High-Dimensional Vector Autoregressive Time Series Modeling via Tensor Decomposition

Ilia Lomasov

UIUC

11/28/2023

Outline

Table of Contents

VAR model (Lütkepohl 2005; Tsay 2010)

Consider the VAR model of the form:

$$y_t = A_1 y_{t-1} + \ldots + A_P y_{t-P} + \epsilon_t \tag{1}$$

$$\{y_t\}_{t=1}^{T} \in \mathbb{R}^N, \quad \{A_j\}_{j=1}^{P} \in \mathbb{R}^{N\times N}$$

$$\epsilon_t \in \mathbb{R}^N \overset{iid}{\sim} (0, \Sigma_\epsilon), \Sigma_\epsilon \succ 0, \Sigma_\epsilon < \text{inf}$$

- Difficult to perform estimation as number of parameters is N^2P - very large.
- Even in low-dimensional setting: $N = 5, P = 2 \Rightarrow npar = 50$

Alternatives to estimating N^2P parameters:

1. PCA & Factor Models – in the STAT556 course
2.
   - **Approach:** Assume sparsity of A_j and apply regularization (e.g., ℓ_1, LASSO or Dantzig selector)
   - **Papers:** Basu & Michailidis 2015; Han, Lu & Liu 2015 etc.
   - **Drawbacks:**
     - Sacrifices temporal and cross-sectional dependencies
     - Average magnitude of parameters is bounded by $O(N^{-1/2})$, which limits sparsity-inducing regularization
3.
   - **Approach:** Reduced-rank regression ($A^{(C)}$ – low-rank):

$$y_t = (A_1 \dots A_P)\left(y'_{t-1} \dots y'_{t-P}\right)' + \epsilon_t =: A^{(C)} x_t + \epsilon_t \tag{2}$$

   - **Papers:** Velu & Reinsel 2013; Carriero, Kapetanios, & Marcellino 2011 (Bayesian extension) etc.
   - **Limitation:** Allows for only 1 type of low-rank structure

Table of Contents

3-Dimensional Tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$

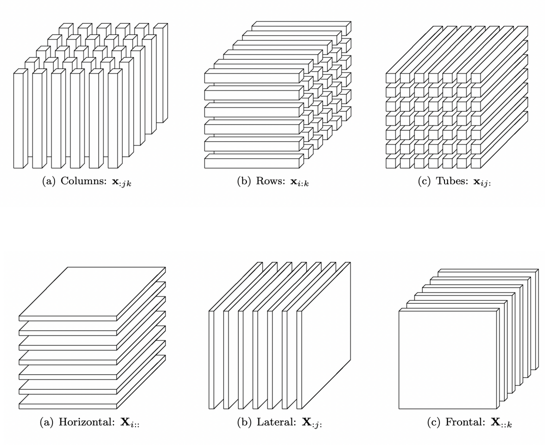

Figure 1: Fibers (top) and Slices (bottom) of $\mathcal{X}$ (Kolda, 2006)

Visualization 1

Let's consider a simple tensor $\mathcal{X} \in \mathbb{R}^{2\times3\times2}$ defined by
$x_{ijk} = 100i + 10j + k$

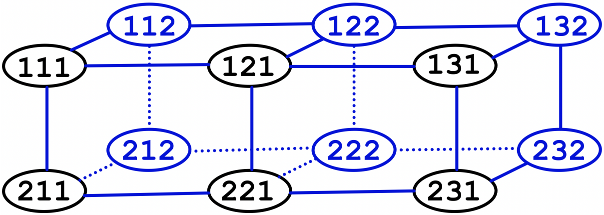

Figure 2: Tensor $\mathcal{X}$

Matricization

$\{\mathcal{X}_{(n)}\}_{n=1}^{3}$ – mode-n matricization of $\mathcal{X}$ – rearrangement of mode-n fibers into a matrix. $\mathcal{X}_{(n)} \in \mathbb{R}^{I_n \times (I_1 I_2 I_3 / I_n)}$.

- $\mathcal{X}_{(1)} = (X_{::1}\ X_{::2}) = \begin{pmatrix} 111 & 121 & 131 & 112 & 122 & 132 \\ 211 & 221 & 231 & 212 & 222 & 232 \end{pmatrix}$
- $\mathcal{X}_{(2)} = (X'_{::1}\ X'_{::2}) = \begin{pmatrix} 111 & 211 & 112 & 212 \\ 121 & 221 & 122 & 222 \\ 131 & 231 & 132 & 232 \end{pmatrix}$
- $\mathcal{X}_{(3)} = (vec(X_{::1})\ vec(X_{::2}))' = \begin{pmatrix} 111 & 211 & 121 & 221 & 131 & 231 \\ 112 & 212 & 122 & 222 & 132 & 232 \end{pmatrix}$
- $r_n = rank_n(\mathcal{X}) = rank(\mathcal{X}_{(n)}) \leq I_n$

