Archetypal Analysis for Machine Learning

Mørup, Morten; Hansen, Lars Kai

Published in:
IEEE International Workshop on Machine Learning for Signal Processing

Link to article, DOI:
10.1109/MLSP.2010.5589222

Publication date:
2010

Document Version
Early version, also known as pre-print

ARCHETYPAL ANALYSIS FOR MACHINE LEARNING

Morten Mørup and Lars Kai Hansen

Cognitive Systems Group, Technical University of Denmark
Richard Petersens Plads, bld 321, 2800 Lyngby, Denmark, e-mail:{mm,lkh}@imm.dtu.dk

ABSTRACT

Archetypal analysis (AA) proposed by Cutler and Breiman in [1] estimates the *principal convex hull* of a data set. As such AA favors features that constitute representative 'corners' of the data, i.e. distinct aspects or archetypes. We will show that AA enjoys the interpretability of clustering - without being limited to hard assignment and the uniqueness of SVD - without being limited to orthogonal representations. In order to do large scale AA, we derive an efficient algorithm based on projected gradient as well as an initialization procedure inspired by the FURTHESTFIRST approach widely used for K-means [2]. We demonstrate that the AA model is relevant for feature extraction and dimensional reduction for a large variety of machine learning problems taken from computer vision, neuroimaging, text mining and collaborative filtering.

1. INTRODUCTION

Decomposition approaches have become a key tool for a wide variety of massive data analysis from modeling of Internet data such as term-document matrices of word occurrences, bio-informatics data such as micro-array data of gene expressions, neuroimaging data such as neural activity measured over space and time to collaborative filtering problems such as the celebrated Netflix problem to mention but a few. The conventional approaches range from low rank approximations such as singular value decomposition (SVD), principal component analysis (PCA) [3], independent component analysis (ICA) [4], sparse coding (SC) [5] and non-negative matrix factorization (NMF) [6] to hard assignment clustering methods such as K-means and K-medoids. Common to these is that they can be understood as a linear mixture or factor analysis type representation of data with various constraints. Thus data $x_{m,n}$, where $m = 1, ..., M$ is the feature index and $n = 1, .., N$ is the sample index, is written in terms of hidden variables $s_{k,n}$ and projections $a_{m,k}$ with $k = 1, ..., K$,

$$x_{m,n} = \sum_k a_{m,k} s_{k,n} + e_{m,n}, \quad e_{m,n} \sim \mathcal{N}(0, \sigma^2), \quad (1.1)$$

typically with a Gaussian noise model. SVD/PCA requires $\boldsymbol{A}$ and $\boldsymbol{S}$ be orthogonal, in ICA statistical independence is assumed for $\boldsymbol{S}$ and in SC a penalty term is introduced that measures deviation from sparsity on $\boldsymbol{S}$, while in NMF all variables are constrained non-negative. In hard clustering by K-means $\boldsymbol{S}$ is constrained to be a binary assignment matrix such that $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{S}^\top(\boldsymbol{S}\boldsymbol{S}^\top)^{-1}$ represents the Euclidean centers of each cluster while for K-medoids $\boldsymbol{a}_k = \boldsymbol{x}_n$ for some n, i.e. the cluster centers have to constitute actual data points.

Despite the similarities of the above approaches their internal representations of the data differ greatly and thus the nature of the interpretations they offer. In SVD/PCA the features constitute the directions of maximal variation, i.e. so-called eigenmaps, for NMF the features are constituent parts, for SC the features are also atoms or dictionary elements while K-means and K-medoids find the most representative prototype objects.

A benefit of clustering approaches is that features are similar to measured data making the results easier to interpret, however, the binary assignments reduce flexibility. Also, clustering typically involves complex combinatorial optimization leading to a plethora of heuristics. On the other hand low rank approximations based on SVD/PCA/NMF have a great degree of flexibility but the features can be harder to interpret both as invariance to rotation of the extracted features can lead to lack of uniqueness, i.e., $\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S} = \boldsymbol{A}\boldsymbol{Q}\boldsymbol{Q}^{-1}\boldsymbol{S} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{S}}$. In addition SVD/PCA/ICA/SC are prone to cancellation effects in which two components both lose meaning because they locally become highly correlated taking positive and negative near-cancelling values (while still being globally orthogonal).

In conclusion, clustering approaches give easy interpretable features but pay a price in terms of modelling flexibility due to the binary assignment of data objects. Approaches such as SVD/PCA/ICA/NMF/SC have added model flexibility and as such can be more efficient in capturing e.g., variance, however, this efficiency can lead to complex representations from which we learn relatively little.

