Mach Learn (2016) 102:85–113
DOI 10.1007/s10994-015-5498-8

Probabilistic archetypal analysis

Sohan Seth[1,2] · **Manuel J. A. Eugster**[1]

Received: 1 April 2014 / Accepted: 13 April 2015 / Published online: 5 June 2015

Abstract Archetypal analysis represents a set of observations as convex combinations of pure patterns, or archetypes. The original geometric formulation of finding archetypes by approximating the convex hull of the observations assumes them to be real–valued. This, unfortunately, is not compatible with many practical situations. In this paper we revisit archetypal analysis from the basic principles, and propose a probabilistic framework that accommodates other observation types such as integers, binary, and probability vectors. We corroborate the proposed methodology with convincing real-world applications on finding archetypal soccer players based on performance data, archetypal winter tourists based on binary survey data, archetypal disaster-affected countries based on disaster count data, and document archetypes based on term-frequency data. We also present an appropriate visualization tool to summarize archetypal analysis solution better.

Keywords Archetypal analysis · Probabilistic modeling · Majorization–minimization · Visualization · Convex hull · Binary observation

1 Introduction

Archetypal analysis (AA) represents observations as composition of pure patterns, i.e., *archetypes*, or equivalently convex combinations of extreme values (Cutler and Breiman 1994). Although AA bears resemblance with many well established prototypical analysis

Editor: Kristian Kersting.

✉ Sohan Seth
sohan.seth@hiit.fi

Manuel J. A. Eugster
manuel.eugster@hiit.fi

[1] Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland

[2] Present Address: Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh, UK

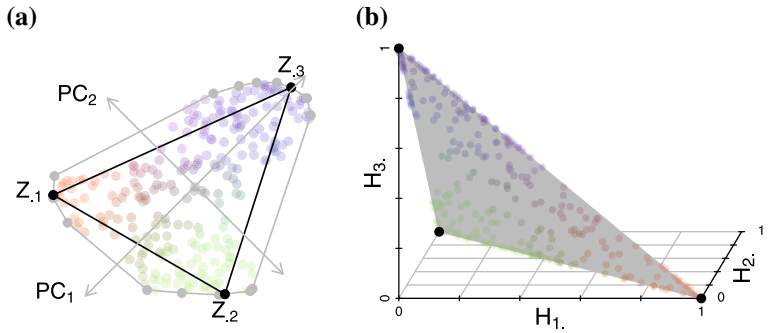

Fig. 1 **a** Illustration of archetypal analysis with three archetypes **Z**, and **b** the corresponding factors **H**, projections of the original observations on the convex hull of the archetypes; **a** also explicates the difference between PCA and AA

tools, such as principal component analysis (PCA, Mohamed et al. 2009), non-negative matrix factorization (NMF, Févotte and Idier 2011), probabilistic latent semantic analysis (Hofmann 1999), and k-means (Steinley 2006); AA is arguably unique, both conceptually and computationally. Conceptually, AA imitates the human tendency of representing a group of objects by its extreme elements (Davis and Love 2010): this makes AA an interesting exploratory tool for applied scientists (e.g., Eugster 2012; Seiler and Wohlrabe 2013). Computationally, AA is *data-driven*, and requires the *factors* to be probability vectors: these make AA a computationally demanding tool, yet brings better interpretability.

The concept of AA was originally formulated by Cutler and Breiman (1994). The authors posed AA as the problem of learning the convex hull of a point-cloud, and solved it using alternating non-negative least squares method. In recent years, different variations and algorithms based on the original geometrical formulation have been presented (Bauckhage and Thurau 2009; Eugster and Leisch 2011; Mørup and Hansen 2012). However, unfortunately, this framework does not tackle many interesting situations. For example, consider the problem of finding archetypal response to a binary questionnaire. This is a potentially useful problem in areas of psychology and marketing research that cannot be addressed in the standard AA formulation, which relies on the observations to exist in a vector space for forming a convex hull. Even when the observations exist in a vector space, standard AA might not be an appropriate tool for analyzing it. For example, in the context of learning archetypal text documents with tf-idf as features, standard AA will be inclined to finding archetypes based on the volume rather than the content of the document.

In this paper we revisit archetypal analysis from the basic principles, and reformulate it to extend its applicability. We admit that the approximation of the convex hull, as in the standard AA, is indeed an elegant solution for capturing the essence of 'archetypes', (see Fig. 1a, b for a basic illustration). Therefore, our objective is to extend the current framework, not to discard it. We propose a probabilistic foundation of AA, where the underlying idea is *to form the convex hull in the parameter space*. The parameter space is often vectorial even if the sample space is not (see Fig. 2 for the plate diagram). We solve the resulting optimization problem using majorization–minimization, and also suggest a visualization tool to help understand the solution of AA better.

The paper is organized as follows. In Sect. 2, we start with an in-depth discussion on what archetypes mean, and how this concept has evolved over the last decade, and has been utilized in different contexts. In Sect. 3, we provide a probabilistic perspective of this concept, and suggest *probabilistic archetypal analysis*. Here we explicitly tackle the cases of

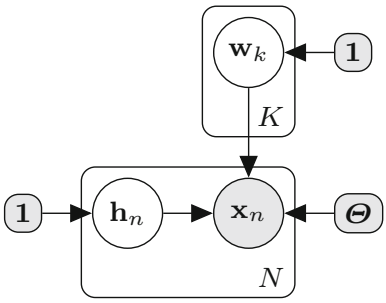

Fig. 2 Plate diagram of probabilistic archetypal analysis

Table 1 Probability distributions used in the paper

Distribution	Notation	Pdf/pmf
Normal	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $\boldsymbol{\mu} \in \mathbb{R}^K$, $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$	$(2\pi)^{-\frac{K}{2}} \lvert\boldsymbol{\Sigma}\rvert^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}$
Dirichlet	$\mathrm{Dir}(\boldsymbol{\alpha})$ $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $K > 1, \alpha_i > 0$	$\frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^K x_i^{\alpha_i - 1}$ where $\mathrm{B}(\alpha) = \frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)}$
Poisson	$\mathrm{Pois}(\lambda)$ $\lambda > 0$	$\frac{\lambda^x}{x!} \exp\{-\lambda\}$
Bernoulli	$\mathrm{Ber}(p)$ $0 < p < 1$	$p^x (1-p)^{1-x}$
Multinomial	$\mathrm{Mult}(n, \mathbf{p})$ $n > 0$, $\boldsymbol{p} = (p_1, \ldots, p_K)$, $\sum_{i=1}^K p_i = 1$	$\frac{n!}{x_1! \cdots x_K!} p_1^{x_1} \cdots p_K^{x_K}$

Bernoulli, Poisson and multinomial probability distributions, and derive the necessary update rules (derivations available as appendix). In Sect. 4, we discuss the connection between AA and other prototypical analysis tools—a connection that has also been partly noted by other researchers (Mørup and Hansen 2012). In Sect. 5 we provide simulations to show the difference betweeen probabilistic and standard archetypal analysis solutions. In Sect. 6, we discuss a visualization method for archetypal analysis, and present several improvements. In Sect. 7, we present an application for each of the above observation models: finding archetypal soccer players based on performance data; finding archetypal winter tourists based on binary survey data; finding archetypal disaster-affected countries based on disaster count data; and finding document archetypes based on term-frequency data. In Sect. 8 we summarize our contribution, and suggest future directions.

Throughout the paper, we represent matrices by boldface uppercase letters, vectors by boldface lowercase letter, and variables by normal lowercase letter. $\mathbf{1}$ denotes the row vector of ones, and $\mathbf{I}$ denotes the identity matrix. Table 1 provides the definitions of the distributions used throughout the paper. Implementations of the presented methods and source code to reproduce the presented examples are available at http://aalab.github.io/.

2 Review

The goal of archetypal analysis is to find archetypes, 'pure' patterns. In Sect. 2.1 we provide some intuition on what these pure patterns imply. In Sect. 2.2, we discuss the mathematical formulation of this archetypal analysis as suggested by Cutler and Breiman (1994). In Sect. 2.3 we discuss how this concept has been utilized since its inception: here, we point out key references, important developments, and convincing applications.

2.1 Intuition

Archetypes are 'ideal example of a type'. The word 'ideal' does not necessarily have a qualitative meaning of being 'good', but this concept is mostly subjective. For example, one can consider ideal example to be a prototype, and other objects to be variations of such prototype. This view is close to the concept of clustering, where the centers of the clusters are the prototypes. For archetypal analysis, however, ideal example has a different meaning. Intuitively it implies that the prototype *can not be surpassed*, its the purest or the most extreme that can be witnessed. A simple example of archetypes are the colors red, blue and green (cf. Fig. 1a) in the RGB color space: any other color can be expressed as combinations of these ideal colors. Another example can be comic book superheros who excel in some unique characteristics, say speed or stamina or intelligence, more than anybody with these abilities: they are the archetypal superheros with that particular ability. The non-archetypal superheroes, on the other hand, possess "many" abilities that are not extreme. It is to be noted that a person with all the abilities to their full realizations, if exists, is an archetype, Similarly, if one considers normal humans alongside super-humans then a person with none of these abilities is also an archetype.

In both these examples, the archetypes are rather trivial. If one represents each color in the RGB space then it is obvious that the unit vectors R, G and B are pure colors or archetypes. Similarly, if one represents every (super-)human in a two dimensional normalized scale of strength and intelligence, then there are four extreme instances, and hence archetypes are: first and second, person with highest score in either of these attributes and none in the other; third and fourth, person with highest/lowest score in both these attributes. However, in reality one may not observe these attributes directly, but some other features. For example, one can describe a person with many personality traits, such as humor, discipline, optimism, etc., but these characteristics cannot be measured directly. However, one can prepare a questionnaire (or observed variables) that explores these (latent) personality traits. From this questionnaire, curious users can attempt to identify archetypal humans, say an archetypal leader or an archetypal jester.

Finding archetypal patterns in the observed space is a non-trivial problem. It is difficult, in particular, since the archetype itself may not be belong to the set of observed samples but *should be inferred*; yet, it should also not be "mythological", but rather something that *might be observed*. Cutler and Breiman (1994) suggested a simple yet elegant solution that finds the approximate convex hull of the observations, and define the vertices as archetypes. This allows individual observations to be best represented by composition (convex combination) of archetypes, while archetypes can only be expressed by themselves, i.e., they are the 'purest form' or 'most extreme' forms. Although, it is certainly not the most desired solution, since, the inferred archetypes are restricted to be on the boundary of the convex hull of the observations, whereas true archetype may be outside; inferring such archetypes outside the observation hull will require strong regularity assumptions. The solution suggested by Cutler

and Breiman (1994) finds a trade off between computational simplicity, and the intuitive nature of archetype.

2.2 Formulation

Cutler and Breiman posed AA as the problem of learning the convex hull of a point-cloud. They assumed the archetypes to be convex combinations of observations, and the observations to be convex combinations of the archetypes. Let $\mathbf{X}$ be a (real-valued) data matrix with each column as an observation. Then, this is equivalent to solving the following optimization problem:

$$\min_{\mathbf{W},\mathbf{H}} ||\mathbf{X} - \mathbf{XWH}||_{\mathrm{F}}^2 \tag{1}$$

with the constraint that both $\mathbf{W}$ and $\mathbf{H}$ are column stochastic matrices. Subscript F denotes Frobenious norm. Given N observations, and K archetypes, $\mathbf{W}$ is $N \times K$ dimensional matrix, and $\mathbf{H}$ is $K \times N$ dimensional matrix. Here, $\mathbf{Z} = \mathbf{XW}$ are the inferred archetypes that exist on the convex hull of the observations due to the stochasticity of $\mathbf{W}$ and for each n-th sample $\mathbf{x}_n$, $\mathbf{Zh}_n$ is its projection on the convex hull of the archetypes.

Cutler and Breiman solved this problem using an alternating non-negative least squares method as follows:

$$\mathbf{H}^{t+1} = \arg\ \min_{\mathbf{H}\geq\mathbf{0}} ||\mathbf{X} - \mathbf{Z}^t\mathbf{H}||_{\mathrm{F}}^2 + \lambda||\mathbf{1H} - \mathbf{1}||^2 \tag{2}$$

and

$$\mathbf{W}^{t+1} = \arg\ \min_{\mathbf{W}\geq\mathbf{0}} ||\mathbf{Z}^t - \mathbf{XW}||_{\mathrm{F}}^2 + \lambda||\mathbf{1W} - \mathbf{1}||^2 \tag{3}$$

where after each alternating step, the archetypes ($\mathbf{Z}$) are updated by solving $\mathbf{X} = \mathbf{Z}^{t+1}\mathbf{H}^t$, and $\mathbf{Z}^{t+1} = \mathbf{XW}^t$, respectively. The algorithm alternates between finding the best composition of observations given a set of archetypes, and then finding the best set of archetypes given a composition of observations. Notice that the stochasticity constraint was cleverly enforced by a suitably strong regularization parameter λ. The authors also proved that $k > 1$ archetypes are located on the boundary of the convex hull of the point-cloud, and $k = 1$ archetype is the mean of the point cloud.

