Efficient Iterative ML Estimation

Nikolaus Hautsch
Ostap Okhrin
Alexander Ristig

University of Vienna
Dresden University of Technology
C.A.S.E. – Center for Applied Statistics and Economics
Humboldt–Universität zu Berlin
http://isor.univie.ac.at
http://tu-dresden.de
http://lvb.wiwi.hu-berlin.de

Vector autoregressive model

Application: Impulse response analysis.

Example 1

Let X_i denote a $(d \times 1)$ vector of random variables, $i = 1, \ldots, n$.

$$X_i = \underbrace{\omega}_{(d\times 1)} + \underbrace{A}_{(d\times d)} X_{i-1} + \varepsilon_i,$$

is known as VAR(1). Efficient estimation is based on $\varepsilon_i \sim \mathrm{N}(0, \Sigma_\varepsilon)$.

Parameter vector $\vartheta = \{\omega, \mathrm{vec}(A), \mathrm{diag}(\Sigma_\varepsilon), \mathrm{vech}(\Sigma_\varepsilon)\}$.

Dynamic conditional correlation model

Application: Value at Risk estimation.

Example 2

Let X_i denote a $(d \times 1)$ vector of returns, $i = 1, \ldots, n$.

$$
\begin{aligned}
& X_i = \mathrm{D}_i \, \varepsilon_i \quad \text{with} \quad \varepsilon_i | \mathcal{F}_{i-1} \sim \mathrm{N}(0, \mathrm{R}_i), \\
\text{with} \quad & \mathrm{R}_i = \mathrm{diag}(\mathrm{Q}_i)^{-1} \, \mathrm{Q}_i \, \mathrm{diag}(\mathrm{Q}_i)^{-1}, \\
& \mathrm{Q}_i = S \odot (1_d 1_d^\top - A - B) + A \odot \varepsilon_{i-1} \varepsilon_{i-1}^\top + B \odot \mathrm{Q}_{i-1}, \\
\text{and} \quad & \mathrm{D}_i^2 = \Omega + K \odot X_{i-1} X_{i-1}^\top + \Lambda \odot \mathrm{D}_{i-1}^2,
\end{aligned}
$$

is known as DCC-model, with $S = n^{-1} \sum_{i=1}^{n} \varepsilon_i \varepsilon_i^\top$.

Parameter vector $\vartheta = \{\mathrm{diag}(K), \mathrm{diag}(\Lambda), \mathrm{vec}(A), \mathrm{vec}(B), \mathrm{diag}(\Omega)\}$.

Multivariate probit model

Applications: Health-care and unemployment analysis.
Example 3

The multivariate probit model has the data generating process

$$Y_{ij} = \mathbf{I}\left\{\varepsilon_{ij} \le \boldsymbol{\beta}_j^\top Z_{ij}\right\}, \quad \text{for} \quad i = 1, \ldots, n, \quad \text{and} \quad j = 1, \ldots, d,$$

where Z_{ij} is a r_j-dimensional vector of covariates including intercept and $(\varepsilon_{i1}, \ldots, \varepsilon_{id})^\top \sim \mathrm{N}(0, \mathrm{R})$ with $\mathrm{diag}(\mathrm{R}) = 1$ for identification.

Parameter vector $\vartheta = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d, \mathrm{vech}(\mathrm{R})\}$.

Stochastic volatility model

Applications: Option pricing.
Example 4

Let X_i denote a $(d \times 1)$ vector of returns, $i = 1, \ldots, n$. The standard stochastic volatility model is

$$X_i = \exp(\sigma_i/2)\varepsilon_i$$
$$\sigma_i = \alpha + \beta\sigma_{i-1} + \gamma\eta_i,$$

where $\varepsilon_i \overset{\text{iid}}{\sim} H(\varepsilon_1, \ldots, \varepsilon_d; \theta)$ denote idiosyncratic shocks, σ_i is the latent log-volatility and $\eta_i \overset{\text{iid}}{\sim} \mathrm{N}(0, I)$.

Parameter vector $\vartheta = \{\theta, \alpha, \beta, \gamma\}$.

Related to practitioners

- Asset and option pricing
- Estimation of VaR and ES
- Forecasting of macroeconomic variables
- Discrete choice models
- ...
- Volatility contagion via connectedness measures

Challenges

- log-likelihood is often complicated in *non*-linear models especially if number of parameters is large.
  - Large-dimensional times series models, see Engle (2002, JBES).

$$\ell(\vartheta_1, \vartheta_2) = -\frac{1}{2}\sum_{i=1}^{n}\Big[d\log(2\pi) + \log\{|\,\mathrm{D}_i(\vartheta_1)\,\mathrm{R}_i(\vartheta_2)\,\mathrm{D}_i(\vartheta_1)|\} + X_i^\top \mathrm{D}_i(\vartheta_1)^{-1}\,\mathrm{R}_i(\vartheta_2)^{-1}\,\mathrm{D}_i(\vartheta_1)^{-1} X_i\Big]$$

where $\vartheta_1 = \mathrm{vec}(A, B)$, $\vartheta_2 = \{\mathrm{diag}(\Omega)^\top, \mathrm{diag}(K)^\top, \mathrm{diag}(\Lambda)^\top\}^\top$

Challenges

- log-likelihood is often complicated in *non*-linear models especially if number of parameters is large.
  - Large-dimensional times series models, see Engle (2002, JBES).
  - High-dimensional copulae, see Aas et al. (2009, IMaE) and Okhrin et al. (2013, JoE).
- Derivatives (numerical) of the entire log-likelihood are not available (unstable) or difficult to derive.