Archetypal analysis (AA) proposed by [1] directly combines the virtues of clustering and the flexibility of matrix factorization. In the original paper on AA [1] the method

We thank Gitte Moos Knudsen and Claus Svarer for providing the PET data set.

was demonstrated useful in the analysis of air pollution and head shape and later also for tracking spatio-temporal dynamics [7]. Recently, Archetypal Analysis has found use in benchmarking and market research identifying typically extreme practices, rather than just good practices [8] as well as in the analysis of astronomy spectra [9] as an approach for the end-member extraction problem [10]. In this paper we demonstrate the following important theoretical properties of AA

- The Archetypal Analysis (AA) model is unique.
- Archetypal Analysis can efficiently be initialized through the proposed FURTHESTSUM method.
- AA can be efficiently computed using a simple projected gradient method.

We further demonstrate that AA is useful for a wide variety of important machine learning problem domains resulting in easy interpretable features that well account for the inherent dynamics in data.

2. ARCHETYPAL ANALYSIS AND THE PRINCIPAL CONVEX HULL

The convex hull also denoted the convex envelope of a data matrix $\boldsymbol{X}$ is the minimal convex set containing $\boldsymbol{X}$. Informally it can be described as a rubber band wrapped around the data points, see also figure 1. While the problem of finding the convex hull is solvable in linear time (i.e., $\mathcal{O}(N)$) [11] the size of the convex set increases dramatically with the dimensionality of the data. The expected size of the convex set for N points in general position in K dimensional space grows exponentially with dimension as $\mathcal{O}(\log^{K-1}(N))$ [12]. As a result, in high dimensional spaces the 'minimal' convex set forming the convex hull does not provide a compact data representation, see also figure 1. Archetypal analysis [1] considers the *principal* convex hull, i.e., the (K-1)-dimensional convex hull that the best account for the data according to some measure of distortion $D(\cdot|\cdot)$. This can formally be stated as the optimal $\boldsymbol{C}$ and $\boldsymbol{S}$ to the problem

$$\begin{aligned} &\arg\min_{\boldsymbol{C},\boldsymbol{S}} D(\boldsymbol{X}|\boldsymbol{XCS}) \\ &\text{s.t.} \quad |\boldsymbol{c}_k|_1 = 1, \quad |\boldsymbol{s}_n|_1 = 1, \\ &\qquad \boldsymbol{C} \geq \boldsymbol{0}, \qquad \boldsymbol{S} \geq \boldsymbol{0}. \end{aligned}$$

The constraint $|\boldsymbol{c}_k|_1 = 1$ together with $\boldsymbol{C} \geq \boldsymbol{0}$ enforces the feature matrix $\boldsymbol{A} = \boldsymbol{XC}$ as given in equation (1.1) to be a weighted average (i.e., convex combination) of the data observations while the constraint $|\boldsymbol{s}_n|_1 = 1$, $\boldsymbol{S} \geq \boldsymbol{0}$ requires the n^{th} data point to be approximated by a weighted average (i.e., convex combination) of the feature vectors $\boldsymbol{XC} \in \mathbb{R}^{M \times K}$. We will presently for brevity consider $D(\boldsymbol{X}|\boldsymbol{XCS}) = \|\boldsymbol{X} - \boldsymbol{XCS}\|_F^2$. As such in line with principal component analysis, the optimal $\boldsymbol{C}, \boldsymbol{S}$ will generate the principal (i.e. dominant) convex hull for the data $\boldsymbol{X}$, see figure 1. Archetypal analysis favors features that constitute representative 'corners' of the data, i.e., distinct aspects or archetypes. Furthermore, the AA model can naturally be considered a model between low-rank factor type approximation and clustering approaches, see table 1. As for K-means and K-medoids Archetypal Analysis is invariant to scale and translation of the data and as noted in [1] the AA problem is non-convex.

