A GREEDY ALGORITHM FOR MODEL SELECTION OF TENSOR DECOMPOSITIONS

Austin J. Brockmeier[*]*, Jose C. Principe*

Electrical and Computer Engineering Department
University of Florida
Gainesville, FL

Anh Huy Phan, Andrzej Cichocki[†]

Brain Science Institute
RIKEN
Wako-shi, Saitama, Japan

ABSTRACT

Various tensor decompositions use different arrangements of factors to explain multi-way data. Components from different decompositions can vary in the number of parameters. Allowing a model to contain components from different decompositions results in a combinatoric number of possible models. Model selection balances approximation error and the number of parameters, but due to the number of possible models, post-hoc model selection is infeasible. Instead, we incrementally build a model. This approach is analogous to sparse coding with a union of dictionaries. The proposed greedy approach can estimate a model consisting of a combination of tensor decompositions.

***Index Terms*—** tensor decompositions, greedy algorithm, model selection

1. INTRODUCTION

Linear synthesis models are fundamental to multivariate signal processing tasks such as denoising, compression, and classification. In this work we explore an approach to approximate data arranged in a tensor, or multiway array, via a combination of data-dependent bases. The bases are chosen from two or more sets each estimated by tensor decomposition yielding orthogonal components.

Truncated singular value decomposition (SVD) finds the optimal reduced rank approximations of data stored in matrices, as proven by Eckart and Young [1]. In multivariate signal processing, there may be multiple ways the signal can be arranged before approximation. For instance, if the signal is arranged as a tensor, then a large number of decomposition/approximation models have been proposed that exploit structure along different modes of the data.

Here we investigate tensor decompositions that can be written as a summation of component tensors each formed as tensor outer products. Decompositions of this nature are of interest because orthogonality can be enforced on any of the factors of the outer product and the resulting components tensors will be orthogonal [2]. This enables us to greedily build tensor models.

Decompositions that are combinations of tensor outer products include the canonical polyadic decomposition (CPD) [3] (also called CANDECOMP [4] or PARAFAC [5]), certain cases of the block tensor decomposition [6], and two-factor outer product expansions such as "Tucker1" models [7, 8, 9]. Notable exceptions are general Tucker models [8]. Additional block-based decompositions can be posed as tensor outer products by leveraging the "tensor Kronecker product" [9, 10, 11].

A general multilinear model can be formed by combining models with different arrangements and ranks. This is what we refer to as a heterogeneous model. Examples of these models include the rank-$(L_r, L_r, 1)$ block term decomposition in [6], and the Kronecker tensor decompositions in [10, 11, 12]. Another example would be using a combination of rank-$(L, L, 1)$, rank-$(M, 1, M)$, and rank-$(1, N, N)$ decompositions.

Depending on the data's structure, multilinear models with different arrangements and ranks may be better approximations. For a user-chosen number of parameters different models can be formed, and the model yielding the lowest approximation error may be chosen. Alternatively, for models with varying number of parameters, model selection criterion can choose the optimal model from a set with various number of parameters [13, 14, 15, 16, 17, 18].

Post-hoc selection works in homogeneous models. However, due to the flexibility in heterogeneous models, model selection is difficult as there is a combinatorial number of models formed from different combinations of arrangements each with different ranks.

When the components are orthogonal and ordered, the number of components can be chosen post-hoc. While tensor components in the same arrangement can be made orthogonal, it is more difficult for components from different arrangements. Without orthogonality, truncating the rank within each arrangement separately does not yield the same solution as running the decomposition constrained to the truncated rank.

We propose using an iterative algorithm, similar to the canonical one proposed by Kolda [2], to greedily select the combination of arrangements and ranks to form a parsimonious heterogeneous model. The greedy selection is in the same manner as sparse coding algorithms that select vectors from a union of dictionaries [19, 20]. Here the data-dependent bases are estimated from the residual on each iteration, exploiting any multilinear structure in the data. The final number of components can be chosen post-hoc, from the models formed at iteration, based on a model selection criteria.

