Multivariate Eigenvalue Problem: Algebraic Theory and Power Method

Moody T. Chu
and
J. Loren Watterson

Department of Mathematics
North Carolina State University
Raleigh, NC 27695-8205

Outline

- Statistics Background
- Homotopy Method and Cardinality
- Power Method and Convergence
- Other Numerical Methods

Multivariate Eigenvalue Problem (MEP)

Find real scalars $\lambda_1, \ldots, \lambda_m$ and a real vector $x \in R^n$ such that

$$\begin{aligned} Ax &= \Lambda x \\ ||x_i|| &= 1, \; i = 1, \ldots, m, \end{aligned}$$

where

- $A \in R^{n \times n}$ is SPD, and is block partitioned into

$$A = \begin{bmatrix} A_{11} & A_{12} & \ldots & A_{1m} \\ A_{21} & A_{22} & \ldots & A_{2m} \\ \vdots & \vdots & & \vdots \\ A_{m1} & A_{m2} & \ldots & A_{mm} \end{bmatrix},$$

  - $A_{ii} \in R^{n_i \times n_i}$.
  - $\Sigma_{i=1}^m n_i = n$.
- $\Lambda = \text{diag}\{\lambda_1 I^{[n_1]}, \ldots, \lambda_m I^{[n_m]}\}$,
  - $I^{[n_i]}$:=The identity matrix of dimension n_i.
- $x = [x_1^T, \ldots, x_m^T]^T$,
  - $x_i \in R^{n_i}$.

Relation to Other Problems

- MEP is a classical symmetric eigenvalue problem when $m = 1$.
- MEP is fundamentally different from the so called multiparameter eigenvalue problem.
- MEP originates from the determination of canonical correlation coefficients for multivariate statistics.

Statistics Background

- $\tilde{X} := (\mathcal{X}_1, \ldots, \mathcal{X}_n)$ denotes an n-dimensional random variable with a certain distribution function.
- Row vectors $[x_{\xi 1}, \ldots, x_{\xi n}]$, $\xi = 1, \ldots, k$ denote a random sample of size k for $\tilde{X}$,
  - $X := [x_{\xi i}]$ is the $k \times n$ sample matrix.
- Assume
  - Sample mean $\mu_i = 0$ for each $\mathcal{X}_i$.
  - No degenerate component and no linear dependence among $\mathcal{X}_i$.

  Then
  - Matrix $\Delta := X^T X$ represents the covariance matrix of the random sample X.
  - Δ is SPD.

Regroup of Variables

- Divide variables into groups $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_m)$,
  - $\tilde{X}_i$ is an n_i-dimensional random variable.
  - Correspondingly, $\Delta = [\Delta_{ij}]$ and $X = [X_1, \ldots, X_m]$.
  - Δ_{ii} is the covariance matrix of the $k \times n_i$ sample block X_i.
- Combining all n_i variables in $\tilde{X}_i$ linearly into a single new variable $\mathcal{Z}_i$ through coefficients $b_i \in R^{n_i}$,
  - Sample matrix X is transformed into $Z := XB := [Z_1, \ldots, Z_m]$ where

$$B := \begin{bmatrix} b_1 & 0 & \ldots & 0 \\ 0 & b_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \ldots & b_m \end{bmatrix}.$$

  - Covariance matrix corresponding to Z is given by

$$\Omega := Z^T Z = B^T \Delta B.$$

An Example ($m = 2$)

- Desire to use $\mathcal{Z}_1$ to predict $\mathcal{Z}_2$,
  - Need to find b_1 and b_2 so that the correlation coefficient between Z_1 and Z_2 is as large as possible.
- Covariance matrix of Z is

$$\Omega = \begin{bmatrix} b_1^T \Delta_{11} b_1 & b_1^T \Delta_{12} b_2 \\ b_2^T \Delta_{21} b_1 & b_2^T \Delta_{11} b_2 \end{bmatrix}.$$

