Fitting and testing vast dimensional time-varying covariance models

Neil Shephard

Oxford-Man Institute
and
Department of Economics
University of Oxford

Joint with
Robert F. Engle
Stern School of Business, NYU
and
Kevin Sheppard
Economics and OMI, Oxford

April 17, 2008

The Problem

- r_t K-dimensional daily returns, $\{\mathcal{F}_t\}$ natural filtration.
- Recent attention towards estimating conditional covariance models

$$\mathrm{E}(r_t|\mathcal{F}_{t-1}) = 0, \quad \mathrm{Cov}(r_t|\mathcal{F}_{t-1}) = H_t,$$

  based on

$$r_1, r_2, ..., r_T.$$

  - Engle (2002); Tse & Tsui (2002); Ledoit, Santa-Clara & Wolf (2003); Cappiello, Engle & Sheppard (2006); Engle & Kelly (2007).
  - H_t is a function of $\mathcal{F}_{t-1}$ through parameters ψ.
- Desire to estimate key dynamic parameters when K is very large. Unbalanced panels.
- $K = 500$?

Relevant Models

- Covariance tracking and scalar dynamics

$$H_t = (1-\alpha-\beta)\,\Sigma + \alpha r_{t-1}r_{t-1}' + \beta H_{t-1}, \quad \alpha,\beta \geq 0, \quad \alpha+\beta<1,$$

  - Special case of Bollerslev, E. & Wooldridge (88) or E. & Kroner(95)
- EWMA:

$$H_t = \alpha r_{t-1}r_{t-1}' + (1-\alpha)\,H_{t-1}, \quad \alpha \in [0,1)$$

  - A simple case of this is RiskMetrics.

Standard Estimation

- Usual assumption

$$\mathrm{E}(r_t|\mathcal{F}_{t-1}) = 0, \quad \mathrm{Cov}(r_t|\mathcal{F}_{t-1}) = H_t,$$

- Usually estimated via Gaussian quasi-likelihood

$$\log L_Q(\psi; r) = \sum_{t=1}^{T} -\frac{1}{2}\log|H_t| - \frac{1}{2} r_t{}' H_t^{-1} r_t$$

- Challanging:
  - the parameter space is typically large — statistical and computational problems;
  - the inversion of H_t takes $O(K^3)$ computations
- Often first can be "dealt with" by concentration.

Think of

$$H_t = (1-\alpha-\beta)\,\Sigma + \alpha r_{t-1}r_{t-1}' + \beta H_{t-1},$$

regard $\lambda = vech(\Sigma)$ as P-dim nuisance, $\theta = (\alpha, \beta)'$ as parameters of interest.

$$\log L_Q(\lambda, \theta; r).$$

Can use a moment estimator to estimate λ,

$$\widehat{\lambda} = vech\left(\frac{1}{T}\sum_{t=1}^{T} r_t r_t'\right).$$

Yields a m-profile likelihood (2-stage estimator)

$$\log L_Q(\widehat{\lambda}, \theta; r).$$

Vast dimensional nuisance parameter (e.g. $K = 100$, over $5,000$)

Using the all S&P 100 stocks, January 2, 1997 - December 31 2006, we quick look at the scaling bias. The first asset is always the market and the other assets are arranged alphabetically by ticker.
The model fit was a scalar *BEKK* using covariance tracking,

$$H_t = (1-\alpha-\beta)\Sigma + \alpha r_{t-1}' r_{t-1} + \beta H_{t-1} \tag{1}$$

	S&P Returns				
	Scalar BEKK		EWMA	DCC	
K	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\alpha}$	$\tilde{\beta}$
5	.0189	.9794	.0134	.0141	.9757
10	.0125	.9865	.0103	.0063	.9895
25	.0081	.9909	.0067	.0036	.9887
50	.0056	.9926	.0045	.0022	.9867
96	.0041	.9932	.0033	.0017	.9711

Table: Parameter estimates from a covariance targeting scalar BEKK, EWMA (estimating H_0) and DCC using maximum m-profile likelihood (MMLE). Based upon a real database built from daily returns from 95 companies plus the index from the S&P100, from 1997 until 2006.