Mode-n multiplication

- Mode-n multiplication:

$$\underset{I_1\times I_2\times I_3}{\mathcal{X}} \times_n \underset{J\times I_n}{U} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 i_3} y_{j i_n}, \quad n \in \{1,2,3\}$$

- Let's take $U = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$, $\mathcal{X}$ defined by $x_{ijk} = 100i + 10j + k$

$$\underset{2\times3\times3}{\mathcal{Y}} = \underset{2\times3\times2}{\mathcal{X}} \times_3 \underset{3\times2}{U} \Rightarrow y_{ijl} = \sum_{k=1}^{3} x_{ijk} u_{lk}$$

- $l \in \{1,2\} \Rightarrow u_{lk} = \mathbb{1}(l=k) \Rightarrow y_{ijl} = x_{ijk}$
- $l = 3 \Rightarrow u_{lk} = 1 \Rightarrow y_{ij3} = x_{ij1} + x_{ij2}$

Visualization 2

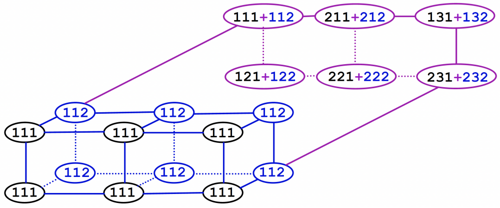

Figure 3: Tensor $\mathcal{Y}$

$$\begin{aligned}\mathcal{Y}_{(3)} &= (vec(Y_{::1})\ vec(Y_{::2})\ vec(Y_{::3}))' \\ &= (vec(X_{::1})\ vec(X_{::2})\quad vec(X_{::1}) + vec(X_{::2}))' \\ &= U\,(vec(X_{::1})\ vec(X_{::2}))' = U\mathcal{X}_{(3)}\end{aligned}$$

In general, $\mathcal{Y} = \mathcal{X} \times_n U \Rightarrow \mathcal{Y}_{(n)} = U\mathcal{X}_{(n)}$

Some facts

1 For $m \neq n$, $\mathcal{X} \times_m A \times_n B = \mathcal{X} \times_n B \times_m A$

$$\text{Proof: } [\mathcal{X} \times_m A \times_m B]_{i_{-n,-m}jk} = \sum_{i_n=1}^{I_n} \left[\sum_{i_m=1}^{I_m} x_{i_1 i_2 i_3} a_{ji_m} \right] b_{ki_n} =$$

$$= \sum_{i_m=1}^{I_m} \left[\sum_{i_n=1}^{I_n} x_{i_1 i_2 i_3} b_{ki_n} \right] a_{ji_m} = [\mathcal{X} \times_n B \times_m A]_{i_{-n,-m}jk}$$

2 $\mathcal{X} \times_n A \times_n B = \mathcal{X} \times_n (BA)$

$$\text{Proof: } [\mathcal{X} \times_n A \times_n B]_{i_{-n},k} = \sum_{j=1}^{k} \left[\sum_{i_n=1}^{I_n} x_{i_1 i_2 i_3} a_{ji_n} \right] b_{kj} =$$

$$= \sum_{i_n=1}^{I_n} x_{i_1 i_2 i_3} \sum_{j=1}^{k} b_{kj} a_{ji_n} = \sum_{i_n=1}^{I_n} x_{i_1 i_2 i_3} (BA)_{ki_n} = [\mathcal{X} \times_n (BA)]_{i_{-n},k}$$

Some facts 2

3 If $rank_n(\mathcal{Y}) < I_n$ for some $n \in \{1, 2, 3\}$, then there exists a Tucker decomposition $\mathcal{Y} = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3$. (Notation: $[\![\mathcal{G}; U_1, U_2, U_3]\!]$) Here, one option is $\mathcal{G} = \mathcal{X}, U_1 = \mathbb{I}_2, U_2 = \mathbb{I}_3, U_3 = U$

4 (Following from 1 & 2): Tucker decomposition is not unique: for any non-singular matrices $\left\{O_n \in \mathbb{R}^{I_n \times I_n}\right\}_{n=1}^{3}$

$$\begin{aligned}&\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3 = \\ =&(\mathcal{G} \times_1 O_1 \times_2 O_2 \times_3 O_3) \times_1 U_1 O_1^{-1} \times_2 U_2 O_2^{-1} \times_3 U_3 O_3^{-1}\end{aligned}$$

5 (Kolda, 2006)

$$(\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3)_{(1)} = U_1 \mathcal{G}_{(1)} (U_3 \otimes U_2)'$$

Table of Contents

Setup

- We can rearrange transition matrices $A_1, \ldots, A_P$ into a tensor $\mathcal{A} \in \mathbb{R}^{N \times N \times P}$, assuming those are its frontal slices (as we can see in Figure 4):