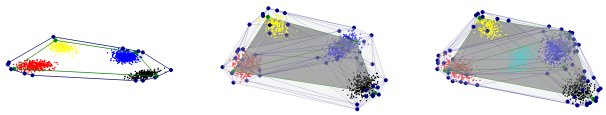

Fig. 1. Illustration of the AA representation of data in 2D (left panel) and 3D (middle and right panel). Blue dots and lines indicate the convex sets and hulls respectively whereas green lines and dark shaded regions indicate the extracted principal convex hulls. Notice how the four component principal convex hull (left and middle panel) and 5 component convex hull (right panel) account for most of the dynamics in the data while the complete convex hull even in 2D is given by a rather large convex set. The end points of the principal convex hull are convex combinations of the data points. These end points constitute the distinct regions of the clusters rather than the central regions as would be extracted by K-means. As such, the 5th cluster given in cyan color (right panel) is modeled as a combination of existing aspects rather than given a feature of its own.

2.1. Uniqueness of AA

Lack of uniqueness in matrix decomposition is a main motivation behind rotational criteria such as varimax in factor analysis as well as imposing statistical independence in ICA [4] and sparsity [5]. We presently prove that AA is in general unique up to permutation of the components.

Theorem 1. *Assume* $\forall k \exists n : c_{n,k} > 0 \wedge c_{n,k'} = 0, k' \neq k$ *then the* AA *model does not suffer from rotational ambiguity, i.e. if* $\boldsymbol{X} \approx \boldsymbol{XCS} = \boldsymbol{XCQ}^{-1}\boldsymbol{QS} = \boldsymbol{X}\tilde{\boldsymbol{C}}\tilde{\boldsymbol{S}}$ *such that both* $\boldsymbol{C}$,$\boldsymbol{S}$ *and* $\widetilde{\boldsymbol{C}}$, $\widetilde{\boldsymbol{S}}$ *are equivalent solutions to* AA *then* $\boldsymbol{Q}$ *is a permutation matrix.*

Proof. Since $\tilde{\boldsymbol{S}} = \boldsymbol{QS}$ and $\boldsymbol{S} \geq 0$ and $\widetilde{\boldsymbol{S}} \geq 0$ both are solutions to AA we have $|\boldsymbol{s}_n|_1 = 1$ and $|\boldsymbol{Q}\boldsymbol{s}_n|_1 = 1 \; \forall n$. For this to be true $\boldsymbol{Q}$ has to be a Markov matrix, i.e. $\sum_{k'} q_{k,k'} = 1$, $\boldsymbol{Q} \geq \boldsymbol{0}$. Since $\boldsymbol{C} \geq 0$ and $\boldsymbol{CQ}^{-1} \geq 0$ and given $\forall k \exists n : c_{n,k} > 0 \wedge c_{n,k'} = 0, k' \neq k$ then $\boldsymbol{Q}^{-1}$ has to be non-negative. Since both $\boldsymbol{Q}$ and $\boldsymbol{Q}^{-1}$ are non-negative it follows that $\boldsymbol{Q}$ can only be a scale and permutation matrix, and as the row sum of $\boldsymbol{Q}$ has to be one it follows that $\boldsymbol{Q}$ can only be a permutation matrix. □

SVD/PCA	NMF	AA/PCH	K-means	K-medoids
$\boldsymbol{C} \in \mathbb{R}$	$\boldsymbol{XC} \geq 0$	$\lvert\boldsymbol{c}_d\rvert_1 = 1,\ \boldsymbol{C} \geq 0$	$\lvert\boldsymbol{c}_d\rvert_1 = 1,\ \boldsymbol{C} \geq 0$	$\lvert\boldsymbol{c}_d\rvert_1 = 1, \boldsymbol{C} \in \mathbb{B}$
$\boldsymbol{S} \in \mathbb{R}$	$\boldsymbol{S} \geq 0$	$\lvert\boldsymbol{s}_n\rvert_1 = 1,\ \boldsymbol{S} \geq 0$	$\lvert\boldsymbol{s}_n\rvert_1 = 1,\ \boldsymbol{S} \in \mathbb{B}$	$\lvert\boldsymbol{s}_n\rvert_1 = 1, \boldsymbol{S} \in \mathbb{B}$

Table 1. Relation between the AA/PCH model and unsupervised methods such as SVD/PCA, NMF, K-means and K-medoids.

The requirement $\forall k \exists n : c_{n,k} > 0 \wedge c_{n,k'} = 0, k' \neq k$ states that for each column of $\boldsymbol{C}$ there has to exist a row where that column element is the only non-zero element. This holds for the AA model as two distinct aspects in general position will not be a convex combination of the same data points. We note that although AA is unique in general there is no guarantee the optimal solution will be identified due to the occurrence of local minima.

