Algorithm Convergence

Jiaxin Hu

05/27/2020

Question: Define a neighbourhood of $\mathcal{A} = (\mathcal{C}, \{M_k\})$:

$$R(\mathcal{A}) = \{\mathcal{A}' | \mathcal{A}' = (\mathcal{C}', \{M_k'\}) = (\mathcal{C} \times_1 P_1 \times_2 \cdots \times_K P_k, \{M_k P_k^T\})\},$$

where P_ks are orthogonal matrices. Is it possible $\mathcal{B}(\mathcal{A}) = \mathcal{B}(\mathcal{A}'')$ if $\mathcal{A}'' \notin R(A)$?
If the answer is NO, the uniqueness of tucker decomposition is valid up to orthogonalization. If the answer is YES, it is hard to guarantee the local uniqueness of tucker decomposition even though we consider the orthogonal problem. justify.
My answer is YES now. However, can we construct some neighbourhoods of $\mathcal{A}$, $N(\mathcal{A}) = \{\mathcal{A}' | \mathcal{A}' \notin R(\mathcal{A}), \mathcal{B}(\mathcal{A}) = \mathcal{B}(\mathcal{A}')\}$ and $L(\mathcal{A}) = R(\mathcal{A}) + N(\mathcal{A})$? The points in the same $L(\mathcal{A})$ are called equivalent points. Therefore, the isolation of stationary point is guaranteed up to equivalence and tensor lipschitz holds for two non-equivalent points.
If the true answer is NO, let $L(\mathcal{A}) = R(A)$.

1 CONVERGENCE PROPERTY

Now, we study the convergence property for the actual estimation sequence $(\mathcal{C}^{(t)}, \{M_k^{(t)}\})$ and $\mathcal{B}^{(t)} = \mathcal{C}^{(t)} \times_1 M_1^{(t)} \times_2 \cdots \times_K M_K^{(t)}$. To simplify the analysis, we set the hyper-parameter α to infinity and define the notation $\mathcal{A} = (\mathcal{C}, \{M_k\})$. Define $\|\mathcal{A} - \mathcal{A}'\|_F = \|\mathcal{C} - \mathcal{C}'\|_F + \sum_k \|M_k - M_k'\|_F$. We propose following assumptions.

A1.(Regularity Condition) The log-likelihood function $\mathcal{L}(\mathcal{A})$ is continuous. The set $\{\mathcal{A} : \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ is compact.
A2.(Strictly local maximum condition) Each block update in Algorithm 1 is well-defined; i.e. the GLM solution for each update exists and is unique, the corresponding diagonal sub-block Hessian of $\mathcal{L}(\mathcal{A})$ is negative definite at the solution.
A3.(Local Uniqueness condition) The stationary points of $\mathcal{L}(\mathcal{A})$ are isolated up to equivalence.
A4.(Tensor Lipschitz condition) The tensor representation $\mathcal{B}(\mathcal{A})$ is tensor Lipschitz at $\mathcal{A}^*$. That means there exists two constant $c_1, c_2 > 0, s.t.$ You need to modify your definition of F-norm in the quotient space A/~, where ~ denotes the equivalent relationship. Otherwise, none of the tensors would satisfy the condition (A4).

$$c_1 \left\|\mathcal{A}' - \mathcal{A}''\right\|_F \leq \left\|\mathcal{B}(\mathcal{A}') - \mathcal{B}(\mathcal{A}'')\right\|_F \leq c_2 \left\|\mathcal{A}' - \mathcal{A}''\right\|_F,$$

where $L(\mathcal{A}') \bigcap L(\mathcal{A}'') = \varnothing$ and $\mathcal{A}', \mathcal{A}''$ are sufficiently close to $\mathcal{A}^*$.

Under your current definition of F-norm, every tensor has the following property:
there exists a sequence of A_n such that Fnorm(B_n-B) —> 0, but Fnorm (A_n - A) >= 0.1 for all n.

