Adaptive LASSO with Coordinate Gradient Descent Algorithm for M-BEKK-ARCH(q) Model

Muhammad Jaffri Mohd Nasir

Universiti Malaysia Kelantan

21 August 2023

Model

m-variate Baba-Engle-Kroner-Kraft Autoregressive Conditional Heteroscedasticity (M-BEKK-ARCH)

$$\varepsilon_t = H_t^{\frac{1}{2}} \eta_t = (\varepsilon_{1,t}, \cdots, \varepsilon_{m,t})^T, \quad \eta_t \overset{iid}{\sim} \mathscr{D}(0, \mathrm{I}_m), \tag{1}$$

$$H_t = C_0 C_0^T + \sum_{i=1}^{q} \sum_{j=1}^{k} \left(A_{ij}^T \varepsilon_{t-i} \varepsilon_{t-i}^T A_{ij} \right) \tag{2}$$

1. Used in estimating individual and cross price volatility in eg., finance and economics.
2. Need to add $\ldots + \sum_{l=1}^{s} \sum_{j=1}^{k} \left(B_{lj}^T H_{t-l} B_{lj} \right)$ in Eqn (2) for the generalized version (GARCH).
3. When $m, k = 1$, it reduces to the univariate GARCH model.
4. Guaranteeing the positive definiteness of H_t.

Motivation

1. Lots of work on estimation and theories, but less so on model selection involving volatility models.
2. Model selection pertains to the estimation of ARCH and GARCH orders, or identifying nonzero coefficients, or both simultaneously.
3. Correct model $\rightarrow$ better understand the variance-covariance structure of the time series $\rightarrow$ improving forecasting.
4. The existing method is computationally cumbersome, slow and having non-convergence issues.
5. This work tries to tackle the issues of the existing method.

Transformation

For the sake of simplicity, consider the case of $k = 1$. For this case, we redefine A_{ij} as $A_i = [\alpha_{ab,i}]$, for $i = 1, \cdots, q$, and $1 \leq a, b \leq m$. Applying the vectorization operator, vec$(.)$, Eqn (2) can be transform to

VEC-type M-BEKK-ARCH

$$\text{vec}(H_t) = \text{vec}(C_0 C_0^T) + \sum_{i=1}^{q} (A_i \otimes A_i)^T \,\text{vec}(\varepsilon_{t-i}\varepsilon_{t-i}^T). \tag{3}$$

Eqn (3) will be used to generate the score and Hessian matrices for the M-BEKK-ARCH(q) model.

QMLE with adaptive LASSO (1)

Denote
$\psi = (\text{vech}(C_0)^T, \text{vec}(A_1)^T, \cdots, \text{vec}(A_q)^T)^T := (\psi_1, \psi_2, \cdots, \psi_{s^*})^T$
where $s^* = m(m+1)/2 + qm^2$ is the total number of parameters.

QMLE-Adaptive LASSO optimization problem

$$\widehat{\psi} = \arg\min_{\psi} \left(l_n(\psi) + \lambda_n \left[\sum_{i=1}^{m(m+1)/2} w_{H,i}|\psi_i| + \sum_{i=m(m+1)/2+1}^{s^*} w_{H,i}|\psi_i| \right] \right) \tag{4}$$

with

$$l_t(\psi) \approx \sum_{t=1}^{n} \frac{1}{2} \log\left[\det(H_t(\psi))\right] + \sum_{t=1}^{n} \frac{1}{2} \varepsilon_t^T H_t^{-1}(\psi) \varepsilon_t$$

QMLE with adaptive LASSO (2)

Adaptive LASSO optimization problem

$$\widehat{\psi} = \arg\min_{\psi} \left(l_n(\psi) + \lambda_n \left[\sum_{i=1}^{m(m+1)/2} w_{H,i} |\psi_i| + \sum_{i=m(m+1)/2+1}^{s^*} w_{H,i} |\psi_i| \right] \right)$$

- $\lambda_n \geq 0$ is a shrinkage or regularization parameter.
- $w_{H,i} \geq 0$ are the adaptive weights.
- Standard way of selecting the adaptive weighting is by setting the weight according to the parameter estimates of non-penalized QMLE, i.e., when $\lambda_n = 0$ with $w_{H,i} = 1/|\widehat{\psi}_i|$.

