Journal of Statistical Theory and Practice (2022) 16:14
https://doi.org/10.1007/s42519-021-00238-4

ORIGINAL ARTICLE

A Literature Review of (Sparse) Exponential Family PCA

Luke Smallman[1] · **Andreas Artemiou**[1]

Accepted: 24 December 2021

Abstract
This is a brief overview of the methodology around exponential family PCA. We revisit classic PCA methodology, and we focus on exponential family PCA due to its applicability on a number of distributions and hence a wide variety of problems. We discuss the applicability of these methods to text data analysis due to the high-dimensional and sparse nature of these data.

Keywords Dimension reduction · Exponential family · Poisson distribution · Gaussian distribution · Text data analysis

1 Introduction

In the era of high-performance computing, researchers across several fields are able to collect massive amount of high-dimensional data. The most challenging part is probably to efficiently analyze and make inference on these high-dimensional datasets. Unfortunately, most statistical methods were developed with small and low-dimensional datasets in mind. Thus, it is rather difficult to extrapolate their use to massive or high-dimensional data, due to the enormous amount of computational time needed for most of these algorithms if the dataset has large number of observations, as well as the singularities that appear in many methods in large p small n problems. To handle the issue of high-dimensional data, researchers have proposed the use of dimension reduction techniques either by means of feature selection or feature extraction. By reducing the dimension of the data, they can bring them to an appropriate dimension and size which can be analyzed efficiently and accurately using classic statistical techniques. Most of these dimension reduction techniques were developed under the assumption that our data follow Gaussian distribution and a lot of work has been done in this framework.

✉ Andreas Artemiou
ArtemiouA@cardiff.ac.uk

[1] School of Mathematics, Cardiff University, Cardiff, UK

Published online: 02 February 2022

One of the research areas that have emerged the last few years, with the explosion of social media as well as due to the digitalization of a lot of sources which record documents, is text data analysis. For example, many websites are based on reviews of customers. If one is interested for a product, a restaurant, a plumber, a hotel or a travel destination, there are dedicated websites that contain the necessary reviews from previous customers who used that product/service/agency, for them to consider before making final decisions. Moreover, text data analytics can be used to analyze documents to identify whether they were written from the same author, whether they discuss a specific topic etc.

The difficulty with text data analysis arises by the fact they are not numeric data. A lot of research has been devoted into quantifying text data. One of the most frequently used methods is to enumerate the number of times a word appears in a document, creating an $n \times p$ matrix where n is the number of observations (might be whole documents, may be paragraphs or even just sentences) being analyzed and p the number of unique words in all n documents. This matrix is known as the *term matrix*. The term matrix is used as the numerical representation of a collection of documents. This numerical representation is mostly filled with small counts, and so it makes sense for someone to model each word/variable using a Poisson distribution. This is obviously very high-dimensional as there is a column for each word even if it is used in only one of the documents in our dataset. This is the reason why the term matrix is usually sparse as there is a limited number of words that appear in each document which implies that most of the words will appear only in some documents/observations and therefore the term matrix will have a lot of 0's.

With the text data analysis in mind and the format of the term matrix, in this paper we focus on exponential family distribution algorithms for principal component analysis. Given the high-dimensional nature of the problems we encounter in text data analysis, it is also sensible to discuss some sparse techniques for dimension reduction in this setting. The rest of the paper is organized as follows: In Sect. 2, we discuss the classic PCA and other important extensions, and in Sect. 3, we discuss other PCA algorithms in the Gaussian setting. In Sect. 4, we discuss exponential family PCA methods with a focus on the Poisson distribution, and in Sect. 5, we discuss sparse extensions. Finally, we close with a discussion section.

2 Principal Component Analysis (PCA)

In this section, we introduce the classic PCA algorithm and then we discuss sparse extensions of it. We use also this section to define some notation we use throughout the paper.