Proposition 1 (Algorithm 1 Convergence)**.** *Suppose A1-A3 holds.* Therefore, c_1 == 0!
1. (Global Convergence) Any sequence $\mathcal{A}^{(t)}$ generated by Algorithm 1 convergences to a stationary point of $\mathcal{L}(\mathcal{A})$.

2.(Local Linear Convergence) Let $\mathcal{A}^*$ be a local maximizer of $\mathcal{L}(\mathcal{A})$. There exists an ϵ-neighborhood of $\mathcal{A}^*$, such that, for any $\mathcal{A}^{(0)}$ in the neighborhood, the iterates $\mathcal{A}^{(t)}$ generated by Algorithm 1 linearly convergent to $\mathcal{A}^*$.

$$\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F \leq \rho^t \left\|\mathcal{A}^{(0)} - \mathcal{A}^*\right\|_F$$

where $\rho \in (0,1)$ is a contraction parameter. Further, if A4 holds, there exists a constant C such that

$$\left\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}(\mathcal{A}^*)\right\|_F \leq C\rho^t \left\|\mathcal{B}(\mathcal{A}^{(0)}) - \mathcal{B}(\mathcal{A}^*)\right\|_F.$$

Assumptions are easy to satisfy. *A1* ensures the maximum of likelihood function exists and log-likelihood is upper bounded due to compactness. Therefore, the stopping rule for Algorithm 1 is well-defined. *A2* ensures the negative-definiteness of sub-block Hessian only for $\mathcal{C}$ or M_k. Note that the full Hessian needs not to be negative definite for all decision variables simultaneously. We only require the solution for one variable($\mathcal{C}$ or M_k) given by GLM is unique when other decision variables are fixed. *A3* guarantees the uniqueness of stationary points up to equivalence. *A4* is mild because the tensor representations are definitely different with non-equivalent points. In our algorithm, the iterates $\mathcal{A}^{(t)}$s are non-equivalent with different t and we can use *A4* to discuss the convergence rate of tensor representation. why? justify.

2 PROOF

For Global Convergence, we need to show every convergent sub-sequence of $\mathcal{A}^{(t)}$ converges to the same limiting point because $\mathcal{A}^{(t)}$ is a bounded sequence by condition *A1*. Suppose an arbitrary convergent sub-sequence $\mathcal{A}^{(t_k)}$ with limiting point $\mathcal{A}^*$. As the objective function $\mathcal{L}(\mathcal{A}^{(t)})$ monotonically increases along with t, $\mathcal{A}^*$ is a stationary point of $\mathcal{L}(\mathcal{A}^{(t)})$. For every convergent sub-sequence, the set of all the limiting points are contained in the set $\{\mathcal{A} : \mathcal{L}(\mathcal{A}) \geq \mathcal{L}(\mathcal{A}^{(0)})\}$ and thus is compact. Compactness implies that the set of limiting points is also connected. The isolation condition by condition *A3* implies the finite number of stationary points. Therefore, the set of all the limiting points becomes a single point and $\mathcal{A}^{(t)}$ converges to a stationary point of $\mathcal{L}(\mathcal{A})$.

To show the Local Convergence, define the differential mapping $S : S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$. Let H be the Hessian matrix of $\mathcal{L}(\mathcal{A})$ at the local maximum $\mathcal{A}^*$. We partition the H:

$$d^2\mathcal{L}(\mathcal{A}^*) = d^2\mathcal{L}(\mathcal{C}^*, M_1^*, \cdots, M_K^*) = \begin{pmatrix} d^2_{CC}\mathcal{L} & d^2_{CM_1}\mathcal{L} & \cdots & d^2_{CM_K}\mathcal{L} \\ d^2_{M_1C}\mathcal{L} & d^2_{M_1M_1}\mathcal{L} & \cdots & d^2_{M_1M_K}\mathcal{L} \\ \vdots & \vdots & \ddots & \vdots \\ d^2_{M_KC}\mathcal{L} & d^2_{M_KM_1}\mathcal{L} & \cdots & d^2_{M_KM_K}\mathcal{L} \end{pmatrix} = L + D + L^\top,$$