2.3 Development

The first publication, to the best of our knowledge, which deals with the idea of "ideal types" and observations related to them, is Woodbury and Clive (1974). There, the authors discuss how to derive estimates of grades of membership of categorical observations, given an a-priori defined set of ideal (or pure) types, in the context of clinical judgment. Twenty years later—in 1994—Cutler and Breiman (1994) formulated archetypal analysis (AA) as the problem of estimating both the membership and the ideal types given a set of real-valued observations. They motivated this new kind of analysis with, among other examples, the estimation of archtyepal head dimensions of Swiss Army soldiers.

One of the original authors continued her work on AA in the fields of physics and applied it on spatio-temporal data (Stone and Cutler 1996). In this line of research, Cutler and Stone (1997) developed *moving archetypes*, by extending the original AA framework with an additional optimization step, which estimates the optimal shift of observations in the spatial domain over time. They applied this method to data gathered from a chemical pulse experiment. Other researches took up the idea of AA and applied it in different fields; the following is

a comprehensive list of problems where other researchers have applied AA: analysis of galaxy spectra (Chan et al. 2003), ethical issues and market segmentation (Li et al. 2003), thermogram sequences (Marinetti et al. 2007), gene expression data (Thøgersen et al. 2013); performance analysis in marketing (Porzio et al. 2008), sports (Eugster 2012), and science (Seiler and Wohlrabe 2013); face recognition (Xiong et al. 2013); and in game AI development (Sifa and Bauckhage 2013).

In recent years, animated by the rise of the non-negative matrix factorization research, various authors have proposed extensions and variations to the original algorithm. The following are a few notable publications. Thurau et al. (2009) introduce the convex-hull non-negative matrix factorization (NMF). Motivated by the convex NMF, the authors make the same assumption as made in AA that observations are convex combinations of specific observations. However, they derive an algorithm which estimates the archetypes not from the entire set of observations but from potential candidates found from 2-dimensional projections on eigenvectors: this leads to a solution also applicable for large data sets. The authors demonstrate their method on a data set consisting of 150 million votes on World of Warcraft® guilds. In Thurau et al. (2010), the authors present an even faster approach by deriving a highly efficient volume maximization algorithm. Eugster and Leisch (2011) tackle the problem of robustness, and that a single outlier can break down the archetype solution. They adapt the original algorithm to be a robust M-estimator and present an iteratively reweighted least squares fitting algorithm. They evaluate there algorithm using the Ozone data from the original AA paper with contaminated observations. Mørup and Hansen (2012) also tackle the problem of deriving an algorithm for large scale AA. They propose a solution based on a simple projected gradient method, in combination with an efficient initialization method for finding candidates of archetypes. The authors demonstrate their method, among other examples, with an analysis of the NIPS bag of words corpus and the Movielens movie rating data set.

3 Probabilistic archetypal analysis

We observe that the original AA formulation implicitly exploits a *simplex latent variable model*, and normal observation model, i.e.,

$$\mathbf{h}_n \sim \mathrm{Dir}(\mathbf{1}),\ \mathbf{x}_n \sim \mathcal{N}(\mathbf{Z}\mathbf{h}_n, \epsilon_1 \mathbf{I}). \tag{4}$$

But, it goes a step further, and generates the loading matrix $\mathbf{Z}$ from a simplex latent model itself with *known* loadings $\boldsymbol{\Theta} \in \mathbb{R}^{M \times N}$, i.e.,

$$\mathbf{w}_k \sim \mathrm{Dir}(\mathbf{1}),\ \mathbf{z}_k \sim \mathcal{N}(\boldsymbol{\Theta}\mathbf{w}_k, \epsilon_2 \mathbf{I}). \tag{5}$$

Thus, the log-likelihood can be written as,

$$\mathbb{LL}(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{Z}, \boldsymbol{\Theta}) = -\frac{\epsilon_1}{2}||\mathbf{X} - \mathbf{Z}\mathbf{H}||_{\mathrm{F}}^2 - \frac{\epsilon_2}{2}||\mathbf{Z} - \boldsymbol{\Theta}\mathbf{W}||_{\mathrm{F}}^2 + C(\epsilon_1, \epsilon_2). \tag{6}$$

The archetypes $\mathbf{Z}$, and corresponding factors $\mathbf{H}$ can then be found by maximizing this log-likelihood (or minimizing the negative log-likelihood) under the constraint that both $\mathbf{W}$ and $\mathbf{H}$ are stochastic: this can be achieved by alternating optimization as Cutler and Breiman did (but with different update rules for $\mathbf{Z}$, and $\epsilon_\cdot$, see Appendix 1 for details).

The equivalence of this approach to the standard formulation requires that $\boldsymbol{\Theta} = \mathbf{X}$. Although unusual in a probabilistic framework, this contributes to the data-driven nature of AA. In the probabilistic framework, $\boldsymbol{\Theta}$ can be viewed as a set of known bases that is defined

by the observations, and the purpose of archetypal analysis is to find a *sparse* set of bases that can explain the observations. These inferred bases are the archetypes, and the stochasticity constraints on $\mathbf{W}$ and $\mathbf{H}$ ensure that they are the extreme values as one desires. It should be noted that $\boldsymbol{\Theta}_n$ does not need to correspond to $\mathbf{X}_n$: more generally, $\boldsymbol{\Theta} = \mathbf{XP}$ where $\mathbf{P}$ is a permutation matrix.

3.1 Exponential family

We describe AA in a probabilistic set-up as follows (see Fig. 2),

$$\mathbf{w}_k \sim \mathrm{Dir}(\mathbf{1}),\ \mathbf{h}_n \sim \mathrm{Dir}(\mathbf{1}),\ \mathbf{x}_n \sim \mathrm{EF}(\mathbf{x}_n; \boldsymbol{\Theta}\mathbf{W}\mathbf{h}_n) \tag{7}$$

where

$$\mathrm{EF}(\mathbf{z}; \boldsymbol{\theta}) = h(\mathbf{z})g(\boldsymbol{\theta})\exp(\eta(\boldsymbol{\theta})^\top s(\mathbf{z})) \tag{8}$$

with standard meaning for the functions g, h, η and s. Notice that, *we employ the normal parameter $\boldsymbol{\theta}$ rather than the natural parameter $\boldsymbol{\eta}(\boldsymbol{\theta})$*, since the former is more interpretable. In fact, the convex combination of $\boldsymbol{\theta}$ is more interpretable than the convex combination of $\boldsymbol{\eta}(\boldsymbol{\theta})$, as a linear combination on $\boldsymbol{\eta}(\boldsymbol{\theta})$ would lead to nonlinear combination of $\boldsymbol{\theta}$. To adhere to the original formulation, we suggest $\boldsymbol{\Theta}_{\cdot n}$ to be the *maximum likelihood point estimate* from observation $\mathbf{X}_{\cdot n}$. Again, the columns of $\boldsymbol{\Theta}$ and $\mathbf{X}$ do not necessarily have to be corresponded. Then, we find archetypes $\mathbf{Z} = \boldsymbol{\Theta}\mathbf{W}$ by solving

$$\arg\min_{\mathbf{W},\mathbf{H}\geq 0} -\mathbb{LL}(\mathbf{X}|\mathbf{W}, \mathbf{H}, \boldsymbol{\Theta}) \text{ such that } \mathbf{1W} = \mathbf{1}, \mathbf{1H} = \mathbf{1}. \tag{9}$$

We call this approach *probabilistic archetypal analysis* (PAA).

The meaning of *archetype* in PAA is different than in the standard AA since the former lies in the parameter space, whereas the latter in the observation space. To differentiate these two aspects, we call the archetypes $\mathbf{Z} = \boldsymbol{\Theta}\mathbf{W}$ found by PAA [solving (9)], *archetypal profiles*: our motivation is that $\boldsymbol{\Theta}_{\cdot n}$ can be seen as the parametric *profile* that best describes the single observation $\mathbf{x}_n$, and thus, $\mathbf{Z}$ are the archetypal profiles that are inferred from them. We generally refer to the set of indices that contribute to the k-th archetypal profile, i.e., $\{i : \mathbf{W}_{ik} > \delta\}$, where δ is a small value, as *generating observations* of that archetype. Notice that, when the observation model is multivariate normal with identity covariance, then this formulation is the same as solving (1). We explore some other examples of EF: multinomial, product of univariate Poisson distributions, and product of Bernoulli distributions.

3.2 Poisson observations

If the observations are integer–valued then they are usually assumed to originate from a Poisson distribution. Then we need to solve the following problem,

$$\arg\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \sum_{mn} \left[-\mathbf{X}_{mn}\log(\boldsymbol{\Lambda}\mathbf{WH})_{mn} + (\boldsymbol{\Lambda}\mathbf{WH})_{mn}\right] \tag{10}$$

such that $\sum_j \mathbf{H}_{jn} = 1$ and $\sum_i \mathbf{W}_{ik} = 1$. Here Λ_{mn} is the maximum likelihood estimate of the Poisson rate parameter from observation $\mathbf{X}_{mn}$.

To solve this problem efficiently, we employ a similar technique used by Cutler and Breiman by relaxing the equality constraint with a suitably strong regularization parameter. However, we employ a multiplicative update rule afterwards instead of an exact method like

the nonnegative least squares. The resulting update rules are (see Appendix 1 for derivation)

$$\mathbf{H}^{t+1} = \mathbf{H}^t \odot \frac{\nabla^-_{\mathbf{H}^t}}{\nabla^+_{\mathbf{H}^t}},\ \nabla^+_{\mathbf{H}_{nj}} = \sum_{im} \boldsymbol{\Lambda}_{im}\mathbf{W}_{mn} + \lambda,$$
$$\nabla^-_{\mathbf{H}_{nj}} = \sum_{i} \frac{\mathbf{X}_{ij}\sum_m \boldsymbol{\Lambda}_{im}\mathbf{W}_{mn}}{\sum_{mn}\boldsymbol{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}} + \frac{\lambda}{\sum_n \mathbf{H}_{nj}} \tag{11}$$

and

$$\mathbf{W}^{t+1} = \mathbf{W}^t \odot \frac{\nabla^-_{\mathbf{W}^t}}{\nabla^+_{\mathbf{W}^t}},\ \nabla^+_{\mathbf{W}_{mn}} = \sum_{ij} \boldsymbol{\Lambda}_{im}\mathbf{H}_{nj} + \lambda,$$
$$\nabla^-_{\mathbf{W}_{mn}} = \sum_{ij} \frac{\mathbf{X}_{ij}\boldsymbol{\Lambda}_{im}\mathbf{H}_{nj}}{\sum_{mn}\boldsymbol{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}} + \frac{\lambda}{\sum_m \mathbf{W}_{mn}}. \tag{12}$$

Here $\odot$ denotes Hadamard product. We choose λ to be 20 times the variance of the samples which is sufficiently large to enforce stochasticity but not too strong to be dominating. A similar value $\lambda^2 = 200$ has been used by Eugster and Leisch (2011).

3.3 Multinomial observations

In many practical problems such as document analysis, the observations can be thought of as originating from a multinomial model. In such cases, PAA expresses the underlying multinomial probability as **PWH** where **P** is the maximum likelihood estimate achieved from word frequency matrix **X**. This decomposition is very similar to PLSA: PLSA estimates a topic by document matrix **H** and a word by topic matrix **Z**, while AA estimates a document by topic matrix (**W**) and a topic by document matrix (**H**) from which the topics can be estimated as archetypes $\mathbf{Z} = \mathbf{PW}$. Therefore, the archetypal profiles are effectively topics, but topics might not always be archetypes. For instance, given three documents {A,B}, {B,C}, {C,A}; the three topics could be {A}, {B}, and {C}, whereas the archetypes can only be the documents themselves. Thus, it can be argued that archetypes are topics with better interpretability.

To find archetypes for this observation model one needs to solve the following problem,

$$\arg\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} -\sum_{mn}\mathbf{X}_{mn}\log\sum_{ij}\mathbf{P}_{mi}\mathbf{W}_{ij}\mathbf{H}_{jn}, \tag{13}$$

such that $\sum_j \mathbf{H}_{jn} = 1$ and $\sum_i \mathbf{W}_{ik} = 1$.

This can be efficiently solved using expectation-maximization (or majorization–minimization) framework with the following update rules (see Appendix 1 for derivation),

$$\mathbf{H}^{t+1}_{ij} = \sum_{kl}\frac{\mathbf{X}_{il}\mathbf{H}_{ij}\mathbf{W}_{jk}\mathbf{P}_{kl}}{(\mathbf{HWP})_{il}},\ \mathbf{H}^{t+1}_{ij} = \frac{\mathbf{H}^{t+1}_{ij}}{\sum_j \mathbf{H}^{t+1}_{ij}} \tag{14}$$

and

$$\mathbf{W}^{t+1}_{jk} = \sum_{il}\frac{\mathbf{X}_{il}\mathbf{H}_{ij}\mathbf{W}_{jk}\mathbf{P}_{kl}}{(\mathbf{HWP})_{il}},\ \mathbf{W}^{t+1}_{jk} = \frac{\mathbf{W}^{t+1}_{jk}}{\sum_k \mathbf{W}^{t+1}_{jk}}. \tag{15}$$

3.4 Bernoulli observations

There are real world applications that deal with binary observations rather than real–valued or integers, e.g., binary questionnaire in marketing research. Such observations can be expressed in terms of the Bernoulli distribution. To find the archetypal representation of binary pattern we need to solve the following problem,

$$\arg\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \sum_{mn} \left[-\mathbf{X}_{mn}\log(\mathbf{PWH})_{mn} - \mathbf{Y}_{mn}\log(\mathbf{QWH})_{mn}\right], \tag{16}$$

such that $\sum_j \mathbf{H}_{jn} = 1$ and $\sum_i \mathbf{W}_{ik} = 1$, where $\mathbf{X}$ is the binary data matrix (with 1 denoting success/true and 0 denoting failure/false), $\mathbf{P}_{mn}$ is the probability of success estimated from $\mathbf{X}_{mn}$ (effectively either 0 or 1), $\mathbf{Y}_{mn} = 1 - \mathbf{X}_{mn}$, and $\mathbf{Q}_{mn} = 1 - \mathbf{P}_{mn}$.

