Sensitivity Analysis for the Problem of Matrix Joint Diagonalization

Bijan Afsari

bijan@math.umd.edu

AMSC Department

Outline

- Introduction and the problem of Exact Joint Diagonalization (EJD)
- A motivating example
- (Approximate) Joint Diagonalization and its cost functions
- A Differential Equation on the GL(n) Group for Joint Diagonalization
- Sensitivity Analysis

Introduction and Motivation

- Consider the set $\{C_i\}_{i=1}^N$ of N symmetric $n \times n$ matrices which are of this form:

(1) $$C_i = A\Lambda_i A^T, \quad 1 \le i \le N$$

where A is a non-singular $n \times n$ matrix and Λ_i are diagonal. Denote j^{th} diagonal element of Λ_i by λ_{ij}.

- Our goal is to find A just by observing C_is
- Problem: assume that we observe the set $\{C_i\}_{i=1}^N$ and we only know that Λ_i's are diagonal, under what conditions can we find A based only on C_i's.
- Only based on diagonally of Λ_i's can not find A beyond permutation and diagonal scaling factors:

(2) $$C_i = A\Lambda_i A^T = \underbrace{A\Lambda^{-1}\Pi^T}_{\text{new}A}\underbrace{\Pi\Lambda\Lambda_i\Lambda\Pi^T}_{\text{new}\Lambda_i}(A\Lambda^{-1}\Pi^T)^T = \hat{A}\hat{\Lambda}_i\hat{A}^T$$

where Λ is a non-singular diagonal matrix and Π is a permutation matrix.

Introduction ...

- Let permutation and scaling be the only indeterminacies, then by fining a non-singular matrix B such that all $BC_iB^T, 1 \leq i \leq N$ are diagonal we can essentially find A, i.e. $A = B^{-1}$ up to scaling and permutation.
- This means that restoring diagonality results in identification.
- Exact Joint Diagonalization (EJD): Find a non-singular matrix B such that all BC_iB^T are diagonal. We say that EJD has a unique solution if scaling and permutation are the only ambiguities in the solution.

How many matrices are needed?Let's count

- We are looking for $A_{n\times n}$ only up to a permutation and scaling by diagonal matrices. So we can assume that the diagonal elements of A are 1. So we have $n^2 - n$ unknowns.
- A symmetric $n \times n$ matrix has $\frac{n(n+1)}{2}$ and a diagonal $n \times n$ matrix has n degrees of freedom.

(3)
$$\underbrace{C_i}_{N\times\frac{n(n+1)}{2}} = \underbrace{A}_{=n^2-n+} \underbrace{\Lambda_i}_{N\times n} A^T, \quad 1 \leq i \leq N$$

(4)
$$\frac{Nn(n+1)}{2} = n^2 - n + nN \Rightarrow N = 2$$

How many matrices are needed?...

- If no degeneracy happens we need $N = 2$ matrices independent of n. Degeneracy means that in the matrix:

(5)
$$L = \begin{bmatrix} \lambda_{11} & \dots & \lambda_{1n} \\ \lambda_{21} & \dots & \lambda_{2n} \\ \dots & \dots & \dots \\ \lambda_{N1} & \dots & \lambda_{Nn} \end{bmatrix}$$

at least two columns are multiple of each other. If $N = 1$ that would be the case.

- Theorem: EJD has a unique solution if and only if no two columns of L are co-linear.
- Sketch of the proof: if C_1 is non-singular then:

$$C_i C_1^{-1} = A\Lambda_i A^T (A^{-1})^T \Lambda_1^{-1} A^{-1} = A\Lambda_i \Lambda_1^{-1} A^{-1}, \;\; 2 \leq i \leq N$$

From these(the eigen decompositions of $C_i C_1^{-1}$) A cannot be found uniquely if and only if there are two columns j and k such that $\frac{\lambda_{ij}}{\lambda_{ik}} = \frac{\lambda_{1j}}{\lambda_{1k}}, \forall 1 < i \leq N$. Which is equivalent to the conditions of the theorem. If all C_i's are singular a modified argument works.

- Let us measure the co-linearity of the columns of L by modulus of uniqueness $|\rho|$ (with $|\rho| \leq 1$, and $|\rho| < 1 \Leftrightarrow$ uniqeness):

(6)
$$|\rho| = \max_{k,l} \frac{|\sum_{i=1}^N \lambda_{ik}\lambda_{il}|}{(\sum_{i=1}^N \lambda_{il}^2)^{\frac{1}{2}} (\sum_{i=1}^N \lambda_{ik}^2)^{\frac{1}{2}}}, \quad 1 \leq k \neq l \leq N$$

A motivating example

- Let $\vec{\mathbf{s}}(t)$ be an unknown n-dimensional vector random non-stationary process. Such that each component of the vector $\vec{\mathbf{s}}$ are uncorrelated process of zero mean, i.e. its correlation matrix is diagonal through time. Suppose this process is mixed by a matrix $A_{n\times n}$ and we observe the mixture:

$$\vec{\mathbf{x}}(t) = A\vec{\mathbf{s}}(t) \tag{7}$$

  Then for the correlation matrix of this process we have that:

$$C(t) = A\Lambda_{\mathbf{s}}(t)A^T \tag{8}$$

  where $\Lambda_{\mathbf{s}}(t)$ is the diagonal correlation of $\vec{\mathbf{s}}(t)$ which is unknown.