2.1 History of PCA

Principal component analysis (PCA) is probably one of the most well-known techniques for dimension reduction. The idea is to find orthogonal projection of the data which maximize the observed variation. Although they were mathematically formu-

lated in [25], they were also discussed in [39]. Most researchers agree that these two papers are the first to formulate PCA as we know them today, but an interesting historical overview in [10] claims that principal components were discussed by researchers back in the nineteenth century.

If we assume that we have a collection of n p-dimensional vectors $\boldsymbol{x}_1, \ldots \boldsymbol{x}_n$, with mean $\bar{\boldsymbol{x}}$ and variance matrix $\boldsymbol{S} = (1/N)\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathsf{T}}$, then the orthogonal projections that maximize the variance on the projected space are the q eigenvectors associated with the largest q eigenvalues of $\boldsymbol{S}$. There are many data-driven methods to decide the value of q. One of the most popular choices is to select the first q components which explain cumulatively at least 80% of the variation. Another common choice is the use of the scree plot, where we plot the eigenvalues and try to identify the point the eigenvalues flatten. (See [28] for a detailed approach to this). We denote the eigenvectors with $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q$, and the principal axes are calculated using $\boldsymbol{z}_i = \boldsymbol{W}^{\mathsf{T}}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})$ where $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_q)$ a $p \times q$ matrix and $\boldsymbol{z}_i \in \mathbb{R}^q$ for $i = 1, \ldots, n$. It was [39] who identified that the principal component projection minimizes the squared loss

$$\sum_{i=1}^{n} \|\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i\|^2 = \sum_{i=1}^{n} \|\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i - \bar{\boldsymbol{x}})\|^2. \tag{1}$$

This formulation of PCA was used also by researchers later in different extensions. See, for example, [30].

In [25] the principal component terminology is introduces and principal component is presented as a variance maximization algorithm. The topic has since expanded in a number of directions making it one of the most well-known feature extraction methods to practitioners.

2.2 Sparse PCA

With the appearance of high-dimensional datasets, there was a bigger need for sparse feature extraction and a number of ideas have appeared in the literature. For example, [53] proposed the use of the L1 (LASSO which stands for Least Absolute Shrinkage and Selection Operator—[47]) penalty as well as the elastic net penalty (combination of LASSO and SCAD which stands for Smoothly Clipped Absolute Deviation—see [52]). Many other methods were also proposed to improve the computational complexity of the algorithm. For example, [11] used semidefinite programming to find sparse PCA and [51] proposed nonnegative sparse PCA, where sparsity is introduced by the use of nonnegative coefficients of the original variables. Finally, in [5] the authors proposed an algorithm to reduce the sparse PCA into a high-dimensional multivariate regression problem. These algorithms enabled the use of PCA to achieve dimension reduction in high-dimensional settings, and it also allowed for the expansion of the scope of PCA (and other dimension reduction techniques) in a number of other areas where high-dimensional data are frequently collected.

2.3 Kernel PCA

Kernel PCA [43], extracted nonlinear features of the predictors by mapping the p-dimensional vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ to a higher-dimensional feature space, using a mapping $\phi : \mathbb{R}^p \to \mathbb{R}^m$ where m is the dimension of the feature space and $m > p$. Since in most cases ϕ is an unknown function, this is achieved by utilizing the "kernel" trick, a well-known procedure in the nonlinear statistical setting where the inner product between two realizations of the function ϕ can be replaced by the kernel matrix, i.e., $\boldsymbol{K}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$. In kernel PCA, the authors assumed the features $\phi(\boldsymbol{x}_j)$, $j = 1, \ldots, p$, to be centered around 0 and proposed the eigenvalue decomposition of the covariance matrix of the mapped features: $\hat{\boldsymbol{C}} = \frac{1}{n} \sum_{i=1}^{n} \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}_j)^{\mathsf{T}}$. This was shown to be equivalent to finding the solution to the following eigenvalue problem: $n\lambda\boldsymbol{\alpha} = \boldsymbol{K}\boldsymbol{\alpha}$ where λ is one of the nonzero eigenvalues and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathsf{T}}$. The coefficients in $\boldsymbol{\alpha}$ are then used to construct the eigenvectors $\boldsymbol{V} = \sum_{i=1}^{n} \alpha_i \phi(\boldsymbol{x}_i)$ and the principal directions $\langle \boldsymbol{V}, \phi(\boldsymbol{x}) \rangle = \sum_{i=1}^{n} \alpha_i \phi(\boldsymbol{x}_i)\phi(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i K(\boldsymbol{x}_i, \boldsymbol{x})$.