This is a more involved form than the previous ones: one cannot use relaxation technique as in the Poisson case, since relaxation over the stochasticity constraint might render the resulting probabilities $\mathbf{PWH}$ greater than 1, thus making the cost function incomputable. Therefore, we take a different approach toward solving this problem by reparameterizing the stochastic vector (say $\mathbf{s}$) by an unnormalized non-negative vector (say $\mathbf{t}$), such that $\mathbf{s} = \mathbf{t}/\sum \mathbf{t}_i$. We show that the structure of the cost allows us to derive efficient update rules over the unnormalized vectors using majorization–minimization. Given $\mathbf{g}_n$ and $\mathbf{v}_k$ to be the reparameterization of $\mathbf{h}_n$ and $\mathbf{w}_k$ respectively, we get the following update equations, (see Appendix 1 for derivation),

$$\mathbf{G}^{t+1} = \mathbf{G}^t \odot \frac{\nabla^n_{\mathbf{G}}}{\nabla^d_{\mathbf{G}}},$$

with

$$\nabla^d_{\mathbf{G}_{nj}} = \sum_i \mathbf{X}_{ij} + \sum_i \mathbf{Y}_{ij}, \tag{17}$$

$$\nabla^n_{\mathbf{G}_{nj}} = \sum_i \frac{\mathbf{X}_{ij}\sum_m \mathbf{P}_{im}\mathbf{W}_{mn}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}} + \sum_i \frac{\mathbf{Y}_{ij}\sum_m \mathbf{Q}_{im}\mathbf{W}_{mn}}{\sum_{mn}\mathbf{Q}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}} \tag{18}$$

and

$$\mathbf{V}^{t+1} = \mathbf{V}^t \odot \frac{\nabla^n_{\mathbf{V}}}{\nabla^d_{\mathbf{V}}},$$

with

$$\nabla^d_{\mathbf{V}_{mn}} = \sum_{ij} \frac{\mathbf{X}_{ij}\sum_m \mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}} + \sum_{ij} \frac{\mathbf{Y}_{ij}\sum_m \mathbf{Q}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{Q}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}, \tag{19}$$

$$\nabla^n_{\mathbf{V}_{mn}} = \sum_{ij} \frac{\mathbf{X}_{ij}\mathbf{P}_{im}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}} + \sum_{ij} \frac{\mathbf{Y}_{ij}\mathbf{Q}_{im}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{Q}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}. \tag{20}$$

4 Related work

Archetypal analysis and its probabilistic extension share close connections with other popular matrix factorization methods. We explore some of these connections in this section, and provide a summary in Fig. 3. We represent the original data matrix by $\mathbf{X} \in \mathcal{X}^{M\times N}$ where

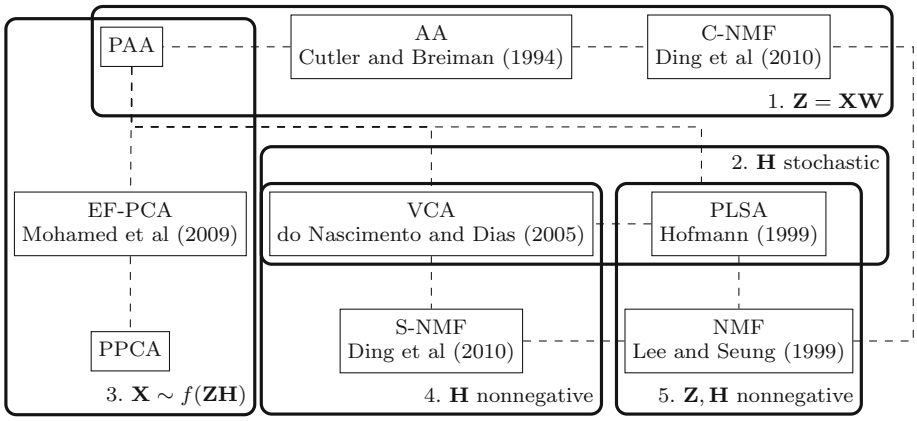

Fig. 3 Relations among factorization methods. *1.* data-driven methods where the loadings depend on input, *2.* simplex factor models with probability vector as factors, *3.* probabilistic methods, *4.* non-negative factors with arbitrary loadings, and *5.* non-negative factors with non-negative loadings. The connections are elaborated in Sect. 4

each column $\mathbf{x}_n$ is an observation; the corresponding latent factor matrix by $\mathbf{H} \in \mathbb{R}^{K \times N}$, and loading matrix by $\mathbf{Z} \in \mathbb{R}^{M \times K}$.

Principal component analysis and extensions: Principal component analysis (PCA) finds an orthogonal transformation of a point-cloud, which projects the observations in a new coordinate system that preserves the variance of the point-cloud the best. The concept of PCA has been extended to a probabilistic as well as a Bayesian framework (Mohamed et al. 2009). Probabilistic PCA (PPCA) assumes that the data originates from a lower dimensional subspace on which it follows a normal distribution ($\mathcal{N}$), i.e.,

$$\mathbf{h}_n \sim \mathcal{N}(0, \mathbf{I}),\ \mathbf{x}_n \sim \mathcal{N}(\mathbf{Z}\mathbf{h}_n, \epsilon\mathbf{I})$$

where $\mathbf{h}_n \in \mathbb{R}^K$, $\mathbf{x}_n \in \mathbb{R}^M$, $K < M$, and $\epsilon > 0$.

Probabilistic principal component analysis explicitly assumes that the observations are *normally distributed*: an assumption that is often violated in practice, and to tackle such situations one extends PPCA to exponential family (EF). The underlying principle here is to change the observation model accordingly:

$$\mathbf{h}_n \sim \mathcal{N}(0, \mathbf{I}),\ \mathbf{x}_n \sim \mathrm{EF}(\mathbf{x}_n; \mathbf{Z}\mathbf{h}_n),$$

i.e., each element of $\mathbf{x}_n$ is generated from the corresponding element of $\mathbf{Z}\mathbf{h}_n$ as $\mathrm{EF}(z; \theta) = h(z)g(\theta)\exp(\theta s(z))$ where $s(z)$ is the sufficient statistic, and θ is the *natural parameter*: PAA utilizes similar approach but with normal parameters.

Similarly, one can also manipulate the latent distribution. A popular choice is the Dirichlet distribution (Dir), which has been widely explored in the literature, e.g., in probabilistic latent semantic analysis (PLSA (Hofmann 1999)), $\mathbf{h}_n \sim \mathrm{Dir}(\mathbf{1})$, $\mathbf{x}_n \sim \mathrm{Mult}(\mathbf{Z}\mathbf{h}_n)$, where $\mathbf{1}\mathbf{Z} = \mathbf{1}$; vertex component analysis (do Nascimento and Dias 2005), $\mathbf{h}_n \sim \mathrm{Dir}(\mathbf{1})$, $\mathbf{x}_n \sim \mathcal{N}(\mathbf{Z}\mathbf{h}_n, \epsilon\mathbf{I})$; and simplex factor analysis (Bhattacharya and Dunson 2012), a generalization of PLSA (or more specifically of latent Dirichlet allocation, LDA (Blei et al. 2003)): PAA additionally decompose the loading in simplex factors with known loading.

Nonnegative matrix factorization and extensions: Non-negative matrix factorization (NMF) decomposes a non-negative matrix $\mathbf{X} \in \mathbb{R}_+^{M\times N}$ in two non-negative matrices $\mathbf{Z} \in \mathbb{R}_+^{M\times K}$ and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ such that $\mathbf{X} \approx \mathbf{ZH}$. (Lee and Seung 1999) applied the celebrated *multiplicative update rule* to solve this problem, and proved that such update rules lead to monotonic decrease in the cost function using the concept of *majorization–minimization* (Lee and Seung 2000). Non-negative matrix factorization has been extended to convex non-negative matrix factorization (C-NMF (Ding et al. 2010)) where $\mathbf{X}$ is not restricted to be non-negative, and $\mathbf{Z}$ is expressed in terms of the $\mathbf{X}$ itself as $\mathbf{Z} = \mathbf{XW}$, where $\mathbf{W}$ is again a non-negative matrix. The motivation for this modification emerges from its similarity to clustering, and C-NMF has been solved using multiplicative update rule as well.

To simulate the exact clustering scenario, however, $\mathbf{H}$ is required to be binary (hard clustering) or at least column stochastic (fuzzy clustering). This leads to a more difficult optimization problem, and is usually solved by proxy constraint $\mathbf{H}^\top\mathbf{H} = \mathbf{I}$ (Ding et al. 2006). Several other alternatives have also been proposed for tackling the stochasticity constraints, e.g., by enforcing it after each iteration (Mørup and Hansen 2012), or by employing a gradient-dependent Lagrangian multiplier (Yang and Oja 2012). However, both these approaches are prone to finding local minima.

5 Simulation

In this section, we provide some simple examples showing the difference between probabilistic and standard archetypal analysis solutions. Since we generate data following the true probabilistic model, it is expected that the solution provided by PAA would be more appropriate compared to the standard AA solution. Therefore, the purpose of this section is to perform sanity check, and provide insight. Notice that generating observations with known archetypes is not straight forward, since $\boldsymbol{\Theta}$ depends on $\mathbf{X}$.

Binary observations: We generate $K = 6$ binary archetypes in $d = 10$ dimensions by sampling $\eta_{ik} \sim \text{Bernoulli}(p_s)$, where $\boldsymbol{\eta}_k$ is an archetype, and $p_s = 0.3$ is the probability of success. Given the archetypes, we generate $n = 100$ observations as $\mathbf{x}_i \sim \text{Bernoulli}(\mathbf{E}\mathbf{h}_i)$, where $\mathbf{E} = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k]$, and each $\mathbf{h}_i$ is a stochastic vector sampled from $\text{Dir}(\alpha)$. To ensure that $\boldsymbol{\eta}$'s are archetypes, we maintain more observations around $\boldsymbol{\eta}_k$s by choosing $\alpha = 0.4$. We find archetypal profiles using both PAA and standard AA, and then binarize them so that they can be matched to the original archetypes using minimum Jaccard distance. We report the results in Fig. 4. We observe that the archetypal profiles found by PAA are more accurate with 15 % more archetypal profiles matching uniquely to the true archetypes than standard AA.

Poisson observations: We generate $K = 6$ count archetypes in $d = 12$ dimensions with one minimal archetype ($\eta_{ik} = 0$), one maximal archetype $\eta_{ik} \sim \text{Unif}\{1, \ldots, 10\}$, and rest of the archetpyes with two nonzero entries $\eta_{ik} \sim \text{Unif}\{1, \ldots, 10\}$. Given the archetypes, we generate $n = 500$ observations as $\mathbf{x}_i \sim \text{Poisson}(\mathbf{E}\mathbf{h}_i)$, where $\mathbf{E} = [\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k]$, and each $\mathbf{h}_i$ is a stochastic vector sampled from $\text{Dir}(\alpha)$. To ensure that $\boldsymbol{\eta}$'s are archetypes, we maintain more observations around $\boldsymbol{\eta}_k$s by choosing $\alpha = 0.4$. We find archetypal profiles using both PAA and standard AA, and match them to the original archetypes using minimum l_1 distance. We report the results in Fig. 5. We observe that the archetypal profiles found by PAA are more accurate with 9 % more archetypal profiles matching uniquely to the true archetypes than standard AA.

