Simultaneous diagonalization: the asymmetric, low-rank, and noisy settings

Volodymyr Kuleshov, Arun Chaganty, and Percy Liang

Stanford University, Stanford CA 94305, USA,
kuleshov@stanford.edu, chaganty@cs.stanford.edu, pliang@cs.stanford.edu

Abstract. Simultaneous matrix diagonalization is used as a subroutine in many machine learning problems, including blind source separation and paramater estimation in latent variable models. Here, we extend algorithms for performing joint diagonalization to low-rank and asymmetric matrices, and we also provide extensions to the perturbation analysis of these methods. Our results allow joint diagonalization to be applied in several new settings.

1   Introduction

Consider a set of $L \geqslant 2$ matrices $\mathcal{M} = \{M_l\}_{l=1}^L$ of the form

$$M_l = U \Lambda_l U^T, \tag{1}$$

where $U \in \mathbb{R}^{d \times k}$ are factors common to all the matrices, and the diagonal $\Lambda_l \in \mathbb{R}^{k \times k}$ contain weights that are specific to each matrix M_l.

Simultaneous matrix diagonalization consists in determining the unknown factors and weights from the matrices M_l. Unlike in traditional single-matrix diagonalization, the U may be non-orthogonal (such factors are identifiable when $L \geqslant 2$; see Afsari [2]), and when the U are orthogonal, simultaneously diagonalizing the entire set $\mathcal{M}$ if often more robust to noise that diagonalizing a single M_l.

Joint diagonalization arises in several machine learning settings, including blind-source separation [12] and latent variable estimation via tensor factorization [7]. However, our understanding of algorithms for jointly diagonalizing matrices is far from complete: even the low-rank ($k < d$) and the asymmetric settings have not been considered in the literature.

Here, we show how to extend existing algorithms — notably the Jacobi [5] and QRJ1D [1] methods — to these two settings. Our extensions enable one to apply simultaneous diagonalization to several new problems. For example, it was recently shown that tensor factorization can be reduced joint matrix diagonalization; our algorithms make this reduction also applicable to low-rank and asymmetric tensors. This in turn leads to accuracy improvements on problems in community detection and for topic modeling.

Finally, we also extend existing perturbation analyses for noisy matrices of the form

$$M_l = U \Lambda_l U^T + \epsilon R_l \tag{2}$$

for some $\epsilon > 0$ and some matrix R_l having unit norm. First, we give a simple generalization of existing bounds to the low-rank and asymmetric settings. These bounds can be used to derive formal guarantees for methods that use simultaneous diagonalization as a subroutine; we again provide tensor factorization as an example. However, current bounds only hold for the true joint diagonalizer of the set $\mathcal{M}$, and there are no guarantees on whether it is attained by joint diagonalization algorithms; this limits the usefulness of theoretical analyses based on these lemmas. In the last section of the paper, we address this shortcoming by showing that for sufficiently small noise, the global minimizer will be attained, as long as we initialize the diagonalization subroutine with factors obtained by diagonalizing a single matrix.

2 Background

2.1 Notation

We establish notation for the classical symmetric simultaneous diagonalization case; we will extend it to the asymmetric case in a subsequent section.

We are given as input a set of matrices $M_1, \ldots, M_L \in \mathbb{R}^{d \times d}$ where each M_l can be expressed as

$$M_l = U \Lambda_l U^T + \epsilon R_l. \tag{3}$$

The diagonal weight matrix $\Lambda_l \in \Re^{d \times d}$ and the noise R_l (satisfying $||R_l|| \leqslant 1$) are individual to each M_l, but the non-singular transform $U \in \mathbb{R}^{d \times d}$ is common to all the matrices.

Our goal is to find an invertible transform V^{-1} such that each $V^{-1} M_l V^{-\top}$ is nearly diagonal. This problem admits a unique solution when there are at least two matrices [2]. There are a number of objective functions that are designed for determining joint diagonalizers [5,9,1], but in this paper, we focus on a popular one that penalizes off-diagonal terms:

$$F(V) = \sum_{l=1}^{L} \text{off}(V^{-1} M_l V^{-\top}) = \sum_{l=1}^{L} \sum_{i \neq j} (V^{-1} M_l V^{-\top})_{ij}^2. \tag{4}$$

An important setting of this problem, which we refer to as the *orthogonal case*, is when U is orthogonal, in which case we will also constrain the optimization to be orthogonal $V^{-1} = V^\top$.

2.2 Previous work

There exist several algorithms for optimizing $F(V)$. In this paper, we will use the Jacobi method [3,5] for the orthogonal case and the QRJ1D algorithm [1] for the non-orthogonal case. Both techniques are based on same idea of iteratively constructing V^{-1} via a product of simple matrices $V^{-1} = B_T \cdots B_2 B_1$, where at each iteration $t = 1, \ldots, T$, we choose B_t to minimize $J(V)$. Typically, this can be done in closed form.

The Jacobi algorithm for the orthogonal case is a simple adaptation of the Jacobi method for diagonalizing a single matrix. Each B_t is chosen to be a *Givens* rotation [3] defined by two of the d axes $i < j \in [d]$: $B_t = \cos\theta(\Delta_{ii} + \Delta_{jj}) + \sin\theta(\Delta_{ij} - \Delta_{ji})$ for some angle θ, with Δ_{ij} being a matrix which is 1 in the (i,j)-th entry and 0 elsewhere. We sweep over all $i < j$, compute the best angle θ in closed form using the formula proposed by [5] to obtain B_t, and then update each M_l by $B_t^\top M_l B_t$. The above can be done in $O(d^3 L)$ time per sweep.

Data: symmetric matrices $(M_l)_{l=1}^L$
Let $U = I$;
while *objective is decreasing* **do**
    for $i = 1, 2, ..., d$ **do**
        for $j = i + 1, i + 2, ..., d$ **do**
            Let $\hat{\theta}$ be the minimizer of $F(G_{ij}(\theta))$;
            $U \leftarrow U G_{ij}(\hat{\theta})$;
            for $l = 1, 2, ..., L$ **do**
                $M_l \leftarrow G_{ij}(\hat{\theta})^T M_l G_{ij}(\hat{\theta})$;
            end
        end
    end
end

Algorithm 1: The Jacobi algorithm for simultaneous diagonalization

For the non-orthogonal case, the QRJ1D algorithm is similar, except that B_t is chosen to be either a lower or upper unit triangular matrix ($B_t = I + a\Delta_{ij}$ for some a and $i \neq j$). The optimal value of a that minimizes $J(V)$ can also be computed in closed form (see [1] for details). The running time per iteration is the same as before.