2. APPROACH

The approach is motivated by tensor decompositions with orthogonal tensor components, orthogonality can be easily enforced for tensors formed by a series of tensor outer products [2]. We use the notation $\boldsymbol{\mathcal{A}}$ to denote a tensor. An order-N tensor has N dimensions (or modes) with the size for the dimensions denoted $I_1 \times I_2 \times \cdots \times I_N$. Given a order-$P$ tensor $\boldsymbol{\mathcal{B}}$ with size $J_1 \times J_2 \times \cdots \times J_P$, the outer product of $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{B}}$ is the order-$(N + P)$ tensor $\boldsymbol{\mathcal{C}} = \boldsymbol{\mathcal{A}} \circ \boldsymbol{\mathcal{B}}$ with size $I_1 \times \cdots \times I_N \times J_1 \times \cdots \times J_P$ with entries such that

[*]This material is based upon work supported by the National Science Foundation under Grant No. OISE-1209922. The work was performed while at RIKEN in Japan with hosting support provided by the Japan Society for the Promotion of Science.

[†]Also affiliated with the EE Dept., Warsaw University of Technology and with Systems Research Institute, Polish Academy of Science, Poland

$\mathcal{C}_{i_1,i_2,\ldots,i_N,j_1,\ldots,j_P} = \mathcal{A}_{i_1,i_2,\ldots,i_N}\mathcal{B}_{j_1,j_2,\ldots,j_P}$. Herein the tensors in a series of outer products $\mathcal{A}^1 \circ \mathcal{A}^2 \circ \cdots \circ \mathcal{A}^M$ are called factors, and each summand $\mathcal{C}_i$ in a linear combinations of tensors $\sum_i \mathcal{C}_i$ are referred to as components.

Consider a series of outer products when each $\mathcal{A}^n \equiv \mathbf{a}^n$ is a vector. An order-N tensor $\mathcal{X}$ is considered to be *rank-1* if it is formed from the outer product of N vectors, i.e., it can written as $\mathcal{X} = \mathbf{a}^1 \circ \mathbf{a}^2 \circ \cdots \circ \mathbf{a}^N$. Otherwise, the rank is R and is equal to the minimal number of rank-1 tensors needed such that $\mathcal{X} = \sum_{r=1}^{R} \mathbf{a}_r^1 \circ \cdots \circ \mathbf{a}_r^N$ [21]; this is the canonical polyadic decomposition (CPD).

Certain block term decompositions [6] consider more general ranks. An order-3 tensor is rank-$(L, L, 1)$ if it can be written as an outer product between a rank-L matrix $\mathbf{A} = \sum_{l=1}^{L} \mathbf{b}_l \mathbf{c}_l^{\mathrm{T}}$ and a vector $\mathbf{a}$, i.e., $\mathcal{X} = \mathcal{A}^1 \circ \mathcal{A}^2 = \mathbf{A} \circ \mathbf{a}$.The rank one term could be assigned to any mode by reordering the modes. Generally, an order-N tensor is rank-$(L, \ldots, L, 1)$ if it can be written as $\mathcal{X} = \mathcal{A}^1 \circ \mathcal{A}^2$ where $\mathcal{A}^1 = \sum_{l=1}^{L} \mathbf{a}_l^1 \circ \cdots \circ \mathbf{a}_l^{N-1}$ is an order-$(N-1)$ tensor with rank L and $\mathcal{A}^2$ is a vector. A tensor can be decomposed into a set of rank-$(L_r, \ldots, L_r, 1)$ tensors as $\mathcal{X} = \sum_{r=1}^{R} \mathcal{A}_r^1 \circ \mathcal{A}_r^2$, where $\mathcal{A}_r^1$ has rank-L_r and $\mathcal{A}_r^2$ is a vector. Under certain conditions this is unique decomposition [6].

The two-factor product expansion [9, 10] takes the form $\mathcal{X} = \sum_{r=1}^{R} \mathcal{A}_r^1 \circ \mathcal{A}_r^2$, where $\mathcal{A}_r^1$ and $\mathcal{A}_r^2$ are general tensors with no specific structure or restriction on rank. The Tucker1 decomposition formed by an outer product between a matrix and vector is of this form [14].The tensor can be reshaped into a matrix, using rearranging and unfolding [9, 10], where each of the factors is treated as vectors and matrix based decompositions (SVD, NMF) are easily applied.