Derivatives (1)

Derivatives of log-likelihood

(a) for $i = 1, 2, \cdots, s^*$,

$$\frac{\partial l_n(\psi)}{\partial \psi_i} = \frac{1}{2} \sum_{t=1}^{n} \mathrm{Tr}\left(\frac{\partial H_t(\psi)}{\partial \psi_i} H_t^{-1}(\psi) - \varepsilon_t \varepsilon_t^T H_t^{-1}(\psi) \frac{\partial H_t(\psi)}{\partial \psi_i} H_t^{-1}(\psi) \right).$$

(b) for $i, j = 1, 2, \cdots, s^*$,

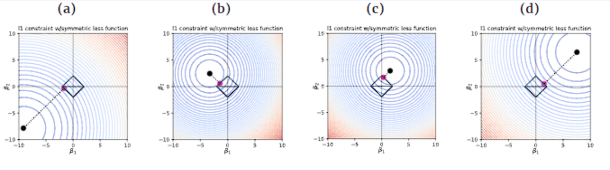
$$\begin{aligned} \frac{\partial^2 l_n(\psi)}{\partial \psi_i \partial \psi_j} = \frac{1}{2} \sum_{t=1}^{n} \mathrm{Tr}\Bigg(& \frac{\partial^2 H_t(\psi)}{\partial \psi_i \partial \psi_j} H_t^{-1}(\psi) - \frac{\partial H_t(\psi)}{\partial \psi_i} H_t^{-1}(\psi) \\ & \times \frac{\partial H_t(\psi)}{\partial \psi_j} H_t^{-1}(\psi) - \varepsilon_t \varepsilon_t^T H_t^{-1}(\psi) \frac{\partial^2 H_t(\psi)}{\partial \psi_i \partial \psi_j} H_t^{-1}(\psi) \\ & + 2 \varepsilon_t \varepsilon_t^T H_t^{-1}(\psi) \frac{\partial H_t(\psi)}{\partial \psi_i} H_t^{-1}(\psi) \frac{\partial H_t(\psi)}{\partial \psi_j} H_t^{-1}(\psi) \Bigg). \end{aligned}$$

Derivatives (2)

Derivatives of LASSO

Let $\widetilde{e}_j$ is the sub-gradient of $|\psi_j| \geq 0$, such that

$$\widetilde{e}_j = \begin{cases} 1, & \text{if } \psi_j > 0, \\ u \in [0,1], & \text{if } \psi_j = 0, \\ -1, & \text{if } \psi_j < 0. \end{cases}$$



Derivatives (3)

Derivatives of M-BEKK-ARCH

K_{mm} and D_m are the commutation and duplication matrices. Let $M_1 = \mathrm{I}_m \otimes K_{mm} \otimes \mathrm{I}_m$, $M_2 = 2(\mathrm{I}_{m^2} \otimes D_m D_m^+)$, and $M_3 = M_2 M_1$.

(a) If $q = 0$ and $k = 1$, then

$$\widetilde{H}_{C_0,t} = 2D_m D_m^+ (C_0 \otimes \mathrm{I}_m) L_m^T, \quad \text{and}$$
$$\widetilde{\boldsymbol{H}}_{C_0 C_0,t} = 2(L_m \otimes D_m D_m^+) M_1 (\mathrm{I}_{m^2} \otimes \mathrm{vec}(\mathrm{I}_m)) L_m^T.$$

(b) If $q > 0$ and $k = 1$, then

$$H_{A_i,t} = 2D_m D_m^+ (\mathrm{I}_m \otimes A_i^T \varepsilon_{t-i} \varepsilon_{t-i}^T), \quad \text{and}$$
$$\boldsymbol{H}_{A_i A_i,t} = M_3 \left[\mathrm{vec}(\mathrm{I}_m) \otimes (\varepsilon_{t-i} \varepsilon_{t-i}^T \otimes \mathrm{I}_m) K_{mm}\right]$$

and $\widetilde{H}_{C_0,t}$ with $\widetilde{\boldsymbol{H}}_{C_0 C_0,t}$ are equal to the one with the case of $q = 0$.