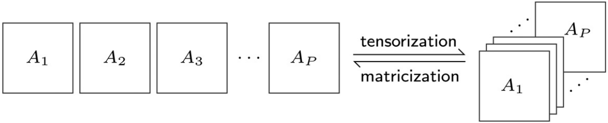

Figure 4: Tensorization from $A_1, \ldots, A_P$ to $\mathcal{A}$

- Recall representation (2): $y_t = (A_1, \ldots, A_P)\, x_t + \epsilon_t$. It's equivalent to $y_t = \mathcal{A}_{(1)} x_t + \epsilon_t$
- **Key idea:** assume $\mathcal{A} = \mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3$

Some equivalent representations (under certain assumptions)

$$y_t = (\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3)_{(1)} x_t + \epsilon_t \tag{3}$$

$$y_t = U_1 \mathcal{G}_{(1)} (U_3 \otimes U_2)' x_t + \epsilon_t = U_1 \mathcal{G}_{(1)} vec(U_2' X_t U_3) + \epsilon_t \tag{4}$$

Where $X_t = (y_{t-1} \ \ldots \ y_{t-P})$

Furthermore, consider a special Tucker decomposition – high-order SVD (HOSVD) (De Lathauwer, DeMoor & Vandewalle 2000):

Given $n \in \{1, 2, 3\}$ and $r_n = rank_n(\mathcal{A}) \leq I_n$, construct $U_n \in \mathbb{R}^{I_n \times r_n}$ as a matrix of top-r_n left singular vectors of $\mathcal{A}_{(n)}$ (i.e., eigenvectors of $\mathcal{A}_{(n)}\mathcal{A}_{(n)}{}'$). $\mathcal{A}_{(n)}\mathcal{A}_{(n)}{}'$ – symmetric $\Rightarrow U_n' U_n = \mathbb{I}_{I_n}, \ U_n U_n' = \mathbb{I}_{r_n}$

Number of parameters comparison

Model	Number of parameters
VAR	N^2P
RRR	$(NP + N - r_1)r_1$ (r_1 independent rows with NP elements each plus $N - r_1$ dependent rows, with r_1 dependency coefficients each)
MLR	$r_1r_2r_3 + (N - r_1)r_1 + (N - r_2)r_2 + (P - r_3)r_3$

* – Since $\mathcal{A}$ is to be estimated, r_n has to be assumed?

Connection with Factor Models

- Recall unknown factor model with r_1 common factors: $Y = F\Lambda' + E$
  Where $Y = (y_1, \dots, y_T)'$, $X = (x_1, \dots, x_T)'$, $E = (e_1, \dots, e_T)'$, $F'F/T = \mathbb{I}_{r_1}$, $\Lambda'\Lambda \in \mathbb{R}^{r_1 \times r_1}$ – full-rank and diagonal.
- Rewrite $y_t = U_1 \mathcal{G}_{(1)} \left(U_3 \otimes U_2\right)' x_t + \epsilon_t$ in matrix form:

$$Y = X\left(U_3 \otimes U_2\right) \mathcal{G}_{(1)}' U_1' + E \tag{5}$$

- Consider SVD: $\left(X\left(U_3 \otimes U_2\right) \mathcal{G}_{(1)}'\right) = U_x D_x V_x'$, where $D_x \in \mathbb{R}^{r_1 \times r_1}$ is diagonal, and U_x and V_x are orthonormal.
- Define $F = \sqrt{T} U_x$ and $\Lambda = U_1 V_x D_x / \sqrt{T}$. Note that $F'F/T = \mathbb{I}_{r_1}$ and that $\Lambda'\Lambda$ is diagonal.
- Thus, $Y = X\left(U_3 \otimes U_2\right) \mathcal{G}_{(1)}' U_1' + E = F\Lambda' + E$

Key difference: MLR can be used directly for predictions

Table of Contents

$\widehat{\mathcal{A}}_{\mathrm{MLR}}, \widehat{A}_{\mathrm{RRR}}, \widehat{A}_{\mathrm{OLS}}$

$$\widehat{\mathcal{A}}_{\mathrm{MLR}} = \widehat{\mathcal{G}} \times_1 \widehat{U}_1 \times_2 \widehat{U}_2 \times_3 \widehat{U}_3 = \arg\min L\left(\mathcal{G}, U_1, U_2, U_3\right), \text{ where}$$

$$L\left(\mathcal{G}, U_1, U_2, U_3\right) = \frac{1}{T} \sum_{t=1}^{T} \left\| y_t - \left(\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3\right)_{(1)} x_t \right\|_2^2$$

$$\left(\widehat{\mathcal{A}}_{\mathrm{OLS}}\right)_{(1)} = \widehat{A}_{\mathrm{OLS}} = \arg \min_{B \in \mathbb{R}^{N \times NP}} \sum_{t=1}^{T} \left\| y_t - B x_t \right\|_2^2$$

$$\left(\widehat{\mathcal{A}}_{\mathrm{RRR}}\right)_{(1)} = \widehat{A}_{\mathrm{RRR}} = \arg \min_{B \in \mathbb{R}^{N \times NP}, \mathrm{rank}(B) \leq r_1} \sum_{t=1}^{T} \left\| y_t - B x_t \right\|_2^2$$

Asymptotic Properties of $\widehat{\mathcal{A}}_{\text{MLR}}, \widehat{A}_{\text{RRR}}, \widehat{A}_{\text{OLS}}$

- Assume true (r_1, r_2, r_3) are known, N, P – fixed **(low-dim. setup)**
- Also assume $\mathbb{E}\left\|\epsilon_t\right\|_2^4 < \inf$, and that all roots of the matrix polynomial $A(z) = \mathbb{I}_N - A_1 z - \ldots - A_P z^P, z \in \mathbb{C}$ lie outside unit circle.