2.1. Constraints

For this approach, we consider combinations of normalized and possibly orthogonal tensors with real numbers as the entries. The inner product of two tensors of the same size is

$$\langle \mathcal{A}^m, \mathcal{A}^n \rangle_F = \mathrm{vec}(\mathcal{A}^m)^{\mathrm{T}} \mathrm{vec}(\mathcal{A}^n). \tag{1}$$

The Frobenius norm of $\mathcal{A}$ can be computed as

$$\|\mathcal{A}\|_F = \sqrt{\langle \mathcal{A}, \mathcal{A} \rangle_F}. \tag{2}$$

For two tensors $\mathcal{C}_1, \mathcal{C}_2$ each formed from a series of outer products of equal sizes $\mathcal{C}_r = \mathcal{A}_r^1 \circ \cdots \circ \mathcal{A}_r^N$, orthogonality of any factor is a sufficient condition for their orthogonality: if $\exists n$ such that $\langle \mathcal{A}_1^n, \mathcal{A}_2^n \rangle_F = 0$ then $\langle \mathcal{C}_1, \mathcal{C}_2 \rangle_F = 0$.

Ensuring this sufficient condition is easy for a set of component tensors (with the same size factors) when at least one factor in the outer product is a rank-1 tensor. The rank-1 factor for each component tensors is vectorized, and all of these vectors are concatenated into a matrix (number of components by the number of elements in the rank-1 tensor). Then the closest set of orthogonal vectors are found via the SVD of this factor matrix. Care should be taken in choosing which mode or modes correspond to the orthogonal factors.

2.2. A general tensor product decomposition

A tensor $\mathcal{X}$ can be represented by a combination of tensors $\mathcal{C}_r$ for $r = 1, \ldots, R$ where each tensor has unit norm $\|\mathcal{C}_r\|_F = 1$ and is formed as the tensor product between N_r factors $\mathcal{A}_r^n$ $n = 1, \ldots, N_r$ [2]. In constraining the norm, the coefficients $s_1, \cdots, s_R$ contain the contributions of each summand

$$\mathcal{X} = \sum_{r=1}^{R} s_r \mathcal{C}_r = \sum_{r=1}^{R} s_r \left(\mathcal{A}_r^1 \circ \cdots \circ \mathcal{A}_r^{N_r} \right). \tag{3}$$

If $N_1 = N_2 = \cdots = N_R$ and all $\mathcal{A}_1^n, \ldots, \mathcal{A}_R^n$ are the same size and rank for $n = 1, \ldots, N_1$, then the decomposition is deemed homogeneous. Alternatively, if the decomposition is heterogeneous, then it can be written as a combination of $P \leq R$ homogeneous models. Let $\mathcal{G}_p$ denote the indexes in the pth group, by definition $\mathcal{G}_1 \cup \cdots \cup \mathcal{G}_P = \{1, \ldots, R\}$ and $\mathcal{G}_i \cap \mathcal{G}_j = \emptyset,\ \forall i \neq j$.

$$\mathcal{X} = \sum_{p=1}^{P} \sum_{r \in \mathcal{G}_p} s_r \mathcal{C}_r = \sum_{p=1}^{P} \sum_{r \in \mathcal{G}_p} s_r \left(\mathcal{A}_r^1 \circ \cdots \circ \mathcal{A}_r^{N_r} \right). \tag{4}$$

We restrict component tensors in the same group to be orthogonal, i.e., $\forall p \quad \langle \mathcal{C}_i, \mathcal{C}_j \rangle_F = 0 \quad \{(i,j) \in \mathcal{G}_p : i \neq j\}$. The proposed algorithm in Section 2.5 uses an iterative method where the candidate component tensors could be forced to be orthogonal to previously added components [2]. We do not explicitly enforce this constraint, but since the components are estimated from the residual, the newly added components are approximately orthogonal to the previously added components. Thus, the result is an orthogonal decomposition [2] where $\forall i \neq j \quad \langle \mathcal{C}_i, \mathcal{C}_j \rangle_F = 0$.

