{t=1}^{T} E[f_i(X_t, \beta_0) f_j(X_t, \beta_0)] + \text{cross terms} \\
&= E[f_i(X_t, \beta_0) f_j(X_t, \beta_0)]
\end{aligned}$$

$$\approx \frac{1}{T}\sum_{t=1}^{T} f_i(X_t,\beta_0)f_j(X_t,\beta_0),$$

which is just the ij^{th} element of the White estimator in Equation (7). In the above, we use the fact that $E(g_i) = E(g_j) = 0$, we denote the i^{th} and j^{th} elements of the vector f as f_i, and f_j respectively, the cross-terms are zero because the moments are assumed uncorrelated (i.e. they have no own- or cross-correlation at any non-zero lag), and at the last step we use the Weak Law of Large Numbers.

If the components of the vector of moments $f(\beta_0)$ do exhibit auto- or cross-correlation (i.e. $E[f_i(X_t,\beta_0)f_j(X_s,\beta_0)] \neq 0$ for some i, j, and some $t \neq s$), and are also potentially heteroscedastic, then the Newey-West (1987) estimator of Ω may be used (Equation (8)). Practical choice of lag length m is discussed in Section C.

$$\begin{aligned}
\hat{\Omega}_{NW} &= \hat{\Phi}_0 + \sum_{j=1}^{m} w(j,m)(\hat{\Phi}_j + \hat{\Phi}_j'), \quad m \ll T, \quad \text{where} \qquad (8)\\
\hat{\Phi}_j &\equiv \frac{1}{T}\sum_{t=j+1}^{T} f(X_t,\hat{\beta}_{GMM})f(X_{t-j},\hat{\beta}_{GMM})', \quad \text{and}\\
w(j,m) &= 1 - \frac{j}{(m+1)}.
\end{aligned}$$

Note that $\hat{\Phi}_0$ in Equation (8) is just White's estimator, which Newey-West extends. Thus the Newey-West estimator is robust to both heteroskedasticity and autocorrelation of the components of the moment vector f. The White and Newey-West estimators are not the only estimators available for Ω (see Ogaki(1993), Hamilton (1994)).

To avoid any misunderstanding, let us emphasize that the White and Newey-West VCV matrix estimators as used here provide standard errors that are robust to heteroskedasticity and autocorrelation in the *moment conditions* f_t, but not necessarily in the *underlying data* X_t. However, if you have autocorrelation or heteroskedasticity in the underlying data, and this generates autocorrelation or heteroskedasticity in the moments, then your standard errors are robust to these latter deviations. Using White and Newey-West are thus analogous to assuming your data are stationary when using OLS, but allowing for non-spherical residuals – the GMM moment conditions are effectively model residuals. Note also that although different moments may have different variances, the variance structure must be stationary else GMM is not valid.

If $q > p$ (so that the parameters are over-identified), then as suggested earlier, multiplying the objective function by the sample size yields a chi-

squared test statistic at the optimum (Equation (9)).

$$T \times Q_T(\hat{\beta}_{GMM}) = T g(\hat{\nu})' \hat{\Omega}^{-1} g(\hat{\nu}) \overset{\text{"a"}}{\sim} \chi^2_{q-p}, \tag{9}$$

The distribution in Equation (9) assumes $Q_T(\beta)$ is minimized using $W = \Omega^{-1}$ (Hansen (1982)). This is a large sample test of whether the sample moments $g_T(\hat{\beta}_{GMM})$ are as close to zero as would be expected if the expectation of the population moments $E[f(X_t, \beta_0)]$ are truly zero. It is a test of model specification and is particularly strong if it rejects (after all, we choose $\hat{\beta}_{GMM}$ specifically to minimize the likelihood that this test will reject).

The chi-squared test statistic in Equation (9) is a quadratic function of the moment conditions. If the test statistic rejects, then the underlying model that generated the system of moment conditions is declared invalid. It is thus important that we select an informative set of moment conditions if the chi-squared test statistic is to truly test the model being estimated. Passing the chi-squared test is no guarantee of statistical significance of individual parameter estimators. Thus the best set of moment conditions will be those that not only pass the chi-squared test, but also admit the least number of possible values for a given model parameter (i.e. a low standard error).

