Computational Algebraic Statistics, Theories and Applications
December 10-11, 2008, Kyoto

CPOPT: OPTIMIZATION FOR FITTING CANDECOMP/PARAFAC MODELS

EVRIM ACAR, TAMARA G. KOLDA, AND DANIEL M. DUNLAVY

1. Introduction

Tensor decompositions (e.g., higher-order analogues of matrix decompositions) are powerful tools for data analysis. In particular, the CANDECOMP/PARAFAC (CP) model [1, 2] has proved useful in many applications such chemometrics, signal processing, and web analysis; see [3] for details. The problem of computing the CP decomposition is typically solved using an alternating least squares (ALS) approach. We discuss the use of optimization-based algorithms for CP, including how to efficiently compute the derivatives necessary for the optimization methods. Numerical studies highlight the positive features of our CPOPT algorithms, as compared with ALS and Gauss-Newton approaches.

2. CANDECOMP/PARAFAC Background

Let $\mathcal{Z}$ be an N-way tensor of size $I_1 \times I_2 \times \cdots \times I_N$. A rank-$R$ CP model [1, 2] is $\mathcal{Z} \approx [\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!] = \sum_{r=1}^{R} \mathbf{a}_r^{(1)} \circ \cdots \circ \mathbf{a}_r^{(N)}$. Here the $\circ$ notation indicates the vector outer product. Each factor matrix $\mathbf{A}^{(n)}$ is of size $I_n \times R$, and $\mathbf{a}_r^{(n)}$ denotes the rth column of $\mathbf{A}^{(n)}$. The corresponding least squares optimization problem is

$$\min_{\mathbf{A}^{(1)},\ldots,\mathbf{A}^{(N)}} f(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}) \equiv \frac{1}{2} \left\| \mathcal{Z} - [\![\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}]\!] \right\|^2. \tag{1}$$

The standard approach for fitting CP models is an alternating least squares (ALS) method which solves for one factor matrix at a time while holding the others constant; that is, the method loops through $n = 1, \ldots, N$ and solves

$$\min_{\mathbf{A}^{(n)}} f(\mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(N)}). \tag{2}$$

The loops continue until there is no further improvement in the objective function. Equation (2) reduces to a linear least squares problem of special form and can be solved efficiently. See [5, 3] for further details.

3. CP via Optimization

We propose to solve (1) using an optimization approach. In this case, f is a mapping from the cross-product of N two-dimensional vectors spaces so that $f : \mathbb{R}^{I_1 \times R} \otimes \mathbb{R}^{I_2 \times R} \otimes \cdots \otimes \mathbb{R}^{I_N \times R} \mapsto \mathbb{R}$. In other words, we have a function of $P = R \sum_{n=1}^{N} I_n$ variables. It is possible to calculate the gradient, i.e., the partial derivative of f with respect to each column of $\mathbf{A}^{(n)}$, as follows

$$\frac{\partial f}{\partial \mathbf{a}_r^{(n)}} = -\left(\mathcal{Z} \underset{\substack{m=1 \\ m \neq n}}{\overset{N}{\times}} \mathbf{a}_r^{(m)} \right) + \sum_{\ell=1}^{R} \left(\prod_{\substack{m=1 \\ m \neq n}}^{N} \mathbf{a}_r^{(m)\mathsf{T}} \mathbf{a}_\ell^{(m)}. \right) \mathbf{a}_\ell^{(n)}.$$

Here $\times$ denotes tensor-times-vector multiplication in the specified modes. Note that this is the same as the result in [4], where it was used as the basis for a

multiplicative update rule. We can use the gradient to solve (1) via a standard optimization method such as the nonlinear conjugate gradient method.

4. Experiments

We compare our optimization approach for solving (1), which we call CPOPT, with ALS and a Gauss-Newton approach called INDAFAC [6]. We ran a series of 600 experiments on simulated datasets of rank three, using the experimental procedures outlined in [7]. The datasets had different levels of both homoscedastic and heteroscedastic noise and we fit a CP model using both three (exact) and four (overfactoring) components. The average performance of each algorithm is presented in Table 1. Because CPOPT does not need to solve the Gauss-Newton equation, it scales better to larger problems and is faster than INDAFAC at the larger sizes without sacrificing its ability to recover the underlying factors (i.e., its accuracy).

	Time (sec)			**Accuracy (%)**		
Size	**ALS**	**INDAFAC**	**CPOPT**	**ALS**	**INDAFAC**	**CPOPT**
$20 \times 20 \times 20$	1.1	1.6	3.4	86.0	100.0	100.0
$50 \times 50 \times 50$	0.8	28.1	8.5	66.8	99.8	100.0
$100 \times 100 \times 100$	2.7	257.1	40.6	57.2	100.0	100.0

Table 1. Comparison of different methods for computing the CP decomposition.

References

[1] J. D. Carroll and J. J. Chang, *Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition*, Psychometrika, 35 (1970), pp. 283–319.

[2] R. A. Harshman, *Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis*, UCLA working papers in phonetics, 16 (1970), pp. 1–84. Available at `http://publish.uwo.ca/~harshman/wpppfac0.pdf`.

[3] T. G. Kolda and B. W. Bader, *Tensor decompositions and applications*, SIAM Review. In press. Preprint available at `http://csmr.ca.sandia.gov/~tgkolda/ref#TensorReview`.

[4] A. Shashua and T. Hazan, *Non-negative tensor factorization with applications to statistics and computer vision*, in ICML 2005: Proceedings of the 22nd International Conference on Machine Learning, 2005, pp. 792–799.

[5] A. Smilde, R. Bro, and P. Geladi, *Multi-Way Analysis: Applications in the Chemical Sciences*, Wiley, West Sussex, England, 2004.

[6] G. Tomasi and R. Bro, *PARAFAC and missing values*, Chemometrics and Intelligent Laboratory Systems, 75 (2005), pp. 163–180.

[7] ———, *A comparison of algorithms for fitting the PARAFAC model*, Computational Statistics & Data Analysis, 50 (2006), pp. 1700–1734.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories, PO Box 969, MS 9159, Livermore, CA 94551 USA
E-mail address: `eacarat@sandia.gov`
URL: `http://csmr.ca.sandia.gov/~eacarat`

Sandia National Laboratories, PO Box 969, MS 9159, Livermore, CA 94551 USA
E-mail address: `tgkolda@sandia.gov`
URL: `http://csmr.ca.sandia.gov/~tgkolda`

Sandia National Laboratories, PO Box 5800, MS 1318, Albuquerque, NM 87185 USA
E-mail address: `dmdunla@sandia.gov`
URL: `http://www.cs.sandia.gov/~dmdunla`