Data array

Move the return vector r_t into a data array $Y_t = \{Y_{1t}, ..., Y_{Nt}\}$ where Y_{jt} is itself a vector containing small subsets of the data (there is no need for the Y_{jt} to have common dimensions).
In our context a leading example would be where we look at all the unique "pairs" of data

$$\begin{aligned} Y_{1t} &= (r_{1t}, r_{2t})', \\ Y_{2t} &= (r_{1t}, r_{3t})', \\ &\vdots \\ Y_{\frac{K(K-1)}{2}} &= (r_{K-1t}, r_{Kt}), \end{aligned}$$

writing $N = K(K-1)/2$. We will continue with this example, in the exposition below, but it is trivial to think about using other subsets of the data in a similiar way.

Let

$$Y_{jt} = S_j r_t, \quad S_j \text{ selection matrix.}$$

Our model trivially implies

$$\mathrm{E}(Y_{jt}|\mathcal{F}_{t-1}) = 0, \quad \mathrm{Cov}(Y_{jt}|\mathcal{F}_{t-1}) = H_{jt} = S_j H_t S_j'. \tag{2}$$

which determined the conditional mean and covariance of each submodel $Y_{jt}|\mathcal{F}_{t-1}$.

$$\log L_j(\psi) = \sum_{t=1}^{T} l_{jt}(\psi), \quad l_{jt}(\psi) = \log f(Y_{jt}; \psi)$$

where

$$l_{jt}(\psi) = -\frac{1}{2}\log|H_{jt}| - \frac{1}{2} Y_{jt}' H_{jt}^{-1} Y_{jt}.$$

$$l_{jt}(\psi) = -\frac{1}{2}\log|H_{jt}| - \frac{1}{2}Y'_{jt}H_{jt}^{-1}Y_{jt}.$$

This quasi-likelihood will have information about ψ but more information can be obtained by averaging the same operation on many submodels

$$c_t(\psi) = \frac{1}{N}\sum_{j=1}^{N}\log L_{jt}(\psi).$$

Of course if the $\{Y_{1t}, ..., Y_{Nt}\}$ were independent this would be the exact likelihood — but this will not be the case for us! Such functions, based on "submodels" or "marginal models", are call composite-likelihoods, following the nomenclature introduced by Lindsay (1988).

Computational points

Previously method was $O(K^3)$.

- Evaluation of $c_t(\psi)$ costs $O(N)$ calculations.
- All distinct pairs — $O(K^2)$ calculations.
- Contiguous pairs — $O(K)$ calculations.
- Choose only $O(1)$ pairs (randomly), which is computationally fast!

We will see in a moment that the efficiency loss of using these subsets compared to computing all possible pairs is extremely small when N is moderately large.
Asymptotically as N increases to infinity "all pairs" and "contiguous" have the same efficiency.
If K is large it is pointless using all pairs.

We now make our main assumption that

$$c_t(\psi) = \frac{1}{N}\sum_{j=1}^{N} \log L_{jt}(\theta, \lambda_j).$$

- Common finite dimensional θ and vector of parameters λ_j which is specific to the j-th subset.
- Our interest is in estimating θ and so the λ_j are nuisances.
- This type of assumption appeared first in the work of Neyman and Scott (1948) — but they had independence. Dependence over j will help us!
- Named a stratified model with a stratum of nuisance parameters and can be analysed by using two-index asymptotics, e.g. Barndorff-Nielsen (1996).

For the j-th submodel we have the common parameter θ and nuisance parameter λ_j. The joint model may imply there are links across the λ_j.

Example

The scalar BEKK model $H_t = (1-\alpha-\beta)\Sigma + \alpha r_{t-1}' r_{t-1} + \beta H_{t-1}$ so

$$Y_{1t} = (r_{1t}, r_{2t})', \quad Y_{2t} = (r_{2t}, r_{3t})',$$

then

$$\lambda_1 = (\Sigma_{11}, \Sigma_{21}, \Sigma_{22})', \quad \lambda_2 = (\Sigma_{22}, \Sigma_{32}, \Sigma_{33})'.$$

Hence, the joint model implies there are common elements across the λ_j.