Then for $\text{method} \in \{\text{"MLR"}, \text{"RRR"}, \text{"OLS"}\}$

$$\sqrt{T}\left\{\text{vec}\left(\left(\widehat{\mathcal{A}}_{\text{method}}\right)_{(1)}\right) - \text{vec}\left(\mathcal{A}_{(1)}\right)\right\} \xrightarrow[T\to\infty]{D} N\left(0, \Sigma_{\text{method}}\right)$$

Where Σ_{MLR} is a function of $\mathcal{G}, U_1, U_2, U_3$, $\Sigma_{\text{OLS}}, \Sigma_{\text{RRR}}$ - of $A_1, \ldots, A_P$

Moreover, $\Sigma_{\text{MLR}} \preceq \Sigma_{\text{RRR}} \preceq \Sigma_{\text{OLS}}$

Alternating Least Squares Estimation

- $L(\mathcal{G}, U_1, U_2, U_3) = \frac{1}{T}\sum_{t=1}^{T} \left\| y_t - (\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3)_{(1)} x_t \right\|_2^2$
- L – convex w.r.t any of $\mathcal{G}$, U_1, U_2, and U_3 when the other three are fixed
- Hence, an ALS algorithm can be implemented. **Idea:**
  1. Initialize $\mathcal{A}^{(0)}$
  2. Perform HOSVD to obtain $U_1^{(0)}, U_2^{(0)}, U_3^{(0)}, \mathcal{G}^{(0)}$
  3. Update individually $U_1^{(k+1)}, U_2^{(k+1)}, U_3^{(k+1)}, \mathcal{G}^{(k+1)}$ (in that order), other 3 fixed
  4. When convergence reached, obtain $\widehat{\mathcal{A}}$
- Authors recommend to initialize $\mathcal{A}^{(0)} = \widehat{\mathcal{A}}_{\text{prelim}} + T^{-1/2}\mathcal{T}$, where $\widehat{\mathcal{A}}_{\text{prelim}}$ is $\widehat{\mathcal{A}}_{\mathrm{OLS}}$ for large T, $\widehat{\mathcal{A}}_{\mathrm{RRR}}$ for small T, and $vec(\mathcal{T}) \sim N(0, \mathbb{I}_{NNP})$. Global minimum is not guaranteed

ALS update equations

$$U_1^{(k+1)} \leftarrow \arg\min_{U_1} \sum_{t=1}^{T} \| y_t - \left(\left(x_t' \left(U_3^{(k)} \otimes U_2^{(k)} \right) \mathcal{G}_{(1)}^{(k)\prime} \right) \otimes I_N \right) \text{vec}\left(U_1 \right) \|_2^2$$

$$U_2^{(k+1)} \leftarrow \arg\min_{U_2} \sum_{t=1}^{T} \| y_t - U_1^{(k+1)} \mathcal{G}_{(1)}^{(k)} \left(\left(X_t U_3^{(k)} \right)' \otimes I_{r_2} \right) \text{vec}\left(U_2' \right) \|_2^2$$

$$U_3^{(k+1)} \leftarrow \arg\min_{U_3} \sum_{t=1}^{T} \| y_t - U_1^{(k+1)} \mathcal{G}_{(1)}^{(k)} \left(I_{r_3} \otimes \left(U_2^{(k+1)\prime} X_t \right) \right) \text{vec}\left(U_3 \right) \|_2^2$$

$$\mathcal{G}^{(k+1)} \leftarrow \arg\min_{\mathcal{G}} \sum_{t=1}^{T} \| y_t - \left(\left(\left(U_3^{(k+1)} \otimes U_2^{(k+1)} \right)' x_t \right)' \otimes U_1^{(k+1)} \right) \text{vec}\left(\mathcal{G}_{(1)} \right) \|_2^2$$

- **Remark:** Let $h(U_1, U_2, U_3, \mathcal{G}) = \text{vec}\left(\left(\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3 \right)_{(1)} x_t \right) =$ $= \text{vec}\left(U_1 \mathcal{G}_{(1)} \left(U_3 \otimes U_2 \right)' x_t \right)$. Consider ${}^{\partial h}/_{\partial \text{vec}(U_i)}$, ${}^{\partial h}/_{\partial \text{vec}(\mathcal{G})}$.