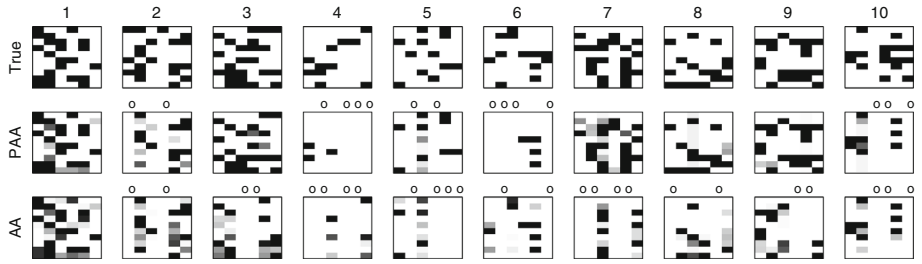

Fig. 4 The figure compares the solutions achieved by standard archetypal analysis and the probabilistic formulation on binary observations. *Each column* is an independent trial, and each algorithm has been run 10 times to find the best archetypal profiles. The archetypal profiles are binarized, and matched with the true archetypes using minimum Jaccard distance. If a unique match is found then the corresponding archetypal profile is displayed. Otherwise they are *left blank*, and tagged by a *circle*. We observe that the probabilistic approach has been able to match archetypes better than the standard solution

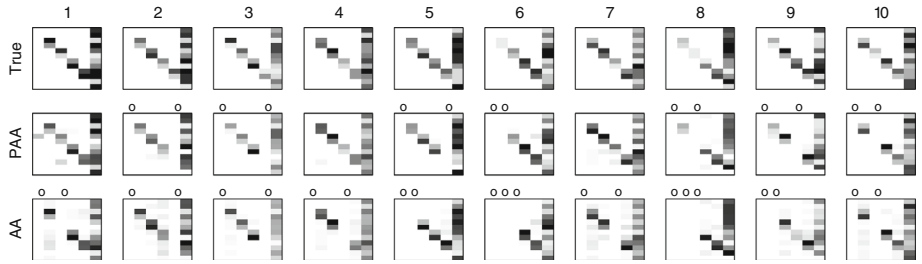

Fig. 5 The figure compares the solutions achieved by standard archetypal analysis and the probabilistic formulation on count observations. *Each column* is an independent trial, and each algorithm has been run 10 times to find the best archetypal profiles. The archetypal profiles are matched with the true archetypes using minimum l_1 distance. If a unique match is found then the corresponding archetypal profile is displayed. Otherwise they are *left blank*, and tagged by a *circle*. We observe that the probabilistic approach has been able to match archetypes better than the standard solution

Term-frequency observations: We generate $K = 5$ archetypes on $d = 3$ dimensional probability simplex by choosing K equidistant points $\mathbf{p}_k$ on a circle in the simplex. Given the archetypes, we generate $n = 500$ observations as $\mathbf{x}_i \sim \text{Mult}(n_i, \mathbf{Ph}_i)$, where $\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_k]$, and each $\mathbf{h}_i$ is a stochastic vector sampled from $\text{Dir}(\alpha)$. To ensure that $\mathbf{p}$'s are archetypes, we maintain more observations around them by choosing $\alpha = 0.5$. We deliberately choose an arbitrary number of occurrences $n_i \sim \text{Uniform}[1000, 2000]$ for each observation: this disrupts the true convex hull structure in the term-frequency observations. We present 10 random runs on these observations for both PAA and standard AA in Fig. 6. We observe that PAA finds the effective archetypes, with occasional local minima. However, standard AA performs poorly since it finds the appropriate archetypes in the term-frequency space, which are different when projected back on the simplex.

To evaluate the difference between PAA and standard AA quantitatively, we generate test samples following the same data generating process for each observation model. For PAA, given a test sample $\mathbf{x}_i$, we compute $\mathbf{h}_i$ by maximizing $p(\mathbf{x}_i|\mathbf{Zh}_i)$. Here $\mathbf{Z}$ are the archetypes inferred by the training samples. This can be done by minimizing the related cost functions with respect to $\mathbf{h}$ while keeping $\mathbf{W}$ fixed (where $\mathbf{Z} = \boldsymbol{\Theta}\mathbf{W}$). For standard AA, we compute the projection $\mathbf{h}_i$ on inferred archetypes $\mathbf{Z}$ by solving (2). For standard AA, we treat $\mathbf{Zh}$ as archetypal profile. We present the negative log-likelihood $-\sum_i \log p(\mathbf{x}_i|\mathbf{Zh}_i)$ over all test

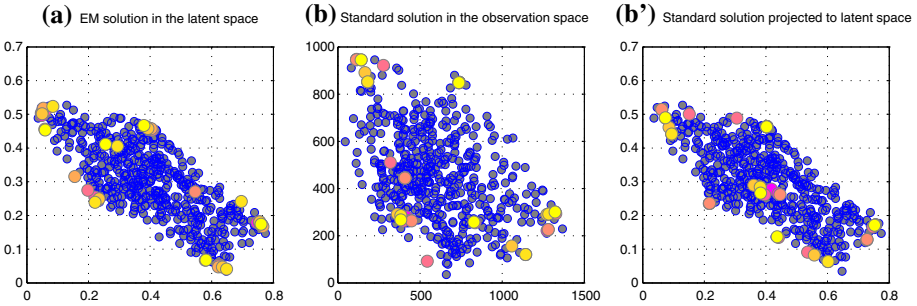

Fig. 6 The figure compares the solutions achieved by standard archetypal analysis and probabilistic archetypal analysis on term-frequency observations. Each observation vector of the term-frequency matrix is generated from a probability vector within a clear convex hull (**a**). The probability vectors are generated such there are 5 archetypes. However, this structure is lost in the term-frequency values due to arbitrary number of occurrences in each term-frequency vector (**b**). The standard AA applied to term-frequency matrix thus, does not capture the true archetypes (**b'**)

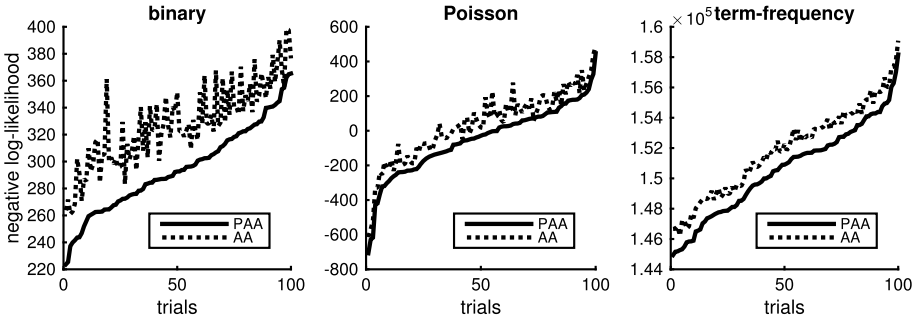

Fig. 7 Comparison of PAA and standard AA on simulated datasets. We evaluate how well the inferred archetypes can represent a new set of samples in terms of negative log-likelihood. We observe that PAA achieves lower error over all trials

samples in Fig. 7. We observe that, as expected, PAA achieves better performance for each observation model.

6 Simplex visualizations

The column stochasticity of $\mathbf{H}$ allows a principled visualization scheme of archetypal analysis solution, referred to as *simplex visualization*. We discuss certain aspects and enhancements of this approach, and show how it can be utilized to better understand the inferred archetypes.

The stochastic nature of $\mathbf{h}_n$ implies that $\mathbf{Z}\mathbf{h}_n$ exists on a standard $(K-1)$-simplex with the K archetypes $\mathbf{Z}$ as the corners, and $\mathbf{h}_n$ as the coordinate with respect to these corners (see Fig. 1b for an illustration). A standard simplex can be projected to two dimensions via a skew orthogonal projection, where all the vertices of the simplex are shown on a circle connected by edges. The individual factors $\mathbf{h}_n$ can be then projected into this circle. Figure 8 illustrates this principle with the simple data set already used in Fig. 1: (a, b) for the three archetypes solution; (g, h) for the four archetypes solution; (j, k) for the five archetypes solution. Color coding is used in the six figures to show the relation between the original observation $\mathbf{x}_n$

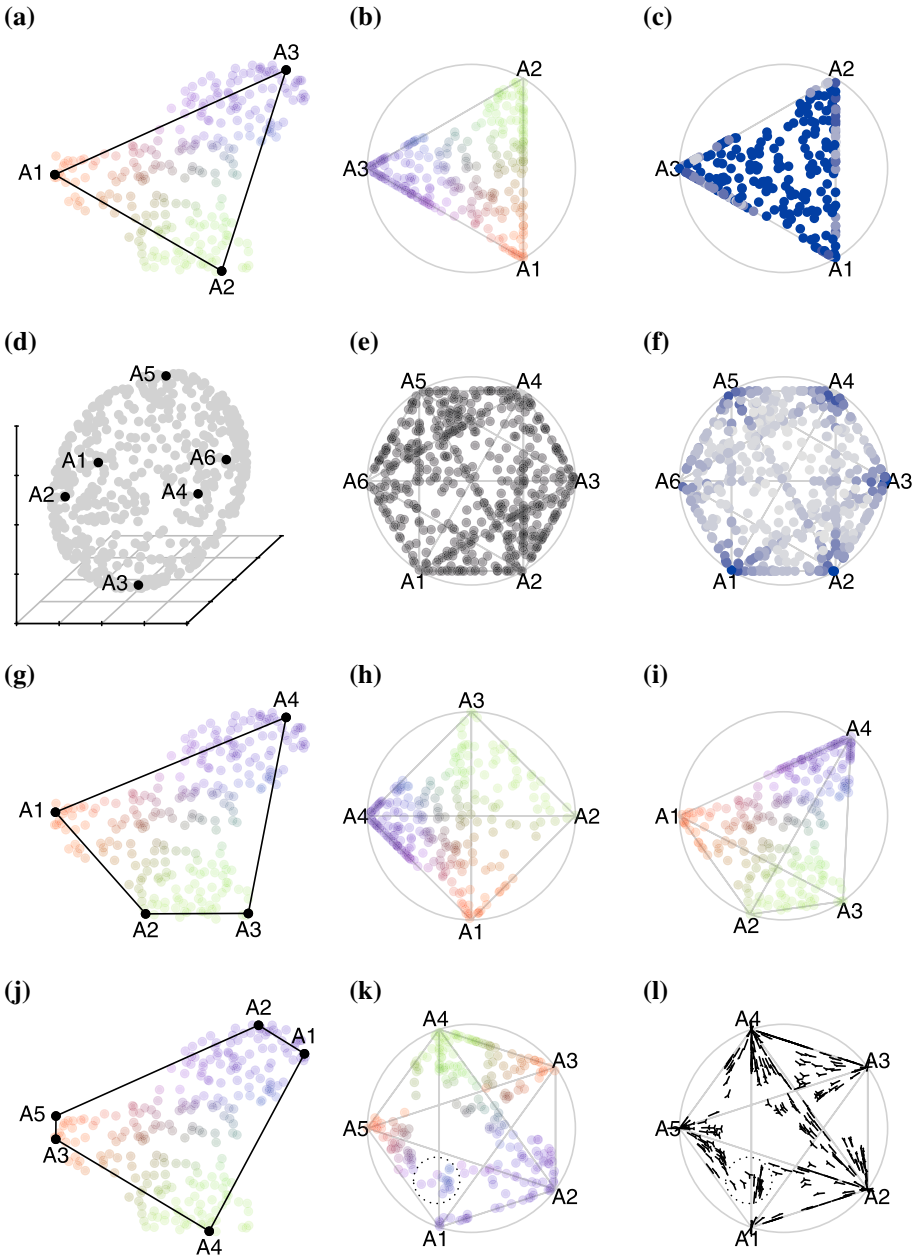

Fig. 8 Simplex visualizations with extensions for different illustrative data sets and AA solutions: **a–c** and **d–f** illustrate *color-coded points* based on the deviance; **g–i** illustrate the ordering of the vertices according to the distances of the archetypes in the original space; and **j–l** illustrate the enhancement of the plot to show the composition of observations (Color figure online)

and its projection $\mathbf{h}_n$. The visualization with three archetypes is known as ternary plot (see, e.g., Friendly 2000), and has been used by Cutler and Breiman (1994). The extension to more than three archetypes has also been used (e.g., Bauckhage and Thurau 2009; Eugster

and Leisch 2013). However, a formal study of this visualization scheme, to the best of our knowledge, remains to be explored. In the following, we present three enhancements of this basic visualization to either highlight certain characteristics of an archetypal analysis solution or to overcome consequences of the one-to-many projection.

In AA, observations which lie outside the approximated convex hull are projected onto its boundary. Figure 8a, b show a simple scenario where these observations are projected onto the corresponding edges. Figure 8d, e show an extreme case of this characteristic: the observations lie on a three-dimensional sphere, and therefore the computed archetypes span a space, which is completely empty. In the corresponding simplex visualization, however, this aspect of the solution is not visible at all. We propose to visualize the 'deviance' $D(\mathbf{x}_n) = 2(\log p(\mathbf{x}_n|\theta_n) - \log p(\mathbf{x}_n|\mathbf{Z}\mathbf{h}_n))$ where θ_n is the maximum likelihood estimate of $\mathbf{x}_n$, as colors of the points. In case of normal observation model the deviance reduces to the residual sum of squares. Figure 8c shows the corresponding simplex visualization with deviance for the three archetypes solution. The color scheme is from blue to white, with blue implying zero deviance and the lighter the color, the higher the deviance. We can now identify how well the original observations have been approximated by the archetypes, and if they are inside the convex hull. This extension is even more insightful in case of the sphere example. In the corresponding simplex visualization in Fig. 8f, we can now clearly see that almost all observations are outside the approximated convex hull; only around the corners the deviance is near to zero.

The basic simplex visualization arranges the vertices, which represents the archetypes, equidistant on the circle. The archetypes in the original space, however, are usually not equidistant to each other. Figure 8g and h illustrate this discrepancy: archetype A2, for example, is much nearer to A3 than A4; in the simplex visualization, however, both are in the same distance. We propose to order the vertices on the circle according to their distances in the original space. This means, we first have to determine an optimal order of the vertices, and then divide the 360° of the circle in relation to the original pairwise distances of the determined neighbor vertices. Here, we solve a Traveling Salesman Problem to get an optimal cyclic order (solved by using, for example, Hahsler and Hornik 2007); and then simply divide the circle proportional to the original distances. Figure 8i shows the result: it is now clearly visible that A2 and A3 are much nearer to each other than A3 and A4.