Classical optimization techniques

- Simulated annealing, genetic algorithm, downhill simplex
  - Robust, non-differentiable functions, ...
  - Slow convergence, few parameters, ...
- Conjugate-gradient
  - Low memory-footprint, large number of parameters, ...
  - Slow convergence, first derivatives, ...
- Newton and quasi-Newton methods
  - Fast convergence, ...
  - First and second derivatives, ...

Proposed solution

- Iterative maximization of the log-likelihood.
- Gauß-Seidel scheme for non-linear equation.
- Decomposition of the parameter space in order to update the estimator.

- Alternatives inappropriate for “large p”.

Outline

1. Motivation ✓
2. Efficient estimation
3. Simulation I
4. Practical issues
5. Simulation II
6. Applications
7. Empirical illustration
8. Summary

An iterative estimation procedure

- Let $X = (X_1^\top, \ldots, X_n^\top)^\top$ be the finite history of the d-dimensional stochastic process $\{X_i\}_{i=1,2,\ldots}$.
- log-likelihood contribution of X_i

$$\ell_i(\vartheta_1, \ldots, \vartheta_G) \quad = \quad \log f_{X_i|\mathcal{F}_{i-1}}(X_{i1}, \ldots, X_{id}; \vartheta),$$

  where $\vartheta = (\vartheta_1^\top, \ldots, \vartheta_G^\top)^\top$.
- Build $\ell(\vartheta) = \ell(\vartheta_1, \ldots, \vartheta_G) = \sum_{i=1}^n \ell_i(\vartheta_1, \ldots, \vartheta_G)$ and use shorthand notation, e.g.,

$$\dot{\ell}(\vartheta_0) = \left.\frac{\partial \ell(\vartheta)}{\partial \vartheta}\right|_{\vartheta=\vartheta_0}.$$

Assumptions on next slide!

Algorithm

$h = 1: \quad \vartheta_n^1 \in \Theta$

$h > 1$:

(1) $\vartheta_{1,n}^h = \arg\max_{\vartheta_1} \ell(\vartheta_1, \vartheta_{2,n}^{h-1}, \ldots, \vartheta_{G,n}^{h-1})$

(2) $\vartheta_{2,n}^h = \arg\max_{\vartheta_2} \ell(\vartheta_{1,n}^h, \vartheta_2, \vartheta_{3,n}^{h-1}, \ldots, \vartheta_{G,n}^{h-1})$

$\vdots$

(G) $\vartheta_{G,n}^h = \arg\max_{\vartheta_G} \ell(\vartheta_{1,n}^h, \ldots, \vartheta_{G-1,n}^h, \vartheta_G)$

Assumptions

(1) Model is identifiable and correctly specified; parameter space Θ is compact, $\vartheta_0 \in \Theta$ and information equality holds.

(2) Asymptotic information matrix and negative Hessian are positive definite.

(3) Starting value is $n^{1/2}$-consistent.

(4) Score converges to a multivariate normal distribution. Appendix

Triangular structure

- Decompose the Hessian $\mathcal{H}(\cdot)$ into $\mathcal{D}(\cdot)$, $\mathcal{L}(\cdot)$ and $\mathcal{U}(\cdot)$, such that $\mathcal{H}(\vartheta) = \mathcal{D}(\vartheta) + \mathcal{L}(\vartheta) + \mathcal{U}(\vartheta)$. Assumptions
- Spectral radius of iteration matrix $\Gamma(\vartheta) = \{-\mathcal{D}(\vartheta) - \mathcal{L}(\vartheta)\}^{-1}\mathcal{U}(\vartheta)$ is strictly smaller than one, i.e., $\rho\{\Gamma(\vartheta)\} < 1$, see Reich (1949) and Ostrowski (1954).
- $\Gamma(\vartheta)$ is a convergent matrix: $\lim_{h\to\infty} \Gamma(\vartheta)^h = 0$.

Asymptotic properties

Theorem

Let the random vectors of the sequence X have an identical conditional density $f_i(\cdot;\vartheta)$ for which Assumptions 1-4 hold. Then,

$$n^{1/2}(\vartheta_n^h - \vartheta_0) \xrightarrow{\mathcal{L}} \mathrm{N}\left\{0, \mathcal{B}_h(\vartheta_0)\mathcal{M}(\vartheta_0)\mathcal{B}_h(\vartheta_0)^\top\right\},$$

$$\mathcal{B}_h(\vartheta) = \left[\Gamma(\vartheta)^{h-1}\left\{-\mathcal{H}^1(\vartheta)\right\}^{-1}, \{-\mathcal{H}(\vartheta)\}^{-1} - \Gamma(\vartheta)^{h-1}\{-\mathcal{H}(\vartheta)\}^{-1}\right].$$

Consistency

Convergence

- $\lim_{n\to\infty} \mathrm{Var}(n^{1/2}\vartheta_n^h)$ iteratively decreases as $h \to \infty$.
- Convergence of ϑ_n^h to the ML estimator ϑ_n as $h \to \infty$.