2.4 Spherical PCA

In [34], the authors recognized that using projections based on distance-based measurement might not be the best option in some application, like information retrieval and signal processing. In those cases, it is considered more appropriate to use similarity-based measurements such as angle distance. To address this, they introduce some constraints in the optimization problem as follows:

$$\min_{\boldsymbol{U},\boldsymbol{V}} \|\boldsymbol{X} - \boldsymbol{V}\boldsymbol{U}^{\mathsf{T}}\|_F^2 = \sum_{i,j} (\boldsymbol{X}_{ij} - (\boldsymbol{V}\boldsymbol{U}^{\mathsf{T}})_{ij})^2$$

under the constraints $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}$ and $\|\boldsymbol{v}_j\| = 1$ for every $j = 1, \ldots, q$ where $\boldsymbol{X}$ is an $n \times p$ where each row is an observation $\boldsymbol{x}_i$, $\boldsymbol{U}$ is a $p \times q$ matrix of the principal directions on each column, and $\boldsymbol{V}$ is an $n \times q$ matrix giving the scores of the principal directions. Also, note that $\|\cdot\|$ denotes the $l2$ norm for vectors and the spectral norm for matrices. The optimization here is nonconvex, and the major contribution of [34] is the use of the proximal approach to approximate a solution.

3 Likelihood-Based PCA in the Gaussian Setting

In recent years, there was an abundance of papers which discussed PCA from a different angle than the squared loss discussed in [39]. These papers derived likelihood-based PCA algorithms. In this section, we will discuss some of these methods.

3.1 Probabilistic PCA

In [48], we find an alternative approach to PCA. Although the authors did not discuss exponential family PCA, their work is considered the first toward this direction, as

they suggested the use of maximum likelihood estimator to estimate the principal axes in a PCA setting. Their work is focused on the Gaussian distribution, but it can be extended to other distributions. (It was extended in the literature by [9].) The method is called probabilistic PCA (or PPCA) to emphasize the probabilistic nature of the algorithm.

In [48], it is assumed that the latent variables $\boldsymbol{z}_i \sim N_q(\boldsymbol{0}, \boldsymbol{I})$. Then, we have the conditional distribution $\boldsymbol{x}|\boldsymbol{z} \sim N_p(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{\mu}, \sigma^2\boldsymbol{I})$. By marginalization, we have that $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} + \sigma^2\boldsymbol{I})$. Probabilistic PCA essentially estimates $\boldsymbol{W}$ and $\boldsymbol{\mu}$ in the model using an EM algorithm. Classical PCA algorithms solve the case when $\sigma^2 \to 0$.

To derive the MLEs of $\boldsymbol{W}$ and $\boldsymbol{\mu}$, we use the following log-likelihood function:

$$-\frac{n}{2}\{p\ln(2\pi) + \ln \boldsymbol{C} + \mathrm{tr}(\boldsymbol{C}^{-1}\boldsymbol{S})\}$$

where $\boldsymbol{C} = \boldsymbol{W}\boldsymbol{W}^{\mathsf{T}} + \sigma^2\boldsymbol{I}$ and $\boldsymbol{S}$ the sample covariance matrix. This will give the maximum likelihood estimators as follows:

- $\hat{\boldsymbol{W}}_{\mathrm{ML}} = \boldsymbol{U}(\boldsymbol{\Lambda} - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R}$, where $\boldsymbol{\Lambda}$ is the $q \times q$ diagonal matrix with the eigenvalues $\lambda_1 > \ldots > \lambda_q$ on the diagonal, $\boldsymbol{U}$ is a $p \times q$ matrix that has as columns the eigenvectors of $\boldsymbol{S}$ corresponding to the largest q eigenvalues, and $\boldsymbol{R}$ is an arbitrary $q \times q$ orthogonal rotation matrix, and
- $\hat{\sigma}_{\mathrm{ML}} = \frac{1}{p-q}\sum_{i=q+1}^{p}\lambda_i$ which can be interpreted as the lost variance in the directions being dropped.

3.2 Bayesian PCA

Bayesian PCA (BPCA) was proposed by [3] as an extension to the Probabilistic PCA idea. Due to the inability of the PPCA to effectively estimate the dimension of the latent space (that is, the reduced dimensional space) and the intractability of the problem in cases where we allowed mixtures of probabilistic PCA with different latent space dimensions, the Bayesian paradigm was used by the authors. This idea addressed both problems as it allowed for direct estimation of the latent space dimension and made the problem tractable in more complex cases. To achieve this, the author proposed the use of a prior distribution on the parameters of the data, denoted by $p(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2)$ and a posterior of the form $p(\boldsymbol{\mu}, \boldsymbol{W}, \sigma^2|\boldsymbol{x})$. Motivated by the work on automatic relevance determination (ARD) by [37], they also use the idea of a normal conditional prior on $\boldsymbol{W}$, that is, $p(\boldsymbol{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{p-1}(\alpha_i/(2\pi))^{p/2}e^{-(1/2)\alpha_i\|\boldsymbol{w}_i\|}$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ are the variances of each direction and $\boldsymbol{w}_i$ is the i^{th} column of $\boldsymbol{W}$. It is this relationship with ARD that allows BPCA to determine the latent space dimensionality by using the number of nonzero elements of the $\boldsymbol{\alpha}$. The authors proposed the use of a simplified version of the algorithm used by [37] which utilizes the EM algorithm to estimate all the parameters.

4 Exponential Family PCA

In this section, we will discuss some of the most well-known exponential family PCA algorithms. The notation is kept similar to the previous section unless otherwise noted. Before giving more details, we give the definition of an exponential family distribution.

Definition 1 In an exponential family, the distribution of the data $\boldsymbol{x}$ given a parameter vector $\boldsymbol{\theta}$ takes the form:

$$\log P(\boldsymbol{x}|\boldsymbol{\theta}) = \log P_0(\boldsymbol{x}) + \boldsymbol{x}\boldsymbol{\theta} - G(\boldsymbol{\theta})$$

where $\boldsymbol{\theta}$ is called the vector of natural parameters of the distribution and $G(\boldsymbol{\theta})$ is a function to ensure that the distribution function integrates to 1.

It is important to note that we first discuss extensions to the exponential family of the methods discussed in the previous section in the Gaussian setting. Those are known as the likelihood-based methods. These extensions allow for the generalization of the Gaussian assumption to any distribution that can be written as an exponential family distribution. Since exponential family PCA is the primary focus of this literature review, we discuss more methods in this section, including semiparametric exponential family PCA and supervised exponential family PCA among others.

4.1 Likelihood-Based Exponential Family PCA

As was mentioned earlier, the first set of methods are primarily extensions of the methodology introduced in earlier sections on the Probabilistic PCA framework for Gaussian data. The main idea behind this methodology is to extend it to allow for data from different distributions and more specifically from distributions that belong to the exponential family.