We may potentially gain by exploiting these links in our estimation. An alternative, is to be **self-denying** and never use these links even if they exist in the data generating process. The latter means the admissible values are

$$(\lambda_1, \lambda_2, ..., \lambda_N) \in \Lambda_1 \times \Lambda_2 \times ... \times \Lambda_N, \tag{3}$$

i.e. they are variation-free.
Throughout we use variation-freeness.

Our estimation strategy can be generically stated as solving

$$\widehat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} \log L_{jt}(\widehat{\theta}, \widehat{\lambda}_j),$$

where $\widehat{\lambda}_j$ solves for each j

$$\sum_{t=1}^{T} g_{jt}(\widehat{\theta}, \widehat{\lambda}_j) = 0.$$

Here g_{jt} is a $\dim(\lambda_j)$-dimensional moment constraint so that for each j

$$\mathrm{E}\left\{g_{jt}(\theta, \lambda_j)\right\} = 0, \quad t = 1, 2, ..., T.$$

This structure has some important special cases.

e.g. Maximum composite likelihood estimator

The maximum composite likelihood estimator (MCLE) follows from writing

$$g_{jt}(\theta, \lambda_j) = \frac{\partial \log L_{jt}(\theta, \lambda_j)}{\partial \lambda_j},$$

so

$$\widehat{\lambda}_j(\theta) = \underset{\lambda_j}{\operatorname{argmax}} \sum_{t=1}^{T} \log L_{jt}(\theta, \widehat{\lambda}_j),$$

which means

$$\frac{1}{N} \sum_{t=1}^{T} \sum_{j=1}^{N} \log L_{jt}(\theta, \widehat{\lambda}_j)$$

is the profile composite likelihood which $\widehat{\theta}$ maximises.

e.g. Maximum m-profile composite-likelihood estimator

Suppose

$$g_{jt}(\theta, \lambda_j) = G_{jt} - \lambda_j, \quad \text{where} \quad \mathrm{E}(G_{jt}) = \lambda_j,$$

then

$$\widehat{\lambda}_j = \frac{1}{T} \sum_{t=1}^{T} G_{jt}.$$

We call the resulting $\widehat{\theta}$ a m-profile composite-likelihood estimator (MMCLE).

Behaviour — no nuisance parameters, no time series

The Cox and Reid (2003): suppose r_t is i.i.d. then we assume

$$\mathcal{I}^{*}_{\theta\theta} = \lim_{T\to\infty} \mathrm{Cov}\left(\frac{1}{N} \sum_{j=1}^{N} \frac{\partial l_{jt}(\theta, \lambda_j)}{\partial \theta} \right) > 0, \; -\mathrm{E}\left\{ \frac{1}{N} \sum_{j=1}^{N} \frac{\partial^2 l_{jt}(\theta, \lambda_j)}{\partial \theta \partial \theta'} \right\} \to \mathcal{J}_{\theta\theta}.$$

The **former assumption is the key** for us: average score does not exhibit a law of large numbers in the cross section. Then we have

$$\sqrt{T} \frac{1}{TN_T} \sum_{t=1}^{T} \sum_{j=1}^{N_T} \frac{\partial l_{jt}(\theta, \lambda_j)}{\partial \theta} \xrightarrow{d} N(0, \mathcal{I}^{*}_{\theta\theta}),$$

and so

$$\sqrt{T}\left(\widehat{\theta} - \theta \right) \xrightarrow{d} N(0, \mathcal{J}_{\theta\theta}^{-1} \mathcal{I}^{*}_{\theta\theta} \mathcal{J}_{\theta\theta}^{-1}).$$

Notice the rate of convergence is now $\sqrt{T}$, so we do not get an improved rate of convergence from the cross-sectional information.