Most popular algorithms both in the orthogonal [3] and the non-orthogonal setting [12,10] are guaranteed to converge locally. The formal question of global convergence is currently open even for methods that are widely used [8,5]. In practice, however, the Jacobi-style methods we adopt are well-known to behave as if they global convergence [3,5,8].

In the asymmetric setting, variants of the Jacobi algorithm have been proposed for orthogonal factors [6]. These variants essentially diagonalize the set of matrices $M_j' = M_j^T M_j$. Notice that when $M_j = U\Lambda V^T$, we have $M_j' = V\Lambda^2 V^T$, and so the assymetric problem can be reduced to one with symmetric M_j'. In the

Data: symmetric matrices $(M_l)_{l=1}^L$
Let $U = I$;
while *objective is decreasing* **do**
 for $i = 1, 2, ..., d$ **do**
  for $j = i + 1, i + 2, ..., d$ **do**
   Let $\hat{\theta}$ be the minimizer of $F(G_{ij}(\theta))$;
   $U \leftarrow U G_{ij}(\hat{\theta})$;
   for $l = 1, 2, ..., L$ **do**
    $M_l \leftarrow G_{ij}(\hat{\theta})^T M_l G_{ij}(\hat{\theta})$;
   end
  end
 end
 for $i = 1, 2, ..., d$ **do**
  for $j = i + 1, i + 2, ..., d$ **do**
   Let $\hat{a}$ be the minimizer of $F(\Delta_{ij}(a))$;
   $U \leftarrow U \Delta_{ij}(\hat{a})$;
   for $l = 1, 2, ..., L$ **do**
    $M_l \leftarrow \Delta_{ij}(\hat{a})^T M_l \Delta_{ij}(\hat{a})$;
   end
  end
 end
end
Algorithm 2: The QRJ1D algorithm for simultaneous non-orthogonal diagonalization

low-rank setting, an extension of the Jacobi algorithm has been proposed for a single matrix [11]. This extension implictely sorts the entries of the matrix and performs only $O(dk^2)$ Givens rotations per sweep. However, the $L \geqslant 2$ setting has not been considered in the literature. Finally, for non-orthogonal matrices, neither the low-rank and asymmetric case has been worked out to the best of our knowledge.

3   Low-rank matrices

In this section, we provide extensions of simultaneous algorithms to low-rank matrices. We start with orthogonal factors; in this setting, we take inspiration from the sorted Jacobi method for a single low-rank matrix and propose a sorted Jacobi method for multiple matrices. Then, we show how to generalize the same sorting idea to the QRJ1D algorithm, which leads to a low-rank algorithm for the non-orthogonal setting as well.

3.1   Orthogonal setting

First, suppose that we applied the Jacobi algorithm to a single matrix M whose eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant ... \geqslant \lambda_k$ (corresponding to the diagonal of the matrix Λ) were sorted. In that case, we would only need Jacobi to zero out the entires of M associated with the first k rows or the first k columns. In other words, we would have to transform the matrix into the following form:

$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix}$$

The two left-most columns of the diagonalizing matrix will correspond to the eigenvectors associated with λ_1 and λ_2 while the other two columns will contain arbitrary numbers. Also, since Givens rotations are essentially independent of each other, it takes only kd Givens rotations per sweep to turn the input matrix into the above form.

If the eigenvalues of the input matrix M are not sorted, then we can sort the diagonal of M after every sweep of Jacobi (while still performing only kd Givens rotations per sweep). When the algorithm terminates, all the k non-zero eigenvalues will find themselves in the top $k \times k$ corner; if that wasn't the case, then they would have been swapped out with another entry from outside that corner.

When there is more than one matrix, the components j over which the rank is positive are ones for which $\sum_{l=1}^{L} |(\Lambda_l)_{jj}| > 0$. This suggests a natural extension of the above idea: choose Givens rotations in a way as to push mass on the sum of the absolute values of the matrix diagonals towards the upper left corner. This idea is implemented in Algorithm 3

The sorting opereation corresponds to a permutation of the columns of the M_j, which does not change the value of the objective function. Thus the sorted Jacobi algorithm also decreases the objective function value at every iteration, which ensures its convergence to a low-rank diagonalizer. However, unlike in the single-matrix setting, we do not provide formal guarantees that the resulting diagonalizer is optimal; even in the full-rank setting convergence properties of the simultaneous Jacobi method remains an open problem. However, in the next section we demonstrate empiracally that just as in the full-rank setting, our low-rank extension appears to converge for any matrix.

Data: symmetric matrices $(M_l)_{l=1}^L$
Let $U = I$;
while *objective is decreasing* **do**
    for $i = 1, 2, ..., k$ **do**
        for $j = i + 1, i + 2, ..., d$ **do**
            Let $\hat{\theta}$ be the minimizer of $F(G_{ij}(\theta))$;
            $U \leftarrow UG_{ij}(\hat{\theta})$;
            for $l = 1, 2, ..., L$ **do**
                $M_l \leftarrow G_{ij}(\hat{\theta})^T M_l G_{ij}(\hat{\theta})$;
            end
            if $\sum_{l=1}^L |(\Lambda_l)_{jj}| > \sum_{l=1}^L |(\Lambda_l)_{jj}|$ **then**
                Flip columns i and j in U and the M_l;
            end
        end
    end
end

Algorithm 3: The low-rank Jacobi algorithm for simultaneous diagonalization

3.2 Non-orthogonal setting

The QRJ1D algorithm has a very similar structure to Jacobi: over the course of a sweep, the matrices M_j are first multiplied by Givens rotation, and then by lower triangular matrices. In each case, a rotation only affects two columns and two rows of M_j.

We can similarly sort the diagonal entries of the matrices and zero-out only the top $k \times k$ square of the M_j. This leads to Algorithm 4.