Another problem of the simplex visualization with more than three archetypes is the *non-uniqueness*. As a result two projections $\mathbf{h}_{n1}$ and $\mathbf{h}_{n2}$ can be close to each other even though they are *composed* of different archetypes. This goes against one's intuition in judging which archetypes the observations belong to. For example, the observations inside the dashed circle in Fig. 8k. We get the idea that these observations are basically composed by A1, A2, A5, and/or A4—but we do not get the exact compositions. We therefore propose to show 'whiskers', which point in the direction of the composing archetypes, Fig. 8l shows the corresponding visualization. We can now easily see the composition of the observations inside the dashed circle: the observations on the right side of the line are composed by A1 and A4; the observations on the left side of the line are composed by A1, A2, A5; and the left-most observation is composed by A1, A5. We vary the length of the 'whiskers' according to the coefficients $\mathbf{h}_n$; the longer the whisker the closer the observation is to the archetype the whisker points toward.

7 Applications

In each application below (except for the first two), we run PAA with 2–15 archetypes, 10 trials with random initializations for each number of archetypes, and choose the solution from

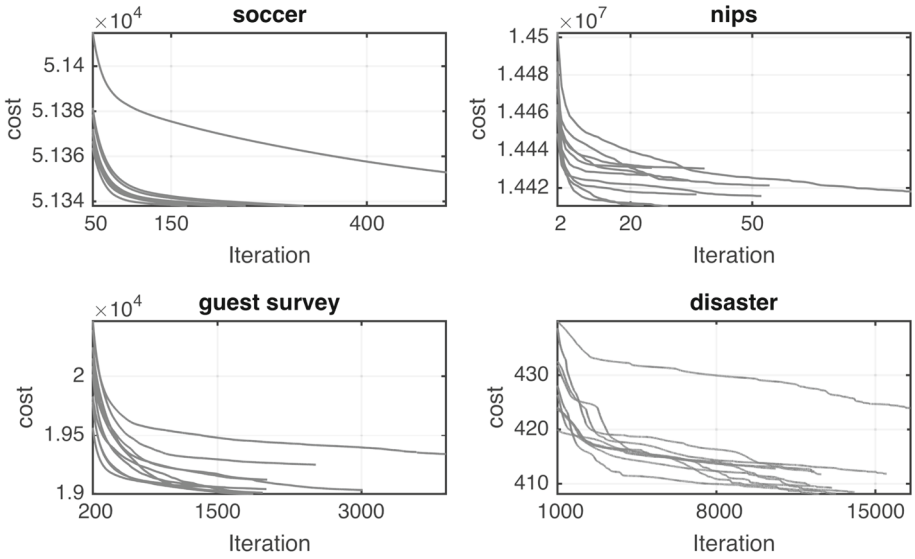

Fig. 9 Convergence curves of each run for the three datasets. Convergence is reached when the relative different between two successive iterations is less than 10^{-6}

the trials with maximum likelihoods according to the "elbow criterion". The elbow criterion is a simple heuristic. Due to the fact that with each additional archetype $\mathbb{LL}$ increases, one can compute solutions with successively increasing number of archetypes, plot $\mathbb{LL}$ against the number or archetypes, and visually pick the solution after which the jump of $\mathbb{LL}$ "is only marginal" (i.e., the elbow). This solution is ad hoc and subjective—but widely used: "Statistical folklore has it that the location of such an 'elbow' indicates the appropriate number of clusters" (Tibshirani and Walther 2005). Figure 9 shows the convergence curves of each run associated with the number of archetypes selected for final analysis.

7.1 Gaussian observations: archetypal soccer players

We use a data set already explored by Eugster (2012) for archetypal analysis, to evaluate the solution provided by maximizing (6). We analyze soccer players playing in four European top leagues. The extracted data set consists of $M = 25$ skills of $N = 1658$ players (all positions—Defender, Midfielder, Forward—except Goalkeepers) from the German Bundesliga, the English Premier League, the Italian Serie A, and the Spanish La Liga. The skills are rated from 0 to 100 and describe different abilities of the players: physical abilities like balance, stamina, and top speed; ball skills like dribble, pass, and shot accuracy and speed; and general skills like attack and defence performance, technique, aggression, and teamwork. Note that Eugster (2012) assumes that the differences are interpretable, i.e., the ratings are on a ratio scale. We compute $K = 4$ archetypes, as in Eugster (2012).

Figure 10 shows the percentile plots of the computed solution. Archetype 1 is the archetypal offensive player with all skills high except the defense, balance, header, and jump. Archetype 2 is the archetypal center forward with high skills in attack, shot, acceleration, header and jump, and low passing skills. Archetype 3 is the archetypal weak soccer player with high skills in running, but low skills in most ball related abilities. Archetype 4 is the archetypal defender with high skills in defense, balance, header, and jump. As expected, this

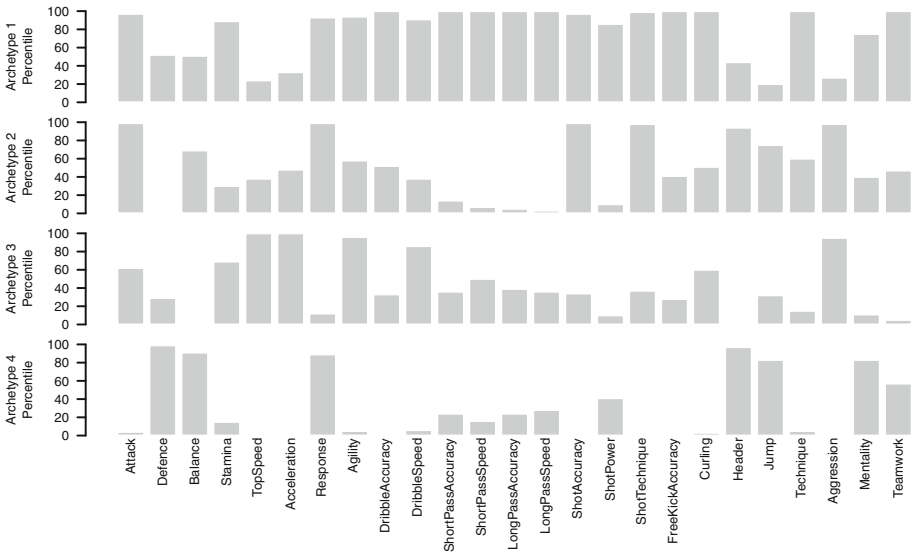

Fig. 10 Percentile plot of the four archetypal profiles for the soccer players example. The result is basically the same as in Eugster (2012). More information in Sect. 7.1

solution is very similar to the solution computed with the classical algorithm and presented by Eugster (2012, Figure 7). We observe only small differences between the two solutions; the biggest difference is that Archetype 3 has a much higher aggression value in the probabilistic solution than in the classical solution.

7.2 Multinomial observations: NIPS bag-of-words

We use a data set already explored by Mørup and Hansen (2012) for archetypal analysis, to qualitatively evaluate the solution provided by PAA with multinomial observation model. We analyze the NIPS bag-of-words corpus consisting of $N = 1500$ documents and $M = 12419$ words (available from (Bache and Lichman 2013)) and compute $K = 10$ document archetypes, as in (Mørup and Hansen 2012). We use the term-frequencies as features without normalizing them by the document frequency as in (Mørup and Hansen 2012) to adhere to the generative nature of the documents. Figure 11 shows the probability for each word available in the corpus to be generated by the corresponding archetype ($\mathbf{Z}$). Following (Mørup and Hansen 2012), we highlight the ten most prominent terms after ignoring the "common" terms, which are present in each of the archetypes in the first 3000 words (with probability values $> 10^{-4}$). We can observe that the prominent terms in a particular archetype have low probability in all the other archetypes: this agrees with our understanding of an archetype. Overall our algorithm finds a similar solution to Mørup and Hansen (2012)—one difference, however, protrudes: we find a "Bayesian Paradigm" archetype (A2), which Mørup and Hansen (2012) finds as a k-means prototype. But, to the best of our knowledge a "Bayesian Paradigm" archetypal document can make sense in a NIPS corpus.

7.3 Bernoulli observations: Austrian national guest survey

Analyzing binary survey data is of utmost importance in social science and marketing research. A binary questionnaire is often preferred over an ordinal multi-category format,

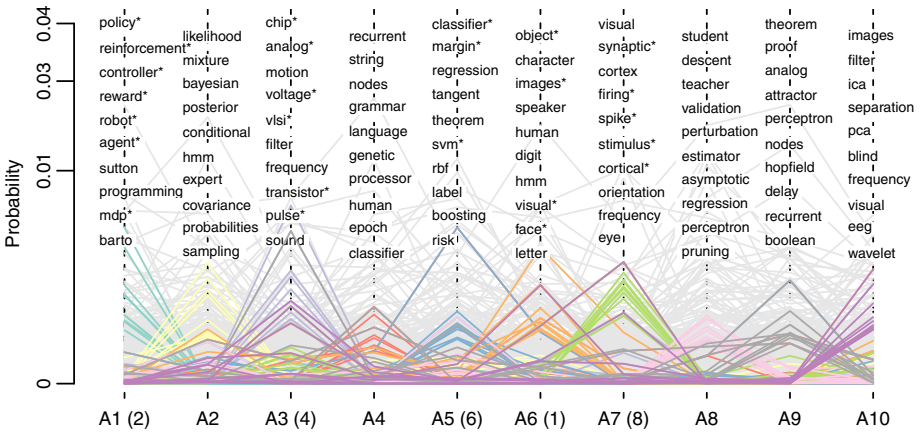

Fig. 11 The probability of the words available in the NIPS corpus for each of the ten archetypal profiles. The number in parentheses refers to the corresponding archetype in (Mørup and Hansen 2012). The *colored lines* show the ten most prominent words (after removing the "common" words), *asterisk indicates* which words appear in both solutions. More information in Sect. 7.2 (Color figure online)

since the former is quicker and easier, whereas both are equally reliable, and the managerial implications derived from them do not substantially differ (Dolnicar et al. 2011). In this application, we analyze binary survey data from the Austrian National Guest Survey conducted in the winter season of 1997 [for a cluster analysis of the summer season of 1997 see (Dolnicar and Leisc 2004)]. The goal is to identify archetypal winter tourists, which may facilitate developing and targeting specific advertising materials. The data consists of 2958 tourists. Each tourist answers 25 binary questions on whether he/she is engaged in a certain winter activity (e.g., alpine skiing, relaxing, or shopping; see row description of Table 2 for the complete list). In addition, a number of descriptive variables are available (e.g., the age and gender of the tourist).

Here we present the six archetypes solution. Table 2 lists the archetypal profiles (i.e., the probability of positive response) and, in parentheses, the corresponding archetypal observations (with maximum w value). Archetype A1 is the maximal winter tourist who is engaged in nearly every sportive and wellness activity with high probability. Archetype A3, on the other hand, is the minimal winter tourist who is only engaged in alpine skiing and having dinner. Both archetypes A5 and A4 are engaged in the basic Austrian winter activities (alpine skiing, indoor swimming, and relaxing). In addition, A5 is engaged in traditional activities (dinner and shopping), whereas A4 is engaged in more modern activities (snowboarding and going to a disco). Finally, A6 and A2 are the non-sportive archetypes. A6 is engaged in wellness activities and A2 with cultural activities. Note that important engagements of the archetypes are missed if one only looks at the archetypal observations rather than the archetypal profiles; e.g., the possible engagement of A2 in hiking. We can now utilize the factors $\mathbf{H}$ for each of the tourists to learn their relations to the archetypal winter tourist profiles. This allows us, for example, to target very specific advertising material to tourists for the next winter season.

To get further insight into the archetypes we explore the simplex visualization. Figure 12 shows four simplex visualizations with the archetypes arranged according to their distance in the original space and the composition of the winter tourists indicated by corresponding whiskers. Figure 12a shows the model deviance normalized from 0 (blue) to 1 (white). We can observe that the tourists mainly explained by the Minimal (A3), Modern (A4),

Table 2 The six archetypal profiles for the winter tourists example

	A1	A3	A5	A4	A6	A2
Alpine Ski	1.00 (1)	1.00 (1)	1.00 (1)	1.00 (1)	0.00 (0)	0.00 (0)
Tour Ski	0.41 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Snowboard	0.00 (0)	0.00 (0)	0.00 (0)	0.59 (1)	0.00 (0)	0.00 (0)
Cross Country	0.75 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Ice Skating	0.60 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Sledge	1.00 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Tennis	0.15 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.20 (0)
Riding	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.08 (0)
Pool Sauna	0.96 (1)	0.00 (0)	0.37 (0)	1.00 (1)	0.82 (1)	0.11 (0)
Spa	0.22 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.79 (1)	0.00 (0)
Hiking	0.95 (1)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (1)	0.18 (0)
Walk	1.00 (1)	0.00 (0)	1.00 (1)	0.00 (0)	1.00 (1)	1.00 (1)
Excursion (org)	0.29 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.41 (1)
Excursion (ind)	0.81 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.94 (1)	1.00 (1)
Relax	0.99 (1)	0.00 (0)	1.00 (1)	1.00 (1)	1.00 (1)	0.81 (0)
Dinner	0.82 (1)	0.53 (0)	0.86 (1)	0.02 (0)	0.00 (0)	1.00 (1)
Shopping	1.00 (1)	0.00 (0)	1.00 (1)	0.01 (0)	0.33 (0)	1.00 (1)
Concert	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.29 (1)
Sightseeing	0.66 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.66 (1)	1.00 (1)
Heimat	0.58 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)
Museum	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	1.00 (1)
Theater	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.30 (1)
Heurigen	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.45 (0)
Local event	0.99 (1)	0.00 (0)	0.00 (0)	0.00 (0)	0.00 (0)	0.10 (0)
Disco	0.68 (1)	0.00 (0)	0.00 (0)	1.00 (1)	0.00 (0)	0.08 (0)
Interpretation of the archetypes	Maximal	Minimal	Basic		Non-sportive	
			Traditional	Modern	Wellness	Cultural

The corresponding archetypal observations are shown in parentheses. For more information, see Sect. 7.3

Traditional (A5), and/or Wellness (A6) archetypes are well represented (darker blue). The Maximal (A1) archetype seems to be an outlier, there are only a few observations near to this archetype, and the deviance is higher for these observations. Figure 12b–d highlight tourists' answers to certain questions (yes/black and no/gray). Figure 12b shows whether a tourist does snowboarding or not. We can see that most of the tourists who do snowboarding are arranged around and point towards A4, which we interpreted as the Modern archetype. In Fig. 12c we highlight whether a tourist visits a museum or not. Here, most of the tourists who go there are arranged around A2, which we interpreted as the Cultural archetype. Figure 12d shows an activity which does not discriminate between archetypes—nearly all tourists do shopping, and no specific pattern is visible in this visualization.