where L is strictly block lower triangle matrix and D is the diagonal part. By condition *A2*, every diagonal block of H is negative definite and thus $L+D$ is invertible. According to Bezdek(2003), we have $dS(\mathcal{A}^*) = -(L+d)^{-1}L$. Let ρ denote the spectral radius of $dS(\mathcal{A}^*)$ and $\rho = max_i|\lambda_i(-(L+d)^{-1}L)|$ is

strictly smaller than 1. According to theorem 2 of Bezdek(2003), we have,

justify.

$$\begin{aligned}\left\|S(\mathcal{A}^{(t)}) - S(\mathcal{A}^*)\right\|_F &= \left\|\int_0^1 S'(t\mathcal{A}^{(t)})\mathcal{A}^{(t)} - S'(t\mathcal{A}^*)\mathcal{A}^* dt\right\|_F \\ &\leq \int_0^1 \left\|S'(t\mathcal{A}^{(t)})\mathcal{A}^{(t)} - S'(t\mathcal{A}^*)\mathcal{A}^*\right\|_F dt \\ &\leq \rho\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F.\end{aligned}$$

That implies S is a contraction mapping with metric $d(,) = \|.\|_F$ because $(\mathcal{A}, \|.\|_F)$ is indeed a complete metric space. By contraction principle, the sequence $S(\mathcal{A}^{(t)}) = \mathcal{A}^{(t+1)}$ linearly converges to the point $\mathcal{A}^*$ that $S(\mathcal{A}^*) = \mathcal{A}^*$,

$$\left\|\mathcal{A}^{(t)} - \mathcal{A}^*\right\|_F \leq \rho^t \left\|\mathcal{A}^{(0)} - \mathcal{A}^*\right\|_F,$$

justify.

for $\mathcal{A}^{(0)}$ sufficiently close to $\mathcal{A}*$. Further, $L(\mathcal{A}^{(t)}) \bigcap L(\mathcal{A}^*) = \varnothing$ and $L(\mathcal{A}^{(0)}) \bigcap L(\mathcal{A}^*) = \varnothing$. According to condition $A4$, there exists a constant C

$$\left\|\mathcal{B}(\mathcal{A}^{(t)}) - \mathcal{B}^*\right\|_F \leq C\rho^t \left\|\mathcal{B}(\mathcal{A}^{(0)}) - \mathcal{B}^*\right\|_F.$$

Now we prove $d(\mathcal{A}, \mathcal{A}') = \|\mathcal{B}(\mathcal{A}) - \mathcal{B}(\mathcal{A}')\|_F$ is not a valid metric for contraction mapping. We only need to show $\|\mathcal{B}(\mathcal{A}) - \mathcal{B}(\mathcal{A}')\|_F$ does not satisfies the identity of indiscernibles; i.e. $\|\mathcal{B}(\mathcal{A}) - \mathcal{B}(\mathcal{A}')\|_F = 0 \Leftrightarrow \mathcal{A} = \mathcal{A}'$. Suppose we have $\mathcal{A}, \mathcal{A}'$ such that

$$\mathcal{A} = (\mathcal{C}, \{M_k\}), \mathcal{A}' = (\mathcal{C} \times_1 P_1 \times_2 \cdots \times_K P_k, \{M_k P_k^T\}),$$

where $P_k \in \mathbb{R}^{r_k \times r_k}$ is an orthogonal matrix. Therefore, $\mathcal{B}(\mathcal{A}) = \mathcal{B}(\mathcal{A}')$ and $\|\mathcal{B}(\mathcal{A}) - \mathcal{B}(\mathcal{A}')\|_F = 0$ but $\mathcal{A} \neq \mathcal{A}'$. The tensor representation is not a valid metric on the space of $\mathcal{A}$.