There is some potential for an unscrupulous econometrician to game the chi-squared test via choice of moment conditions. We illustrate both sensible and non-sensible moment condition choices and discuss gaming the chi-squared statistic in Section IV.

Continuing our Student-t example, Table 1 reports results from several simulations of increasing sample size. In each case the two MOM estimators $\hat{\nu}^{(1)}$, and $\hat{\nu}^{(2)}$ are reported, along with the GMM estimator $\hat{\nu}_{GMM}$, the standard error of the GMM estimator, and the chi-squared goodness-of-fit test (there are $q = 2$ moments, and $p = 1$ parameters, so it is a χ^2_1 statistic). Only in the case $T = 10,000$ does the statistic reject, and that is probably a Type-I error. The standard errors fall as the sample size rises and the GMM estimator is quite close to the true degrees of freedom ($\nu = 10$) in the larger samples, but in each case (at least in this simulation) $\hat{\nu}_{GMM} > \nu = 10$.

III Attractive Features of GMM

The first attractive feature of GMM is that it is distributionally nonparametric. Unlike MLE, it does not place distributional assumptions on the data. However, the GMM moment conditions are certainly *functionally* parametric

– you must impose a specific functional form. GMM's initial assumptions of stationarity and ergodicity are also relatively weak compared to the more traditional assumption that the data are IID. However, the resulting statistics are all asymptotic, so a sizable sample is required

GMM is well-suited to horribly non-linear models such as option pricing models. This is partly because no matter how ugly the option pricing model, it still generates natural moment conditions. Examples include "model price less market price equals zero." These moment conditions can easily be chosen to test particularly interesting phenomena (e.g. the "volatility smile" discussed in Section IV).

If you do not use GMM, and you fit option prices to model prices by minimizing sum of squared errors, then it is not clear how you conduct tests of goodness of fit, or tests for significance of individual parameters (e.g. Bakshi, Cao and Chen (1998) are unable to perform statistical tests). However, for example, if you allow for different implied volatilities across strike prices, then once the parameters are estimated, GMM allows for the traditional Wald-type tests of whether the parameters are the same.

GMM subsumes OLS, 2SLS, 3SLS, and other methods, so it is a very general technique (Ogaki (1993)). Like these methods, it is very easy to incorpo-

rate VCV estimations that allow for heteroskedasticity and autocorrelation in the moment conditions – using White or Newey-West VCV estimators.

The model specifications are used to create the GMM moment conditions. It follows that you do not have to proxy for these elements of the model (which would introduce error). That is, rather than proxying for a parameter and then regressing a left hand side dependent variable on this and other proxies to see if there is a relationship, you set up moment conditions that allow you to deduce the parameter from the data and the functional form of the model. The relationship is then tested using standard errors on individual parameters, and the chi-squared test for the overall model. For example, a consumption-based asset pricing model would not necessarily need a proxy for a consumption parameter.

You can look at the minimizations of specific GMM moment conditions to determine what aspect of the data is not being properly captured by the model. However, you should be careful of data mining in this instance. In fact if you look at the errors in any moment condition over time you may be able to capture specific sections of time where the model is not consistent with the data and this could indicate a regime shift.

IV Implementing GMM

A Choosing Moments

GMM moments should be economically sensible reflections of the model being estimated. For example, elsewhere we fit the Black and Scholes (1973) option pricing model to market prices of SPX (i.e. S&P500 index) options using six moneyness/maturity classes.[7] In this case, the most obvious choice of moments is those that match market prices of options to model prices.

$$f_t(\theta) = \begin{bmatrix} f_{t,1}(\theta) \\ f_{t,2}(\theta) \\ \vdots \\ f_{t,6}(\theta) \end{bmatrix} = \begin{bmatrix} c_{t,1,1}^{(\text{mkt})} - c_{t,1,1}^{(\text{BS})}(\sigma_{1,1}) \\ c_{t,1,2}^{(\text{mkt})} - c_{t,1,2}^{(\text{BS})}(\sigma_{1,2}) \\ c_{t,1,3}^{(\text{mkt})} - c_{t,1,3}^{(\text{BS})}(\sigma_{1,3}) \\ c_{t,2,1}^{(\text{mkt})} - c_{t,2,1}^{(\text{BS})}(\sigma_{2,1}) \\ c_{t,2,2}^{(\text{mkt})} - c_{t,2,2}^{(\text{BS})}(\sigma_{2,2}) \\ c_{t,2,3}^{(\text{mkt})} - c_{t,2,3}^{(\text{BS})}(\sigma_{2,3}) \end{bmatrix}$$