2.2. Efficient initialization of AA by FURTHESTSUM

Cutler and Breiman point out in [1] that careful initialization improves the speed of convergence and lowers the risk of finding insignificant archetypes. For K-means a popular initialization procedure is based on the FURTHESTFIRST method described in [2]. The method proceeds by randomly selecting a data point as centroid for a cluster and selecting subsequent data points the furthest away from already selected points. As such, a new data point j^{new} is selected according to

$$j^{new} = \arg\max_i \{\min_j \|x_i - x_j\|, j \in \mathcal{C}\}, \tag{2.1}$$

where $\|\cdot\|$ is a given norm and $\mathcal{C}$ index current selected data points. For initializing AA we propose the following modification forming our FURTHESTSUM procedure

$$j^{new} = \arg\max_i \{\sum_j \|x_i - x_j\|, j \in \mathcal{C}\}. \tag{2.2}$$

To improve the identified set, $\mathcal{C}$, the first point selected by random is removed and an additional point selected in its place. For the proposed FURTHESTSUM method we have the following important property

Theorem 2. *The points generated by the* FURTHESTSUM *algorithm is guaranteed to lie in the minimal convex set of the unselected data points.*

Proof. We will proof the theorem by contradiction. Assume that there is a point t not in the minimal convex set, i.e. $\boldsymbol{x}_t = \boldsymbol{Xc}$ such that $|\boldsymbol{c}|_1 = 1$, $c_d \geq 0,\ c_t = 0$ while $t = \arg\max_i \sum_j \|x_i - x_j\|, j \in \mathcal{C}$. We then have

$$\begin{aligned}\sum_{j\in\mathcal{C}} \|\boldsymbol{x}_t - \boldsymbol{x}_j\| = \|\boldsymbol{Xc} - \boldsymbol{x}_j\| &< \sum_d c_d \sum_{j\in\mathcal{C}} \|\boldsymbol{x}_d - \boldsymbol{x}_j\| \\ &\leq \max_d \sum_{j\in\mathcal{C}} \|\boldsymbol{x}_d - \boldsymbol{x}_j\|\end{aligned}$$

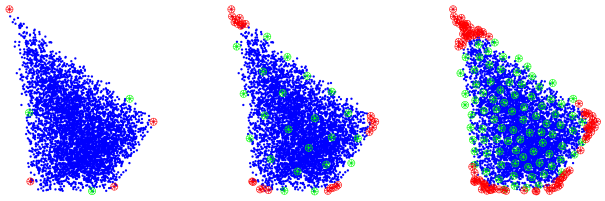

Fig. 2. Illustration of the extracted prototypes by FURTHESTFIRST [2] (green circles) and the proposed FURTHESTSUM initialization procedure (red circles) for 4 (left), 25 (middle) and 100 (right) propotypes. Clearly, the FURTHESTSUM extract points belonging to the convex set of the unselected data (useful for AA) whereas FURTHESTFIRST distribute the prototypes evenly over the data region (useful for K-means).

Where the first inequality follows from the triangular inequality. Hence a better solution is given by a point different from t contradicting that t was the optimal point. □

A comparison between the FURTHESTFIRST and FURTHESTSUM initialization procedures can be found in figure 2 based on the 2-norm, i.e. $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$. The primary computational cost of the FURTHESTSUM procedure for the identification of T candidate points is the evaluation of the distance between all data points and the selected candidate points which has an overall computational complexity of $\mathcal{O}(MNT)$.

2.3. Projected Gradient for AA

We currently propose a simple projected gradient procedure that can be adapted to any proximity measure. In the present analysis we will however without loss of generality consider proximity measured by least squares. In the paper of [1] the model was estimated by non-negative least squares such that the linear constraints were enforced by introducing quadratic penalty terms in the alternating updates of $\boldsymbol{S}$ and $\boldsymbol{C}$, i.e minimizing $\|\boldsymbol{X} - \boldsymbol{XCS}\|_F^2 + M\sum_d(|\boldsymbol{c}_d|_1 - 1)^2 + M\sum_j(|\boldsymbol{s}_j|_1 - 1)^2$, where M is some large number. Alternatively, standard non-negative quadratic programming solvers with linear constraints can be invoked [13] for each alternating subproblem solving $\boldsymbol{S}$ for fixed $\boldsymbol{C}$ and vice versa. We found however, that the following projected gradient method worked efficiently in practice. We recast the AA-problem in the