4 Asymmetric matrices

Suppose now that we have a set of $L \geqslant 2$ matrices $\mathcal{M} = \{M_l\}_{l=1}^L$ of the form

$$M_l = U \Lambda_l V^T, \tag{5}$$

Data: symmetric matrices $(M_l)_{l=1}^L$
Let $U = I$;
while *objective is decreasing* **do**
  for $i = 1, 2, ..., k$ **do**
    for $j = i + 1, i + 2, ..., d$ **do**
      Let $\hat{\theta}$ be the minimizer of $F(G_{ij}(\theta))$;
      $U \leftarrow U G_{ij}(\hat{\theta})$;
      for $l = 1, 2, ..., L$ **do**
        $\mid$ $M_l \leftarrow G_{ij}(\hat{\theta})^T M_l G_{ij}(\hat{\theta})$;
      end
      if $\sum_{l=1}^L |(\Lambda_l)_{jj}| > \sum_{l=1}^L |(\Lambda_l)_{jj}|$ **then**
        $\mid$ Flip columns i and j in U and the M_l;
      end
    end
  end
  for $i = 1, 2, ..., k$ **do**
    for $j = i + 1, i + 2, ..., d$ **do**
      Let $\hat{a}$ be the minimizer of $F(\Delta_{ij}(a))$;
      $U \leftarrow U \Delta_{ij}(\hat{a})$;
      for $l = 1, 2, ..., L$ **do**
        $\mid$ $M_l \leftarrow \Delta_{ij}(\hat{a})^T M_l \Delta_{ij}(\hat{a})$;
      end
      if $\sum_{l=1}^L |(\Lambda_l)_{jj}| > \sum_{l=1}^L |(\Lambda_l)_{jj}|$ **then**
        $\mid$ Flip columns i and j in U and the M_l;
      end
    end
  end
end
Algorithm 4: The QRJ1D algorithm for simultaneous non-orthogonal diagonalization

where $U \in \mathbb{R}^{d_1 \times k}$ and $V \in \mathbb{R}^{d_2 \times k}$ are sets of common factors, possibly non-orthogonal.

When the U and V are orthogonal, there is a well-known procedure that reduces this problem to the symmetric case by instead diagonalizing the matrices $M_l' = M_l^T M_L = V \Lambda^2 V^T$. Unfortunately, this reduction does not work for non-orthogonal matrices; here, we propose another reduction that works in both orthogonal and non-orthogonal cases.

For each M_l, define another matrix $N_l = \begin{pmatrix} 0 & M_l^\top \\ M_l & 0 \end{pmatrix}$ and observe that

$$
\begin{bmatrix} 0 & M_l^\top \\ M_l & 0 \end{bmatrix} = \begin{bmatrix} V & V \\ U & -U \end{bmatrix} \begin{bmatrix} \Lambda_l & 0 \\ 0 & -\Lambda_l \end{bmatrix} \begin{bmatrix} V & V \\ U & -U \end{bmatrix}^\top .
$$

The (N_l) are symmetric matrices with common (in general, non-orthogonal) factors. Therefore, they can be jointly diagonalized and from their components, we can recover the components of the (M_l).

Using the above low-rank algorithms, it is possible to determine the U, V factors in $O(d_1 d_2^2)$ time (assuming $d_2 \geqslant d_1$). This is worse than the $O(d_1^2 d_2)$ time it takes to determine the SVD of a single matrix. It remains to be seen if non-orthogonal joint diagonalization admits algorithms are as fast as ones for the ordinary SVD.
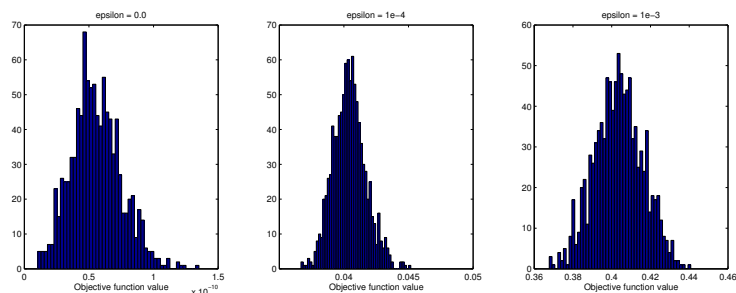
5 Experiments

5.1 Full-rank matrices

Fig. 1: Histogram of objective values attained by the Jacobi simultaneous diagonalization algorithm for 1000 random sets of jointly diagonalizable matrices initialized with various amounts of noise.

To assess the convergence properties of the Jacobi method, we ran the algorithm 1000 times on different sets of L random matrices ($L = d = k = 15$) corrupted with varying amounts of noise ($\epsilon = 0, 1e - 4, 1e - 3$). At each run, we measured the objective value function (the Frobenius norm of the off-diagonal elements), and plotted the resulting histogram (Figure 1).

Overall, we observe that the objective values for a Gaussian distribution for the three settings $\epsilon = 0, 1e - 4, 1e - 3$. All data points are concentrated in an interval that depend either on the convergence threshold ($1e - 12$) in the $\epsilon = 0$ case or on the noise level (when $\epsilon = 1e - 3, 1e - 4$). Most interestingly, when $\epsilon = 0$, the mean of the observations is around $1e - 10$ (which corresponds to the precision of $1e - 12$ multiplied by the ~ 200 entries in the matrices), and $\epsilon > 0$, the mean depends on the noise level. Quite strikingly, multiplying the noise by ten produces essentially the same histogram within the same bounds, except that it has been translated from between 0.035 and 0.045 to between 0.35 and 0.45.

This observation strongly suggests that the Jacobi algorithm attains a globally optimal solutions even at high noise levels.

5.2 Low-rank matrices

5.3 Asymmetric matrices

6 Extension of perturbation analyses

We conclude our discussion with some perturbation results for noisy matrices of the form

$$M_l = U\Lambda_l U^T + \epsilon R_l \tag{6}$$

for some $\epsilon > 0$ and some matrix R_l having unit norm.

A perturbation analyses has been carried out for matrices of this form in both orthogonal and non-orthogonal settings. Our first two lemmas state that these analyses carry over to the low-rank case as well. The fact that they also carry over to the asymmetric case follows trivially from our reduction.