where $\theta = [\sigma_{1,1}\ \sigma_{1,2}\ \ldots\ \sigma_{2,3}]'$ is the vector of implied volatilities to be estimated for the six moneyness/maturity classes over the sample period, $c_{t,i,j}^{(\text{mkt})}$ is the market price of a call option falling in the i^{th} maturity class, and the

j^{th} moneyness class during time window t, and $c^{(\text{BS})}_{t,i,j}(\sigma_{i,j})$ is the dividend-adjusted Black-Scholes call pricing formula for the same option with the implied volatility $\sigma_{i,j}$ as the plug figure. If the $\sigma_{i,j}$ are allowed to differ from one another in the estimation, then the GMM estimation is exactly-identified (number of parameters equals number of moments). If we restrict any of the $\sigma_{i,j}$ to be equal, then the estimation is over-identified.

In the case of the Black-Scholes estimation we may be tempted to append moments to the vector f_t to reflect Black-Scholes assumptions for geometric Brownian motion. These assumptions include that the continuously compounded returns (i.e. log of the price relative) are normally distributed (no

skewness or kurtosis) with no autocorrelation (as in Equation (10)).

$$f_t(\theta) = \begin{bmatrix} f_{t,1}(\theta) \\ f_{t,2}(\theta) \\ \vdots \\ f_{t,6}(\theta) \\ f_{t,7}(\theta) \\ f_{t,8}(\theta) \\ f_{t,9}(\theta) \\ f_{t,10}(\theta) \\ f_{t,11}(\theta) \end{bmatrix} = \begin{bmatrix} c_{t,1,1}^{(\text{mkt})} - c_{t,1,1}^{(\text{BS})}(\sigma_{1,1}) \\ c_{t,1,2}^{(\text{mkt})} - c_{t,1,2}^{(\text{BS})}(\sigma_{1,2}) \\ \vdots \\ c_{t,2,3}^{(\text{mkt})} - c_{t,2,3}^{(\text{BS})}(\sigma_{2,3}) \\ \log_e\left(\frac{S_t}{S_{t-1}}\right) - \mu \\ \left[\log_e\left(\frac{S_t}{S_{t-1}}\right) - \mu\right]^2 - \sigma_h^2 \\ \left[\log_e\left(\frac{S_t}{S_{t-1}}\right) - \mu\right] \cdot \left[\log_e\left(\frac{S_{t-1}}{S_{t-2}}\right) - \mu\right] - \rho \cdot \sigma_h^2 \\ \frac{\left[\log_e\left(\frac{S_t}{S_{t-1}}\right) - \mu\right]^3}{\sigma_h^3} - \psi \\ \frac{\left[\log_e\left(\frac{S_t}{S_{t-1}}\right) - \mu\right]^4}{\sigma_h^4} - \kappa \end{bmatrix} \tag{10}$$

where S_t is the SPX index level at time t, and $\theta = [\sigma_{1,1}\ \sigma_{1,2}\ \ldots\ \sigma_{2,3}\ \mu\ \sigma_h\ \rho\ \psi\ \kappa]'$ where the $\sigma_{i,j}$ are as before, μ is the mean return, σ_h is the historical sample standard deviation of returns (as opposed to the implied volatilities which are forward looking), ρ is the first order autocorrelation of returns, ψ is the skewness of returns, and κ is the kurtosis of returns. However, this expanded set of moment conditions does not make sense. In testing the assumptions of the model rather than the actual quality of the pricing we risk rejecting a

model with good pricing that is robust to assumptions that are not strictly true. Testing assumptions rather than pricing is one step removed from what is economically important. In other words, we do not really care if stock returns are normally distributed in this case as long as the model prices are close to the market prices.

In the over-identified case, adding the above-mentioned or any other "nonsense moments" serves to increase the degrees of freedom of the chi-squared test of over-identifying restrictions and makes it less likely that the test will reject the model. This amounts to a gaming of the chi-squared test and thus we believe that any highly over-identified GMM estimation should be viewed with substantial skepticism. This is so even if the underlying model is rejected because of the potential for both Type I and Type II errors.