- We can sample the correlation matrix of the observed mixture and if we can find instances of times $t_1, ..., t_N$ for which $\{C(t_i)\}_{i=1}^N$ the EJD has a unique solution, then we can find A uniquely.

- The non-stationarity assumption is crucial, otherwise $C(t)$ would remain constant and we only have one matrix.

- Example: n sources in your brain emitting signals and n sensors on your scalp collect the mixture signals from the sources in the brain. Can we recover the source signals without access to the mixing process?

Joint Diagonalization

- (Approximate Joint Diagonalization (JD))In practice we only have:

  (9) $$C_i \approx A\Lambda_i A^T$$

  hence we approximately jointly diagonalize all of them, i.e. find B such that all BC_iB^Ts are "as diagonal as possible".

- We hope that after this $B \approx A^{-1}$. If little error in C_i result in large deviation in B from A^{-1} then the problem is sensitive. The question is what factors make the problem sensitive?

- Sensitivity is closely related to uniqueness as measured by $|\rho|$. In fact if $|\rho| = 1$, then sensitivity is infinity, since even in the case no error in the model there are non-unique solutions different from A^{-1}.

Cost functions for Joint Diagonalization

- Joint diagonalization as minimization of a cost function defined on a suitable Lie group.
- Historically JD was first considered for a case where A (or B) are assumed to be orthogonal.

Orthogonal Joint Diagonalization

- Orthogonal JD: If assume that A is orthogonal then we have $\frac{n(n-1)}{2}$ unknowns and we need $N = 1$ matrix (say C_1) if C_1 has distinct eigen values. $C_1 = A\Lambda_1 A^T$ is the eigen decomposition of C_1. To combat noise and consider the approximate model we look for $\Theta \in \mathrm{SO}(n)$ that minimizes:

$$J_1(\Theta) = \sum_{i=1}^{n} \left\| \Theta C_i \Theta^T - \mathrm{diag}(\Theta C_i \Theta^T) \right\|_F^2 \tag{10}$$

  where $\mathrm{diag}(X)$ and $\|X\|_F$ are the diagonal part the Frobenius norm of matrix X.
- $\mathrm{SO}(n)$ is a compact manifold so the problem of minimization of J_1 over $\mathrm{SO}(n)$ is a well defined problem.

Cost functions for...

(11) $$J_1(B) = \sum_{i=1}^{n} \left\| BC_iB^T - \text{diag}(BC_iB^T) \right\|_F^2$$

- An extension of J_1 from SO(n) to bigger group of non-singular matrices GL(n) is not a suitable, since it is not scale invariant:

(12) $$J_1(\Lambda B) \to 0 \ \text{ as } \ \|\Lambda\| \to 0, \quad \Lambda = \text{diagonal}$$

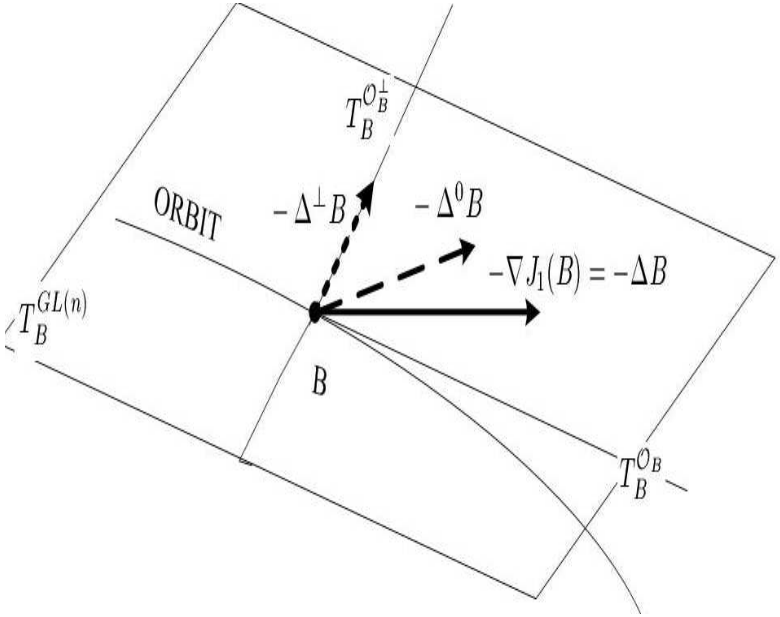
- Reminder: EJD was a scale-invariant problem, i.e. if B is joint diagonalizer ΛB is also another joint diagonalizer.
- Equip GL(n) with a Riemannian metric $< .,. >_B$: $T_BGL(n) \times T_BGL(n) \to \mathbb{R}$:

$$< \xi_1, \xi_2 >_B = \text{tr}(((\xi_1 B^{-1})^T \xi_2 B^{-1}), \ \ \xi_1, \xi_2 \in T_BGL(n)$$