Nuisance parameters: stack the moment constraints

$$\frac{1}{TN_T}\sum_{t=1}^{T}\begin{pmatrix} g_t \\ \sum_{j=1}^{N_T}\frac{\partial l_{jt}}{\partial \theta}\end{pmatrix}, \quad g=\{g_{jt}\}, \quad \widehat{\lambda}-\lambda=\left\{\widehat{\lambda}_j-\lambda_j\right\}.$$

$$\begin{pmatrix} \widehat{\lambda}-\lambda \\ \widehat{\theta}-\theta \end{pmatrix} \simeq \begin{pmatrix} A & c \\ b' & \mathcal{J}_{\theta\theta}\end{pmatrix}^{-1}\left\{\frac{1}{TN_T}\sum_{t=1}^{T}\begin{pmatrix} g_t \\ \sum_{j=1}^{N_T}\frac{\partial l_{jt}}{\partial \theta}\end{pmatrix}\right\},$$

$$\begin{aligned} A &= N^{-1}diag(\mathcal{J}_{\lambda_1\lambda_1},...,\mathcal{J}_{\lambda_N\lambda_N}),\ b=N^{-1}\left\{\mathcal{J}_{\theta\lambda_j}\right\},\ c=N^{-1}\left\{\mathcal{J}_{\lambda_j\theta}\right\}, \\ \mathcal{J}_{\lambda_j\lambda_j} &= -p\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial \lambda_j'}, \quad \mathcal{J}_{\lambda_j\theta}=-p\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial g_{jt}}{\partial\theta'}, \\ \mathcal{J}_{\theta\lambda_j} &= -p\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 l_{jt}}{\partial\theta\partial\lambda_j'}, \quad \mathcal{J}_{\theta\theta}=-\left(p\lim_{T\to\infty}\frac{1}{TN_T}\sum_{t=1}^{T}\sum_{j=1}^{N_T}\frac{\partial^2 l_{jt}}{\partial\theta\partial\theta'}\right). \end{aligned}$$

Then

$$\widehat{\theta}\simeq\theta+\mathcal{D}_{\theta\theta}^{-1}\frac{1}{T}\sum_{t=1}^{T}Z_{t,T}, \quad \mathcal{D}_{\theta\theta}=\lim_{N_T\to\infty}\frac{1}{N_T}\sum_{j=1}^{N_T}\left(\mathcal{J}_{\theta\theta}-\mathcal{J}_{\theta\lambda_j}\mathcal{J}_{\lambda_j\lambda_j}^{-1}\mathcal{J}_{\lambda_j\theta}\right),$$

where

$$Z_{t,T}=\frac{1}{N}\sum_{j=1}^{N}\left(\frac{\partial l_{jt}(\theta,\lambda_j)}{\partial\theta}-\mathcal{J}_{\theta\lambda j}\mathcal{J}_{\lambda_j\lambda_j}^{-1}g_{jt}\right).$$

We assume as $T\to\infty$

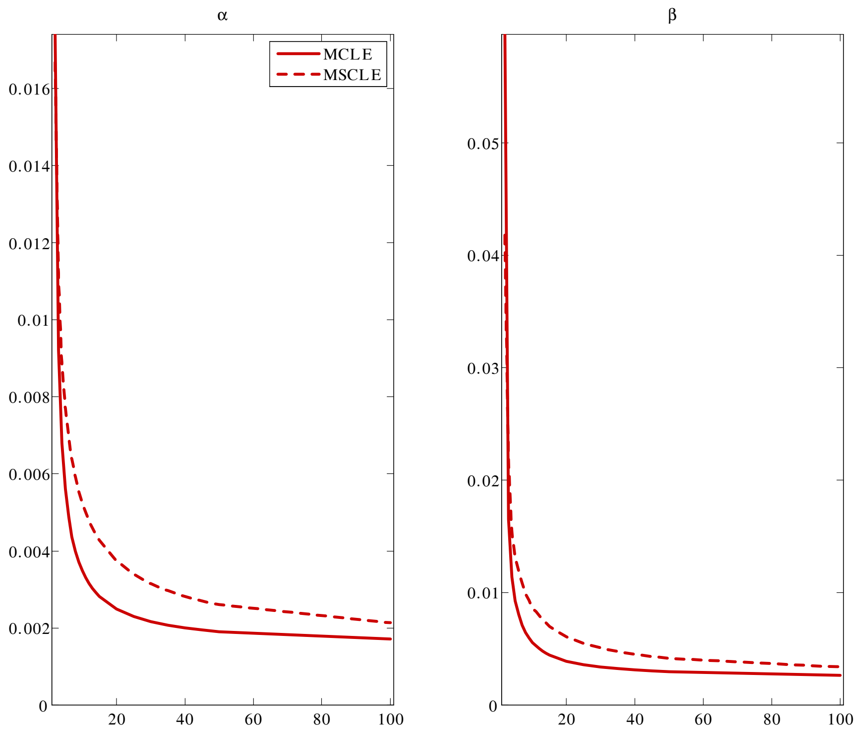
$$\mathrm{Cov}\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T}Z_{t,T}\right)\to\mathcal{I}_{\theta\theta},$$