2.3. Approximation

For linear model approximation, the objective is to minimize the distance between a model, here $\hat{\mathcal{X}} = \sum_{r=1}^{\hat{R}} s_r \left(\mathcal{A}_r^1 \circ \cdots \circ \mathcal{A}_r^{N_r} \right)$ and the original tensor $\mathcal{X}$, say in terms of Frobenius norm

$$D(\mathcal{X} \| \hat{\mathcal{X}}) = \|\mathcal{X} - \hat{\mathcal{X}}\|_F = \|\mathcal{X} - \sum_{r=1}^{\hat{R}} s_r \left(\mathcal{A}_r^1 \circ \cdots \circ \mathcal{A}_r^{N_r} \right)\|_F. \tag{5}$$

A special case is the homogeneous decomposition with $\mathcal{X} = \sum_{r=1}^{R} \mathcal{A}_r^1 \circ \mathcal{A}_r^2$, this decomposition is unique up to an ordering by applying the SVD to a particular unfolding of the tensor. If one factor is a vector this is a Tucker1 model [7, 8, 9]. As in classic SVD, the optimal approximation $\hat{\mathcal{X}}$ is found by taking only the largest $\hat{R}$ singular values and their corresponding singular vectors [9]. Polyadic or rank constrained block term decompositions [6] require using alternating least square algorithms [2, 22] to estimate the factors.

In practice, a model selection criterion is often needed to systematically choose how many tensors of each size are needed in the model.

2.4. Model Selection

Increasing the number of parameters in the linear model will always decrease the approximation error; however, to identify underlying structure in the tensor, a model selection criterion is needed to balance the number parameters with the approximation error. A number of heuristics and criteria have been proposed for model selection of Tucker or CP based decompositions of tensors [13, 14, 15, 16, 17]. Some require calculations using both the current models and "gradually augmented models" [16]. Others use criteria for post-hoc analysis between many models of different arrangements and ranks. Mørup and Hansen [17] propose a Bayesian approach to shrink a CP or Tucker model large enough to include any desired ranks; consequently, their approach avoids computing many models.

Here we need a straightforward criterion that uses only the approximation error or each model and number of parameters as input [23]. Initial investigation showed that Akaike Information Criterion (AIC) provides an adequate measure but tends to choose complex models. Here we used the Bayesian Information Criterion (BIC) [24]. Assuming i.i.d. zero-mean Gaussian distributed errors, the model selection problem is to minimize BIC

$$\arg\min_{\hat{\boldsymbol{\mathcal{X}}}} BIC(\hat{\boldsymbol{\mathcal{X}}}) = 2M \ln\left(D(\boldsymbol{\mathcal{X}} \| \hat{\boldsymbol{\mathcal{X}}})\right) + L \ln(M) + C \quad (6)$$

where $\boldsymbol{\mathcal{X}}$ has M elements, L is the number of free parameters in $\hat{\boldsymbol{\mathcal{X}}}$, and C is a constant that is independent of $\hat{\boldsymbol{\mathcal{X}}}$.

This criterion can be used to chose the model structure for the general outer product decomposition, as it allows different models with different number of parameters to be explored. For homogeneous models where all factors have the same size, it is straightforward to select the rank and arrangement that optimizes a model selection criterion. For heterogenous models, it is computationally prohibitive to solve (5) across all of the combinations of different model structures and then pick the best one based on a criterion. One alternative is a greedy approach.

Furthermore, in homogenous models, the number of parameters is proportional to the number of component tenors, but this is not the case in heterogenous models where a large number of low-rank component tenors may provide a more parsimonious fit than a few high-rank components. To handle this we propose a selection criterion for the iterative approach that approximates BIC.

2.5. Greedy Selection Algorithm

A greedy approach is based on the observation that the approximation problem (5) and the model selection criterion are related to sparse coding of vectors in that the error term is equivalent $\|\mathrm{vec}(\boldsymbol{\mathcal{X}}) - \mathrm{vec}(\hat{\boldsymbol{\mathcal{X}}})\|_2 = \|\boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}\|_F$. Additionally, the model selection criterion can be related to the l_0 cost of sparse coding by incorporating a penalty for adding components that correspond to a large number of parameters. Thus, we propose using an iterative algorithm for sparse coding which chooses vectors from a union of dictionaries—morphological component analysis (MCA) [19, 20]—as an approach to solve the model selection of the heterogeneous product decomposition. We consider MCA with a hard threshold selection criterion and an adaptive threshold value to selectively add groups of components into the model [20]. A parsimonious model is built by selectively adding a weighted combination of tensor components each formed from factors along different modes of the tensor. This allows the model to exploit any multilinear structure in approximating the tensor.