- Reminder: every tangent vector at B is of the form $\xi = \Delta B$, where $\Delta \in \mathfrak{gl}(n)$.
- The gradient of $J_1 : GL(n) \to \mathbb{R}$:

(13) $$\nabla J_1(B) = \Delta B = \sum_{i=1}^{N} \left(BC_iB^T - \text{diag}(BC_iB^T) \right) BC_iB^TB$$

A Differential Equation for Joint Diagonalization



- We want to reduce $J_1(B)$ without going along directions that correspond to reduction by diagonal matrices.
- At each point B the orbit of the action of the group of non-singular matrices is a manifold and the tangent space to this manifold is $T_B\mathcal{O}_B \subset T_B GL(n)$.
- A flow for JD:

(14) $$\dot{B} = -\nabla J_1^\perp(B) = -\Delta^\perp B = -(\Delta - \text{diag}(\Delta))B$$

- Define the joint diagonalizer of $\{C_i\}_{i=1}^N$ as the stationary points of this flow:

(15) $$\sum_{i=1}^{N} \left((BC_iB^T)^\perp BC_iB^T \right)^\perp = 0$$

Sensitivity Analysis

(16) $$\sum_{i=1}^{N} \left((BC_iB^T)^{\perp} BC_iB^T \right)^{\perp} = 0$$

- Let us model the noise or error in C_i as:

  (17) $$C_i(t) = A\Lambda_i A^T + tN_i, \quad 1 \leq i \leq N, \quad t \in [-\delta, \delta]$$

  where t measures the contribution of noise and N_i's are symmetric noise or error matrices.

- Now as t changes $B(t)$ satisfies (16) and if δ is small (from the Implicit Function Theorem) we have that:

  (18) $$B(t) = (I + t\Delta)A^{-1} + o(t)$$

  where I is the $n \times n$ identity and Δ is a matrix with zero diagonal and $\frac{\|o(t)\|}{t} \to 0$ as $t \to 0$.

- $\|\Delta\|$ measures the sensitivity to noise.

Sensitivity...

Theorem 1 *Let* $C_i = A\Lambda_i A^T + tN_i, 1 \leq i \leq N$ *(*$t \in [-\delta, \delta]$*). Let us define* $B(t)$ *the non-orthogonal joint diagonalizer for* $\{C_i\}_{i=1}^N$ *as (16). Then for small enough* δ *the joint diagonalizer can be written as:* $B(t) = (I + t\Delta)A^{-1} + o(t)$ *where* Δ *(with* diag$(\Delta) = 0$*) satisfies*

(19)
$$\|\Delta\|_F < \frac{\alpha}{(1-\rho^2)} \|\mathcal{T}\|_F$$

with

(20)
$$\mathcal{T} = -\sum_{i=1}^{N} (A^{-1} N_i (A^{-1})^T)^{\perp} \Lambda_i$$

and

(21)
$$\gamma_{kl} = (\sum_{i=1}^{N} \lambda_{ik}^2)^{\frac{1}{2}} (\sum_{i=1}^{N} \lambda_{il}^2)^{\frac{1}{2}}, \quad \eta_{kl} = \frac{(\sum_{i=1}^{N} \lambda_{ik}^2)^{\frac{1}{2}}}{(\sum_{i=1}^{N} \lambda_{il}^2)^{\frac{1}{2}}}$$

and

(22)
$$\alpha = \max_{k \neq l} \frac{\eta_{kl} + \frac{1}{\eta_{kl}}}{\gamma_{kl}}$$

- The derivation of this bound shows that this is not a loose bound
- Corollary: Modulus of uniqueness and $\|A^{-1}\|$ (and consequently the condition number of A) affect the sensitivity of JD.

On Number of Matrices

- $|\rho|$ is the cosine of the smallest angle between n points in the real projective plane $\mathbf{RP}^N$. These vectors are columns of the matrix L formed from Λ_is.
- If N is small this angle can be very small.
- For $N = 2$, we can think of putting n points on the unit semi-circle. Which means that $\rho \geq \cos \frac{\pi}{n-1}$. Or if λ_{ij} are non-negative then $\rho \geq \cos \frac{\pi}{2(n-1)}$. If n is large these bounds can be close to 1 which makes the JD problem very sensitive.
- How to find a similar bound for $N \geq 3$?(I haven't done it!)
- Also if N is small we will not see much cancelation of the effect of noise in $\mathcal{T}$ (averaging effect) and that also contributes to more sensitivity.
- These results can be confirmed numerically.

Conclusion

- Introduced EJD and JD and the motivation for these problems
- Defined the joint diagonalizer as the stationary points of a flow on the group GL(n)
- Defined the sensitivity of the problem and the factors that affect it: especially the modulus of uniqueness and the condition of A