where $\mathcal{I}_{\theta\theta}$ has diagonal elements which are bounded from above and $\mathcal{I}_{\theta\theta}>0$ (estimate by low dimensional HAC estimator!). Then

$$\sqrt{T}\left(\widehat{\theta}-\theta\right)\to N(0,\mathcal{D}_{\theta\theta}^{-1}\mathcal{I}_{\theta\theta}\mathcal{D}_{\theta\theta}^{-1}).$$

Monte Carlo

Plot: s.e. of estimator against K for the maximized MCLE and MSCLE. e.g. $K = 50$, MCLE is based on $1,225$ submodels while MSCLE uses 49.



Empirical Application

- S&P 100 components
- January 2, 1997 - December 31 2006
  - 2516 daily observations
- Also include the S&P 100 index
- Asset had to be continually available for including
  - MCLE is well suited to the case where assets are added or dropped
- 97 assets in total, incl. the index

Same model, Different Estimates

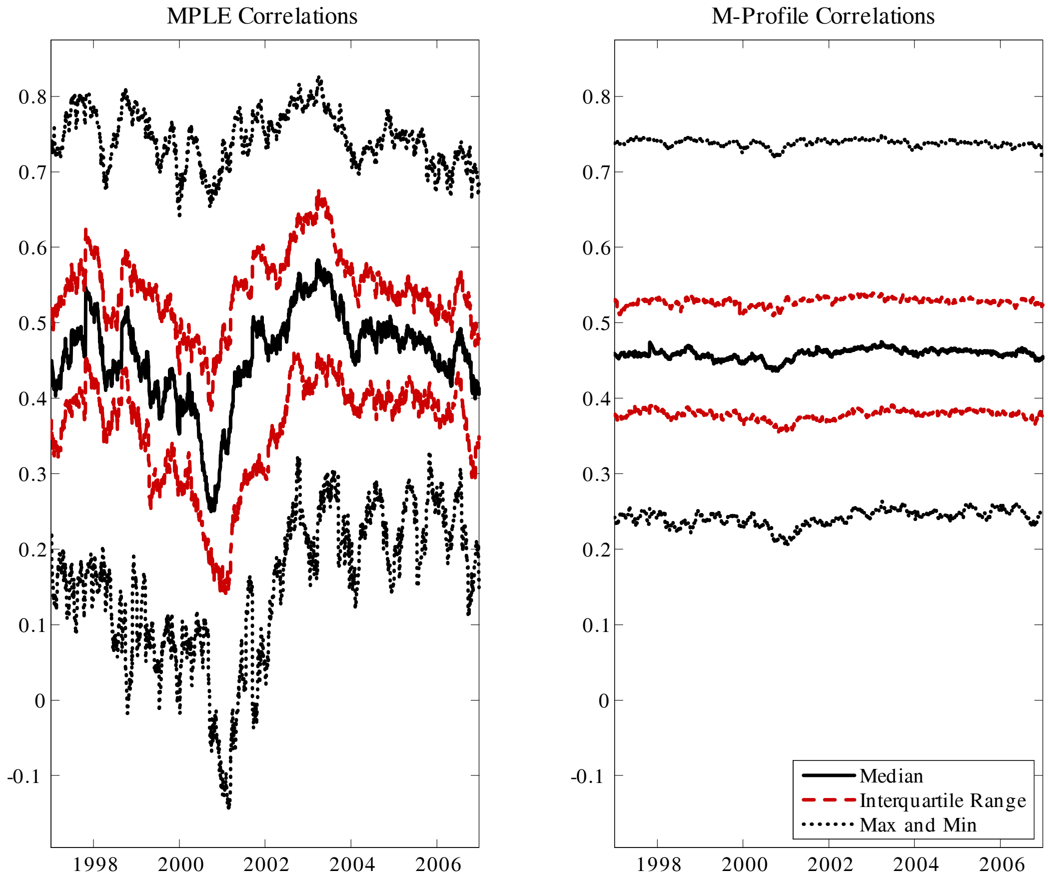
	m-profile					maximised			
	Scalar BEKK		EWMA	DCC		Scalar BEKK		DCC	
K	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\alpha}$	$\tilde{\beta}$
All Pairs									
5	.0287	.9692	.0205	.0143	.9829	.0288	.9692	.0116	.9873
	(.0081)	(.0092)	(.0037)	(.0487)	(.0846)	(.0073)	(.0082)	(.0048)	(.0056)
10	.0281	.9699	.0211	.0107	.9881	.0276	.9705	.0107	.9875
	(.0055)	(.0063)	(.0027)	(.0012)	(.0016)	(.0050)	(.0057)	(.0013)	(.0021)
25	.0308	.9667	.0234	.0100	.9871	.0327	.9646	.0102	.9866
	(.0047)	(.0055)	(.0023)	(.0009)	(.0017)	(.0043)	(.0047)	(.0010)	(.0021)
50	.0319	.9645	.0225	.0101	.9856	.0345	.9615	.0104	.9848
	(.0046)	(.0056)	(.0026)	(.0008)	(.0018)	(.0037)	(.0042)	(.0009)	(.0017)