Table of Contents

Sparse Higher-Order Reduced-Rank VAR (SHORR)

- Same $L = \frac{1}{T}\sum_{t=1}^{T}\left\|y_t - (\mathcal{G} \times_1 U_1 \times_2 U_2 \times_3 U_3)_{(1)}\, x_t\right\|_2^2$
- Introduce regularization and all-orthogonality constraint:

$$\widehat{\mathcal{A}}_{\mathrm{SHORR}} \equiv [\![\widehat{\mathcal{G}}; \widehat{U}_1, \widehat{U}_2, \widehat{U}_3]\!] = \underset{\mathcal{G}, U_1, U_2, U_3}{\arg\min} \left\{L\left(\mathcal{G}, U_1, U_2, U_3\right) + \lambda \left\|U_3 \otimes U_2 \otimes U_1\right\|_1\right\} \quad \text{subject to } U_i' U_i = \mathbb{I}_{r_i} \text{ and}$$

$$\mathcal{G} \in \left\{\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3} : \left(\mathcal{G}_{(i)}\right)_{i=1}^{3} \text{ – row-orthogonal}\right\}$$

- Sparsity assumption: each column of U_i has at most s_i nonzero entries
- Under these and certain extra assumptions, non-asymptotic UB's for $\left\|\widehat{\mathcal{A}}_{\mathrm{SHORR}} - \mathcal{A}\right\|_{\mathrm{F}}$ and $T^{-1}\sum_{t=1}^{T}\left\|\left(\widehat{\mathcal{A}}_{\mathrm{SHORR}} - \mathcal{A}\right)_{(1)} \boldsymbol{x}_t\right\|_2^2$ were derived by the authors
- Difference with LASSO: $\mathcal{A}$ – not necessarily sparse

Developing an algorithm

- **Issue:** ℓ_1 regularization – non-smooth, orthogonality constraint – non-convex
- **Solution:** alternating direction method of multipliers (ADMM) algorithm (Boyd et al. 2011)
- **Idea:** assume a decomposition $\mathcal{G}_{(i)} = D_i V_i'$ exists, where $D_i \in \mathbb{R}^{r_i \times r_i}$, $V_i \in \mathbb{R}^{(r_1 r_2 r_3 / r_i) \times r_i}$, $V_i' V_i = \mathbb{I}_{r_i}$
- Augmented Lagrangian:

$$\mathcal{L}_\varrho\left(\mathcal{G}, \{U_i\}, \{D_i\}, \{V_i\}; \{\mathcal{C}_i\}\right) = L\left(\mathcal{G}, U_1, U_2, U_3\right) + \lambda \left\| U_3 \otimes U_2 \otimes U_1 \right\|_1$$

$$+ 2 \sum_{i=1}^{3} \varrho_i \left\langle (\mathcal{C}_i)_{(i)}, \mathcal{G}_{(i)} - D_i V_i' \right\rangle + \sum_{i=1}^{3} \varrho_i \left\| \mathcal{G}_{(i)} - D_i V_i' \right\|_{\mathrm{F}}^2$$

Where ϱ_i – regularization constants, $\mathcal{C}_i \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ – dual variables.

ADMM algorithm for SHORR

Initialize: $\mathcal{A}^{(0)}$

HOSVD: $\mathcal{A}^{(0)} \approx \mathcal{G}^{(0)} \times_1 \boldsymbol{U}_1^{(0)} \times_2 \boldsymbol{U}_2^{(0)} \times_3 \boldsymbol{U}_3^{(0)}$ with multilinear ranks (r_1, r_2, r_3).

repeat

$\quad \boldsymbol{U}_1^{(k+1)} \leftarrow \underset{\boldsymbol{U}_1'\boldsymbol{U}_1=\boldsymbol{I}_{r_1}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \boldsymbol{U}_1, \boldsymbol{U}_2^{(k)}, \boldsymbol{U}_3^{(k)}) + \lambda \|\boldsymbol{U}_1\|_1 \|\boldsymbol{U}_2^{(k)}\|_1 \|\boldsymbol{U}_3^{(k)}\|_1 \right\}$

$\quad \boldsymbol{U}_2^{(k+1)} \leftarrow \underset{\boldsymbol{U}_2'\boldsymbol{U}_2=\boldsymbol{I}_{r_2}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \boldsymbol{U}_1^{(k+1)}, \boldsymbol{U}_2, \boldsymbol{U}_3^{(k)}) + \lambda \|\boldsymbol{U}_1^{(k+1)}\|_1 \|\boldsymbol{U}_2\|_1 \|\boldsymbol{U}_3^{(k)}\|_1 \right\}$

$\quad \boldsymbol{U}_3^{(k+1)} \leftarrow \underset{\boldsymbol{U}_3'\boldsymbol{U}_3=\boldsymbol{I}_{r_3}}{\arg\min} \left\{ L(\mathcal{G}^{(k)}, \boldsymbol{U}_1^{(k+1)}, \boldsymbol{U}_2^{(k+1)}, \boldsymbol{U}_3) + \lambda \|\boldsymbol{U}_1^{(k+1)}\|_1 \|\boldsymbol{U}_2^{(k+1)}\|_1 \|\boldsymbol{U}_3\|_1 \right\}$