Derivatives (4)

More details regarding structure for the matrices is given to facilitate computations.

$$\widetilde{H}_{C_0,t} = \begin{bmatrix} \boldsymbol{h}^{C_0}_{c_{11,0},t} & \cdots & \boldsymbol{h}^{C_0}_{c_{m1,0},t} & \boldsymbol{h}^{C_0}_{c_{22,0},t} & \cdots & \boldsymbol{h}^{C_0}_{c_{m2,0},t} & \cdots & \boldsymbol{h}^{C_0}_{c_{mm,0},t} \end{bmatrix}, \quad \text{and}$$

$$H_{A_i,t} = \begin{bmatrix} \boldsymbol{h}^{A_i}_{\alpha_{11,i},t} & \cdots & \boldsymbol{h}^{A_i}_{\alpha_{m1,i},t} & \boldsymbol{h}^{A_i}_{\alpha_{12,i},t} & \cdots & \boldsymbol{h}^{A_i}_{\alpha_{m2,i},t} & \cdots & \boldsymbol{h}^{A_i}_{\alpha_{1m,i},t} & \cdots & \boldsymbol{h}^{A_i}_{\alpha_{mm,i},t} \end{bmatrix}$$

with

$$\widetilde{\boldsymbol{H}}_{C_0C_0,t} = \begin{bmatrix} \boldsymbol{h}^{C_0}_{c_{11,0}c_{11,0},t} & \cdots & \boldsymbol{h}^{C_0}_{c_{11,0}c_{mm,0},t} \\ \vdots & \ddots & \vdots \\ \boldsymbol{h}^{C_0}_{c_{11,0}c_{mm,0},t} & \cdots & \boldsymbol{h}^{C_0}_{c_{mm,0}c_{mm,0},t} \end{bmatrix} \text{ and } \boldsymbol{H}_{A_iA_i,t} = \begin{bmatrix} \boldsymbol{h}^{A_i}_{\alpha_{11,i}\alpha_{11,i},t} & \cdots & \boldsymbol{h}^{A_i}_{\alpha_{11,i}\alpha_{mm,i},t} \\ \vdots & \ddots & \vdots \\ \boldsymbol{h}^{A_i}_{\alpha_{11,i}\alpha_{mm,i},t} & \cdots & \boldsymbol{h}^{A_i}_{\alpha_{mm,i}\alpha_{mm,i},t} \end{bmatrix}$$

where $\boldsymbol{h}^{\Gamma_k}_{\gamma_{ij,k},t}$ and $\boldsymbol{h}^{\Gamma_k}_{\gamma_{ij,k}\gamma_{lm,k},t}$, $\gamma_{ij,k}, \gamma_{lm,k} \in \Gamma_k$ are respectively the m^2-dimensional column vectors of gradient corresponding to $\gamma_{ij,k,t}$ and Hessian matrix corresponding to $\gamma_{ij,k,t}$ and $\gamma_{lm,k,t}$.