Theorem

Let the random vectors of the sequence X have an identical conditional density $f_i(\cdot;\vartheta)$ for which Assumptions 1-4 hold. Then,

$$h \geq 1 + \left\lceil \frac{\log(n^{1/2}\epsilon)}{\log\{\rho(\Gamma_n)\}} \right\rceil \quad \textit{with} \quad n^{1/2}\epsilon \in (0,1).$$

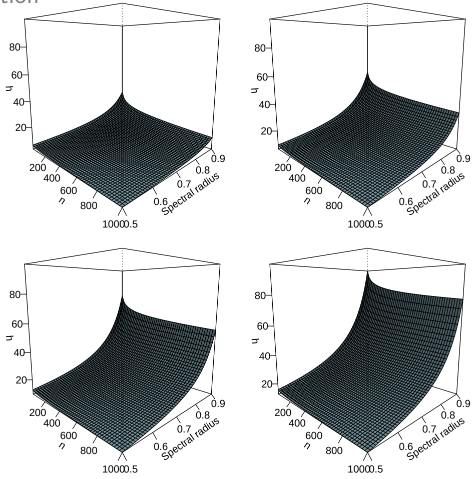


Figure 1: Approximate h until convergence for pre specified precision $\epsilon \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$, (u. left, u. right, l. left, l. right), sample size n and spectral radius $\rho(\Gamma_n)$.

Setup I

Similar to Kascha (2012, Econometric Reviews):

$$X_i = A\,X_{i-1} + \varepsilon_i + B\,\varepsilon_{i-1}.$$

- $d = 5$, $n = 100$, $r = 17$ and $\varepsilon_i \sim \mathrm{N}(0, \Sigma)$.
- Consistent & inconsistent starting values.
- Replication: 5000.
- 20 decomposition, e.g., $\vartheta_1 = \mathrm{vec}(A)$, $\vartheta_2 = \mathrm{vec}(B)$, $\vartheta_3 = \mathrm{vech}(\Sigma)$.

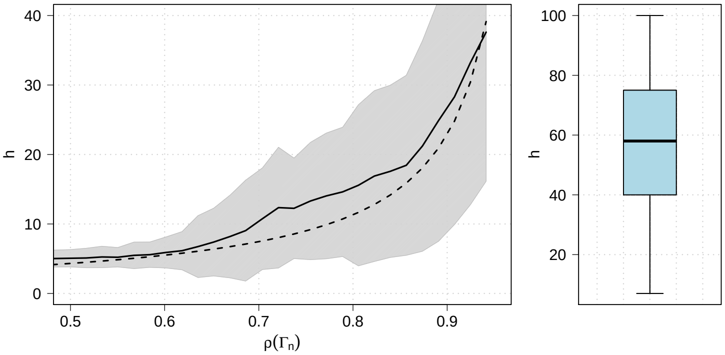


Figure 2: Based on *consistent estimates as starting values*, graphic shows the average number of iterations h until $\|\vartheta_n^h - \vartheta_n\|_1 \leq 0.1$. Gray area refers to the empirical sd of h. Boxplot shows the average number of iterations until $\ell(\vartheta_n^h) = \ell(\vartheta_n^{h+1})$, if $\|\vartheta_n^h - \vartheta_n\|_1 > 0.1$.

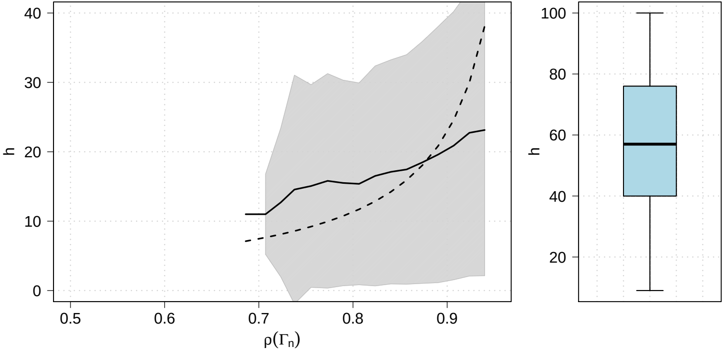


Figure 3: Based on *inconsistent estimates as starting values*, graphic shows the average number of iterations h until $\|\vartheta_n^h - \vartheta_n\|_1 \leq 0.1$. Gray area refers to the empirical sd of h. Boxplot shows the average number of iterations until $\ell(\vartheta_n^h) = \ell(\vartheta_n^{h+1})$, if $\|\vartheta_n^h - \vartheta_n\|_1 > 0.1$.

Boosting convergence

- Increasing n helps merely marginally to speed up the algorithm.
- Reduce $\rho(\Gamma_n)$ by
  - ruling out dependence among the estimators $\vartheta_{g,n}^h$.
  - simplifying the model.

Example 5

For $G = 2$ and $\mathcal{H}_{11}(\vartheta) = I_{r_1}$, estimator $\vartheta_{1,n}^h$ obeys the recursion

$$(\vartheta_{1,n}^h - \vartheta_1) = n^{-1}\dot{\ell}_{\vartheta_1}(\vartheta) + n^{-1}\ddot{\ell}_{\vartheta_1,\vartheta_2}(\vartheta_1,\vartheta_2)(\vartheta_{2,n}^{h-1} - \vartheta_2).$$

Assume a model simplification such that $\vartheta_{1,n}^1 = 0$.

Algorithm

Iteration $h > 1$*:*

(1) *{blank step}*

(2) $\vartheta_{2,n}^h = \arg\max_{\vartheta_2} \ell(0, \vartheta_2, \vartheta_{3,n}^{h-1}, \ldots, \vartheta_{G,n}^{h-1})$

$\vdots$

(G) $\vartheta_{G,n}^h = \arg\max_{\vartheta_G} \ell(0, \vartheta_{2,n}^h, \ldots, \vartheta_{G-1,n}^h, \vartheta_G)$