Consider the case where the decomposition uses P groups of different sized factors. Within a group all component tensors are restricted to be orthogonal such that each group of components form a set of data-dependent orthonormal tensors. Let L_p denote the number of parameters associated with adding a single component to the model from the pth group.

Initially, the components for the pth group are estimated by an applicable algorithm from the original tensor $\boldsymbol{\mathcal{X}}$ to independently minimize (5). On each iteration of the selection algorithm, a criterion is used to select which components are added. After every selection step, a new set of components are estimated for each group based on approximating the new residual tensor.

Let $\boldsymbol{\mathcal{R}}_t = \boldsymbol{\mathcal{X}} - \hat{\boldsymbol{\mathcal{X}}}_t$ denote the residual after t iterations, $\boldsymbol{\mathcal{R}}_0 = \boldsymbol{\mathcal{X}}$. For each group $p = 1, \ldots, P$, compute the inner product between each component in the group $r \in \mathcal{G}_p$ and the residual $s_r = \langle \boldsymbol{\mathcal{R}}_t, \boldsymbol{\mathcal{C}}_r \rangle_F$. Vector selection in MCA only considers the magnitude of this inner product, but for model selection based on (2.5) we use the term $\sigma_r = 2M \ln(|s_r|) - L_p \ln(M)$, since $-\ln(|s_r|) \propto \ln(D(\boldsymbol{\mathcal{R}}_t \| \boldsymbol{\mathcal{C}}_r))$. At each iteration if σ_r is larger than λ_t then $s_r, \boldsymbol{\mathcal{A}}_r^1, \ldots, \boldsymbol{\mathcal{A}}_r^{N_r}$ will be added to the basis of the model and $\hat{\boldsymbol{\mathcal{X}}}_t$ updated by

$$\hat{\boldsymbol{\mathcal{X}}}_{t+1} = \hat{\boldsymbol{\mathcal{X}}}_t + \sum_{\{r:\sigma_r > \lambda_t\}} s_r \left(\boldsymbol{\mathcal{A}}_r^1 \circ \cdots \circ \boldsymbol{\mathcal{A}}_r^{N_r}\right). \quad (7)$$

After each iteration $\{\boldsymbol{\mathcal{A}}_r^1, \ldots, \boldsymbol{\mathcal{A}}_r^{N_r}\}_r$ are estimated from $\boldsymbol{\mathcal{R}}_{t+1}$. Because the components are estimated from the residual, the components added between two different iterations are nearly orthogonal even though they are not necessarily of the same arrangement. Generally, the iterative approaches cannot find true tensor rank, because the space of tensors with a certain canonical rank is not closed with respect to addition [25].

The choice of the threshold λ_t remains. As described by Bobin et al. [20], the threshold should allow multiple components from the same group to be added at once since they are orthogonal, but it should avoid adding components from different groups since within the same iteration they are not necessarily orthogonal. Let $\sigma^\star = \max_r \sigma_r$ and $p^\star$ be the group where $\sigma^\star$ is achieved; the value for the next best group is $\sigma^\circ = \max_{r \notin \mathcal{G}_{p^\star}} \sigma_r$. By setting $\lambda_t = 0.5(\sigma^\star + \sigma^\circ)$, a so-called mean of max approach [20], multiple components may be added at once but only if they are from the same group. Empirically this approach works well as seen in the next section.