Lemma 1 ([4]). *Let $M_l = U\Lambda_l U^\top + \epsilon R_l$, $l \in [L]$, be matrices with common factors $U \in \mathbb{R}^{d \times k}$ and diagonal $\Lambda_l \in \mathbb{R}^{k \times k}$. Let $\bar{U} \in \Re^{d \times d}$ be a full-rank extension of U with columns $u_1, u_2, \ldots, u_d$ and let $\tilde{U} \in \Re^{d \times d}$ be the orthogonal minimizer of the joint diagonalization objective $F(\cdot)$. Then, for all u_j, $j \in [k]$, there exists a column $\tilde{u}_j$ of $\tilde{U}$ such that*

$$\|\tilde{u}_j - u_j\|_2 \leqslant \epsilon \sqrt{\sum_{i=1}^{d} E_{ij}^2} + o(\epsilon), \tag{7}$$

where $E \in \mathbb{R}^{d \times k}$ is

$$E_{ij} = \frac{\sum_{l=1}^{L} (\lambda_{il} - \lambda_{jl}) u_j^\top R_l u_i}{\sum_{l=1}^{L} (\lambda_{il} - \lambda_{jl})^2} \tag{8}$$

when $i \neq j$ and $i \leqslant k$ or $j \leqslant k$. We define $E_{ij} = 0$ when $i = j$ and $\lambda_{il} = 0$ when $i > k$.

Proof. See Proposition 1 in [4]. Note that in the low rank setting, the entries of E_{ij} (Equation 15 in [4]) where $i, j > k$ are not defined, however, these terms only effect the last $d - k$ columns of $\tilde{U}$. The bounds for vectors $u_1, \ldots, u_k$ only depend on E_{ij} where $i \in [d]$ and $j \in [k]$, and these are derived in the low-rank setting in the same way as they are derived in the full-rank proof of [4]. ∎

We now present the corresponding perturbation bounds in [2] to the low rank setting.

Lemma 2 ([2]). *Let $M_l = U\Lambda_l U^\top + \epsilon R_l$, $l \in [L]$, be matrices with common factors $U \in \mathbb{R}^{d \times k}$ and diagonal $\Lambda_l \in \mathbb{R}^{k \times k}$. Let $\bar{U} \in \Re^{d \times d}$ be a full-rank extension of U with columns $u_1, u_2, \ldots, u_d$ and let $\bar{V} = \bar{U}^{-1}$, with rows $v_1, v_2, \ldots, v_d$. Let $\tilde{V} \in \Re^{d \times d}$ be the minimizer of the joint diagonalization objective $F(\cdot)$ and let $\tilde{U} = \tilde{V}^{-1}$.*

Then, for all u_j, $j \in [k]$, there exists a column $\tilde{u}_j$ of $\tilde{U}$ such that

$$\|\tilde{u}_j - u_j\|_2 \leqslant \epsilon \sqrt{\sum_{i=1}^{d} E_{ij}^2} + o(\epsilon), \tag{9}$$

where the entries of $E \in \mathbb{R}^{d \times k}$ satisfy the equation

$$\begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix} = \frac{-1}{\gamma_{ij}(1 - \rho_{ij}^2)} \begin{bmatrix} \eta_{ij} & -\rho_{ij} \\ -\rho_{ij} & \eta_{ij}^{-1} \end{bmatrix} \begin{bmatrix} T_{ij} \\ T_{ji} \end{bmatrix}.$$

when $i \neq j$ and either $i \leqslant k$ or $j \leqslant k$. When $i = j$, $E_{ij} = 0$. The matrix T has zero on-diagonal elements, and is defined as

$$T_{ij} = \sum_l v_i^\top R_l v_j \lambda_{jl}, \qquad \qquad for 1 \leqslant j \neq i \leqslant d$$

and the other parameters are

$$\gamma_{ij} = \|\lambda_i\|_2 \|\lambda_j\|_2, \qquad \eta_{ij} = \frac{\|\lambda_i\|_2}{\|\lambda_j\|_2}, \qquad \rho_{ij} = \frac{\lambda_i^\top \lambda_j}{\|\lambda_j\|_2 \|\lambda_i\|_2}, \qquad (\lambda_i)_k = \lambda_{ik}.$$

We define $\lambda_{il} = 0$ when $i > k$.

Proof. In Theorem 3 in [2] it is shown that $\tilde{V} = (I + \epsilon E)V + o(\epsilon)$, where E_{ij} is defined for $i, j \in [d]$ (Equation 36 in [2]). Then,

$$\tilde{U} = \tilde{U}(I + \epsilon E)^{-1} + o(\epsilon)$$
$$= \tilde{U}(I - \epsilon E) + o(\epsilon).$$

Note that, once again, in the low rank setting, the entries of E_{ij} when $i, j > k$ are not characterized by Afsari's results; however, these terms only effect the last $d - k$ columns of $\tilde{U}$. ∎

Lemma 3. *Let $M_l = U \Lambda_l U^\top + \epsilon R_l$, $l \in [L]$, be matrices with common factors $U \in \mathbb{R}^{d \times k}$ and diagonal $\Lambda_l \in \mathbb{R}^{k \times k}$. Let $\bar{U} \in \Re^{d \times d}$ be a full-rank extension of U with columns $u_1, u_2, \ldots, u_d$ and let $\bar{V} = \bar{U}^{-1}$, with rows $v_1, v_2, \ldots, v_d$. Let $\tilde{V} \in \Re^{d \times d}$ be the minimizer of the joint diagonalization objective $F(\cdot)$ and let $\tilde{U} = \tilde{V}^{-1}$.*

Then, for all u_j, $j \in [k]$, there exists a column $\tilde{u}_j$ of $\tilde{U}$ such that

$$\|\tilde{u}_j - u_j\|_2 \leqslant \epsilon \sqrt{\sum_{i=1}^{d} E_{ij}^2} + o(\epsilon), \tag{10}$$

where the entries of $E \in \mathbb{R}^{d \times k}$ are bounded by

$$|E_{ij}| \leqslant \frac{1}{1 - \rho_{ij}^2} \left(\frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2} \right) \left(\left| \sum_{l=1}^{L} v_i^\top R_l v_j \lambda_{jl} \right| + \left| \sum_{l=1}^{L} v_i^\top R_l v_j \lambda_{il} \right| \right),$$

when $i \neq j$ and $E_{ij} = 0$ when $i = j$ and $\lambda_{il} = 0$ when $i > k$. Here $\lambda_i = (\lambda_{i1}, \lambda_{i2}, ..., \lambda_{iL}) \in \Re^L$ and $\rho_{ij} = \frac{\lambda_i^\top \lambda_j}{\|\lambda_i\|_2 \|\lambda_j\|_2}$ is the modulus of uniqueness, a measure of how ill-conditioned the problem is.