- Correlation coefficient ρ to be maximized is

$$\rho = \frac{b_1^T \Delta_{12} b_2}{\sqrt{b_1^T \Delta_{11} b_1}\sqrt{b_2^T \Delta_{11} b_2}}.$$

- Normalize the variances of Z_1 and Z_2 to unity. The maximal correlation problem (MCP) becomes

$$\begin{aligned} \text{Maximize} \quad & b^T \Delta b \\ \text{Subject to} \quad & b_i^T \Delta_{ii} b_i = 1, \text{ for } i = 1, 2 \end{aligned}$$

  where $b := [b_1^T, b_2^T]^T$.

Change MCP to MEP

- Each Δ_{ii} is SPD, so the Cholesky decomposition $\Delta_{ii} = T_i^T T_i$ exists.
- Define

$$\begin{aligned} T &:= \mathrm{diag}\{T_1, T_2\} \\ x &:= Tb := [x_1^T, x_2^T]^T, \\ A &:= T^{-T} \Delta T^{-1}. \end{aligned}$$

  MCP is transformed into

$$\begin{array}{rl} \text{Maximize} & x^T A x \\ \text{Subject to} & x_i^T x_i = 1, \ i = 1, 2. \end{array}$$

- MEP follows from differentiating the Lagrangian function

$$\phi(x, \lambda_1, \lambda_2) := x^T A x - \sum_{i=1}^{2} \lambda_i (x_i^T x_i - 1)$$

  with λ_1 and λ_2 as the Lagrange multipliers.

Maximal Correlation When $m > 2$

- Heuristic observation:
  - The closer to orthogonality any two sample vectors Z_i, Z_j are, the closer to zero the correlation coefficients will be.
  - The more *similar* the vectors $Z_1, \ldots, Z_m$ are to each other, the more closely the correlation coefficients will approach unity.
- MCP requires the sum of all off-diagonal elements of Ω be maximized subject to the condition that the diagonal elements of Ω be unity.
- Same procedure as fo $m = 2$ to leads MCP to

$$\begin{array}{rl} \text{Maximize} & x^T A x \\ \text{Subject to} & x_i^T x_i = 1, \ i = 1, m, \end{array}$$

and, then, to MEP.

What Are the Difficulties?

- MEP represents a non-linear algebraic system in $n + m$ unknowns.
- When $m = 1$,
  - Counting multiplicity, there are exactly n eigenvalues.
  - Counting negative signs, there are exactly $2n$ eigenvectors.
- When $m > 1$, no discussion on the cardinality of solutions to MEP,
  - Characteristic polynomial is not applicable to MEP.
- How to compute a multivariate eigenvalue and the corresponding eigenvector?
  - Horst proposed an iterative procedure without a proof.
  - Not heard of any other numerical method since Horst

Homotopy Method and Cardinality

- MEP is a nonlinear system:

$$F(x,\Lambda) := \begin{bmatrix} \Lambda x - Ax \\ \frac{x_1^T x_1 - 1}{2} \\ \vdots \\ \frac{x_m^T x_m - 1}{2} \end{bmatrix} = 0,$$

  – $F : R^n \times R^m \longrightarrow R^n \times R^m$.
- The simple MEP:

$$\begin{aligned} Dx &= \Lambda x \\ ||x_i|| &= 1,\ i = 1, \ldots, m. \end{aligned}$$

  with $D = \mathrm{diag}\{d_1^{(1)}, \ldots, d_{n_1}^{(1)}, \ldots, d_1^{(m)}, \ldots, d_{n_m}^{(m)}\}$ has exactly $\Pi_{i=1}^m 2n_i$ solutions,
  – For $i = 1, \ldots, m$,

$$\begin{aligned} \lambda_i &= d_{j_i}^{(i)}, \\ x_i &= \pm e_{j_i}^{[n_i]}, \end{aligned}$$

    * $j_i = 1, \ldots, n_i$.
    * $e_s^{[t]} := s^{th}$ column of $I^{[t]}$.