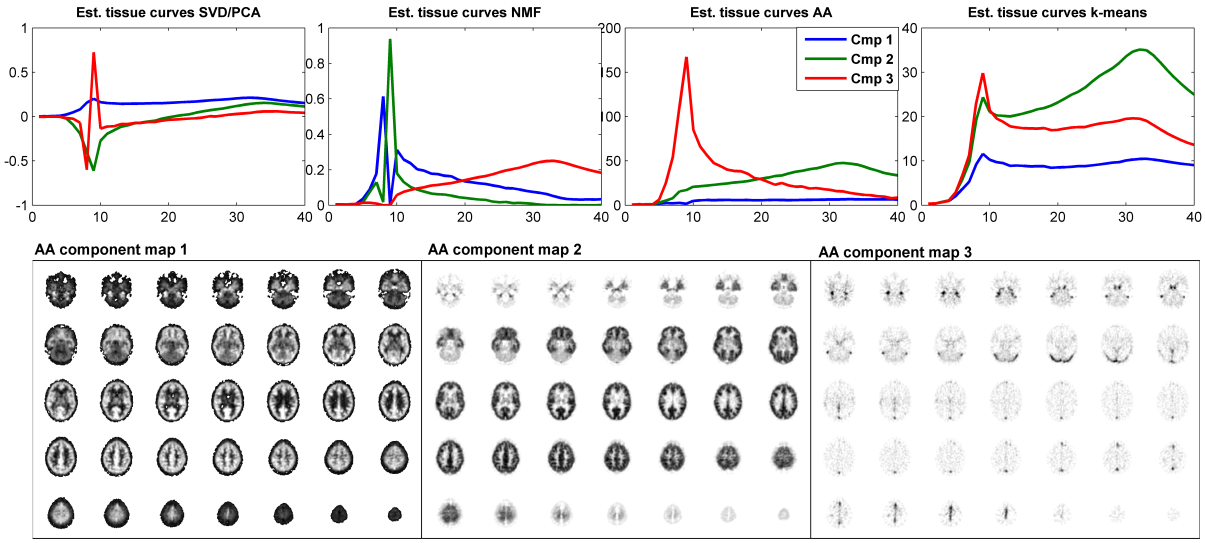

Fig. 4. Analysis of an Altanserin tracer Positron Emission Tomography data set. Top panel: The extracted temporal profiles for a three component SVD/PCA, NMF, AA and K-means models. Of the four models only the AA model has correctly extracted the region specific temporal profiles that from the spatial maps in the bottom panel correspond well to non-binding, high-binding and vascular regions respectively. As there is no scaling ambiguity the estimated arterial tissue curves are in the correct units.

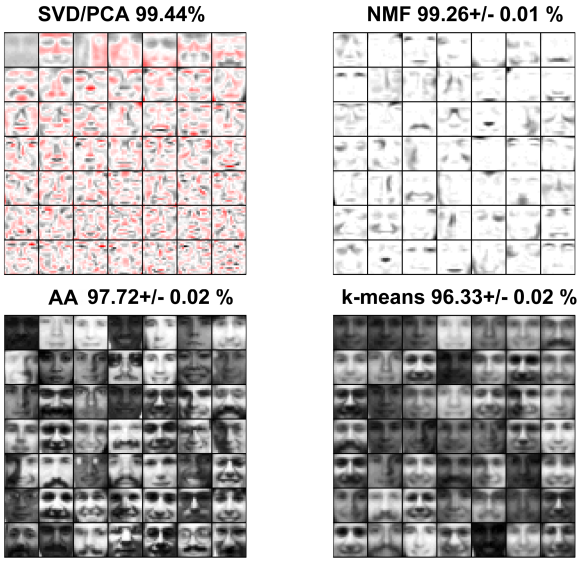

Fig. 3. Properties of features extracted based on SVD/PCA, NMF, AA and K-means. Whereas SVD/PCA extract low to high spatial frequency components and NMF decompose the data into an atomic mixture of constituting parts the AA model extracts notably more distinct aspects than K-means.

