Tensor Approximation using fibre cross
Espig, Grasedyek and Hackbusch

1 introduction

ALS algorithm workhorse in computing low rank approximations and decomps of higher-order tensors. In this paper investigate the ALS algorithm for low-rank approx of tensors in canonical format. Majority of lit focuses on global propteries of the iteration like existance of convergent subsequences and critial points or the occurrence of swamps. Any algorithm should also be backed by a local convergence theory. Local convergence theory means a theory for the parameters (iterates) of an algorithm, not for the residuals. Few works on rank one approximation source [27]. For higher ranks most of the difficulties with the global behavior of ALS seem to be intimately related to the fact the the approximation problem can be ill-posed or ill-conditioned. Not discussed in this paper. Will assume that a local minimum exists.
The ALS algorithm is an alternating optimization scheme so called nonlinear block Gauss-Seidel method. Well developed local theory for this type of method. It has been shown that, up to higher order terms, locally equals the linear block Gauss-Seidel iteration applied to the Hessian matrix at the solution. Thus it is locally linearly convergent provided that this Hessian matrix is positive definite.
The problem is that local minima in canonical low-rank tensor approximations do not have this property due to the nonuniqueness of the representation caused by the scaling indeterminacy. The Gauss-Seidel method for only semidefinite linear systems is still convergent but only up to elements in the null space of the system matrix. One has to remove the null space of the hessian from the iteration.
This paper present a local convergence result for ALS under the assumption that the Hessian of the loss function at the solution is positive definite execpt on a trivial null space caused by scaling. This assumption requires local uniqueness of the CP decomposition which is reasonable for tensors of order higher than three. The main idea of the proof is to show that an inbuild normalization procedure in the ALS algorithm acts as a projection onto a subspace complementary to the null space of the Hessian. It turns out that the global minimization in one ALS direction can be replaced by a single newton step to obtain an approximative scheme which is still convergent.
The approach provided avoids the explicit use of lagrange multipliers. Can by applied to delicate types of redundancies as they occur in the Tucker format or newly developed TT format (looking at these sources soon)

f(x) function f'(x) derivative of function at x and f"(x) for the hessian. By (.,.) and —— . —— denote frobenius inner product and norm. write f"(x)[h,h] instead of (h,f"(x)h) for the application of the second derivitive since the formar notation is also meaningful for vector valued functions such as tau introduced later. note that it holds for a rank one tensor that

$$(a \otimes b \otimes c, a' \otimes b' \otimes c') = (a, a')(b, b')(c, c')$$

2 The ALS algorithm

Let n1,n2,n3 $\in \mathbb{N}$ {1} and let T $\in \mathbb{R}^{n1xn2xn3}$ T $\neq 0$ be a real third order tensor treated here as a three-D array given $r \in \mathbb{N}$ let

$$X = \mathbb{R}^{n1xr} \times \mathbb{R}^{n2xr} \times \mathbb{R}^{n3xr}$$

The elements of X will be denoted by x = (A,B,C). consider the function. The matrices are called factor matrices in the literature. function

$$f : X \to \mathbb{R} : x = (A, B, C) \to 1/2\|T - \sum_{j=1}^{r} a_j \otimes b_j \otimes c_j\|^2$$

seak to find a solution of

$$f(A, B, C) = min.$$

it assumed that at least one local min exists. this is denoted by x* = (A*, B*, C*).

ALS algorithm is a simple method for solving the minimization. Given a starting point $x^{(0)}$ it consists of iterating the cycle

$$A^{n+1} = argmin_{A \in \mathbb{R}^{n1xr}} f(A, B^{(n)}, C^{(n)}),$$
$$B^{n+1} = argmin_{B \in \mathbb{R}^{n2xr}} f(A^{(n+1)}, B, C^{(n)}),$$
$$C^{n+1} = argmin_{C \in \mathbb{R}^{n3xr}} f(A^{(n+1)}, B^{(n+1)}, C),$$

This algorithm is a particular example of the nonlinear block Gauss-Seidel (relaxation) method [18,23]
Each step can be defined using an operator S so $S(n^{(n)})$ is the set of updated factor matrices.
Consider only local minima in the open subset

$$\hat{X} = \{(A, B, C) \in X | a_j \neq 0, b_j \neq 0 c_j \neq 0 for j = 1, 2, ..., r\}$$