Homotopy Function

- Define $H : R^n \times R^m \times R \longrightarrow R^n \times R^m$ by

$$H(x, \Lambda, t; D) := \begin{bmatrix} \Lambda x - [D + t(A - D)]x \\ \frac{x_1^T x_1 - 1}{2} \\ \vdots \\ \frac{x_m^T x_m - 1}{2} \end{bmatrix},$$

  - D is a diagonal matrix whose elements will be specified.
- Basic Ideas:
  - The set $\Gamma := \{(x, \Lambda, t) | H(x, \Lambda, t) = 0\}$ is a one dimensional smooth submanifold in $R^n \times R^m \times R$.
  - No homotopy curve will escape to infinity or turn back.

Major Theorem

For each $(x, \Lambda, t) \in R^n \times R^m \times R$ such that $H(x, \Lambda, t) = 0$, the matrix

$$D_{(x,\Lambda)}H = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{B}^T & 0 \end{bmatrix}$$

where

$$\mathcal{A} = \mathcal{A}(\Lambda, t, D) := \Lambda - (D + t(A - D)).$$

and

$$\mathcal{B} := \begin{bmatrix} x_1 & 0 & \dots & 0 \\ 0 & x_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & & \dots & x_m \end{bmatrix}$$

has rank $n + m$.

- If $H(x, \Lambda, t) = 0$, then $\mathcal{A}x = 0$.
- None of the m columns of $\mathcal{B}$ can be in the range of $\mathcal{A}$.
- $\mathcal{A}$ is at least of rank $n - m$.

Auxiliary Lemma

The set of D such that the matrix $\Lambda - (D + t(A - D))$ is of rank less than $n - m$ for some Λ and some $t \in (0, 1)$ is of measure zero.

- If the rank of $\mathcal{A}$ is less than $n - m$, then one of the diagonal blocks

  $$\mathcal{A}_{ii} = \lambda_i I^{[n_i]} - (1 - t)\mathrm{diag}(d_1^{(i)}, \ldots, d_{n_i}^{(i)}) - tA_{ii}.$$

  must be rank deficient by at least two.
- For some $\tau \in (0, \infty)$ the matrix $\tau D^{(i)} + A_{ii}$ has an eigenvalue with multiplicity at least two,
  – $D^{(i)} := \mathrm{diag}(d_1^{(i)}, \ldots, d_{n_i}^{(i)})$.
- For any symmetric matrix M, the set

  $$E_r := \{\text{diagonal } D | M + D \text{ rank deficient by } r\}$$

  has dimension $\leq \dim(M) - 3$ for $r \geq 2$.
- The set

  $$\bigcup_{\tau \in (0,\infty)} \{D^{(i)} | \tau D^{(i)} + A_{ii} \text{ has multiple eigenvalues}\}$$

  is of dimension at most $n_i - 1$.

Homotopy Curves

For $i = 1, \ldots, m$, the solution to the initial value problem

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ \vdots \\ x_m \\ \lambda_1 \\ \vdots \\ \lambda_m \end{bmatrix} = \begin{bmatrix} \mathcal{A} & \mathcal{B} \\ \mathcal{B}^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} -\mathcal{C} \\ 0 \end{bmatrix}$$

$$x_i(0) = \pm e_{j_i}^{[n_i]}$$

$$\lambda_i(0) = d_{j_i}^{(i)}$$

is a curve in $R^n \times R^m$ that extends from $t = 0$ to $t = 1$.

- The (MEP) has exactly $\Pi_{i=1}^{m} 2n_i$ solutions.
- The positive definiteness of the matrix A is not needed in the proof.