l_1-normalization invariant variables $\widetilde{s}_{k,n} = \frac{s_{k,n}}{\sum_k s_{k,n}}$ and $\widetilde{c}_{n,k} = \frac{c_{n,k}}{\sum_n c_{n,k}}$ such that the equality constraints are explicitly satisfied. Noticing that $\frac{\partial \widetilde{s}_{k',n}}{\partial s_{k,n}} = \frac{\delta_{k',k}}{\sum_k s_{k,n}} - \frac{s_{k',n}}{(\sum_k s_{k,n})^2}$ and differentiating by parts, we find the following updates for the AA parameters recast in the above normalization invariant variables

$$s_{k,n} \leftarrow \max\{\widetilde{s}_{k,n} + \mu_{\widetilde{\boldsymbol{S}}}(g^{\widetilde{\boldsymbol{S}}}_{k,n} - \sum_{k'} g^{\widetilde{\boldsymbol{S}}}_{k',n}\widetilde{s}_{k',n}), 0\},$$

$$\widetilde{s}_{k,n} = \frac{s_{k,n}}{\sum_k s_{k,n}}, \quad \boldsymbol{G}^{\widetilde{\boldsymbol{S}}} = \widetilde{\boldsymbol{C}}^\top \boldsymbol{X}^\top \boldsymbol{X} - \widetilde{\boldsymbol{C}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{S}}$$

$$c_{n,k} \leftarrow \max\{\widetilde{c}_{n,k} + \mu_{\widetilde{\boldsymbol{C}}}(g^{\widetilde{\boldsymbol{C}}}_{n,k} - \sum_{n'} g^{\widetilde{\boldsymbol{C}}}_{n',k}\widetilde{c}_{n',k}), 0\},$$

$$\widetilde{c}_{n,k} = \frac{c_{n,k}}{\sum_n c_{n,k}}, \quad \boldsymbol{G}^{\widetilde{\boldsymbol{C}}} = \boldsymbol{X}^\top \boldsymbol{X}\widetilde{\boldsymbol{S}}^\top - \boldsymbol{X}^\top \boldsymbol{X}\widetilde{\boldsymbol{C}}\widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{S}}^\top$$

Each alternating update is performed on all elements simultaneously and μ is a step-size parameter that we tuned by line-search. In the update of $\boldsymbol{S}$, $\widetilde{\boldsymbol{C}}^\top \boldsymbol{X}^\top \boldsymbol{X}$ and $\widetilde{\boldsymbol{C}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widetilde{\boldsymbol{C}}$ can be pre-computed having a cost of $\mathcal{O}(KMN)$ while the computation of the gradient as well as the evaluation of least squares objective given by

$$const. - 2\langle \widetilde{\boldsymbol{C}}^\top \boldsymbol{X}^\top \boldsymbol{X}, \widetilde{\boldsymbol{S}}\rangle + \langle \widetilde{\boldsymbol{C}}^\top \boldsymbol{X}^\top \boldsymbol{X}\widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{S}}^\top\rangle$$

has computational cost $\mathcal{O}(K^2N)$. In the update of $\boldsymbol{C}$, $\boldsymbol{X}^\top \boldsymbol{X}\widetilde{\boldsymbol{S}}^\top$ and $\widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{S}}^\top$ can be precomputed having a computational cost of $\mathcal{O}(KMN)$ and $\mathcal{O}(K^2N)$ respectively while calculating the gradient as well as the evaluation of the least squares

objective given by

$$const. - 2\langle \boldsymbol{X}^\top \boldsymbol{X} \widetilde{\boldsymbol{S}}^\top, \widetilde{\boldsymbol{C}} \rangle + \langle \widetilde{\boldsymbol{C}}^\top \boldsymbol{X}^\top \boldsymbol{X} \widetilde{\boldsymbol{C}}, \widetilde{\boldsymbol{S}} \widetilde{\boldsymbol{S}}^\top \rangle$$

also has a computational complexity of $\mathcal{O}(KMN)$. When only considering T candidate points used to define the archetypes as proposed in [13] this latter complexity can be further reduced to $\mathcal{O}(KMT)$. In [13] it was suggested to identify these candidates point by outlying data points found through projections of the eigenvectors of the covariance matrix of $\boldsymbol{X}$. We found that the proposed FURTHESTSUM algorithm form an efficient alternative approach to the identification of these T candidate points. In our implementation of the projected gradient method we carried out 10 line-search updates for each alternating update of $\boldsymbol{S}$ and $\boldsymbol{C}$.

2.4. kernel-AA

From the above updates it can be seen that the estimation of the parameters only depend on the pairwise relations, i.e. the inner products Kernel $\boldsymbol{K} = \boldsymbol{X}^\top \boldsymbol{X}$. As such, the AA model trivially generalize to kernel representations based on pairwise relations between the data points (kernel-AA) and can here be interpreted as extracting the principal convex hull in a potentially infinite Hilbert space (i.e., similar to the corresponding interpretation of kernel-K-means and kernel-PCA). These types of analysis are however out of the scope of the current paper.