If we denote $\mathcal{X}$ the domain of $\boldsymbol{x}$, one can show that $G(\boldsymbol{\theta}) = \log \sum_{\boldsymbol{x}\in\mathcal{X}} P_0(\boldsymbol{x}) e^{\boldsymbol{x}\boldsymbol{\theta}}$ and it is the form of this function that changes between different exponential family distribution functions. (Note that if we have a continuous exponential family distribution, the sum can be replaced with an integral.) For example, for the one-dimensional Gaussian distribution with mean θ and variance 1, $G(\theta) = \theta^2/2$, while for the Bernoulli distribution $G(\theta) = \log(1+e^{\theta})$ for $\theta = \log(p/(1-p))$. Also, note that $P_0(\boldsymbol{x})$ depends only on $\boldsymbol{x}$ and can be treated as constant when optimal values are being sought. Also, note that we denote with $g(\boldsymbol{\theta})$ the derivative of $G(\boldsymbol{\theta})$.

4.1.1 Exponential PCA

In [9], the authors noticed that the work by [48] can be further extended to accommodate any distribution in the exponential family. Actually, their proposal on using exponential family allowed for different attributes in the data to be modeled using different exponential family distribution. To make things clear of what [9] suggest, we give the definition of Bregman distance. Let $F : A \to R$ be a differentiable and

strictly convex function defined on a closed and convex set $B \subseteq R$. Then, for $a, b \in B$ one can define the Bregman distance associated with F as follows:

$$B_F(a\|b) \doteq F(a) - F(b) - f(b)(a - b)$$

where $f(x)$ is the derivative of $F(x)$. One can show that the negative log likelihood can be written as a function of Bregman distances:

$$-\log p(x|\boldsymbol{\theta}) = -\log p(x_0) - F(x) + B_F(x\|g(\boldsymbol{\theta}))$$

Having defined some important details, we can now discuss the idea of [9] on exponential family PCA. Let $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_q$ be a basis in $\mathbb{R}^p$, and therefore, each $\boldsymbol{\theta}_i$ can be represented as a linear combination of the $\boldsymbol{v}_k$'s ($k = 1, \ldots, q$), that is, $\boldsymbol{\theta}_i = \sum_k a_{ik}\boldsymbol{v}_k$. Define now the matrix $\boldsymbol{X}$ to be the $n \times p$ matrix with $\boldsymbol{x}_i$ on the i^{th} row, $\boldsymbol{V}$ the $q \times p$ matrix with $\boldsymbol{v}_i$ on the i^{th} row and $\boldsymbol{A}$ an $n \times q$ matrix with entries a_{ik}. Then, $\boldsymbol{\Theta} = \boldsymbol{AV}$ is the matrix with $\boldsymbol{\theta}_i$ on the i^{th} row.

From the negative log likelihood, we have the loss function:

$$L(\boldsymbol{V}, \boldsymbol{A}) = -\log P(\boldsymbol{X}|\boldsymbol{A}, \boldsymbol{V}) = C + \sum_i \sum_j (-x_{ij}\theta_{ij} + G(\theta_{ij}))$$

where C is a constant we will ignore and $G(\theta_{ij})$ can take different forms, i.e., $\log(1 + e^{\theta})$ if we have the Bernoulli distribution. Using the relationship between the negative log likelihood and Bregman distance, one can show that

$$L(\boldsymbol{V}, \boldsymbol{A}) = \sum_i B_F(\boldsymbol{x}_i\|g(\boldsymbol{\theta}_i))$$

Therefore, the generalized PCA can be seen as a search for low-dimensional basis from matrix $\boldsymbol{V}$, which defines the surface that is close to all the data points. Thus, the optimal value of $\boldsymbol{V}$ is given by $\arg\min_{\boldsymbol{V}} \sum_i \min_{\boldsymbol{q}} B_F(\boldsymbol{x}_i\|\boldsymbol{q})$ where $\boldsymbol{q}$ is a member of the set $\{g(\boldsymbol{aV})|\boldsymbol{a} \in \mathbb{R}^q\}$.