96	.0334	.9636	.0249	.0103	.9846	.0361	.9601	.0106	.9841
	(.0041)	(.0049)	(.0019)	(.0009)	(.0019)	(.0031)	(.0034)	(.0009)	(.0018)
Contiguous Pairs									
5	.0284	.9696	.0189	.0099	.9885	.0251	.9733	.0078	.9917
	(.0083)	(.0094)	(.0037)	(.0033)	(.0045)	(.0070)	(.0079)	(.0055)	(.0059)
10	.0272	.9709	.0201	.0093	.9886	.0266	.9717	.0088	.9900
	(.0054)	(.0062)	(.0027)	(.0016)	(.0018)	(.0049)	(.0055)	(.0018)	(.0020)
25	.0307	.9668	.0227	.0089	.9889	.0315	.9660	.0088	.9894
	(.0049)	(.0056)	(.0024)	(.0011)	(.0012)	(.0044)	(.0050)	(.0012)	(.0013)
50	.0316	.9647	.0220	.0092	.9869	.0347	.9612	.0095	.9864
	(.0047)	(.0057)	(.0029)	(.0010)	(.0019)	(.0038)	(.0043)	(.0011)	(.0019)
96	.0335	.9634	.0247	.0094	.9860	.0364	.9598	.0095	.9863
	(.0043)	(.0051)	(.0020)	(.0009)	(.0014)	(.0032)	(.0035)	(.0009)	(.0012)

Visualizing the Differences

- Do these parameter values make any qualitative difference?
  - Yes!
- Construct a plot based on Quasi-βs
  - Correlation of standardized return on asset j with the standardized return on the market
- Still 95 series
  - Median
  - Interquartile range
  - 95% interval

One model, DCC, two estimators

Correlation of returns with the market.



Testing the Differences

- Do these differences matter for application?
  - Yes!
- High dimension parameter space rules out in-sample testing
- composite-out-of sample experiment from January 2, 2003 - December 31, 2006
- All parameters estimated using data January 2, 1997 - December 31, 2002
  - Dynamic Correlation parameters largely similar to full sample
  - QMLE estimate somewhat less persistent