Restricting to $\hat{X}$ is reasonable to avoid pseudoinverse since only if $x* \in \hat{X}$ we can hope that each iteration of factor matrices has a unique solution. We consider local minima in $X left divides \hat{X}$ to be too degenerate for our framework.
The major difficulty in the analysis of the algorithm lies in the fact that x* cannot be an isolated local minimum of the minimization function, since every rank one term can be replaced by a scaling as long as the multiplication of the scaling =1. If the scaling factors are positive called rescaling of x*. in fact every such rescaled solution is a local minimm and a fixed point of the iteration S.
There is on reason why the iteration should converge to a particular prescribed solution x*.
Additionally it can happen what a component can tend to infinity while another tends to zero such that the product remains bounded. This deteriorates the condition of each microstep.
for both reasons a normalization strategy invoked. represented in typical cp form.
to avoid the parameter consider tensors in the equilibrated format

$$\sum_{j=1}^{r} a_j \otimes b_j \otimes c_j with \|a_j\| = \|b_j\| = \|c_j\| \forall j$$

removing possibilities of rescaling
the function R is used to put the tensors into equilibriated format without changing the signs of the vectors.
Fixed points of this algorithm are necessarily equilibrated. Variants of ALS are possible include normalization instead of equilibration. To increase numerical stability it may be reasonable to equilibrate after each microstep (S). They are all the same in the sense that given a single starting point they produce the same sequence of iterates up to rescaling.

3 convergence analysis

3.1 positive definiteness assumption.

return attention to the scaling indeterminacy. the function f is constant on the 2r-dimensional submanifold
Assumption 1: f"(x*) shall be positive definite in every direction except those tangent to the rescalings.
This implies that the paramatrization x* is locally essentially unique. Disucss if realistic later.

Lemma 3.1: R'(x*) the equilibration function is a projector whose null space is precisely TM^*_{x*} . Proof on page 5/6

Lemma 3.2: Assume that Assumption 1 holds. Then the ALS operator S is well-defined and continuously differentiable in some neighborhood of x*. x* is a fixed point of S and we have $S'(x^*) = I - M^{-1} f''(x^*)$ where M is the lower block triangular part of f"(x*) corresponding to the partition x = (A,B,C)

In a possibly smaller neigborhood the composition $R ø S$ is well defined and continuously differentiable. One loop of algorithm 1 is feasible if the current iterate $x^{(n)}$ is close enough to x*
Lemma 3.2 holds under weaker assumption that the diagonal blocks of f”(x*) are positive definite.

3.2 Convergence theorem

To stay at local minimum, one could replace in each microstep the function f by tis second order expansion at the current mircroiterate and minimize that one. This means perfoming a single newton step wrt current block variable. This procedure is called approximate nonlinear relaxation in [23] and the nonlinear Gauss-Seidel-Newton method in [18]

$$A^{n+1)} = A^{(n)} - [f''_A(A^{(n)}, B^{(n)}, C^{(n)},)]^{-1} * \nabla_A f(A^{(n)}, B^{(n)}, C^{(n)})$$

and similar for B and C n+1 with properly updated previous factor matrices.
Where $\nabla_A f$ and f''_A the gradient and Hessian (diagonal block) with regards to variable A.

4 Important point from numerical tests

The normalized factor matrices of the calculated solutions were randomly perturbed by different orders of magnitude and then used as a starting point for a restart of the ALS algorithm. The initial solution would be recovered if the perturbation was of magnitude at most 10^0 Larger perturbations result in an almost random starting point. So the als iteration almost surely will converge to a different critical point. Usually the convergence rate decreased for larger r but not in all experiments. Example of convergence rate for differnt ranks is plotted in figure 4.1. The convergence rate depends on the rank. FOr special tensors such as hyperdiagonal or more genrally complete or orthogonal tensors one can exactly determine the best rank-r approximation by truncation of a complete orthogonal representation. The tensors converged very fast (at most 4 iterations independent of r).

5 Summary

The easiest way to prove local linear convergence of a nonlinear fixed point iteration consists of showing a contraction property of its linearization at the fixed point.
The idea of convergence proof was to realize that a suitable chosen equilibration operaton, which fixed the canonical representation and is employed in practice to stabilize the iteration, removes the parts belonging to this tangent space from the first order terms of the error iteration. After assuming the Hessian is positive in all directions, local convergence of algorithm 1 was routinely established.