B Exactly- versus Over-Identified Estimation

When a GMM estimation is exactly-identified there is a $\hat{\theta}$ that sets $g(\hat{\theta}) \equiv \frac{1}{T}\sum_{t=1}^{T} f_t(\hat{\theta})$ identically to zero. This means that whatever the weighting matrix used, the same parameter estimates will be obtained. In some exactly-identified cases (e.g. our Black-Scholes estimation) a numerical optimization

is needed to locate the parameter estimates that set g to zero, but in other cases no numerical technique is required.[8] The chi-squared goodness of fit test no longer applies in the exactly-identified case.

Testing whether Black-Scholes describes the data reduces to testing whether the "volatility smile" is flat or not ($H_0 : \sigma_{i,1} = \sigma_{i,2} = \sigma_{i,3}$ for each maturity class i). This may be achieved using a standard Wald test in the exactly-identified case.[9] An alternative to the exactly identified approach is to keep the same moments, but impose $\sigma_{i,1} = \sigma_{i,2} = \sigma_{i,3} = \sigma_i$, say, for each maturity class i during the estimation. The GMM estimation is then over-identified and we cannot conduct a standard Wald test of whether the smile is flat, but if we have falsely forced the smile to be flat, then the moments will be non-zero, and the chi-squared test of over-identifying restrictions will reject the model. Our experience with GMM estimation of Black-Scholes leaves us strongly in favor of exactly-identified estimation for reasons summarized in Table 2. We concede that in some other applications the trade-off between exactly- and over-identified estimation might favor over-identified estimation. Hints for handling multiple local optima appear in Table 3. Pseudo code for GMM implementation of our Black-Scholes estimation appears in Appendix A.

C Newey-West Lag Lengths

When implementing GMM using Newey-West standard errors, the choice of lag length m in Equation (8) is important for two reasons. Firstly, insufficient lag length can lead to inappropriate standard errors (either too large or too small depending upon the nature of the dependence in the data). Secondly, insufficient lag length can cause a lack of smoothness in the GMM objective function that can slow numerical optimization (recall that $W = \Omega^{-1}$ is estimated via Newey-West). A practical method for choosing the lag length m in the Newey-West estimator is to estimate the parameters and their standard errors using $m = 5$ and then re-estimate everything using increasing lag lengths until the lag length has negligible effect on the standard errors of the parameters to be estimated. In our Black-Scholes estimation we noticed substantial differences in both standard errors and quality of optimization as our lag lengths increased up to about 35, but beyond lag 50 (which we used) there was no change. One rule of thumb is to use $m = \sqrt{T} + 5$ where T is the sample size. In our case, $T \approx 2500$, so 55 lags is suggested (in agreement with our empirical findings).

V Conclusion

We believe the GMM estimation technique is underutilized in empirical finance. The goal of this paper is to provide a practical guide to GMM in order to promote wider use. We use a simple example to explain how and why Generalized Method of Moments (GMM) works. We identify problem areas in implementation and we give tactical estimation advice and troubleshooting tips. We pay particular attention to proper choice of moment conditions, exactly-identified versus over-identified estimation, estimation of Newey-West standard errors, and numerical optimization in the presence of multiple local extrema.