Proof. From Lemma 2, we have that

$$\left\| \begin{bmatrix} E_{ij} \\ E_{ji} \end{bmatrix} \right\| \leqslant \frac{\eta_{ij} + \eta_{ji}}{\gamma_{ji}(1 - \rho_{ij}^2)} \left\| \begin{bmatrix} T_{ij} \\ T_{ji} \end{bmatrix} \right\|,$$

where

$$\gamma_{ij} = \|\lambda_i\|_2 \|\lambda_j\|_2, \qquad \eta_{ij} = \frac{\|\lambda_i\|_2}{\|\lambda_j\|_2}, \qquad \rho_{ij} = \frac{\lambda_i^\top \lambda_j}{\|\lambda_j\|_2 \|\lambda_i\|_2},$$

and the matrix T is defined to be zero on the diagonal and for $i \neq j$ defined as

$$T_{ij} = \sum_{l=1}^{L} v_i^\top R_l v_j \lambda_{jl}, \qquad \text{for} 1 \leqslant j \neq i \leqslant d$$

Taking $\| \cdot \|$ to be the l_1-norm in the above expression, we have that

$$|E_{ij}| \leqslant |E_{ij}| + |E_{ji}| \leqslant \frac{\eta_{ij} + \eta_{ji}}{\gamma_{ji}(1 - \rho_{ij}^2)} \left(|T_{ij}| + |T_{ji}| \right).$$

Since

$$\frac{\eta_{ij} + \eta_{ji}}{\gamma_{ji}} = \frac{\|\lambda_i\|_2^2 + \|\lambda_j\|_2^2}{\|\lambda_i\|_2^2 \|\lambda_j\|_2^2} = \frac{1}{\|\lambda_i\|_2^2} + \frac{1}{\|\lambda_j\|_2^2}$$

and

$$T_{ij} = \sum_{l=1}^{L} v_i^\top R_l v_j \lambda_{jl},$$

the claim follows. ∎

These results hold for the joint matrix diagonalizer, i.e. the global optimum of the objective F. It is not clear whether this optimum can be attained by existing algorithms such as Jacobi, which limits the theoretical applicability of the above lemmas. Although empirical results strongly indicate that the global minimizer is indeed always found in the orthogonal case, we also complement these results with the lemma below.

Lemma 4. *Suppose the Jacobi simultaneous diagonalization algorthim is initialized with $\hat{U}$, the set eigenvectors obtained from diagonalizing a single matrix. Then for $\epsilon > 0$ small enough, Jacobi will converge to a point at which Lemmas 1 and 2 hold.*

Proof. See appendix. ∎

The above lemmas can be used to derive theoretical guarantees for algorithms that use them as a subroutine.

References

1. Afsari, B.: Simple LU and QR based non-orthogonal matrix joint diagonalization. In: Independent Component Analysis and Blind Signal Separation. pp. 1–7 (2006)
2. Afsari, B.: Sensitivity analysis for the problem of matrix joint diagonalization. SIAM Journal on Matrix Analysis and Applications 30(3), 1148–1171 (2008)
3. Bunse-Gerstner, A., Byers, R., Mehrmann, V.: Numerical methods for simultaneous diagonalization. SIAM Journal on Matrix Analysis and Applications 14(4), 927–949 (1993)
4. Cardoso, J.: Perturbation of joint diagonalizers. Tech. rep., T'el'ecom Paris (1994)
5. Cardoso, J., Souloumiac, A.: Jacobi angles for simultaneous diagonalization. SIAM Journal on Matrix Analysis and Applications 17(1), 161–164 (1996)
6. Congedo, M., Phlypo, R., Pham, D.T.: Approximate joint singular value decomposition of an asymmetric rectangular matrix set. Signal Processing, IEEE Transactions on 59(1), 415–424 (Jan 2011)
7. Kuleshov, V., Chaganty, A., Liang, P.: Tensor factorization via random projection and simultaneous matrix diagonalization. In: Artificial Intelligence and Statistics (AISTATS) (2015)
8. Lathauwer, L.D., Moor, B.D., Vandewalle, J.: Independent component analysis and (simultaneous) third-order tensor diagonalization. Signal Processing, IEEE Transactions on 49(10), 2262–2271 (2001)
9. Yeredor, A.: Non-orthogonal joint diagonalization in the least-squares sense with application in blind source separation. IEEE Transactions on Signal Processing 50(7), 1545–1553 (2002)
10. Yeredor, A., Ziehe, A., Müller, K.: Approximate joint diagonalization using a natural gradient approach. Independent Component Analysis and Blind Signal Separation 1, 86–96 (2004)
11. Zha, H., Zhang, Z.: A sorted partial jacobi method and its convergence analysis. Linear Algebra and its Applications 270(13), 79 – 108 (1998), http://www.sciencedirect.com/science/article/pii/S0024379597002279
12. Ziehe, A., Laskov, P., Nolte, G., Müller, K.: A fast algorithm for joint diagonalization with non-orthogonal transformations and its application to blind source separation. Journal of Machine Learning Research (JMLR) 5, 777–800 (2004)

A Extension of perturbation analyses

In this section, we analyze the convergence of the Jacobi method. We argue that for small enough $\epsilon > 0$, Jacobi will converge to a point at which our perturbation lemma holds. The results in this section complement our empirical assessment of the convergence of the method.

A.1 Notation

Let $F(V, \mathcal{M}) : O(n) \to \mathbb{R}^+$ be the orthogonal joint diagonalization criterion for a set of matrices $\mathcal{M} = \{M_1, \cdots, M_L\}$,

$$F(V, \mathcal{M}) = \sum_{i=1}^{L} \| \operatorname{off}(V^\top M_l V) \|_F^2$$

$$= \sum_{i=1}^{L} \| V^\top M_l V \|_F^2 - \| \mathbf{diag}(V^\top M_l V) \|_F^2.$$

Let $\mathcal{M}_\epsilon = \{U^\top D_l U + \epsilon R_l\}_{l=1}^L$ be the set of matrices given as input. Note that when $\epsilon = 0$, the set of matrices commute and are jointly diagonalizable, with diagonalizer U. For simplicity, we will let the set of matrices $U^\top D_l U$ be fixed and use the shorthand $F(V, \epsilon)$ to denote $F(V, \mathcal{M}_\epsilon)$.