Horst's Algorithm

- Block form:
  Given $x^{(0)} = (x_1^{(0)^T}, \ldots, x_m^{(0)^T})^T$ with $||x_i^{(0)}|| = 1$, do
  for $k = 1, 2, \ldots$
  for $i = 1, \ldots, m$

$$
\begin{aligned}
y_i^{(k)} &:= \sum_{j=1}^{m} A_{ij} x_j^{(k)}, \\
\lambda_i^{(k)} &:= ||y_i^{(k)}||, \\
x_i^{(k+1)} &:= \frac{y_i^{(k)}}{\lambda_i^{(k)}}.
\end{aligned}
$$

  end
  end
  – Define $x_i^{(k+1)} := x_i^{(k)}$ in case $||y_i^{(k)}|| = 0$.
- Compact form:

$$
Ax^{(k)} = \Lambda^{(k)} x^{(k+1)},
$$

  – $x^{(k)} := [x_1^{(k)^T}, \ldots, x_m^{(k)^T}]^T$
  – $\Lambda^{(k)} := \mathrm{diag}\{\lambda_1^{(k)} I^{[n_1]}, \ldots, \lambda_m^{(k)} I^{[n_m]}\}$.

Is This a Power Method?

- Horst's iteration may be viewed as a generalization of the classical power method.
- The convergence property of this method is not nearly obvious,
  - Without the positive definiteness, the method may fail to converge.
  - A limit point of the method may depend upon the starting point.
  - MCP may have multiple local solutions.

Convergence!

- The sequence $\{r(x^{(k)})\}$ where
$$r(x) := x^T A x.$$
is a monotonically increasing sequence and converges.
  – Key fact,
$$\begin{aligned} & r(x^{(k+1)}) - r(x^{(k)}) \\ = \; & (x^{(k+1)} - x^{(k)})^T (A + \Lambda^{(k)})(x^{(k+1)} - x^{(k)}). \end{aligned}$$
- The residual $\{\delta x^{(k)}\}$ where
$$\delta(x) := Ax - \Lambda x$$
converges to zero.
  – Key fact,
$$\delta(x^{(k)}) = \Lambda^{(k)}(x^{(k+1)} - x^{(k)}).$$

Convergence?

- Is this enough to prove convergence of $\{x^{(k)}\}$?

$$r(x^{(k+1)}) - r(x^{(k)}) \geq \kappa ||x^{(k+1)} - x^{(k)}||^2$$

- The sequence $\{x^{(k)}\}$ does have have cluster point(s),
  - Every cluster point x^* solves MEP with eigenvalues $\lambda_i^* := || \Sigma_{j=1}^m A_{ij} x_j^* ||$.
- A lemma from real analysis,
  - Let $\{a_k\}$ be a bounded sequence of real numbers with the proper $|a_{k+1} - a_k| \longrightarrow 0$ as $k \longrightarrow \infty$. If there are only finitely many limit points for the sequence, then $\{a_k\}$ converges to a unique limit point.
- The fact of finite number of solutions of MEP proves that
  - The sequence $\{\Lambda^{(k)}\}$ converges.
  - The sequence $\{x^{(k)}\}$ converges.

Dependence on Starting Points

Consider the positive definite matrix

$$A = \begin{bmatrix} 4.3299 & 2.3230 & -1.3711 & -0.0084 & -0.7414 \\ 2.3230 & 3.1181 & 1.0959 & 0.1285 & 0.0727 \\ -1.3711 & 1.0959 & 6.4920 & -1.9883 & -0.1878 \\ -0.0084 & 0.1285 & -1.9883 & 2.4591 & 1.8463 \\ -0.7414 & 0.0727 & -0.1878 & 1.8463 & 5.8875 \end{bmatrix}$$

with $m = 2$, $n_1 = 2$ and $n_2 = 3$.