A Pseudo-Code for GMM

```
% INITIAL DATA MANIPULATION
% load options and t-bill data
  load data=[day time-stamp call put S X tau r]
% infer dividend yields
  q=(1/tau).*log(S./(call-put+X.*exp(-r.*tau))
% screen for no-arbitrage violation
  F=S.*exp((r-q).*tau)
  failure=call<exp(-r.*tau).*(F-X).*(sign(log(F./X))+1)/2
  data(find(failure))=void

% FORM GMM MOMENTS AND OPTIMIZE
  theta=initialguess
  while update./theta>sqrt(my computer's floating point precision)[10]
    d=(log(S./X)+(r-q+(theta.^2)/2).*tau)./(theta.*sqrt(tau))
    f=(c-(S.*exp(-q.*tau).*N(d)-...
    X.*exp(-r.*tau).*N(d-theta.*sqrt(tau))))'
    g=mean(f')'
    % FIND DERIVATIVE OF g W.R.T. THETA
    dfdth=S.*exp(-q.*tau).*(sqrt(tau)/sqrt(2*pi)).*exp(-((d.^2)/2))
    dg=diag(mean(dfdth)) % same as Gamma
    % USE WHITE AND NEWEY-WEST TO FIND W=OMEGA
    WHITE=(1/T)*f*f'; NWEST=0
    m=50
    for j=1:m
       phij=(1/T)*f(:,j+1:T)*f(:,1:T-j)'
       NWEST=NWEST+(1-(j/(m+1)))*(phij+phij')
    end
    NWEST=NWEST+WHITE
    W=inverse(NWEST)
    update=inverse(dg'*W*dg)*(dg'*W*g)
    theta=theta+update
  end while loop

% CALCULATE STANDARD ERRORS AND DO TESTS
```

```
  OMEGA=WHITE
  VGMM=inv(Gamma'*inv(OMEGA)*Gamma)
  SEWHITE=sqrt(diag(VGMM)/T)

  OMEGA=NWEST
  VGMM=inv(Gamma'*inv(OMEGA)*Gamma)
  SENWEST=sqrt(diag(VGMM)/T)
  print [theta SEWHITE SENWEST]

% TEST VOLATILITY SMILE
  R1=[...
  1 -1 0 0 0 0
  0 1 -1 0 0 0
  0 0 0 1 -1 0
  0 0 0 0 1 -1]
  testsmile=(R1*theta)'*inv(R1*VGMM*R1'/T)*(R1*theta)
  pvalue=1-cdf('chi2',testskew,#rows(R))]
  print [testsmile pvalue]

% TEST VOLATILITY TERM STRUCTURE
  R2=[...
  1 0 0 -1 0 0
  0 1 0 0 -1 0
  0 0 1 0 0 -1]
  testterm=(R2*theta)'*inv(R2*VGMM*R2'/T)*(R2*theta)
  pvalue=1-cdf('chi2',testterm,#rows(R2))
  print [testterm pvalue]
```

B Footnotes

1. The existence of multiple roots to Equation (2) could cause problems for an optimizing algorithm. However, comparison of $\hat{\nu}^{(2)}$ to $\hat{\nu}^{(1)}$ should enable the correct root to be located. Thus, the use of both moments becomes necessary in this case.

2. We shall see in Section IV that you sometimes need to begin a GMM optimization routine with an identity weighting matrix to get initial values for the full estimation.

3. The reader can confirm that Equation (5) shares the same property: multiply it by T and it is asymptotically chi-squared – with one degree of freedom.

4. Note also a local minimum visible in Figure 1 at approximately $\nu = 3$. We discuss over-identified estimation and the existence of local extrema in Section IV.

5. Stationarity is stronger than "identically distributed," but weaker than IID, since stationarity does not imply independence. Ergodicity is weaker than independence – it is a form of average asymptotic indepen-

dence that restricts dependence or memory in a sequence. Stationary and ergodicity together are strictly weaker than IID. See White (1984, pages 41-46) for details.

6. The ij^{th} element of Ω is given by $E\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T} f_i(X_t,\beta_0)f_j(X_s,\beta_0)\right]$, where f_i, and f_j are the i^{th} and j^{th} elements respectively of the vector f.

7. A much-expanded version of the current paper titled "A Practical Guide to GMM (with Applications to Option Pricing)" is available upon request from the authors.

8. If you are estimating the mean, variance, skewness, kurtosis, and other simple parameters then not only is the GMM estimation exactly identified, but no numerical optimization technique is required because the traditional estimators set the mean of the GMM moments to zero. In this case, the optimal weighting matrix is not needed in the optimization, but is still used to calculate standard errors.

9. For $\hat{\theta} = [\sigma_{1,1}\ \sigma_{1,2}\ \sigma_{1,3}\ \sigma_{2,1}\ \sigma_{2,2}\ \sigma_{2,3}\ ]'$, we test the volatility smile for

both sets of maturities by forming the restriction matrix

$$R_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix},$$

and constructing the test statistic $(R_1\hat{\theta})'[R_1\widehat{V_{GMM}}R_1'/T]^{-1}(R_1\hat{\theta})$ where $\widehat{V_{GMM}}$ is the estimated asymptotic VCV of $\hat{\theta}$ (hence our division by T). The test statistic provides a Wald test asymptotically chi-squared with four degrees of freedom.