Several observations should be made about $F(V, \epsilon)$:

- $F(V, \epsilon)$ is continuous in V, ϵ. Since $O(n)$ is compact, for any fixed ϵ, $F(V, \epsilon)$ is uniformly continuous.
- If we restrict ϵ to lie on a closed interval I, then $F(V, \epsilon) : O(n) \times I \to \mathbb{R}^+$ is uniformly continuous with both variables and with the metric $|| \cdot ||'$ on (V, ϵ) defined as $||(V, \epsilon)||' = ||V|| + |\epsilon|$.
- Each global minimizer U of $F(V, 0)$ is isolated. We let $\gamma > 0$ denote the radius of the ball
$$\Gamma := \{V : ||V - U|| < \gamma\}$$
around U on which $F(V, 0) > F(U, 0)$ for all $V \in \Gamma$.

Finally we introduce a few additional pieces of notation. We let $V(\epsilon)$ minimizers of $F(V, \epsilon)$ on the set Γ; such minimizers always exist by compactness of $F(V, \epsilon)$. We also define $U(\epsilon)$ to be a global minimizer of $F(V, \epsilon)$; note that $U(0) = U$. Let $d(V)$ denote the Jacobi step taken from V: if $V^+ = \operatorname{Jacobi}(V)$, then $d(V) = (V^+ - V)$. Finally, let $W(\epsilon)$ to be the joint diagonalizer obtained from diagonalizing a single matrix in $\mathcal{M}_\epsilon$.

A.2 Outline

Our goal is to show that if Jacobi is initialized with $W(\epsilon)$, then all sufficiently small ϵ, it will will converge to a point at which Cardoso's lemma holds. This involves several steps:

1. Showing that Jacobi converges to a point close to an unperturbed local minimum U.
2. For all such points that are close to U, Cardoso's lemma holds.

A.3 Convergence of Jacobi to a neighborhood of U

Our first lemma states that the steps taken by Jacobi get arbitrarily small when the objective value is close to optimal.

Lemma 5 (Vanishing steps). *Let $U(\epsilon)$ be a global minimizer of $F(V, \epsilon)$. For all $\delta > 0$, there is a $t > 0$ such for all $\epsilon > 0$, whenever $|F(U(\epsilon), \epsilon) - F(V, \epsilon)| < t$, $||d(V)|| < \delta$.*

Proof. Follows from some algebra (see Maleko, 2003). ∎

Next, we want to establish local convergence results for Jacobi: when it is started close to a local minimizer, it converges to that minimizer.

First, we will need to show that the step sizes Jacobi takes are small around certain local minimizers of $F(V, \epsilon)$. In particular, when a local minimizer $V(\epsilon)$ is close to U, we can use the fact that the steps around U are small to show that steps around V are small as well. This is what the following lemma establishes.

Lemma 6 (Vanishing steps around local minima). *For all $\gamma > \delta > 0$, there exists an $\epsilon_0 > 0$ and an $r > 0$ such that for all $0 \leqslant \epsilon \leqslant \epsilon_0$, if $V(\epsilon)$ is a minimizer of $F(V, \epsilon)$ on the set $\Gamma = \{V : |V - U| \leqslant \gamma\}$ satisfying $||U - V(\epsilon)|| < r$, then we have*

$$||V - V(\epsilon)|| + ||d(V)|| < \delta$$

whenever $||V - V(\epsilon)|| < r$.

Proof. By Lemma 5, there exists a $t > 0$ such that for any $\epsilon \geqslant 0$, $|F(U(\epsilon), \epsilon) - F(V, \epsilon)| < t$ implies that $||d(V)|| < \delta/2$.

Choose $r > 0$ such that $r < \delta/2$ and such that for all $||V - U|| < 2r$, we have $|F(U(0), 0) - F(V, 0)| < t/4$. Choose $\epsilon_0 > 0$ such that for all $V \in O(n)$ and all $0 \leqslant \epsilon \leqslant \epsilon_0$, we have $||F(V, 0) - F(V, \epsilon)|| < t/4$.

Then, when $||V - V(\epsilon)|| < r$, we have $||V - U|| \leqslant ||V - V(\epsilon)|| + ||V(\epsilon) - U|| \leqslant 2r$, and thus $|F(U(0), 0) - F(V, 0)| < t/4$.

Next, for $0 \leqslant \epsilon \leqslant \epsilon_0$, and for the same V as above we have

$$
\begin{aligned}
F(V, \epsilon) - F(U(\epsilon), \epsilon) &\leqslant |F(V, \epsilon) - F(V, 0)| + |F(U(\epsilon), \epsilon) - F(U(\epsilon), 0)| \\
&\quad + |F(U(0), 0) - F(U(\epsilon), 0)| + |F(V, 0) - F(U(0), 0)| \\
&< t/4 + t/4 + t/4 + t/4 = t.
\end{aligned}
$$

By definition of t, $||d(V)|| < \delta/2$ and

$$||V - V(\epsilon)|| + ||d(V)|| < \delta/2 + \delta/2 \leqslant \delta.$$

∎

Next, we want to show for minima $V(\epsilon)$ close to U, there is an open set around each $V(\epsilon)$ that "captures" the Jacobi iterates such that they never leave that set.

Here is the intuition behind the proof. Suppose $V(\epsilon)$ is a local minimum. Consider a connected level set L around $V(\epsilon)$. Because our algorithm always decreases the objective function, if we start in L then we shouldn't leave it. The only way we can leave L is if we take a step size large enough to move us far from the local minimum $V(\epsilon)$ to an region where the objective function value is smaller than in L. However, if we define a set S that is contained in L and in which the step sizes are always sufficiently small, then it is possible to show that the iterates will never leave S.

Lemma 7 (Convergence ball). *For all $\gamma > \delta > 0$, there exist $\epsilon_0, r, s > 0$ such that for all $0 \leqslant \epsilon \leqslant \epsilon_0$, and for all local minima $V(\epsilon)$ of $F(V, \epsilon)$ on the set $\Gamma = \{V : |V - U| < \gamma\}$ satisfying $\|U - V(\epsilon)\| < r$, if we start the algorithm at a point V_k in the set*

$$D(V(\epsilon), \epsilon, \delta) = \{\|V - V(\epsilon)\| < \delta; \ F(V, \epsilon) < F(V(\epsilon), \epsilon) + s\}.$$

then we will also have $V_{k+1} \in D(V(\epsilon), \epsilon, \delta)$.