Theory for simplified models

Parameter shrinkage via nonconcave penalized likelihood, see Fan and Li (2001, JASA). Formulate the penalized log-likelihood

$$\mathcal{Q}(\vartheta) = \ell(\vartheta) - n \sum_{k=1}^{r_1+r_2} p_{\lambda_n}\left(|\vartheta_k|\right),$$

where $p_{\lambda_n}(|\cdot|)$ is the SCAD penalty with

$$p'_{\lambda,a}(x) = \lambda \mathbf{I}(x \leq \lambda) + \max(a\lambda - x, 0) / (a-1) \mathbf{I}(x > \lambda).$$

with $a > 2$ and $x > 0$. Assumptions

Corollary

Let the random vectors of the sequence X have an identical conditional density $f_i(\cdot;\vartheta)$ for which Assumptions 1–2, 4–6 hold. Then,

$$n^{1/2}\mathcal{B}_{h,n}^{-1}(\tilde\vartheta_0)\Big[(\tilde\vartheta_n^h - \tilde\vartheta_0) + \Gamma(\tilde\vartheta_0)^{h-1} \{\mathrm{B}_n(\tilde\vartheta_0) - \mathcal{H}^1(\tilde\vartheta_0)\}^{-1}\,\mathrm{b}_n(\tilde\vartheta_0)\Big] \xrightarrow{\mathcal{L}} \mathrm{N}\left\{0, \mathcal{M}(\tilde\vartheta_0)\right\},$$

$$\mathcal{B}_{h,n}(\tilde\vartheta) = \left[\Gamma(\tilde\vartheta)^{h-1}\{\mathrm{B}_n(\tilde\vartheta) - \mathcal{H}^1(\tilde\vartheta)\}^{-1}, \Gamma(\tilde\vartheta)^{h-1}\mathcal{H}(\tilde\vartheta)^{-1} - \mathcal{H}(\tilde\vartheta)^{-1}\right]$$

▸ Consistency ▸ $\mathrm{B}_n = \ldots, \mathrm{b}_n = \ldots$

Setup II

- R-vine, see Kurowicka and Joe (2011).
  - Decomposition of a d-dimensional copula density into $d(d-1)/2$ (conditional) bivariate copula densities.
- Natural decomposition ϑ.
- $d = 15$, $n = 250$, $r = 105$.
- Replications: 5000.

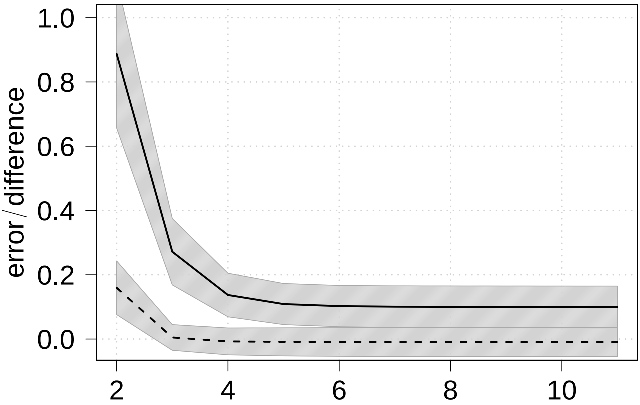


Figure 4: *R-vine*: Solid line shows the average error $\|\vartheta_n - \vartheta_n^h\|_1$ and the dashed line the difference $\ell(\vartheta_n) - \ell(\vartheta_n^h)$. The gray area refers to the respective empirical standard deviation.

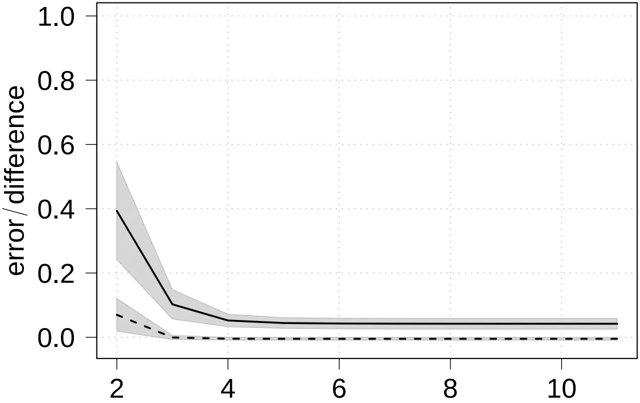


Figure 5: *Simplified R-vine*: Solid line shows the average error $\|\tilde{\vartheta}_n - \tilde{\vartheta}_n^h\|_1$ and the dashed line the difference $\ell(\tilde{\vartheta}_n) - \ell(\tilde{\vartheta}_n^h)$. The gray area refers to the respective empirical standard deviation.

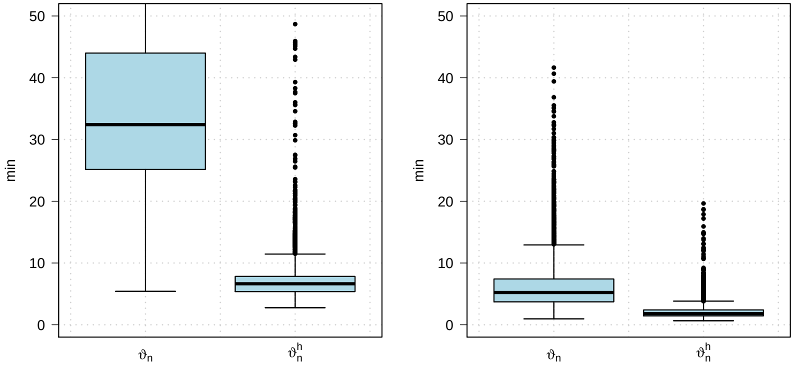


Figure 6: Left boxplots illustrate the computational time (in minutes) needed to compute the ML estimator ϑ_n and our estimator ϑ_n^h. Right boxplots refer to the computational times for the simplified R-vine model.