For comparison purpose, we also compute the six archetypes solution with the original archetypal analysis algorithm (details and results can be found online). The classical algorithm finds archetypes with a similar intrepretation. However, we observe that the probabilistic archetypal profiles are "more extreme" and therefore easier to interpret.

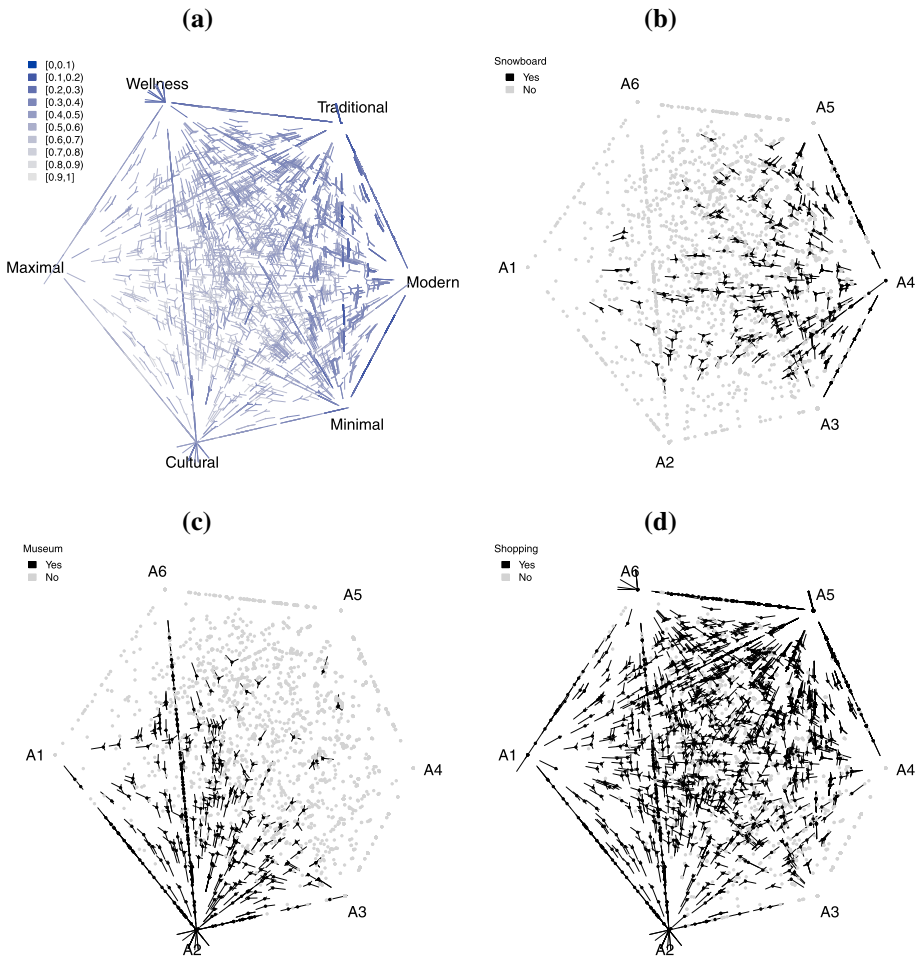

Fig. 12 Simplex visualizations for the the Austrian national guest survey example. The archetypes are arranged according to their distance in the original space and the composition of the winter tourists is indicated by corresponding whiskers. **a** shows the model deviance normalized from 0 (*blue*) to 1 (*white*); figures **b**–**d** highlight tourists' answers to certain questions (yes/*black* and no/*gray*). See Sect. 7.3 for detailed interpretations (Color figure online)

7.4 Poisson observations: disasters worldwide from 1900–2008

In this application, the goal is to identify archetypal countries that are affected by a particular disaster or a combination of disasters. This may be helpful in emergency management and to facilitate devising disaster prevention plans for countries based on prevention plans designed for the archetypal disaster-affected countries. We compile a dataset with disaster counts for 227 countries (historical and present countries) in 15 categories from the EM-DAT database (EM-DAT 2013). This is a global database on natural and technological disasters between 1900–present. The criteria to be a disaster are: ten or more reported casualty; hundred or more people reported affected; declaration of a state of emergency; or call for international assistance. The list of disaster categories is provided in Figure 13; see the EM-DAT website for specific details.

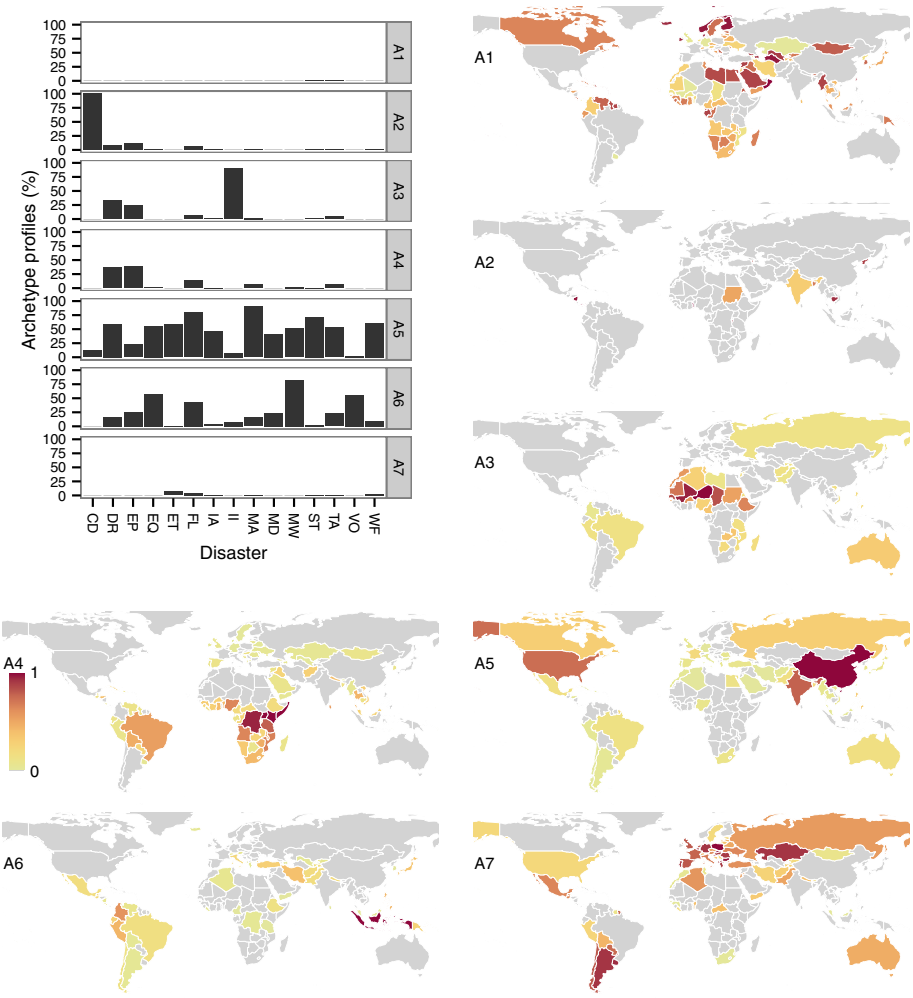

Fig. 13 The seven archetypal profiles for the disaster example: (*top left*) Plot of archetypal profiles (% of maximum value); (world maps) Factors **H** for each archetype. Disasters: complex disasters (CD), drought (DR), earthquake (EQ), epidemic (EP), extreme temperature (ET), flood (FL), industrial accident (IA), insect infestation (II), mass movement dry (MD), mass movement wet (MW), miscellaneous accident (MA), storm (ST), transport accident (TA), volcano (VO), and wildfire (WF). See Sect. 7.4 for details

We present the seven archetypes solution; Figure 13 shows a summary. There are two minimal profiles A1 and A7 with small differences in the categories extreme temperature/flood and storm. A1 can be considered as the archetypal profile for safe country where the corresponding archetypal observations include Malta and the Cayman Islands (other close observations are the Nordic countries). Archetype A5 is the maximal archetypal profile with counts in every category, and the corresponding archetypal observations include China and United States. This can be expected from the size and population of the countries; China (third and first), USA (fourth and third). Other countries with high factor **H** for this archetypal profile are India (seventh and second), and Russia (first and ninth). A3 and A4 are the archetypes

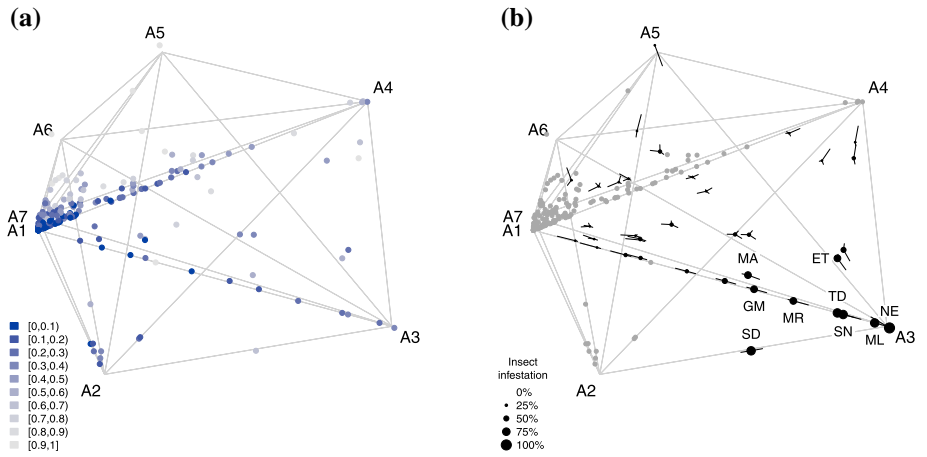

Fig. 14 Simplex visualizations for the disasters example: The archetypes are arranged according to their distance in the original space. **a** shows the model deviance normalized from 0 (*blue*) to 1 (*white*); **b** shows the projected countries scaled according to the number of Insect infestations (ISO2 country codes for the top countries). See Sect. 7.4 for detailed interpretations (Color figure online)

that are affected by drought and epidemic; where A3 additionally has a high insect infestation count. A2 is the archetype that is susceptible to complex disasters (where neither nature nor human is the definitive cause) only, whereas A6 has high counts in the categories earthquake, flood, mass movement wet, and volcano. Here the archetypal countries include Indonesia and Colombia.

Figure 14 shows two simplex visualizations with the archetypes arranged according to their distance in the original space. This arrangement shows that A1 is very near to A7, which is in line with the archetypal profiles plot shown in Figure 13. Figure 14a shows the model deviance normalized from 0 (blue) to 1 (white). We can see that A2, A3, A4, A5 and A6 are basically outliers with only one or very few observations are around them. Most of the observations with low deviance are near A1 and A7. Figure 14b highlights the countries' insect infestation (the higher the count the bigger the point). We can see a clear pattern around A3, with Niger (NE), Chad (TD), Mali (ML), Senegal (SN), Sudan (SD), Ethiopia (ET), Gambia (GM), Mauritania (MR), and Morocco (MA) as the top countries affected by insect infestation.

For comparison purpose, we also compute the seven archetypes solution with the original archetypal analysis algorithm (details and results can be found online). Both solutions are similar. However, we observe that the probabilistic solution is more robust towards outliers. This can be observed from the two simplex visualizations (available online). Both solutions go towards the outliers, however, the classical solution is more sensitive towards it compared to the probabilistic solution (e.g., PAA6 vs. AA6 and PAA5 vs. AA7). The robustness provides more freedom to explain the other observations and increases the interpretability.