4.1.2 Bayesian Exponential PCA

In [38], the authors identified the limitation of BPCA ([3]) being defined only for the Gaussian distribution, and they proposed the extension of the above algorithm to the case exponential family of distributions. Their approach uses the same idea as the work by [9] where we use the product $\boldsymbol{\Theta} = \boldsymbol{AV}$ to represent the natural parameters of the distribution over the data. Instead of using the maximization of Bregman distances though, the authors suggest a completely probabilistic approach where a prior distribution is used on all parameters. This prior distribution can be from any exponential family distribution rather than just the Gaussian distribution, and all of these parameters can be integrated out of the model using Markov chain Monte Carlo (MCMC) methods. Also, the authors claim that their method is not prone to data overfitting as the method proposal by [9].

4.1.3 Simple Exponential PCA

Simple exponential PCA was proposed by [32], and the main idea is that their proposal is a much simpler algorithm than the one proposed by [38]. The main benefit of their proposal is that it uses the Bayesian formulation to automatically estimate the dimension of the latent space. In a few words and as [32] put it simple exponential PCA automatically determines q given observations $\boldsymbol{X}$ by using automatic relevance determination (ARD—[37]) in a similar way [3] does it for Bayesian PCA. Therefore, it finds q vectors of loadings and constructs matrix $\boldsymbol{W}$ and q principal component score vectors and construct matrix $\boldsymbol{Y}$. Using this, it creates $\boldsymbol{\Theta} = \boldsymbol{WY}$ and each column of $\boldsymbol{\Theta}$ specifies the natural parameters of an exponential family distribution used to generate the corresponding column of $\boldsymbol{X}$. More specifically, the algorithm is as follows:

- Draw q scores as $\boldsymbol{y}_i$ which is the lower-dimensional representation of $\boldsymbol{x}_i$ from a standard Gaussian prior $\boldsymbol{y}_n \sim N_p(\boldsymbol{0}, \boldsymbol{I})$.
- Using a vector of precision parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)$ draw $\boldsymbol{w}_j \sim N_p(\boldsymbol{0}, a_j^{-1}\boldsymbol{I})$ where $\boldsymbol{w}_j$ denotes the j^{th} principal component. One can define then $\boldsymbol{W}$ to be the matrix with $\boldsymbol{w}_j$ as it's j^{th} column.
- Using $\boldsymbol{\theta}_n = \boldsymbol{W}\boldsymbol{y}_n$, one can get the distribution of $\boldsymbol{x}_n|\boldsymbol{\theta}_n$ to be from any exponential family distribution with natural parameter $\boldsymbol{\theta}_n$.

This is a relatively simple process and hence the name of the algorithm. To accommodate for any exponential distribution to be used, the authors proposed an alternating approach for inference of $\boldsymbol{W}$ and $\boldsymbol{Y}$. They use an EM approach for inference of $\boldsymbol{W}$ and a maximum a posteriori (MAP) approach for inference on $\boldsymbol{Y}$.

4.1.4 Generalized PCA

[30] used a similar approach to the one identified by [39] and which we explained in Sect. 2 [see Eq. (1)]. They tried to find the optimal projection matrix $\boldsymbol{U}$ which minimizes the objective function:

$$\sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{\mu}_i - \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\|^2 \tag{2}$$

where $\boldsymbol{\mu}_i \in \mathbb{R}^P$. Using the above formulation, the authors identified that the approximation of $\boldsymbol{x}_i$ by $\boldsymbol{\mu}_i - \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{\mu}_i)$ is equivalent to looking for a deviance optimal approximation to the saturated model parameters $\tilde{\boldsymbol{\theta}}_i$ with $\boldsymbol{\mu}_i - \boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu}_i)$. Then, one uses the following form to calculate deviance:

$$D(\boldsymbol{U}, \boldsymbol{\mu}) = \sum_{i=1}^{n}\sum_{j=1}^{p}\left[b_j\left(\boldsymbol{\mu}_i - (\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu}_i))_j\right) - x_{ij}\left(\boldsymbol{\mu}_i - (\boldsymbol{U}\boldsymbol{U}^{\mathsf{T}}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\mu}_i))_j\right)\right] \tag{3}$$

If $(\boldsymbol{U}^*, \boldsymbol{\mu}^*)$ minimizes the deviance above, then the generalized PCA projection is given by

$$(g(\boldsymbol{X}) - \mathbf{1}^{\mathsf{T}}(\boldsymbol{\mu}^*)^{\mathsf{T}})\boldsymbol{U}^*$$

where g is the canonical link function and $\mathbf{1}$ is the vector with all entries equal to 1.