VAR model

Consider the time series model

$$X_i = c + \sum_{l=1}^{q} A_l X_{i-l} + \varepsilon_i,$$

where $c = (c_1, \ldots, c_d)^\top$ and A_l is a $(d \times d)$ matrix. Given standard assumptions like

- $\mathrm{E}(\varepsilon_i \varepsilon_i^\top) = \Sigma_\varepsilon$ and $\mathrm{E}(\varepsilon_i \varepsilon_{i-l}^\top) = 0_{dd}$ for $l > 0$
- $\varepsilon = \mathrm{vec}(\varepsilon_1, \ldots, \varepsilon_d) \sim \mathrm{N}(0, I_n \otimes \Sigma_\varepsilon)$

the parameters can be efficiently estimated by OLS. But

- $r > n$ especially for a large q!

Define $\mathsf{Y} = \text{vec}(X_1, \dots, X_n)$, $Z_i = (1, X_{i-1}^\top, \dots, X_{i-q}^\top)^\top$ and $\mathsf{Z} = (Z_1, \dots, Z_n)$ and rewrite the model in matrix notation

$$\mathsf{Y} = (\mathsf{Z}^\top \otimes I_d)\beta + \varepsilon,$$

where $\beta = \text{vec}(c, A_1, \dots, A_q)$. We assume $\varepsilon \sim \mathsf{N}(0, \Sigma)$, with $\Sigma \neq I_n \otimes \Sigma_\varepsilon$, but the GLS estimator

$$\beta_n = \left\{(\mathsf{Z} \otimes I_d)\Sigma^{-1}(\mathsf{Z}^\top \otimes I_d)\right\}^{-1}(\mathsf{Z} \otimes I_d)\Sigma^{-1}\mathsf{Y}$$

is not feasible.

Algorithm

Iteration $h = 1$*:*

(1) $\Sigma_n^1 = I_n \otimes \Sigma_\varepsilon$

(2) $\beta_n^1 = \left\{(\mathrm{Z}\,\mathrm{Z}^\top)^{-1}\,\mathrm{Z} \otimes I_d\right\} \mathrm{Y}$

Iteration $h > 1$*:*

(1) $\Sigma_n^h = \left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\beta_n^{h-1}\right\}\left\{\mathrm{Y} - (\mathrm{Z}^\top \otimes I_d)\beta_n^{h-1}\right\}^\top$

(2) $\beta_n^h = \left\{(\mathrm{Z} \otimes I_d)(\Sigma_n^h)^{-1}(\mathrm{Z}^\top \otimes I_d)\right\}^{-1}(\mathrm{Z} \otimes I_d)(\Sigma_n^h)^{-1}\,\mathrm{Y}$

Penalization of β can be embedded in *Iteration 1*!

DCC model

For a d-dimensional vector of returns X_i, the DCC model follows

$$
\begin{aligned}
& X_i = \mathrm{D}_i \, \varepsilon_i \quad \text{with} \quad \varepsilon_i | \mathcal{F}_{i-1} \sim \mathrm{N}(0, \mathrm{R}_i), \\
\text{with} \quad & \mathrm{R}_i = \mathrm{diag}(\mathrm{Q}_i)^{-1} \, \mathrm{Q}_i \, \mathrm{diag}(\mathrm{Q}_i)^{-1}, \\
& \mathrm{Q}_i = S \odot (1_d 1_d^\top - A - B) + A \odot \varepsilon_{i-1}\varepsilon_{i-1}^\top + B \odot \mathrm{Q}_{i-1}, \\
\text{and} \quad & \mathrm{D}_i^2 = \Omega + K \odot X_{i-1}X_{i-1}^\top + \Lambda \odot \mathrm{D}_{i-1}^2,
\end{aligned}
$$

where A and B are $(d \times d)$-matrices, 1_d is a d-dimensional vector of ones, Ω, K and Λ are quadratic diagonal matrices, $S = n^{-1} \sum_{i=1}^{n} \varepsilon_i \varepsilon_i^\top$.

log-likelihood can be decomposed into a correlation part $\ell^C(\vartheta_1, \vartheta_2)$ and a volatility part $\ell^V(\vartheta_2)$, such that $\ell(\vartheta_1, \vartheta_2) = \ell^V(\vartheta_2) + \ell^C(\vartheta_1, \vartheta_2)$, with

$$\ell^C(\vartheta_1, \vartheta_2) = -\frac{1}{2} \sum_{i=1}^{n} \left\{ \log(|\, \mathrm{R}_i \,|) + \varepsilon_i^\top \mathrm{R}_i^{-1} \varepsilon_i - \varepsilon_i^\top \varepsilon_i \right\}$$

where $|\cdot|$ computes the determinant, $\vartheta_1 = \mathrm{vec}(A, B)$, $\vartheta_2 = \{\mathrm{diag}(\Omega)^\top, \mathrm{diag}(K)^\top, \mathrm{diag}(\Lambda)^\top\}^\top$ and

$$\ell^V(\vartheta_2) = -\frac{1}{2} \sum_{i=1}^{n} \left\{ d \log(2\pi) + \log(|\, \mathrm{D}_i \,|^2) + X_i^\top \mathrm{D}_i^{-2} X_i \right\}.$$

Algorithm

Iteration $h = 1$*:*

(1) $\vartheta_{1,n}^{1} = 0$

(2) $\vartheta_{2,n}^{1} = \arg\max_{\vartheta_2} \ell^{V}(\vartheta_2)$

Iteration $h > 1$*:*

(1) $\vartheta_{1,n}^{h} = \arg\max_{\vartheta_1} \ell(\vartheta_1, \vartheta_{2,n}^{h-1})$

(2) $\vartheta_{2,n}^{h} = \arg\max_{\vartheta_2} \ell(\vartheta_1^{h}, \vartheta_2)$

Bivariate probit model

The bivariate probit model has the data generating process

$$Y_{ij} = \mathbf{I}\left\{\varepsilon_{ij} \leq \boldsymbol{\beta}_j^\top Z_{ij}\right\}, \quad \text{for} \quad i = 1, \ldots, n, \quad \text{and} \quad j = 1, 2,$$

where Z_{ij} is a r_j-dimensional vector of covariates including intercept and $(\varepsilon_{i1}, \varepsilon_{i2})^\top \sim \Phi(x_1, x_2; \rho)$.