$\quad \mathcal{G}^{(k+1)} \leftarrow \arg\min \left\{ L(\mathcal{G}, \boldsymbol{U}_1^{(k+1)}, \boldsymbol{U}_2^{(k+1)}, \boldsymbol{U}_3^{(k+1)}) + \sum_{i=1}^3 \varrho_i \|\mathcal{G}_{(i)} - \boldsymbol{D}_i^{(k)} \boldsymbol{V}_i^{(k)\prime} + (\mathcal{C}_i^{(k)})_{(i)}\|_{\mathrm{F}}^2 \right\}$

$\quad$ **for** $i \in \{1, 2, 3\}$ **do**

$\qquad \boldsymbol{D}_i^{(k+1)} \leftarrow \underset{\boldsymbol{D}_i=\mathrm{diag}(\boldsymbol{d}_i)}{\arg\min} \|\mathcal{G}_{(i)}^{(k+1)} - \boldsymbol{D}_i \boldsymbol{V}_i^{(k)\prime} + (\mathcal{C}_i^{(k)})_{(i)}\|_{\mathrm{F}}^2$

$\qquad \boldsymbol{V}_i^{(k+1)} \leftarrow \underset{\boldsymbol{V}_i'\boldsymbol{V}_i=\boldsymbol{I}_{r_i}}{\arg\min} \|\mathcal{G}_{(i)}^{(k+1)} - \boldsymbol{D}_i^{(k+1)} \boldsymbol{V}_i' + (\mathcal{C}_i^{(k)})_{(i)}\|_{\mathrm{F}}^2$

$\qquad (\mathcal{C}_i^{(k+1)})_{(i)} \leftarrow (\mathcal{C}_i^{(k)})_{(i)} + \mathcal{G}_{(i)}^{(k+1)} - \boldsymbol{D}_i^{(k+1)} \boldsymbol{V}_i^{(k+1)\prime}$

$\quad$ **end for**

$\quad \mathcal{A}^{(k+1)} \leftarrow \mathcal{G}^{(k+1)} \times_1 \boldsymbol{U}_1^{(k+1)} \times_2 \boldsymbol{U}_2^{(k+1)} \times_3 \boldsymbol{U}_3^{(k+1)}$

until convergence

Updating $\mathcal{G}$, D_i, V_i – LS problem, updating U_i – very complicated

Updating U_i

- Original problem:

$$U_i = \arg\min_B \left\{ n^{-1} \|y - X\,\mathrm{vec}(B)\|_2^2 + \lambda \|B\|_1 \right\} \quad \text{s.t.} \quad B'B = \mathbb{I}$$

- Idea: separate orthogonality and regularization

$$\min_B \left\{ n^{-1} \|y - X\,\mathrm{vec}(B)\|_2^2 + \lambda \|W\|_1 \right\} \quad \text{s.t.} \quad B'B = \mathbb{I},\ B = W$$

- Augmented Lagrangian (M – dual variable)

$$n^{-1} \|y - X\,\mathrm{vec}(B)\|_2^2 + \lambda \|W\|_1 + 2\kappa \langle M, B - W \rangle + \kappa \|B - W\|_{\mathrm{F}}^2$$

- Apply ADMM to find $B = W = U_i$:

1: Initialize: $\boldsymbol{B}^{(0)} = \boldsymbol{W}^{(0)}, \boldsymbol{M}^{(0)} = \mathbf{0}$
2: **repeat**
3: $\boldsymbol{B}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{B}'\boldsymbol{B}=\boldsymbol{I}} \left\{ n^{-1} \|\boldsymbol{y} - \boldsymbol{X}\mathrm{vec}(\boldsymbol{B})\|_2^2 + \kappa \|\boldsymbol{B} - \boldsymbol{W}^{(k)} + \boldsymbol{M}^{(k)}\|_{\mathrm{F}}^2 \right\}$
4: $\boldsymbol{W}^{(k+1)} \leftarrow \arg\min_{\boldsymbol{W}} \left\{ \kappa \|\boldsymbol{B}^{(k+1)} - \boldsymbol{W} + \boldsymbol{M}^{(k)}\|_{\mathrm{F}}^2 + \lambda \|\boldsymbol{W}\|_1 \right\}$
5: $\boldsymbol{M}^{(k+1)} \leftarrow \boldsymbol{M}^{(k)} + \boldsymbol{B}^{(k+1)} - \boldsymbol{W}^{(k+1)}$
6: **until convergence**

Convergence and initialization

- Under certain conditions on $\mathcal{L}_\varrho$, algorithm converges to local minimum of our objective function:

$$L\left(\mathcal{G}, U_1, U_2, U_3\right) + \lambda \left\| U_3 \otimes U_2 \otimes U_1 \right\|_1$$