- Examined the hedging errors of a conditional CAPM where the S&P 100 index proxied for the market. Using one-step ahead forecasts, the conditional time-varying market betas were computed as

$$\widehat{\beta}_{j,t} = \frac{\widehat{h}_{j,t}^{1/2}\widehat{\rho}_{jm,t}}{\widehat{h}_{m,t}^{1/2}}, \quad j = 1, 2, \ldots, N, \tag{4}$$

and the corresponding hedging errors were computed as

$$\widehat{\nu}_{j,t} = r_{j,t} - \widehat{\beta}_{j,t} r_{m,t}. \tag{5}$$

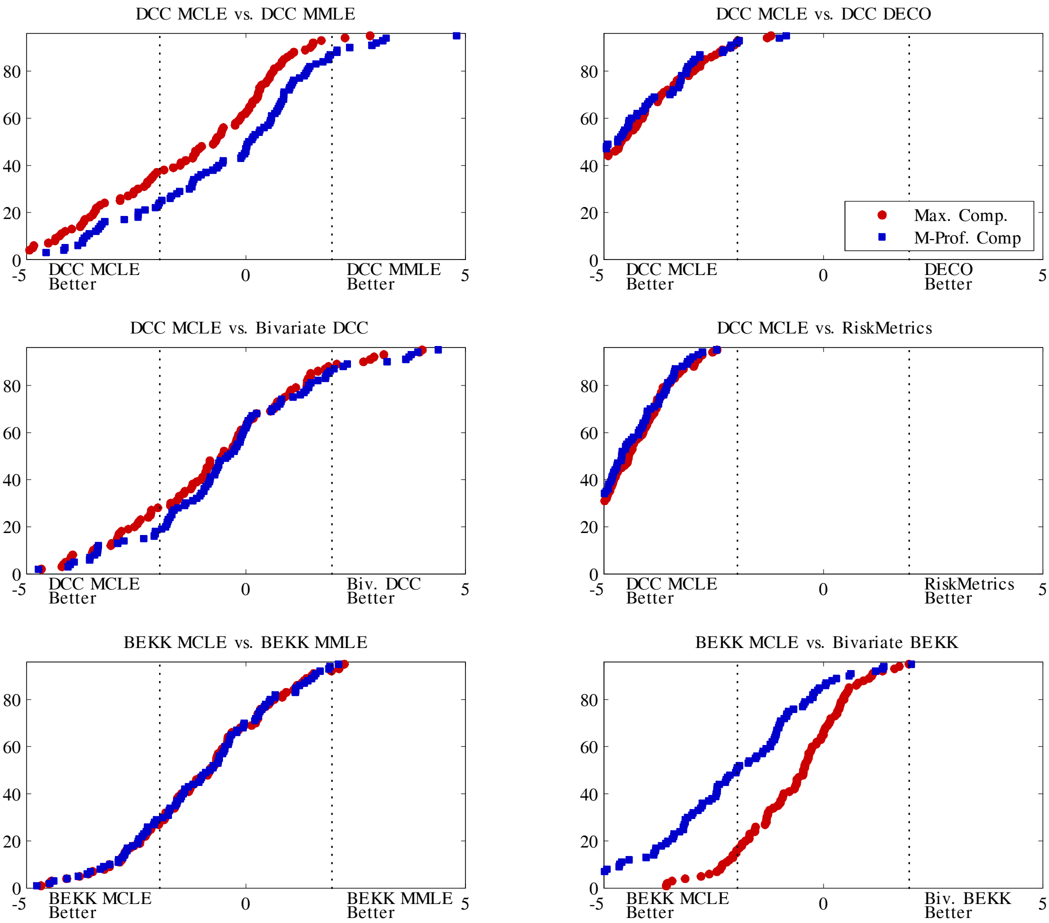
Testing for Superior Predictive Ability

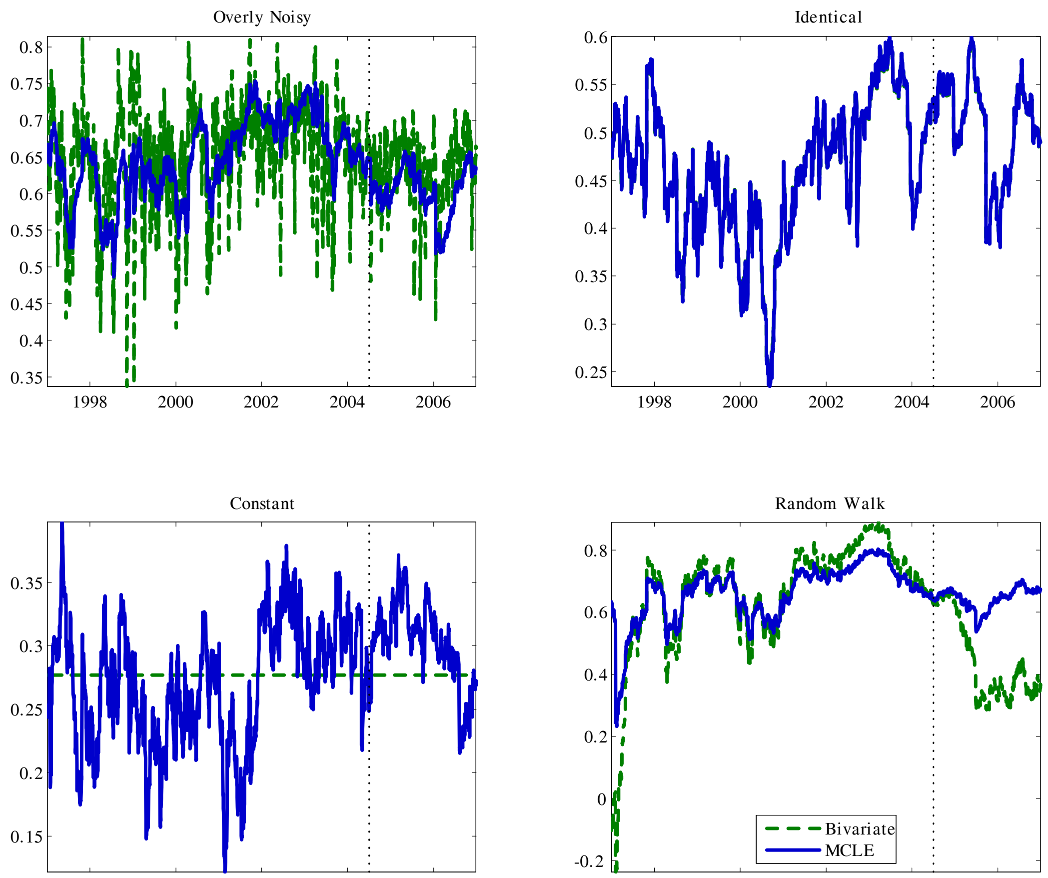
- Comparisons via Giacomini-White(06) tests

$$\widehat{\delta}_{j,t} = \left(\widehat{\nu}_{j,t}\left(\widehat{\rho}_{j,t}^{MCLE}\right)\right)^2 - \left(\widehat{\nu}_{j,t}\left(\widehat{\rho}_{j,t}^{MMLE}\right)\right)^2$$

- Test statistic is

$$\frac{\bar{\delta}_j}{\text{avar}\left(\sqrt{T}\bar{\delta}_j\right)}$$


DCC MCLE vs. DCC MMLE
DCC MCLE Better
DCC MMLE Better
DCC MCLE vs. DCC DECO
Max. Comp.
M-Prof. Comp
DCC MCLE Better
DECO Better
DCC MCLE vs. Bivariate DCC
DCC MCLE Better
Biv. DCC Better
DCC MCLE vs. RiskMetrics
DCC MCLE Better
RiskMetrics Better
BEKK MCLE vs. BEKK MMLE
BEKK MCLE Better
BEKK MMLE Better
BEKK MCLE vs. Bivariate BEKK
BEKK MCLE Better
Biv. BEKK Better


Overly Noisy
Identical
Constant
Random Walk
Bivariate
MCLE

Conclusions I

- Paper proposes a new estimator for time varying-covariance models
- Can provide moderate to large improvements in computation time
  - Or equivalently increases in feasible cross-section sizes
- Estimator is *more* accurate in large models
- Composite structure looks similar to Neyman-Scott problem, but has some differences which are key.
- Relatively easy to carry out statistical inference on these models
- Same problems arise when we estimate copulas!

Conclusions II

- We love composite likelihoods!