Assume sparse model, i.e.,

$$\boldsymbol{\beta}_{j,0} = (\beta_{j1,0}, \ldots, \beta_{jr_j,0})^\top = (\boldsymbol{\beta}_{j1,0}^\top, \boldsymbol{\beta}_{j2,0}^\top)^\top \quad \text{with} \quad \boldsymbol{\beta}_{j2,0} = 0.$$

- Full log-likelihood: $\ell(\rho, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$.
- "Sparse" log-likelihood:

$$\tilde{\ell}(\rho, \boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{21}) = \ell\left\{\rho, (\boldsymbol{\beta}_{11}, 0), (\boldsymbol{\beta}_{21}, 0)\right\}.$$

Ignoring the dependence between Y_{i1} and Y_{i2}, i.e., $\rho = 0$, the marginal penalized log-likelihoods are

$$\mathcal{Q}_j(\boldsymbol{\beta}_j) = \sum_{i=1}^{n} \left[Y_{ij} \log\left\{\Phi(\boldsymbol{\beta}_j^\top Z_{ij})\right\} + (1 - Y_{ij}) \log\left\{1 - \Phi(\boldsymbol{\beta}_j^\top Z_{ij})\right\}\right] - n \sum_{k_j=1}^{r_j} p_{\lambda_{j,n}}(|\beta_{jk_j}|) \quad \text{for} \quad j = 1, 2.$$

Algorithm

Iteration $h = 1$*:*

(1) $\rho_n^1 = 0$

(2) $\boldsymbol{\beta}_{1,n}^1 = \arg\max_{\boldsymbol{\beta}_1} \mathcal{Q}_1(\boldsymbol{\beta}_1)$

(3) $\boldsymbol{\beta}_{2,n}^1 = \arg\max_{\boldsymbol{\beta}_2} \mathcal{Q}_2(\boldsymbol{\beta}_2)$

Iteration $h > 1$*:*

(1) $\rho_n^h = \arg\max_{\rho} \tilde{\ell}(\rho, \boldsymbol{\beta}_{11,n}^{h-1}, \boldsymbol{\beta}_{21,n}^{h-1})$

(2) $\boldsymbol{\beta}_{11,n}^h = \arg\max_{\boldsymbol{\beta}_{11}} \tilde{\ell}(\rho_n^h, \boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{21,n}^{h-1})$

(3) $\boldsymbol{\beta}_{21,n}^h = \arg\max_{\boldsymbol{\beta}_{21}} \tilde{\ell}(\rho_n^h, \boldsymbol{\beta}_{11,n}^h, \boldsymbol{\beta}_{21})$

SV model

The standard stochastic volatility model is discrete-time counterpart of continuous-time models and given by

$$X_i = \exp(\sigma_i/2)\varepsilon_i$$
$$\sigma_i = \alpha + \beta\sigma_{i-1} + \gamma\eta_i,$$

where $\varepsilon_i \overset{\text{iid}}{\sim} H(\varepsilon_1, \dots, \varepsilon_d)$ denote idiosyncratic shocks, $X = (X_1, \dots, X_n)^\top$ is the return process, $\sigma = (\sigma_1, \dots, \sigma_n)^\top$ is the univariate *latent* log-volatility process and $\eta_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$.

- Full log-likelihood: $\ell^f(\vartheta_1, \vartheta_2, \vartheta_3, \sigma) = \log\{f_{X,\sigma}(X, \sigma; \vartheta)\}$.
- "Observed" log-likelihood:

$$L^o(\vartheta_1, \vartheta_2, \vartheta_3) = \int f_{X|\sigma}(X, s; \vartheta_1, \vartheta_2, \vartheta_3) g_\sigma(s; \vartheta_3) ds,$$

- $f_{X,\sigma}(\cdot; \vartheta_1, \vartheta_2, \vartheta_3)$ equals the density of a Gaussian model $g_{X,\sigma}(\cdot; \vartheta_2, \vartheta_3)$ for a specific ϑ_1^*.
- $\vartheta_2 = \mathrm{vech}(\mathrm{R})$ and $\vartheta_3 = (\alpha, \beta, \gamma)^\top$.

Rewrite $L^o(\cdot)$ as

$$L^o(\vartheta_1, \vartheta_2, \vartheta_3) = L^g(\vartheta_2, \vartheta_3) \int \frac{f_{X|\sigma}(X, s; \vartheta_1, \vartheta_2, \vartheta_3)}{g_{X|\sigma}(X, s; \vartheta_2, \vartheta_3)} g_{\sigma|X}(X, s; \vartheta_2, \vartheta_3) ds.$$

log-likelihood under Gaussian assumption $\ell^g(\vartheta_2, \vartheta_3)$.