8 Discussion

Archetypal analysis expresses observations as composition of extreme values, or archetypes. Archetypes can be thought of as ideal or pure characteristics, and the goal of archetypal analysis is to find these characteristics, and to explain the available observations as combination of these characteristics. The standard formulation of archetypal analysis was suggested by

Cutler and Breiman and is based on finding the approximate convex hull of the observations. Over the last decade this approach has been extensively used by researchers. But, their applications have mostly been limited to real–valued observations. In this paper, we have proposed a probabilistic formulation of archetypal analysis, which enjoys several crucial advantages over the geometric approach, including but not limited to the extension to other observation models: Bernoulli, Poisson and multinomial. We have achieved this by approximating the convex hull in the parameter space under a suitable observation model. Our contribution lies in formally extending the standard AA framework, suggesting efficient optimization tools based on majorization–minimization method, and demonstrating the applicability of such approaches in practical applications. We have also suggested improvements of the standard simplex visualization tool to better show the intricacies in the archetypal analysis solution.

The probabilistic framework provides further advantages that remain to be explored in their entirety. For example, it provides a theoretically sound approach for choosing the number of archetypes. This can be done by imposing appropriate priors over $\mathbf{W}$ and $\mathbf{H}$ matrices, such as a symmetric Dirichlet distribution with coefficient < 1. The prior can be used to effectively shrink and expand the convex hull to fit the observations. Since the Dirichlet distribution is a natural prior for multinomial distribution, this solution can be approximated relatively easily using variational Bayes' approach, and initial results show that this is indeed and effective approach for choosing the number of archetypes. However, this becomes a slightly trickier problem when applied to other observation models, such as normal and Poisson. We are currently working on suitable methods to solve the related optimization problems efficiently. It is worth mentioning that the convergence of the suggested algorithms is usually slower than standard factor models due to the additional constraint imposed on the loading matrix through $\mathbf{W}$. Additionally, multiplicative update itself can be slow, and therefore faster algorithms should be investigated to scale up probabilistic archetypal analysis to larger datasets.

Another potential extension of the probabilistic framework is to tackle ordinal or Likert scale variables. Since ordinal variables lack additivity, they must be addressed through a probabilistic set-up with suitable observation model. Given the fact that survey data is often in Likert scale, archetypal analysis of such observations can have a large impact on social science and marketing: describe, e.g., the personality of consumers in terms of the personality of the most "extreme", i.e., archetypal, consumers (using, e.g., the Likert scaled items defined by the Big Five Inventory). We believe that these suggested improvements will make archetypal analysis more robust and accessible to non-scientific users.

Acknowledgments The calculations presented above were performed using computer resources within the Aalto University School of Science "Science-IT" project.

Appendix 1: Alternate update rule for standard archetypal analysis

A different set of update rules can be derived if standard archetypal analysis is viewed in a probabilistic framework, i.e., if we wish to maximize,

$$\mathbb{LL}(\mathbf{X}|\mathbf{W}, \mathbf{H}, \mathbf{Z}, \boldsymbol{\Theta}) = -\frac{\epsilon_1}{2}||\mathbf{X} - \mathbf{ZH}||_{\mathrm{F}}^2 - \frac{\epsilon_2}{2}||\mathbf{Z} - \boldsymbol{\Theta}\mathbf{W}||_{\mathrm{F}}^2 + \frac{NM}{2}\log\epsilon_1 + \frac{MK}{2}\log\epsilon_2$$

where the constant terms have been ignored. We additionally set priors over ϵ_1 and ϵ_2 to be Gamma distributed with parameters (α_0, β_0), and assume the same family, i.e., Gamma(α, β) for the variational distributions. The variational distributions are then approximated as

$$q(\epsilon_1) = \text{Gamma}\left(\alpha_0 + \frac{NM}{2}, \beta_0 + \frac{1}{2}||\mathbf{X} - \mathbf{ZH}||_{\text{F}}^2\right)$$
$$q(\epsilon_2) = \text{Gamma}\left(\alpha_0 + \frac{MK}{2}, \beta_0 + \frac{1}{2}||\mathbf{Z} - \boldsymbol{\Theta}\mathbf{W}||_{\text{F}}^2\right).$$

We find the point estimates of $\mathbf{W}$ and $\mathbf{H}$ by maximizing the variational lower bound, and impose the stochasticity constraint in the same way as in (3) and (2). Finally, we find the point estimate of $\mathbf{Z}$ by maximizing the variational lower bound as

$$\mathbf{Z} = (\langle\epsilon_1\rangle\mathbf{XH}^\top + \langle\epsilon_2\rangle\boldsymbol{\Theta}\mathbf{W})(\langle\epsilon_1\rangle\mathbf{HH}^\top + \langle\epsilon_2\rangle)^{-1} \tag{21}$$

where for Gamma distribution $\langle\epsilon\rangle = \alpha/\beta$, and $\langle\cdot\rangle$ denotes the expectation operator. We use $\alpha_0 = \beta_0 = 1$ such that the prior mean is 1.

Appendix 2: Update rules for multinomial observations

For simplicity, let us consider the following problem of finding, $\mathbf{X} = \mathbf{HWP}$ where $\mathbf{X}$ is now $n \times m$ matrix instead of $m \times n$ matrix in the earlier sections. Then this problem can be viewed as given a document choose a topic following $\mathbf{H}$, then given a topic choose a document (subtopic) following $\mathbf{W}$, and finally given a document (subtopic) choose a word following $\mathbf{P}$. Let z_{il}^{jk} be the indicator variable for selecting topic j and document (subtopic) k. Then the log-likelihood of the observations $\mathbf{X}$ is given by

$$\begin{aligned}\mathbb{LL}(\mathbf{X}|\mathbf{H}, \mathbf{W}, \mathbf{P}, \mathbf{R}) &= \sum_{il} \mathbf{X}_{il} \log\left(\prod_{jk} \mathbf{H}_{ij}\mathbf{W}_{jk}\mathbf{P}_{kl}\right)^{z_{il}^{jk}} + C_0 \\ &= \sum_{ijkl} \mathbf{X}_{il} z_{il}^{jk} \log\left(\mathbf{H}_{ij}\mathbf{W}_{jk}\mathbf{P}_{kl}\right) + C_0\end{aligned}$$

At each expectation step we need to evaluate,

$$\mathbb{E}\left[z_{il}^{jk}|\mathbf{X}, \mathbf{H}, \mathbf{W}, \mathbf{P}\right] = \text{P}(z_{ik}^{jk} = 1|\mathbf{X}, \mathbf{H}, \mathbf{W}, \mathbf{P}) = \frac{\mathbf{H}_{ij}\mathbf{W}_{jk}\mathbf{P}_{kl}}{(\mathbf{HWP})_{il}}$$

then the maximization step gives us the final update equation.

Appendix 3: Update rule for Poisson observations

We show that the update rule discussed in the article leads to monotonic decrease in the cost function using majorization–minimization. We reformulate the problem as

$$\begin{aligned}\min_{\mathbf{W},\mathbf{H}\geq 0} \sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_{mn}\boldsymbol{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj} + \sum_{mn}\boldsymbol{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right] \\ + \lambda\sum_j\left(-\log\sum_n\mathbf{H}_{nj} + \sum_n\mathbf{H}_{nj}\right) + \lambda\sum_n\left(-\log\sum_m\mathbf{W}_{mn} + \sum_m\mathbf{W}_{mn}\right)\end{aligned}$$

where $\lambda > 0$ is a suitably large regularization parameter to enforce the equality constraint in a relaxed fashion.

Given $\phi_{inj} = \frac{\sum_m \mathbf{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn} \mathbf{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}$, and $\psi_{nj} = \frac{\mathbf{H}_{nj}}{\sum_n \mathbf{H}_{nj}}$, we can construct the auxiliary function for $\mathbf{H}$ as,

$$
\begin{aligned}
&\sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_{mn}\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj} + \sum_{mn}\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}\right] \\
&\quad + \lambda\sum_j\left(-\log\sum_n\tilde{\mathbf{H}}_{nj} + \sum_n\tilde{\mathbf{H}}_{nj}\right) \\
&\quad = \sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_n\frac{\phi_{inj}}{\phi_{inj}}\sum_m\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj} + \sum_{mn}\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}\right] \\
&\quad + \lambda\sum_j\left(-\log\sum_n\frac{\psi_{nj}}{\psi_{nj}}\tilde{\mathbf{H}}_{nj} + \sum_n\tilde{\mathbf{H}}_{nj}\right) \\
&\quad \leq \sum_{ij}\left[-\mathbf{X}_{ij}\sum_n\phi_{inj}\log\frac{\sum_m\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}}{\phi_{inj}} + \sum_{mn}\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}\right] \\
&\quad + \lambda\left(\sum_{nj}-\psi_{nj}\log\frac{\tilde{\mathbf{H}}_{nj}}{\psi_{nj}} + \sum_{nj}\tilde{\mathbf{H}}_{nj}\right) \\
&\quad = \sum_{ij}\left[-\mathbf{X}_{ij}\sum_n\phi_{inj}\log\tilde{\mathbf{H}}_{nj} + \sum_{mn}\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}\right] \\
&\quad + \lambda\left(\sum_{nj}-\psi_{nj}\log\tilde{\mathbf{H}}_{nj} + \sum_{nj}\tilde{\mathbf{H}}_{nj}\right) + C
\end{aligned}
$$

where irrelevant terms not related to $\tilde{\mathbf{H}}$ have been included in the constant C. The derivative is then given by

$$
\begin{aligned}
\frac{\partial\cdot}{\partial\tilde{\mathbf{H}}_{nj}} &= -\frac{\sum_i\mathbf{X}_{ij}\phi_{inj}}{\tilde{\mathbf{H}}_{nj}} + \sum_{im}\mathbf{\Lambda}_{im}\mathbf{W}_{mn} - \frac{\lambda\psi_{nj}}{\tilde{\mathbf{H}}_{nj}} + \lambda \\
&= -\frac{\mathbf{H}_{nj}}{\tilde{\mathbf{H}}_{nj}}\left(\sum_i\frac{\mathbf{X}_{ij}\sum_m\mathbf{\Lambda}_{im}\mathbf{W}_{mn}}{\sum_{mn}\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}} + \frac{\lambda}{\sum_n\mathbf{H}_{nj}}\right) + \left(\sum_{im}\mathbf{\Lambda}_{im}\mathbf{W}_{mn} + \lambda\right)
\end{aligned}
$$

Equating the derivative to zero provides the update rule.

Given $\phi_{imnj} = \frac{\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}$, and $\psi_{mn} = \frac{\mathbf{W}_{mn}}{\sum_m\mathbf{W}_{mn}}$, we can construct the auxiliary function for $\mathbf{W}$ as,

$$
\begin{aligned}
&\sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_{mn}\mathbf{\Lambda}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj} + \sum_{mn}\mathbf{\Lambda}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}\right] \\
&\quad + \lambda\sum_n\left(-\log\sum_m\tilde{\mathbf{W}}_{mn} + \sum_m\tilde{\mathbf{W}}_{mn}\right) \\
&\quad = \sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_{mn}\frac{\phi_{imnj}}{\phi_{imnj}}\mathbf{\Lambda}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj} + \sum_{mn}\mathbf{\Lambda}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\lambda\sum_n\left(-\log\sum_m\frac{\psi_{mn}}{\psi_{mn}}\tilde{\mathbf{W}}_{mn}+\sum_m\tilde{\mathbf{W}}_{mn}\right)\\
&\leq\sum_{ij}\left[-\mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\frac{\boldsymbol{\Lambda}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}}{\phi_{imnj}}+\sum_{mn}\boldsymbol{\Lambda}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}\right]\\
&+\lambda\left(\sum_{mn}-\psi_{mn}\log\frac{\tilde{\mathbf{W}}_{mn}}{\psi_{mn}}+\sum_{mn}\tilde{\mathbf{W}}_{mn}\right)\\
&=\sum_{ij}\left[-\mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\tilde{\mathbf{W}}_{mn}+\sum_{mn}\boldsymbol{\Lambda}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}\right]\\
&+\lambda\left(\sum_{mn}-\psi_{mn}\log\tilde{\mathbf{W}}_{mn}+\sum_{mn}\tilde{\mathbf{W}}_{mn}\right)+C
\end{aligned}
$$

where irrelevant terms not related to $\tilde{\mathbf{W}}$ have been included in the constant C.

The derivative is then given by

$$
\begin{aligned}
\frac{\partial\cdot}{\partial\tilde{\mathbf{W}}_{mn}}&=-\frac{\sum_{ij}\mathbf{X}_{ij}\phi_{imnj}}{\tilde{\mathbf{W}}_{mn}}+\sum_{ij}\boldsymbol{\Lambda}_{im}\mathbf{H}_{nj}-\frac{\lambda\psi_{mn}}{\tilde{\mathbf{W}}_{mn}}+\lambda\\
&=-\frac{\mathbf{W}_{mn}}{\tilde{\mathbf{W}}_{mn}}\left(\sum_{ij}\frac{\mathbf{X}_{ij}\boldsymbol{\Lambda}_{im}\mathbf{H}_{nj}}{\sum_{mn}\boldsymbol{\Lambda}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}+\frac{\lambda}{\sum_m\mathbf{W}_{mn}}\right)+\left(\sum_{ij}\boldsymbol{\Lambda}_{im}\mathbf{H}_{nj}+\lambda\right)
\end{aligned}
$$

Equating the derivative to zero provides the update rule.