- Authors recommend to choose $\mathcal{A}^{(0)} = \widehat{\mathcal{A}}_{\mathrm{NN}} + (NP/T)^{1/2}\mathcal{T}$, where:
  - $vec(\mathcal{T}) \sim N(0, \mathbb{I}_{NNP}),\ \|\mathcal{T}\|_{\mathrm{F}} = O_p(1)$
  - $\widehat{\mathcal{A}}_{\mathrm{NN}} = \arg\min \frac{1}{T}\sum_{t=1}^{T} \left\| y_t - \mathcal{A}_{(1)} x_t \right\|_2^2 + \lambda \left\| \mathcal{A}_{(1)} \right\|_*$
  - $\left\| \mathcal{A}_{(1)} \right\|_*$ – nuclear norm, or sum of all singular values of $\mathcal{A}_{(1)}$

Rank selection

- Let $\widehat{\mathcal{A}}$ be a consistent initial estimator of $\mathcal{A}$ (e.g., $\widehat{\mathcal{A}}_{\mathrm{NN}}$)
- Ridge-type ratio estimator (Xia, Xu, and Zhu 2015):

$$\widehat{r}_i = \arg\min_{1 \le j \le p_i - 1} \frac{\sigma_{j+1}\left(\widehat{\mathcal{A}}_{(i)}\right) + c}{\sigma_j\left(\widehat{\mathcal{A}}_{(i)}\right) + c} \quad \text{where } p_1 = p_2 = N, p_3 = P$$

- Denote $\zeta_i = \frac{1}{\sigma_{r_i}\left(\mathcal{A}_{(i)}\right)} \cdot \max_{1 \le j < r_i} \frac{\sigma_j\left(\mathcal{A}_{(i)}\right)}{\sigma_{j+1}\left(\mathcal{A}_{(i)}\right)}$
- $c > 0$ is chosen such that:
  1. $\|\widehat{\mathcal{A}} - \mathcal{A}\|_{\mathrm{F}} = o_p(c)$
  2. $\max_{1 \le i \le 3} \zeta_i = o(1/c)$
- Authors recommend $c = \sqrt{NP \ln(T)/(10T)}$

Table of Contents

Rank selection consistency – simulation setup

- $(N, P) = (10, 5)$, $(r_1, r_2, r_3) = (3, 3, 3)$, and $\epsilon_t \overset{\text{iid}}{\sim} N(0, \mathbb{I}_N)$
- $\mathcal{G}$ – a diagonal cube with $(\mathcal{G}_{111}, \mathcal{G}_{222}, \mathcal{G}_{333}) = (2, 2, 2)$ (case a), $(4, 3, 2)$ (case b), $(1, 1, 1)$ (case c), or $(2, 1, 0.5)$ (case d).
- Then nonzero singular values of $\mathcal{A}_{(i)}$ are $\mathcal{G}_{111}, \mathcal{G}_{222}$, and $\mathcal{G}_{333}$
- Generate U_i's as the first r_i left singular vectors of Gaussian random matrices while ensuring the stationarity.
- $c = \sqrt{NP\ln(T)/(10T)}$ was used
- 1000 replications for each $T \in \{50, 100, \ldots, 400\}$

Rank selection consistency – simulation results

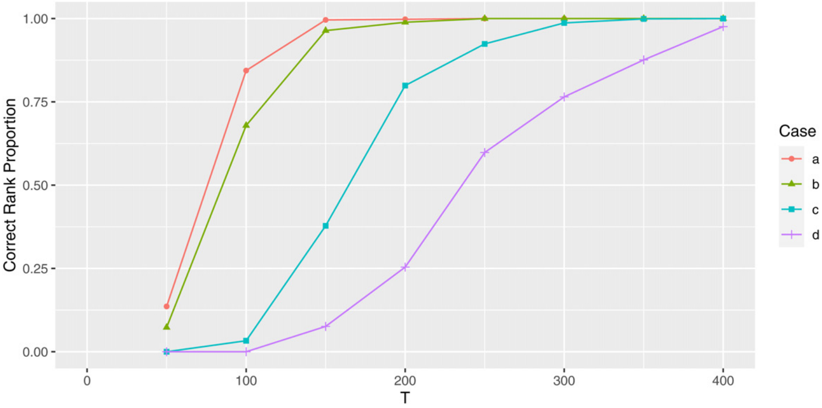

Figure 5: Proportion of correct rank selection when the nonzero singular values of each $\mathscr{A}_{(i)}$ are $(2, 2, 2)$ (case a), $(4, 3, 2)$ (case b), $(1, 1, 1)$ (case c), or $(2, 1, 0.5)$ (case d)

OLS vs. RRR vs. MLR – setup

- N, P, U_i – same. Number of replications – same
- $r_1 = r_2 = 3$, and $r_3 \in \{2, 3, 4\}$
- Generate $\mathcal{G}$ by scaling a random iid Gaussian tensor s.t. $\min_{1 \le i \le 3} \sigma_{r_i}\left(\mathcal{G}_{(i))}\right) = 1$