3. RESULTS

We demonstrate the utility of the AA model on four data sets taken from a variety of important machine learning problem domains.

Computer Vision: To illustrate the properties of the AA model we compared the extracted model representation to the representations obtained by SVD/PCA, NMF and K-means on the CBCL face data base of 361 pixels $\times$ 2429 images used in [6]. The results of the analysis is given in figure 3. SVD/PCA extracts features that have low to high spatial frequency while NMF gives a part based representation as reported in [6]. K-means extracts cluster centers of anonymous 'typical' faces while features extracted by AA represents more distinct (archetypal) face prototypes exposing variability and face diversity. Thus AA accounts for more variation than K-means but less than SVD/PCA and NMF as indicated in Table 1, while the distinct facial aspects are efficiently extracted.

NeuroImaging: We analyzed a Positron Emission Tomography data set containing 40 time points $\times$ 157244 voxels based on $[^{18}F]$-Altanserin as radioligand in order to measure serotonin-2A neuroreceptors. Each recorded voxel is a mixture of vascular regions, non-binding regions and high-binding regions. The AA model should ideally extract these profiles as well as how each voxel is a convex mixture of these regions. This holds provided a convex combination of the observations are able to generate the pure tissue profiles, i.e. can extract the distinct profiles. From figure 4 the AA model have indeed extracted three components that well correspond to non-binding, high-binding and vascular regions respectively. No such profiles are clearly extracted by SVD/PCA, NMF or K-means that all extract components that are mixtures of the three different region types.

Text mining: Latent Semantic Analysis has become an important tool for extracting word and document associations in text corpora. In figure 5 we demonstrate the utility of the AA-model for this task by analyzing the NIPS bag of words corpus consisting of 1,500 documents and 12,419 words with approximately 6,400,000 word occurrences (see also `http://archive.ics.uci.edu/ml/datasets/Bag+of+Words`). Each word was normalized by the inverse document frequency (IDF). Clearly, the 10 component model has well extracted distinct categories of the NIPS papers. For the AA model $\boldsymbol{C}$ indicate which documents constitute the extracted distinct aspects $\boldsymbol{XC}$ given in the figure while $\boldsymbol{S}$ (not shown) gives the fraction by which each document resemble these aspects. Indeed the extracted archetypes correspond well to distinct types of papers in the NIPS corpora.

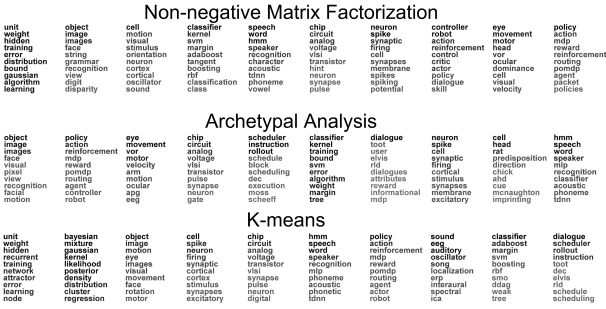

Fig. 5. Analysis of the NIPS bag of words corpus. The 10 most prominent words for each feature component of NMF, AA and K-means respectively of the extracted 10 component models. The average cosine angle (averaged over all the $10(10-1)/2$ feature comparisons) are $84.16°$, $80.31°$ and $73.33°$ respectively, thus, the AA model has extracted more distinct aspects than K-means. Clearly each of the components correspond well to aspects of the documents in the NIPS corpus. The components are ordered according to importance and in gray scale are given the relative strength of each of the 10 top words for each of the extracted term groups.

Collaborative filtering: Collaborative filtering has become an important machine learning problem of which the celebrated Netflix prize is widely known. Given the preferences of users the aim is to infer preferences of products the

user has not yet rated based on the ratings given by other users. We analyzed the medium size and large size Movie lens data given by 1,000,209 ratings of 3,952 movies by 6,040 users and 10,000,054 ratings of 10,677 movies given by 71,567 users with ratings from $\{1, 2, 3, 4, 5\}$ (`http://www.grouplens.org/`). The AA model extracts idealized users (extreme user behaviors) while at the same time relating the users to these archetypal preference types. Movies that are left unrated by the users we treated as missing values based on the following extension of the AA objective to accommodate missing information, i.e.