Appendix 4: Update rule for Bernoulli observations

Since the cost function consists of two similar terms, we show how to establish the auxiliary function for one of them.

For $\mathbf{H}$ we have, $\phi_{inj}=\frac{\sum_m\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}$, and $\sum_n\phi_{inj}=1$, then

$$
\begin{aligned}
&\sum_{ij}\left[-\mathbf{X}_{ij}\log(\mathbf{PW}\tilde{\mathbf{H}})_{ij}\right]\\
&=\sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}\right]\\
&=\sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_n\frac{\phi_{inj}}{\phi_{inj}}\sum_m\mathbf{P}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}\right]\\
&\leq\sum_{ij}\left[-\mathbf{X}_{ij}\sum_n\phi_{inj}\log\frac{\sum_m\mathbf{P}_{im}\mathbf{W}_{mn}\tilde{\mathbf{H}}_{nj}}{\phi_{inj}}\right]\\
&=\sum_{ij}\left[-\mathbf{X}_{ij}\sum_n\phi_{inj}\log\frac{\tilde{\mathbf{H}}_{nj}}{\mathbf{H}_{nj}}-\mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{ij}\left[-\mathbf{X}_{ij}\sum_n \phi_{inj}\log\left(\frac{\tilde{\mathbf{G}}_{nj}}{\sum_p \tilde{\mathbf{G}}_{pj}}\frac{\sum_p \mathbf{G}_{pj}}{\mathbf{G}_{nj}}\right) - \mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]\\
&= \sum_{ij}\left[-\mathbf{X}_{ij}\sum_n \phi_{inj}\log\frac{\tilde{\mathbf{G}}_{nj}}{\mathbf{G}_{nj}} + \mathbf{X}_{ij}\sum_n \phi_{inj}\log\frac{\sum_p \tilde{\mathbf{G}}_{pj}}{\sum_p \mathbf{G}_{pj}} - \mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]\\
&\le \sum_{ij}\left[-\mathbf{X}_{ij}\sum_n \phi_{inj}\log\frac{\tilde{\mathbf{G}}_{nj}}{\mathbf{G}_{nj}} + \mathbf{X}_{ij}\left(\frac{\sum_p \tilde{\mathbf{G}}_{pj}}{\sum_p \mathbf{G}_{pj}} - 1\right) - \mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]
\end{aligned}
$$

Taking derivative we get,

$$
\begin{aligned}
&\sum_i\left[-\mathbf{X}_{ij}\phi_{inj}\frac{1}{\tilde{\mathbf{G}}_{nj}} + \mathbf{X}_{ij}\frac{1}{\sum_p \mathbf{G}_{pj}}\right]\\
&\Rightarrow \tilde{\mathbf{G}}_{nj} = \frac{\sum_p \mathbf{G}_{pj}}{\sum_i \mathbf{X}_{ij}}\left(\sum_i \frac{\mathbf{X}_{ij}\sum_m \mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}\right) = \frac{\mathbf{G}_{nj}}{\sum_i \mathbf{X}_{ij}}\left(\sum_i \frac{\mathbf{X}_{ij}\sum_m \mathbf{P}_{im}\mathbf{W}_{mn}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}\right)
\end{aligned}
$$

For $\mathbf{W}$ we have, $\phi_{imnj} = \frac{\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}$, and $\sum_{mn}\phi_{imnj} = 1$, then ter

$$
\begin{aligned}
&\sum_{ij}\left[-\mathbf{X}_{ij}\log(\mathbf{P}\tilde{\mathbf{W}}\mathbf{H})_{ij}\right]\\
&= \sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}\right]\\
&= \sum_{ij}\left[-\mathbf{X}_{ij}\log\sum_{mn}\frac{\phi_{imnj}}{\phi_{imnj}}\mathbf{P}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}\right]\\
&\le \sum_{ij}\left[-\mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\frac{\mathbf{P}_{im}\tilde{\mathbf{W}}_{mn}\mathbf{H}_{nj}}{\phi_{imnj}}\right]\\
&= \sum_{ij}\left[-\mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\frac{\tilde{\mathbf{W}}_{mn}}{\mathbf{W}_{mn}} - \mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]\\
&= \sum_{ij}\left[-\mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\left(\frac{\tilde{\mathbf{V}}_{mn}}{\sum_p \tilde{\mathbf{V}}_{pn}}\frac{\sum_p \mathbf{V}_{pn}}{\mathbf{V}_{mn}}\right) - \mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]\\
&= \sum_{ij}\left[-\mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\frac{\tilde{\mathbf{V}}_{mn}}{\mathbf{V}_{mn}} + \mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\frac{\sum_p \tilde{\mathbf{V}}_{pn}}{\sum_p \mathbf{V}_{pn}} - \mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]\\
&\le \sum_{ij}\left[-\mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\log\frac{\tilde{\mathbf{V}}_{mn}}{\mathbf{V}_{mn}} + \mathbf{X}_{ij}\sum_{mn}\phi_{imnj}\left(\frac{\sum_p \tilde{\mathbf{V}}_{pn}}{\sum_p \mathbf{V}_{pn}} - 1\right) - \mathbf{X}_{ij}\log\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}\right]
\end{aligned}
$$

Taking derivative we get,

$$\sum_{ij}\left[-\mathbf{X}_{ij}\phi_{imnj}\frac{1}{\tilde{\mathbf{V}}_{mn}}+\mathbf{X}_{ij}\sum_{m}\phi_{imnj}\frac{1}{\sum_{p}\mathbf{V}_{pn}}\right]$$

$$\Rightarrow \tilde{\mathbf{V}}_{mn}=\frac{\sum_{p}\mathbf{V}_{pn}}{\sum_{ij}\mathbf{X}_{ij}\sum_{m}\phi_{imnj}}\left(\sum_{ij}\frac{\mathbf{X}_{ij}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}\right)$$

$$=\frac{\mathbf{V}_{mn}}{\sum_{ij}\frac{\mathbf{X}_{ij}\sum_{m}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}}\left(\sum_{ij}\frac{\mathbf{X}_{ij}\mathbf{P}_{im}\mathbf{H}_{nj}}{\sum_{mn}\mathbf{P}_{im}\mathbf{W}_{mn}\mathbf{H}_{nj}}\right)$$

The update rule can be derived from these equations after including the other term with $\mathbf{Q}$.

References

Bache, K., & Lichman, M. (2013). UCI machine learning repository. http://archive.ics.uci.edu/ml.

Bauckhage, C., & Thurau, C. (2009). Making archetypal analysis practical. In *Pattern recognition, lecture notes in computer science, vol. 5748*, Springer, Berlin Heidelberg, pp. 272–281. doi:10.1007/978-3-642-03798-6_28.

Bhattacharya, A., & Dunson, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *Journal of the American Statistical Association*, *107*(497), 362–377. doi:10.1080/01621459.2011.646934.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Chan, B. H. P., Mitchell, D. A., & Cram, L. E. (2003). Archetypal analysis of galaxy spectra. *Monthly Notices of the Royal Astronomical Society*, *338*(3), 790–795. doi:10.1046/j.1365-8711.2003.06099.x.

Cutler, A., & Stone, E. (1997). Moving archetypes. *Physica D: Nonlinear Phenomena, 107*(1), 1–16. doi:10.1016/S0167-2789(97)84209-1, http://www.sciencedirect.com/science/article/pii/S0167278997842091.

Cutler, A., & Breiman, L. (1994). Archetypal analysis. *Technometrics*, *36*(4), 338–347.

Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, *21*(2), 234–242. doi:10.1177/0956797609357712.

Ding, C., Li, T., Peng, W., & Park, H. (2006). Orthogonal nonnegative matrix tri-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 126–135. doi:10.1145/1150402.1150420.

Ding, C. H. Q., Li, T., & Jordan, M. I. (2010). Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 45–55.

Dolnicar, S., & Leisc, F. (2004). Segmenting markets by bagged clustering. *Australasian Marketing Journal*, *12*(1), 51–65.

do Nascimento, J. M. P., & Dias, J. M. B. (2005). Vertex component analysis: A fast algorithm to unmix hyperspectral data. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(4), 898–910. doi:10.1109/TGRS.2005.844293.

Dolnicar, S., Grün, B., & Leisch, F. (2011). Quick, simple and reliable: Forced binary survey questions. *International Journal of Market Research*, *53*(2), 231–252. doi:10.2501/IJMR-53-2-231-252.

EM-DAT (2013). The OFDA/CRED international disaster database. Universite catholique de Louvain, Brussels, Belgium; http://www.emdat.net.

Eugster, M. J. A., & Leisch, F. (2013). archetypes: Archetypal analysis. http://CRAN.R-project.org/package=archetypes, R package version 2.1-2.

Eugster, M. J. A., & Leisch, F. (2011). Weighted and robust archetypal analysis. *Computational Statistics and Data Analysis*, *55*(3), 1215–1225. doi:10.1016/j.csda.2010.10.017.

Eugster, M. J. A. (2012). Performance profiles based on archetypal athletes. *International Journal of Performance Analysis in Sport*, *12*(1), 166–187.

Févotte, C., & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, *23*(9), 2421–2456.

Friendly, M. (2000). *Visualizing categorical data*. Cary, NC: SAS Institute.

Hahsler, M., & Hornik, K. (2007). TSP—infrastructure for the traveling salesperson problem. *Journal of Statistical Software*, *23*(2), 1–21. http://www.jstatsoft.org/v23/i02/.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*. ACM, New York, NY, USA, SIGIR '99, pp. 50–57. doi:10.1145/312624.312649

Lee, D. D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems, vol. 13*, pp 556–562.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. doi:10.1038/44565.

Li, S., Louviere, J., Carson, R., & Wang, P. (2003). Archetypal analysis: A new way to segment markets based on extreme individuals. In *A celebration of ehrenberg and bass: Marketing knowledge, discoveries and contribution. Proceedings of the ANZMAC 2003 conference*. http://epress.lib.uts.edu.au/research/handle/10453/2183.

Marinetti, S., Finesso, L., & Marsilio, E. (2007). Archetypes and principal components of an IR image sequence. *Infrared Physics & Technology*, *49*(3), 272–276. doi:10.1016/j.infrared.2006.06.017, http://www.sciencedirect.com/science/article/pii/S1350449506000910.

Mohamed, S., Heller, K. A., & Ghahramani, Z. (2009). Bayesian exponential family PCA. In *Advances in Neural Information Processing Systems, vol. 21*, pp 1089–1096.

Mørup, M., & Hansen, L. K. (2012). Archetypal analysis for machine learning and data mining. *Neurocomputing*, *80*, 54–63. doi:10.1016/j.neucom.2011.06.033.

Porzio, G. C., Ragozini, G., Vistocco, D. (2008). On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry*, *24*(5), 419–437. doi:10.1002/asmb.727, http://onlinelibrary.wiley.com/doi/10.1002/asmb.727/abstract.

Seiler, C., & Wohlrabe, K. (2013). Archetypal scientists. *Journal of Informetrics*, *7*(2), 345–356. doi:10.1016/j.joi.2012.11.013.

Sifa, R., & Bauckhage, C. (2013). Archetypical motion: Supervised game behavior learning with archetypal analysis. In: *2013 IEEE conference on computational intelligence in games (CIG)*, pp. 1–8. doi:10.1109/CIG.2013.6633609.

Steinley, D. (2006). K-means clustering: A half-century synthesis. *British Journal of Mathematical and Statistical Psychology*, *59*(1), 1–34. doi:10.1348/000711005X48266.

Stone, E., & Cutler, A. (1996). Archetypal analysis of spatio-temporal dynamics. *Physica D: Nonlinear Phenomena*, *90*(3), 209–224. doi:10.1016/0167-2789(95)00244-8.

Thøgersen, J. C., Mørup, M., Damkiær, S., Molin, S., & Jelsbak, L. (2013). Archetypal analysis of diverse pseudomonas aeruginosa transcriptomes reveals adaptation in cystic fibrosis airways. *BMC Bioinformatics*, *14*(1), 279. doi:10.1186/1471-2105-14-279, http://www.biomedcentral.com/1471-2105/14/279/abstract.

Thurau, C., Kersting, K., & Bauckhage, C. (2009). Convex non-negative matrix factorization in the wild. In *Ninth IEEE international conference on data mining, 2009. ICDM '09*, pp. 523–532. doi:10.1109/ICDM.2009.55.

Thurau, C., Kersting, K., & Bauckhage, C. (2010). Yes we can: Simplex volume maximization for descriptive web-scale matrix factorization. In: *Proceedings of the 19th ACM international conference on information and knowledge management*, ACM, New York, NY, USA, CIKM '10, pp. 1785–1788. doi:10.1145/1871437.1871729.

Tibshirani, R., & Walther, G. (2005). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*, 511–528.

Woodbury, M. A., & Clive, J. (1974). Clinical pure types as a fuzzy partition. *Journal of Cybernetics*, *4*(3), 111–121. doi:10.1080/01969727408621685.

Xiong, Y., Liu, W., Zhao, D., & Tang, X. (2013). Face recognition via archetype hull ranking. In *2013 IEEE international conference on computer vision (ICCV)*, pp. 585–592. doi:10.1109/ICCV.2013.78.

Yang, Z., & Oja, E. (2012). Clustering by low-rank doubly stochastic matrix decomposition.arXiv:12064676 http://arxiv.org/abs/1206.4676.