- If $x^{(0)} = [0.9777, 0.2098, 0.5066, 0.5069, 0.6975]^T$, then

$$\begin{aligned} x^* &= [0.9357, 0.3528, -0.9341, 0.3508, 0.0667]^T, \\ \lambda_1^* &= 6.5186, \\ \lambda_2^* &= 8.2116. \end{aligned}$$

- If $x^{(0)} = [0.7914, 0.6114, 0.4753, 0.2517, -0.8431]^T$, then

$$\begin{aligned} x^{**} &= [0.7166, 0.6975, 0.5654, -0.4327, -0.7022]^T, \\ \lambda_1^{**} &= 6.2405, \\ \lambda_2^{**} &= 7.8607. \end{aligned}$$

Random Test

- Approximately 60% randomly generated starting points converge to x^* while all the remaining converge to x^{**},
  - Out of the 24 solutions, there are *two* local maxima to the maximal correlation problem.
  - Horst's algorithm has a substantial possibility of not converging to the absolute maximal correlation.

Multivariate Shifting

- $(A - \Gamma)x = \Lambda x$ if and only if $Ax = (\Gamma + \Lambda)x$,
  - Shifting is a possible strategy to find other solution of MEP.
- How do limit points depend on the starting value and on the shift parameters?
- For what Γ will the matrix $A - \Gamma$ become positive semi-definite?

Gauss-Seidel Algorithm

- Block form:
  Given $x^{(0)} = ({x_1^{(0)}}^T, \ldots, {x_m^{(0)}}^T)^T$ with $||x_i^{(0)}|| = 1$, do
  for $k = 1, 2, \ldots$
  for $i = 1, \ldots, m$

$$
\begin{aligned}
y_i^{(k)} &:= \sum_{j=1}^{i-1} A_{ij} x_j^{(k+1)} + \sum_{j=i}^{m} A_{ij} x_j^{(k)} \\
\lambda_i^{(k)} &:= ||y_i^{(k)}||, \\
x_i^{(k+1)} &:= \frac{y_i^{(k)}}{\lambda_i^{(k)}}.
\end{aligned}
$$

  end
  end
- Compact form:

$$(D + U)x^{(k)} = (\Lambda^{(k)} - U^T)x^{(k+1)},$$

$- A = D + U^T + U.$

SOR algorithm

- Block form:

  Given $x^{(0)} = ({x_1^{(0)}}^T, \ldots, {x_m^{(0)}}^T)^T$ with $||x_i^{(0)}|| = 1$, do

  for $k = 1, 2, \ldots$

  for $i = 1, \ldots, m$

$$\begin{aligned}
\overline{y}_i^{(k)} &:= \sum_{j=1}^{i-1} A_{ij} x_j^{(k+1)} + \sum_{j=i}^{m} A_{ij} x_j^{(k)} \\
\xi_i^{(k)} &:= ||\overline{y}_i^{(k)}||, \\
\overline{z}_i^{(k+1)} &:= \frac{\overline{y}_i^{(k)}}{\xi_i^{(k)}}. \\
y_i^{(k)} &:= \omega_i \overline{z}_i^{(k+1)} + (1 - \omega_i) x_i^{(k)} \\
\lambda_i^{(k)} &:= ||y_i^{(k)}||, \\
x_i^{(k+1)} &:= \frac{y_i^{(k)}}{\lambda_i^{(k)}}.
\end{aligned}$$

  end

  end

  – Relaxation parameters ω_i may be different.

  – The scaling may be done differently.

- Compact form:

$$[(I - \Omega)\Xi^{(k)} + \Omega(D + U)]x^{(k)} = (\Xi^{(k)}\Lambda^{(k)} - \Omega U^T)x^{(k+1)},$$

  – $\Xi^{(k)} := \text{diag}\{\xi_1^{(k)} I^{[n_1]}, \ldots, \xi_m^{(k)} I^{[n_m]}\}$.

  – $\Omega := \text{diag}\{\omega_1 I^{[n_1]} \ldots, \omega_m I^{[n_m]}\}$.

Future Research

A partial list of problems includes

- Proof of convergence,
  - This can be done.
- Rate of convergence.
- Acceleration of convergence.