3. SIMULATIONS

3.1. Synthetic Example

We test the proposed algorithm on a synthetic tensor. The tensor $\boldsymbol{\mathcal{X}}$ has size $512 \times 16 \times 32$ and is formed as the sum of two components and noise $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{C}}_1 + \boldsymbol{\mathcal{C}}_2 + \boldsymbol{\mathcal{E}} = \sum_{r=1}^{8} \boldsymbol{\mathcal{A}}_r^1 \circ \boldsymbol{\mathcal{A}}_r^2 + \sum_{r=1}^{12} \boldsymbol{\mathcal{A}}_r^3 \circ \boldsymbol{\mathcal{A}}_r^4 + \boldsymbol{\mathcal{E}}$ where $\boldsymbol{\mathcal{A}}_r^1, \boldsymbol{\mathcal{A}}_r^2, \boldsymbol{\mathcal{A}}_r^3, \boldsymbol{\mathcal{A}}_r^4, \boldsymbol{\mathcal{E}}$ have i.i.d. zero-mean unit-variance Gaussian entries and sizes 512×16, 32×1, 512×1, 16×32, and $512 \times 16 \times 32$, respectively.

$\boldsymbol{\mathcal{C}}_1, \boldsymbol{\mathcal{C}}_2$ are individually Tucker1 models aligned to two different modes; whereas, $\boldsymbol{\mathcal{C}}_1 + \boldsymbol{\mathcal{C}}_2$ (and the best approximation of $\boldsymbol{\mathcal{X}}$) is a case of the Kronecker tensor decomposition [10]. By unfolding the tensor along a single mode the basis of one arrangement can be estimated, and the same for the other arrangement; however, a large number of components from a single unfolding are required to explain the structure along the other mode of the tensor.

As shown in Fig. 1, the greedy algorithm iteratively builds the model, and the model selection criterion identifies the correct number of components. Across 20 Monte Carlo runs, the mean and standard deviation of the correlation coefficients between the estimated components $\hat{\boldsymbol{\mathcal{C}}}_1, \hat{\boldsymbol{\mathcal{C}}}_2$ and the true components $\boldsymbol{\mathcal{C}}_1, \boldsymbol{\mathcal{C}}_2$ were $(0.91, 0.01)$ and $(0.87, 0.01)$, and $(0.98, 0.01)$ between $\hat{\boldsymbol{\mathcal{C}}}_1 + \hat{\boldsymbol{\mathcal{C}}}_2$ and $\boldsymbol{\mathcal{C}}_1 + \boldsymbol{\mathcal{C}}_2$.

3.2. Face Images

To test the approach's ability to succinctly describe data we use AT&T Laboratories Cambridge database of faces commonly known as ORL[1]. We use a a subset of the images and down-sample them to 56×46 grayscale pixels with pixel values in the range $[0, 255]$. We

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

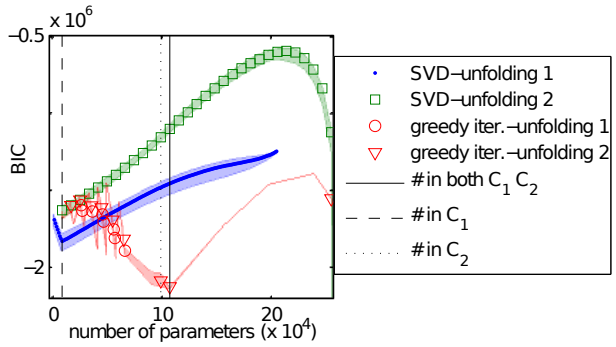

Fig. 1. Model selection for a synthetic model when the underlying model is a combination of components with two different multilinear arrangements. We compare two Tucker1 models, which use truncated SVD along a single unfolding. Alternatively, the proposed algorithm can iteratively select components from both unfoldings. The markers indicate a single run, and the filled area indicates the range across 20 Monte Carlo realizations of the random tensor. The correct model was built on all 20 realizations and was selected as the one that has minimizes the BIC .

use all the images (40 people, 10 photos per person) and form a $56 \times 46 \times 400$ tensor. We run the algorithm to select components from a set of different block-term decompositions [6], see Fig. 2. The orthogonality of the decomposition was empirically confirmed: the maximum correlation coefficient between component tensors was 0.0585.