Algorithm

Iteration $h = 1$:

$$(2)-(3)\ \ (\vartheta_{2,n}^1, \vartheta_{3,n}^1) = \arg \max_{(\vartheta_2,\vartheta_3)} \ell^g(\vartheta_2, \vartheta_3)$$

Iteration $h > 1$:

$$(1)\ \ \vartheta_{1,n}^h = \arg \max_{\vartheta_1} \ell^f(\vartheta_1, \vartheta_{2,n}^{h-1}, \vartheta_{3,n}^{h-1}, \sigma_n^{h-1})$$

$$(2)\ \ \vartheta_{2,n}^h = \arg \max_{\vartheta_2} \ell^f(\vartheta_{1,n}^h, \vartheta_2, \vartheta_{3,n}^{h-1}, \sigma_n^{h-1})$$

$$(3)\ \ \vartheta_{3,n}^h = \arg \max_{\vartheta_3} \ell^o(\vartheta_{1,n}^h, \vartheta_{2,n}^h, \vartheta_3)$$

Measuring volatility connectedness

- Daily realized volatilities (RVs) from January 2007 - December 2008.
- 30 U.S. blue chip companies similar to the DJIA.

- VMEM$(1, 1)$ with R-vine based on bivariate t-copulae.
- $r/n \approx 1.7$

Assuming a stationary VMEM$(1,1)$ for the RVs $\{x_i\}_{i=1}^n$, whose zero-mean MA(∞) representation is

$$y_i = \eta_i + \sum_{l=1}^{\infty} \Psi_l \eta_{i-l},$$

with $\mathsf{E}(\eta_i) = 0, \mathsf{E}(\eta_i \eta_i^\top) = \Sigma_\eta$ and $y_i = x_i - \{I_d - (A+B)\}^{-1} \omega$.

(Un)conditional H-step prediction error:

- $\nu_i(H) = \sum_{l=0}^{H-1} \Psi_l \eta_{i+H-l}$ and
- $\nu_{i,\ell}(H) = \sum_{l=0}^{H-1} \Psi_l \left\{\eta_{i+H-l} - \mathsf{E}(\eta_{i+H-l} | \eta_{\ell,i+H-l} = \delta)\right\}$.

Connectedness measures

Diebold and Yilmaz (2014, JoE) suggest aggregating elements $v_{k\ell,H}$ of the generalized variance decomposition matrix V_H to

- the effect from others to k by $C_{k\leftarrow\bullet,H} = \sum_{\ell\neq k} v_{\ell\bullet,H}$,
- the effect to others from ℓ by $C_{\bullet\leftarrow\ell,H} = \sum_{\ell\neq k} v_{\bullet\ell,H}$,
- the total connectedness $C_H = \sum_{k\neq\ell} v_{k\ell,H}$.

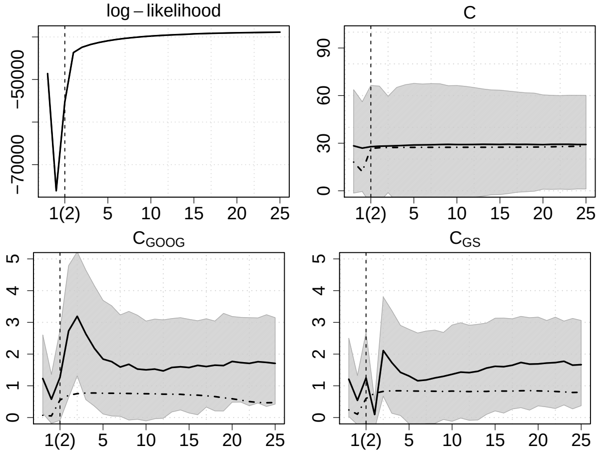


Figure 7: Upper panel: log-likelihood values and total systemic connectedness C_{12} in dependence of h. Lower panel: volatility contagion from Google $C_{\bullet\leftarrow \text{GOOG},12}$ and Goldman Sachs $C_{\bullet\leftarrow \text{GS},12}$ in dependence of h.

Conclusion

- Maximization strategy for complicated and high-parameterized log-likelihood functions.
- Asymptotic properties of the estimator are established.
- Accuracy of the procedure is illustrated in a simulation study.
- Algorithm is broadly applicable.
- Application emphasizes the importance of efficiency.

Future research:

- Non-parametric components

Efficient Iterative ML Estimation

Nikolaus Hautsch

Ostap Okhrin

Alexander Ristig

University of Vienna

Dresden University of Technology

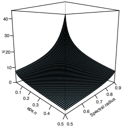
C.A.S.E. – Center for Applied Statistics and Economics

Humboldt–Universität zu Berlin

http://isor.univie.ac.at

http://tu-dresden.de

http://lvb.wiwi.hu-berlin.de



References

Aas, K., Czado, C., Frigessi, A. and H. Bakken
Pair-copula Constructions of Multiple Dependence
Insurance: Mathematics and economics 44 (2), 182–198, 2009

Diebold, F. X. and K. Yilmaz
On the Network Topology of Variance Decompositions: Measuring the Connectedness of Financial Firms
Journal of Econometrics, 182(1), 119–134, 2014

Engle, R.
Dynamic Conditional Correlation – a Simple Class of Multivariate GARCH Models
Journal of Business and Economic Statistics, 20(3), 339–350, 2002

- Fan, J. and R. Li
  Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties
  Journal of the American Statistical Association, 96(456), 1348–1360, 2001
- Kascha, C.
  A Comparison of Estimation Methods for Vector Autoregressive Moving-Average Models
  Econometric Reviews 31(3), 297–324, 2012
- Kurowicka, D. and H. Joe
  Dependence Modeling: Vine Copula Handbook
  World Scientific Publishing Company, Incorporated, 2011

Okhrin, O., Okhrin, Y. and W. Schmid
Determining the Structure and Estimation of Hierarchical Archimedean Copulas
Journal of Econometrics 173(2), 189–204, 2013