Derivatives (5)

Denote mat(.) as the inverse vectorization operator, which transform a m^2-dimensional column vector into a $(m \times m)$ matrix. We denote the first derivatives of $H_t(\psi)$ with respect of each M-BEKK-ARCH(q) coefficient by

$$\frac{\partial H_t(\psi)}{\partial c_{ab,0}} = \mathrm{mat}(\boldsymbol{h}^{C_0}_{c_{ab,0},t}), \frac{\partial H_t(\psi)}{\partial \alpha_{ij,f}} = \mathrm{mat}(\boldsymbol{h}^{A_i}_{\alpha_{ij,i},t}), \tag{5}$$

and the second derivatives of $H_t(\psi)$ with respect of each M-BEKK-ARCH(q) coefficient by

$$\frac{\partial^2 H_t(\psi)}{\partial c^2_{ab,0}} = \mathrm{mat}(\boldsymbol{h}^{C_0}_{c_{ab,0}c_{ab,0},t}), \frac{\partial^2 H_t(\psi)}{\partial \alpha^2_{ij,f}} = \mathrm{mat}(\boldsymbol{h}^{A_i}_{\alpha_{ij,i}\alpha_{ij,i},t}), \tag{6}$$

for $1 \leq b \leq a \leq J$, $J = 1, \cdots, m$, $1 \leq i,j \leq m$ and $1 \leq f \leq q$.

Algorithm

Define $\bigtriangledown \boldsymbol{l}(\psi) = \left(\frac{\partial l_n(\boldsymbol{\psi})}{\partial \psi_1}, \frac{\partial l_n(\boldsymbol{\psi})}{\partial \psi_2}, \cdots, \frac{\partial l_n(\boldsymbol{\psi})}{\partial \psi_{s^*}}\right)^T$ as the vector of the first derivatives, and $\boldsymbol{q}(\psi) = \left(\frac{\partial^2 l_n(\boldsymbol{\psi})}{\partial \psi_1^2}, \frac{\partial^2 l_n(\boldsymbol{\psi})}{\partial \psi_2^2}, \cdots, \frac{\partial^2 l_n(\boldsymbol{\psi})}{\partial \psi_{s^*}^2}\right)^T$ as the vector of the second derivatives of the negative log-likelihood.

Coordinate Gradient Descent

$$\widehat{d}_j = \arg\min_{d_j} \left(\bigtriangledown \boldsymbol{l}(\psi)_j d_j + \frac{1}{2}\boldsymbol{q}(\psi)_j d_j^2 + \lambda_n w_{H,j}(\psi_j + d_j) \middle| d_i = 0, \forall i \notin j \right). \tag{7}$$

Algorithm (1)

Algorithm 1: Coordinate Gradient Descent Algorithm (*CGD*) for Adaptive LASSO of M-BEKK-ARCH

Data: $E = [\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n] \in \mathbb{R}^{m\times n}$, $\lambda_n \geq 0$, $0 < \varepsilon \leq 10^{-4}$, $(w_{H,1}, w_{H,2}, \cdots, w_{H,s^*})$, $0 < \delta_*, \alpha_* < 1$, $0 < \widetilde{h} < 1$ and $0 < \widetilde{\gamma} < 1$.

Result: ψ satisfying (10).

Initialize: $\psi = (\psi_1, \cdots, \psi_{m^*}, \psi_{m^*+1}, \cdots, \psi_{s^*})^T = (1, \cdots, 1, 0.01, \cdots, 0.01)^T$.