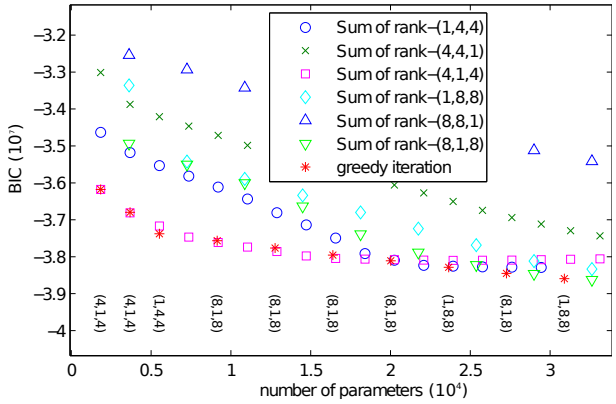

Fig. 2. Model selection for compression of the ORL face dataset as a $56 \times 46 \times 400$ tensor. Different permutations of block-term decompositions [6] with different interior ranks are used. In each decomposition, the factors corresponding to the rank-1 dimension are orthogonal. At each iteration the algorithm greedily chooses one or more components from one type of decomposition; the rank structure of the component added at each iteration is shown.

In terms of classifying the faces, the number of features is not exactly equal to the number of tensor components used in the model, but instead is equal to the sum of the ranks in the image index mode. For instance, a combination of 10 rank-(4,1,4) and 5 rank-(4,4,1) tensors components would correspond to 45 features in the pixel space. The first two modes form the basis for the linear projection onto the feature space. Using PCA as a preprocessor for face classification vectorizes the first two modes [26, 27]; whereas, a heterogenous model allows a combination of varying rank components across the image plane [11].

We performed a very simple classification problem to compare the performance of the heterogenous model to the use of PCA/SVD to form eigenfaces as feature bases [26, 27]. For both approaches, the features were only found once without the labels and using all samples. 100 Monte Carlo divisions of the samples were performed with 50% of the samples for training and the rest for testing. Samples in the test set were labeled by their nearest neighbor (using the Euclidean metric) in the feature space. For the best number of features, the average test performance was not different between using SVD and the heterogenous model 92.24 ± 1.65 and 91.59 ± 2.04, respectively. However, in terms of the number of coefficients in the linear projection, the heterogenous model is far more parsimonious, as in Fig. 3. This indicates that most of the eigenfaces are themselves low-rank in image plane.

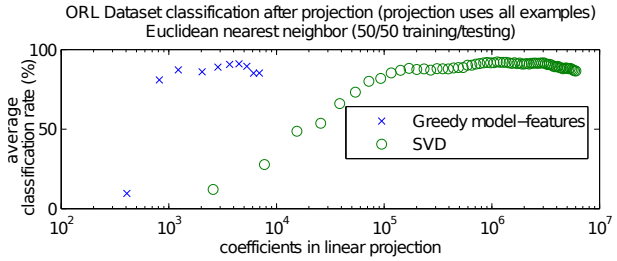

Fig. 3. Classification based on features extracted via the greedy iterations in the heterogeneous model versus those of a single SVD. The x-axis is the coefficients used in forming the linear projection to the feature space.

4. CONCLUSION

We proposed a greedy algorithm for model selection of heterogeneous tensor decompositions. At each iteration the algorithm selects one or more components from a single type of tensor decomposition to be added to the model. Across iterations, the model may consist of a mixture of decompositions. The flexibility in the synthesis model yields parsimonious approximations, which do not require a priori selection of the number of each type of component.

The approach can be extended with the tensor Kronecker product [9] to build heterogeneous block-wise approximations [10, 11, 12]. The iterative algorithm would replace the alternating least squares approach for the heterogenous case where the ranks of each type of decomposition are chosen a priori. The model selection criterion is useful in choosing a block size or combination of block sizes.

Previous work on tensors decompositions [3, 8, 6, 10, 28] has set the foundation for this approach. Our approach considers model selection [13, 14, 15, 16, 17] but in the greedy orthogonal decomposition proposed by Kolda [2]. To our knowledge this is the first iterative algorithm to build a heterogeneous model.

The thresholding approach comes from the MCA algorithm [19, 20] for sparse coding of vectors with a union of dictionaries. Here we have used data-dependent bases, which are estimated from the residual on every iteration, yielding a nearly orthogonal decomposition. As signal processing is a called upon to process multiway signals, there is the potential for more cross-fertilization between algorithms for sparse vector representations and those for tensor analysis.