Proof. By Lemma 6, there exists an $\epsilon_0 > 0$ and an $r > 0$ such that for all $0 \leqslant \epsilon \leqslant \epsilon_0$ and for $\|V - V(\epsilon)\| < r$, where $V(\epsilon)$ is a local minimizer of $F(V, \epsilon)$ satisfying $\|U - V(\epsilon)\| < r$, we have

$$\|V - V(\epsilon)\| + \|d(V)\| < \delta.$$

Let $\phi(t, \epsilon) = \inf_{t < \|V - V(\epsilon)\|; \ V \in \Gamma} F(V, \epsilon) - F(V(\epsilon), \epsilon) \geqslant 0$ and observe that $\phi(t, \epsilon)$ is monotonically increasing in t. Let $s = \inf_{0 \leqslant \epsilon \leqslant \epsilon_0} \phi(r, \epsilon)$ and consider the set

$$D(V(\epsilon), \epsilon, \delta) = \{\|V - V(\epsilon)\| < \delta; \ F(V, \epsilon) < F(V(\epsilon), \epsilon) + s\}.$$

Suppose that $V_k \in D(V(\epsilon), \epsilon, \delta)$. Since

$$\phi(\|V_k - V(\epsilon)\|) \leqslant F(V_k, \epsilon) - F(V(\epsilon), \epsilon) < s \leqslant \phi(r, \epsilon)$$

and ϕ is monotonically increasing, we have $\|V_k - V(\epsilon)\| < r$, and thus $\|V_{k+1} - V(\epsilon)\| < \delta$. Also, $F(V_{k+1}, \epsilon) \leqslant F(V_k, \epsilon)$, so

$$F(V_{k+1}, \epsilon) - F(V(\epsilon), \epsilon) < s.$$

Thus $V_{k+1} \in D(V(\epsilon), \epsilon, \delta)$. ∎

The next lemma says that for small enough ϵ, most of these "capture" sets also contain the unperturbed minimizer U.

Lemma 8 (Convergence balls in the vicinity of global optima). *For all $\gamma \geqslant \delta \geqslant 0$, there exists an open set C around U, as well as $\epsilon_0, t > 0$ such that for all $0 \leqslant \epsilon \leqslant \epsilon_0$, for all $0 \leqslant \delta \leqslant \delta_0$, and for all local minimizers $V(\epsilon)$ such that $\|U - V(\epsilon)\| < t$, $C \subseteq D(V(\epsilon), \epsilon, \delta)$.*

Proof. Let $\delta > 0$ be fixed. Recall from Lemma 7, that the sets $D(V(\epsilon), \epsilon, \delta)$ have the form

$$D(V(\epsilon), \epsilon, \delta) = \{||V - V(\epsilon)|| < \delta; \; F(V, \epsilon) < F(V(\epsilon), \epsilon) + s\},$$

where s is a constant independent of ϵ. These sets are defined for all local minima $V(\epsilon)$ such that $||V(\epsilon) - U|| < r$, for some $r > 0$. Let ϵ_1 be the upper bound on ϵ that is given by the lemma.

Let t' be such that by the uniform continuity of f, we have $|F(U, \epsilon) - F(V(\epsilon), \epsilon)| < s/4$ for $||U - V(\epsilon)|| < t'$ and for all $0 \leqslant \epsilon \leqslant \epsilon_1$. Let ϵ_2 be such that for $0 \leqslant \epsilon, \epsilon' \leqslant \epsilon_2$, $||F(V, \epsilon) - F(V, \epsilon')|| < s/4$ for all V such that $||V - U|| < \delta$. Finally let $\epsilon_0 = \min(\epsilon_1, \epsilon_2)$ and let $t = \min(t', r, \delta/2)$.

Define the set C as

$$C = \{||V - U|| < \delta/2; \; F(V, 0) < F(U, 0) + s/4\}.$$

For $V \in C$ and for $0 \leqslant \epsilon \leqslant e_0$, we have

$$\begin{aligned} |F(V, \epsilon) - F(V(\epsilon), \epsilon)| &\leqslant |F(V, 0) - F(V, \epsilon)| + |F(U, \epsilon) - F(V(\epsilon), \epsilon)| \\ &\quad + |F(U, 0) - F(U, \epsilon)| + |F(V, 0) - F(U, 0)| \\ &< s/4 + s/4 + s/4 + s/4 = s. \end{aligned}$$

Moreover, for $V \in C$,

$$||V - V(\epsilon)|| \leqslant ||V - U|| + ||U - V(\epsilon)|| < \delta/2 + \delta/2 = \delta.$$

This establishes that $C \subseteq D(V(\epsilon), \epsilon, \delta)$. ∎

Thus the sets $D(V(\epsilon), \epsilon, \delta)$ have non-empty interior and the size of that interior is bounded from below. Next, we would like to show that for small enough perturbations, there exist local minima $V(\epsilon)$ of $F(V, \epsilon)$ that are arbitrarily close to U.

Lemma 9 (Existence of close local minima). *For all $\gamma > \delta > 0$, there exists an $\epsilon_0 > 0$ such that for $0 \leqslant \epsilon \leqslant \epsilon_0$, there is a $V(\epsilon)$ such that $||V(\epsilon) - U|| < \delta$ and $V(\epsilon)$ is a local minimizer of $F(V, \epsilon)$ in the sense that $F(V(\epsilon), \epsilon) \geqslant F(V, \epsilon)$ on the set $\{V : ||V - V(\epsilon)|| < \gamma\}$.*

Proof. By compactness, for every ϵ, there exists a minimum $V(\epsilon)$ of $F(V, \epsilon)$ on the set $\{V : ||U - V|| \leqslant \gamma\}$. Suppose that the claim of the lemma does not hold; then there is a $\gamma > \delta' > 0$ and a sequence of $V(\epsilon_n), \epsilon_n$ with $\epsilon_n \to 0$ and $\delta' \leqslant ||V(\epsilon_n) - U|| \leqslant \gamma$ such that

$$F(V(\epsilon_n), \epsilon_n) - F(U, \epsilon_n) \leqslant 0$$

for all n. Taking limits, we find that $F(Y, 0) \leqslant F(U, 0)$, where $Y = \lim_{n \to \infty} V(\epsilon_n)$ and $\delta' \leqslant ||Y - U|| \leqslant \gamma$, contradicting the fact that U was the only minimum on Γ. ∎

The next lemma deals with the initializer $W(\epsilon)$.