repeat
- **for** $j = 1, 2, \cdots, s^*$, **do**
  - **Compute** $\bigtriangledown \boldsymbol{l}(\psi)$ and $\boldsymbol{q}(\psi)$.
  - **Set** $\alpha_* = 1$.
  - **if** $|(-\bigtriangledown \boldsymbol{l}(\psi)_j) - \boldsymbol{q}(\psi)_j(-\psi_j)| \leq \lambda_n w_{H,j}$, **then**
    - **Set** $\widehat{d}_j = -\psi_j$, $\widehat{d}_j \in \widehat{\boldsymbol{d}}$.
  - **else if** $-\bigtriangledown \boldsymbol{l}(\psi)_j - \boldsymbol{q}(\psi)_j(-\psi_j) > \lambda_n w_{H,j}$, **then**
    - **Compute** $\widehat{d}_j = (-\lambda_n w_{H,j} - \bigtriangledown \boldsymbol{l}(\psi)_j)/(\boldsymbol{q}(\psi)_j)$.
  - **else if** $-\bigtriangledown \boldsymbol{l}(\psi)_j - \boldsymbol{q}(\psi)_j(-\psi_j) < -\lambda_n w_{H,j}$, **then**
    - **Compute** $\widehat{d}_j = (\lambda_n w_{H,j} - \bigtriangledown \boldsymbol{l}(\psi)_j)/(\boldsymbol{q}(\psi)_j)$.
  - **end**
  - **Compute** $\zeta = (\widetilde{\gamma} - 1)\widehat{d}_j^2 \boldsymbol{q}(\psi)_j$.
  - **repeat**
    - **if** $f_\lambda(\psi + \alpha_* \widehat{\boldsymbol{d}}) > f_\lambda(\psi) + \alpha_* \widetilde{h}\zeta$, where $f_\lambda(\psi) = l_n(\psi) + \lambda_n \sum_{i=1}^{s^*} w_{H,i}|\psi_i|$, **then**
      - $\alpha_* \leftarrow \delta_* \alpha_*$.
    - **end**
  - **until** $\alpha_* \leq \varepsilon$, or $f_\lambda(\psi + \alpha_* \widehat{\boldsymbol{d}}) \leq f_\lambda(\psi) + \alpha_* \widetilde{h}\zeta$ is met.
  - **Update** $\psi_j \leftarrow \psi_j + \alpha_* \widehat{d}_j$, $\quad \psi_j \in \psi$.
- **end**

until some convergence criterion of parameters is met.

Algorithm (2)

Algorithm 2: Full algorithm for the adaptive LASSO of M-BEKK-ARCH

Data: $E = [\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n] \in \mathbb{R}^{m \times n}$, c_λ and $k_0 \geq 0$.

Result: The adaptive LASSO estimator, corresponding to the lowest BIC, $\widehat{\psi}_{BIC}$.

Initialize: Setup a grid of shrinkage parameter: $\lambda_1 = \lambda_{\max}, \lambda_2, \cdots, \lambda_{k_0} = 0$, such that $\lambda_1 > \lambda_2 > \cdots > \lambda_{k_0}$ (default $\lambda_{\max} = 10$).

Apply Algorithm 1 with $\lambda_n = 0$, **output** $\widetilde{\psi} \leftarrow \psi$.

Obtain the adaptive weights $w_{H,i}$ from $\widetilde{\psi}$ as follows: set $w_{H,i} = 1/|\widetilde{\psi}_i|$, for $i = 1, \cdots, s^*$.

for $j = 1, 2, \cdots, k_0$, **do**

  Apply Algorithm 1 with λ_j to obtain $\widehat{\psi}$.

  Compute $v_j = \text{BIC}_\lambda$ using $\widehat{\psi}$, where BIC_λ is given in (19).

end

Choose $\widehat{\psi}_{BIC} \leftarrow \widehat{\psi}$ which corresponds to the lowest v_j, for $j = 1, \cdots, k_0$.

Simulation Study 1 (a)

Sparse Bivariate BEKK-ARCH(2) model, for $t = 1, \cdots, n$, (1000 replication, n=1000)

$$\varepsilon_t = H_t^{1/2}\boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(0, \mathrm{I}_2),$$
$$H_t = C_0C_0^T + A_1^T\varepsilon_{t-1}\varepsilon_{t-1}^T A_1 + A_2^T\varepsilon_{t-1}\varepsilon_{t-1}^T A_2$$

with the matrices of parameters

$$C_0 = \begin{bmatrix} 1.1 & 0 \\ 0 & 0.9 \end{bmatrix}, A_1 = \begin{bmatrix} 0.35 & 0 \\ 0.4 & 0.8 \end{bmatrix} \text{ and } A_2 = \begin{bmatrix} 0.4 & 0 \\ 0 & 0.35 \end{bmatrix}.$$