OLS vs. RRR vs. MLR – results

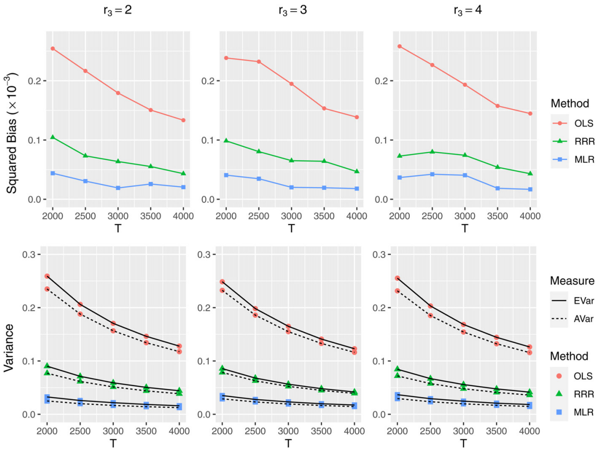

Figure 6: Squared bias, empirical variance (EVar) and asymptotic variance (AVar) for $\widehat{\mathcal{A}}_{\text{OLS}}\widehat{\mathcal{A}}_{\text{RRR}}$, and $\widehat{\mathcal{A}}_{\text{MLR}}$ under various multilinear ranks.

Comparison with existing methods – setup

- $(N, P) = (10, 5)$ (case a), $(15, 8)$ (case b)
- $(r_1, r_2, r_3) = (3, 3, 3), (s_1, s_2, s_3) = (3, 3, 2)$
- For case a, $\mathcal{G}$ and U_i 's are generated by the same methods as in RRR vs. MLR vs. OLS
- For case b, zeros rows are added below the U_i 's in case a
- In both cases, $\|\mathcal{A}\|_0 = 500$. Hence, $\mathcal{A}$ is not sparse in case a, but is sparse in case b

Comparison with existing methods – results

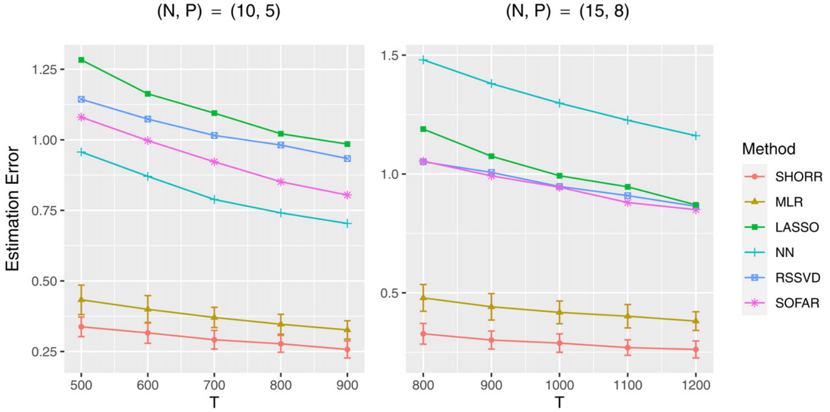

Figure 7: Plots of the estimation error $\|\widehat{\mathcal{A}} - \mathcal{A}\|_{\mathrm{F}}$ against T for six estimation methods under two settings of (N, P).

- **Issue:** NN performes the worst in the sparse case

Modeling real data

- **Data:** $N = 40$ quarterly macroeconomic sequences of the United States from 1959 to 2007 (from Koop, 2013)
- Lag $P = 4$ for the VAR model is suggested by Koop (2013). $N >> P \Rightarrow$ penalty on U_3 is not needed. New penalty – $\|U_2 \otimes U_1\|_1$
- $(r_1, r_2, r_3) = (4, 3, 2)$ are selected by the ridge-type ratio estimator
- Tuning parameter λ is selected by BIC

	Unregularized methods				Regularized methods				
Criterion	OLS	RRR	DFM	MLR	SHORR	LASSO	NN	RSSVD	SOFAR
ℓ_2 norm	20.16	13.31	6.36	**5.81**	5.35	6.72	8.16	6.33	6.28
ℓ_∞ norm	8.32	4.55	2.85	**2.56**	2.44	3.06	3.36	3.02	3.02

Figure 8: Forecasting errors for different methods

- Again, NN performs the worst

Table of Contents

Conclusions, issues and improvements

- The novelty of the approach is in its ability to jointly enforce three different reduced-rank structures at the same time
- order P of VAR is not estimated. Possible solution: IC-based selection
- Selecting r_i is dependent on initialization $\widehat{\mathcal{A}_0}$, derived from other methods and which can even be consistent but biased/inefficient and hence make low-T estimation incorrect.
  IC-based selection or hypothesis testing – problematic (3 parameters, too many combinations)
- NN estimator perform the worst in both simulations and real data
  Possible solution: use other estimators at initialization $\widehat{\mathcal{A}_0}$ (e.g., SOFAR or RSSVD)
- Despite those limitations, MLR and SHORR perform the best on real macroeconomic data