10. See discussion of multiplicative tolerance factors in Press *et al.* (1996, p398, p410).

C References

Bakshi, Gurdip, Charles Cao, and Zhiwu Chen, "Option pricing and hedging performance under stochastic volatility and stochastic interest rates," *Journal of Finance* 52 (December 1997), 2003-2049.

Berndt, Ernst K., Bronwyn Hall, Robert Hall, and Jerry A. Hausman, "Estimation and Inference in Nonlinear Structural Models," *Annals of Economic and Social Measurement* Vol 3 No 4 (October 1974), 653-665.

Black, F. and M. Scholes, "The pricing of options and corporate liabilities," *Journal of Political Economy* Vol 81 No 3 (May/June 1973), 637-659.

Evans, Merran, Nicholas Hastings, and Brian Peacock, *Statistical Distributions*, (New York: John Wiley and Sons, 1993).

Greene, William H., *Econometric Analysis*, Third Edition, (Upper Saddle River: Prentice Hall, 1997).

Hamilton, James D., *Time Series Analysis*, (Princeton: Princeton University Press, 1994).

Hansen, L, "Large sample properties of generalized method of moments estimators," *Econometrica* 50 (July 1982), 1029-1054.

Newey, W., and K. West, "A simple positive semi-definite heteroscedasticity and autocorrelation consistent covariance matrix," *Econometrica* 55 (May 1987), 703-708.

Ogaki, Masao, Generalized method of moments: Econometric applications, Ch 17 in *Handbook of Statistics*, Vol 11, (1993), Ed. Maddala, Rao and Vinod.

Press, William H., Saul Teukolsky, William T. Vetterling, and Brian P. Flannery, *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition, (Cambridge: Cambridge University Press, 1996).

White, H., "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity," *Econometrica* 48 (May 1980), 817-838.

D Tables

Parameter	Simulation Results			
T	100	1,000	10,000	100,000
$\hat{\nu}^{(1)}$	6.030	9.252	9.475	10.491
$\hat{\nu}^{(2)}$	9.164	10.408	10.561	10.392
$\hat{\nu}_{GMM}$	14.899	11.337	10.439	10.481
SE_{WHITE}	9.422	1.745	0.620	0.240
SE_{NW}	9.146	1.742	0.621	0.239
χ^2_1	3.075	1.184	**7.259**	0.142
p-value	0.080	0.277	0.007	0.706

Table 1: Simulated Student-t Data.
For sample size T we simulate IID Student-t data with $\nu = 10$ degrees of freedom. The MOM estimators $\hat{\nu}^{(1)}$, and $\hat{\nu}^{(2)}$ are reported along with the GMM estimator $\hat{\nu}_{GMM}$. Both White and Newey-West standard errors are reported along with the chi-squared goodness of fit test (critical 5% value 3.842).

Why we Prefer Exactly- to Over-Identified Estimation
• Exactly-identified estimation yields more information (e.g. a $\sigma_{i,j}$ for each moneyness/maturity class).
• Exactly-identified estimation allows tests of many different specific hypotheses (e.g. $H_0: \ \sigma_{i,1} = \sigma_{i,2} = \sigma_{i,3}$).
• Over-identified estimation produces parameters that are not necessarily economically meaningful if model rejects, and this goes hand-in-hand with the existence of multiple local minima in the objective function.
• Exactly-identified estimation converges very quickly using any hill climber. Over-identified estimation converges very slowly with all hill climbers.
• Exactly-identified estimation can be run with $W = I$ because solution is independent of weighting matrix. You save time by not re-computing W via Newey-West at each iteration. Also decreases complexity of code.
• You know if solution to exactly-identified estimation is global minimum because you know *a priori* that objective function is zero at global optimum – though uniqueness is not guaranteed.
• You cannot game the exactly-identified estimation as you can the over-identified estimation.