5. REFERENCES

[1] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[2] T. G. Kolda, "Orthogonal tensor decompositions," *SIAM Journal on Matrix Analysis and Applications*, vol. 23, no. 1, pp. 243–255, 2001.

[3] F. L. Hitchcock, "Multiple invariants and generalized rank of a p-way matrix or tensor," *Journal of Mathematics and Physics*, vol. 7, no. 1, pp. 39–79, 1927.

[4] J. D. Carroll and J.-J. Chang, "Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.

[5] R. A. Harshman, "Foundations of the PARAFAC procedure: models and conditions for an "explanatory" multimodal factor analysis," *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[6] L. De Lathauwer, "Decompositions of a higher-order tensor in block terms-Part II: Definitions and uniqueness," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1033–1066, 2008.

[7] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[8] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, no. 3, pp. 279–311, 1966.

[9] S. Ragnarsson, *Structured Tensor Computations: Blocking, Symmetries and Kronecker Factorizations*, Ph.D. thesis, Cornell University, 2012.

[10] A.-H. Phan, A. Cichocki, P. Tichavský, D. P. Mandic, and K. Matsuoka, "On revealing replicating structures in multiway data: a novel tensor decomposition approach," *Latent Variable Analysis and Signal Separation*, pp. 297–305, 2012.

[11] A.-H. Phan, A. Cichocki, P. Tichavský, R. Zdunek, and S. Lehky, "From basis components to complex structural patterns," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013.

[12] A.-H. Phan, A. Cichocki, P. Tichavský, G. Luta, and A. Brockmeier, "Tensor completion through multiple kronecker product decomposition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013.

[13] M. E. Timmerman and H. A. L. Kiers, "Three-mode principal components analysis: Choosing the numbers of components and sensitivity to local optima," *British Journal of Mathematical and Statistical Psychology*, vol. 53, no. 1, pp. 1–16, 2000.

[14] E. Ceulemans and H. A. L. Kiers, "Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 133–150, 2006.

[15] E. Ceulemans and H. A. L. Kiers, "Discriminating between strong and weak structures in three-mode principal component analysis," *British Journal of Mathematical and Statistical Psychology*, vol. 62, no. 3, pp. 601–620, 2009.

[16] R. Bro and H. A. L. Kiers, "A new efficient method for determining the number of components in PARAFAC models," *Journal of Chemometrics*, vol. 17, no. 5, pp. 274–286, 2003.

[17] M. Mørup and L. K. Hansen, "Automatic relevance determination for multi-way models," *Journal of Chemometrics*, vol. 23, no. 7-8, pp. 352–363, 2009.

[18] Z. He, A. Cichocki, and S. Xie, "Efficient method for Tucker3 model selection," *Electronics Letters*, vol. 45, no. 15, pp. 805–806, 16 2009.

[19] M. Elad, J.-L. Starck, P. Querre, and D. L. Donoho, "Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA)," *Applied and Computational Harmonic Analysis*, vol. 19, no. 3, pp. 340– 358, 2005.

[20] J. Bobin, J.-L. Starck, J. M. Fadili, Y. Moudden, and D. L. Donoho, "Morphological component analysis: An adaptive thresholding strategy," *Image Processing, IEEE Transactions on*, vol. 16, no. 11, pp. 2675–2681, 2007.

[21] J. B. Kruskal, "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics," *Linear Algebra and its Applications*, vol. 18, no. 2, pp. 95–138, 1977.

[22] L. De Lathauwer and D. Nion, "Decompositions of a higher-order tensor in block terms-Part III: Alternating least squares algorithms," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 3, pp. 1067–1083, 2008.

[23] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *Signal Processing Magazine, IEEE*, vol. 21, no. 4, pp. 36–47, July 2004.

[24] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[25] A. Stegeman and P. Comon, "Subtracting a best rank-1 approximation may increase tensor rank," *Linear Algebra and its Applications*, vol. 433, no. 7, pp. 1276–1300, 2010.

[26] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[27] X. Jiang, B. Mandal, and A. Kot, "Eigenfeature regularization and extraction in face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 3, pp. 383–394, 2008.

[28] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari, *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*, Wiley, 2009.