Lemma 10 (Diagonalization of a single matrix). *Suppose the eigenvalues of the $\mathcal{M}_0$ are distinct. Let $\epsilon > 0$, and let $W(\epsilon)$ be the estimate of the eigenvalues obtained by diagonalizing one of the $\mathcal{M}_\epsilon$. As $\epsilon \to 0$, $W(\epsilon) \to U$.*

Proof. Follows from the fact that eigenvectors and eigenvalues of $M + \epsilon D$ are continuous in ϵ for small enough ϵ. ∎

We use all of the above results to establish the following important lemma says that we can converge arbitrarily close to a global optimum for small amounts of noise.

Lemma 11 (Convergence to vicinity of global minimum). *There exist $\delta_0 > 0$ and $\epsilon_0 > 0$ such that for all $0 \leqslant \delta < \delta_0$, there exists an $0 \leqslant \epsilon < \epsilon_0$ such that when the Jacobi algorithm is initialized at a point $W(\epsilon)$ obtained from diagonalizing one matrix will almost surely converge to a local minimum V of $F(\cdot, \mathcal{M}_\epsilon)$ with $||V - U|| < 2\delta$.*

Proof. By Lemma 8, for all $\gamma \geqslant \delta \geqslant 0$, there exists an open set C around U, as well as $\epsilon_1, t > 0$ such that for all $0 \leqslant \epsilon \leqslant \epsilon_1$, for all $0 \leqslant \delta \leqslant \delta_0$, and for all local minimizers $V(\epsilon)$ such that $||U - V(\epsilon)|| < t$, $C \subseteq D(V(\epsilon), \epsilon, \delta)$, where $D(V(\epsilon), \epsilon, \delta)$ is a set defined in Lemma 7.

Let one such $\delta > 0$ be fixed. By Lemma 9, there exists an $\epsilon_2 > 0$ such that for $0 < \epsilon \leqslant \epsilon_2$, there exists a local minimizer $V(\epsilon)$ such that $||V(\epsilon) - U|| < \min(t, \delta)$.

Now let $\beta > 0$ be the radius of a ball around U that is contained in C. By Lemma 10, there exists an $\epsilon_3 > 0$ such that for $0 < \epsilon \leqslant \epsilon_3$, $||W(\epsilon) - U|| < \beta$. Note that the lemma holds because we obtain eigenvalues by random projection, and hence they will be distinct almost surely.

Since $||W(\epsilon) - U|| < \beta$, $W(\epsilon) \in C$. Suppose that we start the algorithm on the function $F(V, \epsilon)$ at $W(\epsilon)$. Then since $S \subseteq D(V(\epsilon), \epsilon, \delta)$, all subsequent iterates will be in $D(V(\epsilon), \epsilon, \delta)$ as well. By compactness, the iterates will converge to a point Y satisfying $||V(\epsilon) - Y|| < \delta$. Furthermore, since $||U - V(\epsilon)|| < \min(t, \delta)$, we have

$$||U - Y|| < 2\delta.$$

Thus the lemma holds with $\delta_0 := \gamma$ and $\epsilon_0 = \min(\epsilon_1, \epsilon_2, \epsilon_3)$. ∎

A.4 Cardoso's lemma holds in a neighborhood of U

Finally, we will make the argument that Cardoso's lemma holds for the points to which the Jacobi method will converge.

First, recall that any orthogonal matrix V can be written as $U \exp(E)$ for some skew-symmetric matrix E. The fact that $V = (I + E)U + o(E)$ follows from this observation by using the series form of the matrix exponential $\exp(E)$.

Cardoso's result holds for any $V = (I + \epsilon E')U + o(\epsilon E')$, where $E = O(1)$. Thus, it is enough to show that the E defined above satisfies $E = O(\epsilon)$. To prove this, we will use the following lemma by Cardoso:

Lemma 12 (Cardoso). *If V is critical point of $F(\cdot, \mathcal{M})$, then $S(V, \mathcal{M}) = S^T(V, \mathcal{M})$, where*

$$S(V, \mathcal{M}) := \sum_k \sum_{i \neq j} (e_i^T V^T M_k V e_j)(e_i e_j^T V^T M_k^T V - V^T M_k^T V e_i e_j^T)$$

Lemma 13. *Let $V = (I + E)U + o(E)$ be the expanded expression of a local minimizer $V(\epsilon)$ of $F(\cdot, \mathcal{M}_\epsilon)$. Then $||E|| = O(\epsilon)$ as $\epsilon \to 0$.*

Proof. Before we proceed, observe that the faction that V converges to a global minimizer U by our previous lemma and thus $E \to 0$ as $\epsilon \to 0$. All we need to establish is the rate. For this, we will use Lemma 12. Recall that V is a critical point of F whenever $S(V, \mathcal{M}) = S(V, \mathcal{M})^T$, where

$$S(V, \mathcal{M}) := \sum_k \sum_{i \neq j} (e_i^T V^T M_k V e_j)(e_i e_j^T V^T M_k^T V - V^T M_k^T V e_i e_j^T).$$

It is shown by Cardoso (1994) that

$$(S(V, \mathcal{M}_\epsilon) - S(V, \mathcal{M}_\epsilon)^T)_{ij} = 2\epsilon \sum_k (d_i(k) - d_j(k)) e_i^T U^T R_k U e_j + 2E_{ij} \sum_k (d_i(k) - d_j(k))^2 + o(\epsilon) + o(E_{ij}).$$

Thus, we can define

$$f_{ij}(E_{ij}, \epsilon) = (S(V, \mathcal{M}_\epsilon) - S(V, \mathcal{M}_\epsilon)^T)_{ij}.$$

Let e_n be any sequence such that $V(e_n) \to U$ as $e_n \to 0$ and let $E_{ij}^{(n)}$ be the (i, j)-th element of the matrix E_n defined in $V_n = (I + E_n) + o(E_n)$. Combining this with Cardoso's expansion and using the fact that $V(e_{(n)})$ is a minimizer, we have that

$$0 = A_{ij} E_{ij}^{(n)} + B_{ij} \epsilon_n + o(E_{ij}) + o(\epsilon),$$

for all n, where A_{ij} and B_{ij} are the constants from Cardoso's expansion. This immediately establishes that $E_{ij} = O(\epsilon)$ (with constant B_{ij}/A_{ij}). The fact that $||E|| = O(\epsilon)$ follows by combining the result for each (i, j).
∎

The above lemma establishes that Lemma 1 to holds for points attained by the Jacobi algorithm.