Ostrowski, A.
On the Linear Iteration Procedures for Symmetric Matrices
Rend. Mat. Appl. 14(1), 140–163, 1954

Reich, E.
On the Convergence of the Classical Iterative Procedures for Symmetric Matrices
Annals of Mathematical Statistics, 20(1), 448–451, 1949

Smith, M., Min, A., Almeida, C. and C. Czado
Modelling Longitudinal Data Using a Pair-copula Decomposition of Serial Dependence
Journal of the American Statistcal Association, 105(492), 1467–1479, 2010

White, H.
Estimation, Inference and Specification Analysis
Cambride University Press (Cambridge), 1st Edition, 1994

Assumptions

(1) The model is identifiable and the true value ϑ_0 is an interior point of the compact parameter space Θ. We assume that the model is correctly specified in the sense that $\mathsf{E}_\vartheta\{\dot{\ell}_{i,\vartheta_g}(\vartheta)\} = 0$ and information equality holds,

$$\mathcal{I}_{i,gl}(\vartheta) \stackrel{\text{def}}{=} \mathsf{E}_\vartheta\left\{\dot{\ell}_{\vartheta_g,i}(\vartheta)\dot{\ell}_{\vartheta_l,i}(\vartheta)^\top\right\} = -\,\mathsf{E}_\vartheta\left\{\ddot{\ell}_{\vartheta_g\vartheta_l,i}(\vartheta)\right\},$$

for $g, l = 1, \ldots, G$ and $i = 1, \ldots, n$.

(2) The information matrix is $\mathcal{I}(\vartheta) = \sum_{i=1}^{n} \mathcal{I}_i(\vartheta)$, with $\mathcal{I}_i(\vartheta) = \{\mathcal{I}_{i,gl}(\vartheta)\}_{g,l=1}^{G}$. Let the limit of $n^{-1}\mathcal{I}(\vartheta) \stackrel{P}{\to} \mathcal{J}(\vartheta)$ be the asymptotic information matrix, which is finite and positive definite at ϑ_0 and $n^{-1}\ddot{\ell}(\vartheta) \stackrel{P}{\to} \mathcal{H}(\vartheta)$ be the asymptotic Hessian, which is finite and negative definite for $\vartheta \in \{\vartheta : ||\vartheta - \vartheta_0|| < \delta\}$, $\delta > 0$. ▸ Decomposition

(3) The starting value is a consistent estimator $\vartheta_n^1 - \vartheta_0 = \mathcal{O}_p(1)$ with $\vartheta_n^1 = \arg\max_{\vartheta} \ell^1(\vartheta)$ and $\dot{\ell}^1(\vartheta) \neq \dot{\ell}(\vartheta)$.

(4) The "joint" score $s(\vartheta) = \{\dot{\ell}^1(\vartheta)^\top, \dot{\ell}(\vartheta)^\top\}^\top$ obeys $n^{-1/2} s(\vartheta_0) \xrightarrow{\mathcal{L}} \mathrm{N}\{0, \mathcal{M}(\vartheta_0)\}$, where

$$\mathcal{M}(\vartheta) = \begin{Bmatrix} \mathcal{J}^1(\vartheta) & \mathcal{J}^{1\star}(\vartheta) \\ \mathcal{J}^{\star 1}(\vartheta) & \mathcal{J}(\vartheta) \end{Bmatrix}.$$

▸ Assumptions

(5) The starting value of $\tilde{\vartheta} \stackrel{\text{def}}{=} (\vartheta_2^\top, \ldots, \vartheta_G^\top)^\top$ is a consistent estimator $\tilde{\vartheta}_n^1 - \tilde{\vartheta}_0 = \mathcal{O}_p(1)$, for $\lambda_n \to 0$ as $n \to \infty$, with $\vartheta_n^1 = \arg\max_{\vartheta} \mathcal{Q}(\vartheta)$ and $\dot{\ell}^1(\vartheta) \neq \dot{\ell}(\vartheta)$.

(6) If $\vartheta_{1,0} = 0$, $\lambda_n \to 0$ and $n^{1/2}\lambda_n \to \infty$ as $n \to \infty$, the estimator $\vartheta_{1,n}^1$ satisfies $\vartheta_{1,n}^1 = 0$ with probability tending to one.

Asymptotic Normality

Lemma

Let the random vectors of the sequence X have an identical conditional density $f_{X_i|\mathcal{F}_{i-1}}(\cdot;\vartheta)$ for which Assumptions 1-2 hold. If $\vartheta_n^1 \xrightarrow{P} \vartheta_0$, then $\vartheta_n^h \xrightarrow{P} \vartheta_0$, $\forall\ h = 2, 3, \ldots$.

▸ Asymptotic Normality

Lemma

Under the assumptions of Corollary 1, if $\lambda_n \to 0$ as $n \to \infty$, $\tilde{\vartheta}_n^h \xrightarrow{P} \tilde{\vartheta}_0\ \forall\ h = 2, 3, \ldots$.

▸ Asymptotic Normality

Definitions

$$\mathrm{b}_n(\tilde{\vartheta}) = \left\{ p'_{\lambda_n}(|\vartheta_{21}|)\,\mathrm{sign}(\vartheta_{21}), \ldots, p'_{\lambda_n}(|\vartheta_{2r_2}|)\,\mathrm{sign}(\vartheta_{2r_2}), 0 \right\}^\top,$$
$$\mathrm{B}_n(\tilde{\vartheta}) = \mathrm{diag}\left\{ p''_{\lambda_n}(|\vartheta_{21}|), \ldots, p''_{\lambda_n}(|\vartheta_{2r_2}|), 0 \right\}.$$

Asymptotic Normality