Table 2: Why we Prefer Exactly- to Over-Identified Estimation
This is a summary of why we prefer exactly- to over-identified estimation in an option pricing context. In the table "W" is the GMM weighting matrix. The null hypothesis $H_0: \ \sigma_{i,1} = \sigma_{i,2} = \sigma_{i,3}$ states that the volatility smile is flat. That is, $\sigma_{i,j}$ is constant across moneyness classes j for each maturity class i. Two strikes against exactly-identified estimation is that it eats degrees of freedom and that over-identifying may increase power by optimally combining multiple moments to estimate a parameter. The former is not a big problem because GMM uses asymptotic results so you need plenty of data anyway.

Hints for Handling Local Extrema
• Use many starting points
• Upon convergence restart routines like DFP or BFGS that start with $W = I$ and update W.
• Allow for unorthodox step lengths that look far beyond nearby extrema during line searches.
• We prefer a simple Newton routine with unorthodox step lengths (slow but sure) to a canned higher-tech BFGS (which is fast but finds local extrema).

Table 3: Advice for Handling Local Extrema in Numerical Optimization of Objective Functions
In the table "DFP" is the Davidon-Fletcher-Powell hill climber and "BFGS" is the Broyden-Fletcher-Goldfarb-Shanno hill climber (Press (1996, pp425-428)).

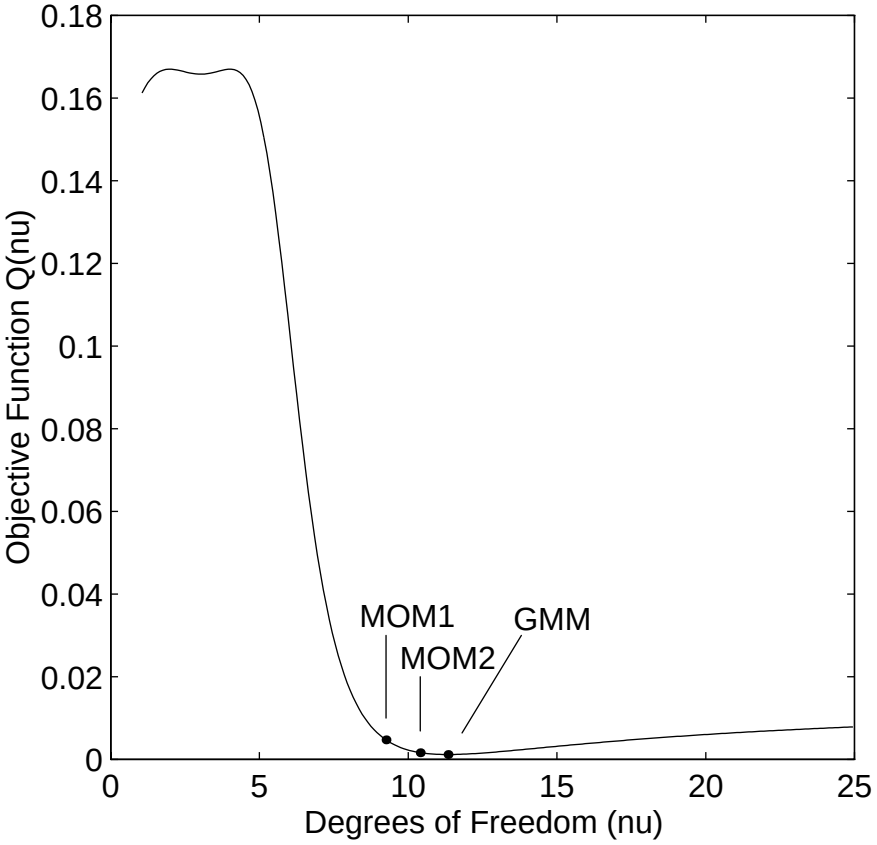

Figure 1: Objective Function for Simulated Student-t data.
For 1,000 independent drawings from a Student-t with $\nu = 10$ degrees of freedom, we calculate the objective function $Q(\hat{\nu}) = g(\hat{\nu})' W g(\hat{\nu})$ where $W = \hat{\Omega}^{-1}$ is the estimated asymptotic VCV of $g(\nu)$ (estimated using the Newey-West technique described later in this paper). Our original two MOM estimators are $\hat{\nu}^{(1)} = 9.252$ and $\hat{\nu}^{(2)} = 10.408$ (labelled "MOM1," and "MOM2"). The objective function is minimized at the GMM estimator $\hat{\nu} = 11.337$ (labelled "GMM").