$$\min_{\boldsymbol{S},\boldsymbol{C}} \sum_{n,m:q_{n,m}=1} (x_{n,m} - \sum_k \frac{\sum_{m'} x_{n,m'} c_{m',k}}{\sum_{m'} q_{n,m'} c_{m',k}} s_{k,m})^2,$$

where $\boldsymbol{Q}$ is an indicator matrix such that $q_{n,m} = 1$ if the n^{th} movies was rated by the m^{th} user and zero otherwise. Since the SVD and NMF model turned out to be more prone to local minima than the AA model these methods were initialized by the AA solution obtained. From figure 6 it can be seen that despite that the AA model is more restricted than NMF and SVD it has a very similar test error performance and extracts features that are much more useful for predicting the ratings than K-means.

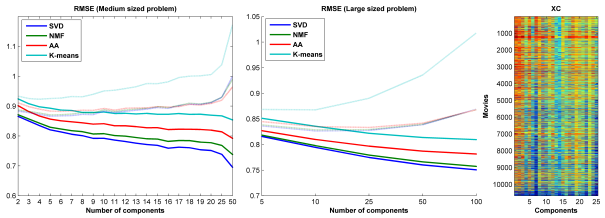

Fig. 6. Left and middle panel: Root mean square error (rmse) for the various models as a function of the number of components when analyzing the medium size Movielens (left) and large sized (middle) Movielens data. Training error is given by solid lines and test error performance by dotted lines based on removing 10% of the data for testing. Clearly, the AA model has added flexibility over K-means in accounting for the data. Right panel: Illustration of the extracted archetypal users-types, $\boldsymbol{XC}$, and their distinct preferences for the 25 component AA models with lowest validation error based on the large sized problem.

4. DISCUSSION

We demonstrated how the Archetypal Analysis model of [1] is useful for a large variety of machine learning problems. A simple algorithm for fitting the AA model was derived as well as the FURTHESTSUM initialization procedure to extract end-members for initial archetypes. The utility of AA over clustering methods is that it focuses more on distinct or discriminative aspects yet has additional modeling flexibility by using soft assignment. We saw examples of improved interpretability in the AA representations over existing matrix factorization and clustering methods. An open problem is to determine the number of components used. This problem is no different from the problem of choosing the number of components in approaches such as SVD/PCA/NMF/SC/K-means, thus, methods based on approximating the model evidence or generalization error can be invoked. AA is a promising unsupervised learning tool for many machine learning problems and as the representation is unique in general we believe the method holds particularly great promise for data mining applications.

5. REFERENCES

[1] Adele Cutler and Leo Breiman, "Archetypal analysis," *Technometrics*, vol. 36, no. 4, pp. 338–347, Nov 1994.

[2] D. S. Hochbaum and D. B. Shmoys., "A best possible heuristic for the k-center problem.," *Mathematics of Operational Research*, vol. 10, no. 2, pp. 180–184, 1985.

[3] Gene H. Golub and Charles F. Van Loan, *Matrix Computation*, Johns Hopkins Studies in Mathematical Sciences, 3 edition, 1996.

[4] Pierre Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[5] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[6] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[7] Emily Stone and Adele Cutler, "Introduction to archetypal analysis of spatio-temporal dynamics," *Phys. D*, vol. 96, no. 1-4, pp. 110–131, 1996.

[8] Giovanni C. Porzio, Giancarlo Ragozini, and Domenico Vistocco, "On the use of archetypes as benchmarks," *Appl. Stoch. Model. Bus. Ind.*, vol. 24, no. 5, pp. 419–437, 2008.

[9] B. H. P. Chan, D. A. Mitchell, and L. E. Cram, "Archetypal analysis of galaxy spectra," *MON.NOT.ROY.ASTRON.SOC.*, vol. 338, pp. 790, 2003.

[10] P. Perez R. Plaza J. Plaza, A. Martinez, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 42, no. 3, pp. 650–663, 2004.

[11] D. McCallum and D. Avis, "A linear algorithm for finding the convex hull of a simple polygon," *Information Processing Letters*, vol. 9, pp. 201–206, 1979.

[12] Rex A. Dwyer, "On the convex hull of random points in a polytope," *Journal of Applied Probability*, vol. 25, no. 4, pp. 688–699, 1988.

[13] Christian Bauckhage and Christian Thurau, "Making archetypal analysis practical," in *Proceedings of the 31st DAGM Symposium on Pattern Recognition*, Berlin, Heidelberg, 2009, pp. 272–281, Springer-Verlag.