Simulation Study 1 (b)

Table: Summary statistics of the *nQMLE* and *AdLassoBIC* for 1000 replications. The true, minimum, mode, average and maximum cardinality of the estimated non-zero parameters are denoted as TRUE, Min., Mode, Avg. and Max. respectively. %Rel and %Match are the percentages of samples that pick all significant variables and select the correct model, respectively.

n=1000	TRUE	Min.	Mode	Avg.	Max.	%Rel	%Match
nQMLE	7	11	11	11	11	100%	0%
AdLassoBIC	7	5	7	7	9	97%	87.5%

Simulation Study 1 (c)

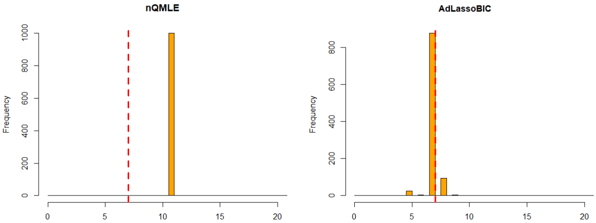


Figure: Bar charts of cardinality of non-zero parameter estimate via *nQMLE* and *AdLassoBIC*, for 1000 samples. The red, vertical dashed lines are the true cardinality.

Simulation Study 2 (a)

Sparse trivariate BEKK-ARCH(2) model, for $t = 1, \cdots, n$, (1000 replication, n=1000)

$$\varepsilon_t = H_t^{1/2}\eta, \quad \eta \sim \mathcal{N}(0, \mathrm{I}_3),$$
$$H_t = C_0 C_0^T + A_1^T \varepsilon_{t-1}\varepsilon_{t-1}^T A_1 + A_2^T \varepsilon_{t-2}\varepsilon_{t-2}^T A_2$$

with

$$C_0 = \begin{bmatrix} 0.75 & 0 & 0 \\ 0.16 & 0.68 & 0 \\ 0.34 & 0 & 0.47 \end{bmatrix}, A_1 = \begin{bmatrix} 0.32 & 0 & 0.35 \\ 0 & 0.27 & 0 \\ 0.18 & 0 & 0.45 \end{bmatrix}$$

$$\text{and } A_2 = \begin{bmatrix} 0.23 & 0.25 & 0.46 \\ 0.14 & 0.31 & 0 \\ 0 & 0 & 0.35 \end{bmatrix}.$$

Simulation Study 2 (b)

Table: Summary statistics of the *nQMLE* and *AdLassoBIC* for 1000 replications. The true, minimum, mode, average and maximum cardinality of the estimated non-zero parameters are denoted as TRUE, Min., Mode, Avg. and Max. respectively. %Rel and %Match are the percentages of samples that pick all significant variables and select the correct model, respectively.

n=1000	TRUE	Min.	Mode	Avg.	Max.	%Rel	%Match
nQMLE	16	24	24	24	24	100%	0%
AdLassoBIC	16	9	16	16	18	72%	60.3%

Simulation Study 2 (c)

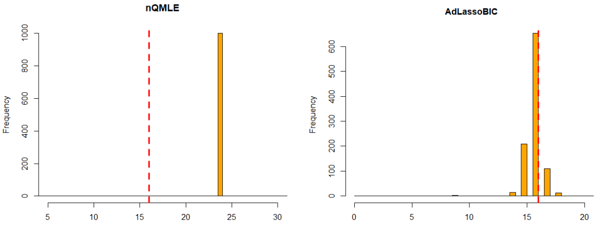


Figure: Bar charts of cardinality of non-zero parameter estimate via *nQMLE* and *AdLassoBIC*, for 1000 samples. The red, vertical dashed lines are the true cardinality.

Possible Future Works

- Extension to M-BEKK-GARCH and its families.
- Application of sparse group adaptive LASSO (estimating Models's orders)
- Non-convex penalization for better performance.
- Robust estimators (account for fatter tail distribution, outliers).