4.2 Semiparametric Approaches to Exponential Family PCA

In this section, we discuss semiparametric approaches and relevant methodology that was introduced in exponential family distribution PCA. Although we have not discussed this earlier in the Gaussian setting, similar proposals to this methodology exist in the Gaussian setting, but in the interest of space we will give just the references to them in this section and we will focus in explaining the methodology for PCA in the exponential family distribution setting.

4.2.1 Semiparametric Exponential Family PCA

In [41], the authors proposed a semiparametric approach to the problem. First, they identify that in the PPCA by [48] it is assumed that there is a latent distribution model which implies that we have a latent or mixing distribution $P(\boldsymbol{\theta})$ and a conditional or component distribution $P(\boldsymbol{x}|\boldsymbol{\theta})$. They also recognize the limitations that is introduced by assuming this latent distribution to be Gaussian.

To formulate their semiparametric PCA (SP-PCA) method, they make no assumptions on the distribution of the latent random variable $\boldsymbol{\theta}$. This allow for multimodality to be preserved in the projection space. Therefore, in cases where the data form clusters, this allows for simultaneous clustering and dimension reduction. Probably one of the most impressive features of this algorithm is the fact that it uses the nonparametric likelihood estimation by [33] which allows the use of one prior distribution which is conjugate to all exponential family distributions. Therefore, this allows for a unified approach for all exponential family distributions.

4.2.2 Copula PCA

[20], the authors proposed the use of a semiparametric model which can be applied to data from the nonparanormal family. By definition of the nonparanormal family, these are data that can be transformed through a possibly unknown monotone transformation to Gaussian. Copula PCA is then proposed to estimate the leading eigenvectors of the Gaussian distribution. To be more specific, copula PCA solves the following problem to find $\hat{\boldsymbol{u}}_1$ the leading eigenvector of the covariance matrix $\boldsymbol{\Sigma}$:

$$\hat{\boldsymbol{u}}_1 = \arg\max_{\boldsymbol{u}\in\mathbb{R}^p} \boldsymbol{u}^{\mathsf{T}}\hat{\boldsymbol{S}}\boldsymbol{u}, \quad \text{subject to } \boldsymbol{u} \in \mathbb{S}^{p-1} \cap \mathbb{B}_q(R_q)$$

where $\mathbb{S}^{p-1} = \{\boldsymbol{v} \in \mathbb{R}^p : \|\boldsymbol{v}\|_2 = 1\}$ is the p-dimensional l_2 sphere, $\mathbb{B}_q(R_q) = \{\boldsymbol{v} \in \mathbb{R}^d : \|\boldsymbol{v}\|_q^q \leq R_q\}$ is the l_q ball, and $\hat{\boldsymbol{S}}$ is the estimated Spearman's rho covariance coefficient matrix based on copulas.

4.3 Supervised Exponential Family PCA

In this section, we discuss a set of methods in the exponential family PCA which are a bit different than classic PCA. It is well known that classic PCA is an unsupervised dimension reduction method; that is, there is no response (or label) variable involved in the process. There are methods though, which have been proposed within the exponential family PCA framework and work in a supervised setting.

4.3.1 Supervised Exponential Probabilistic PCA

[50] proposed the supervised probabilistic PCA (SPPCA) approach, which incorporates label information into the projection. First, they assume that the data are modeled as:

$$\begin{aligned} \boldsymbol{x} &= \boldsymbol{W}_{\boldsymbol{x}}\boldsymbol{z} + \boldsymbol{\mu}_{\boldsymbol{x}} + \epsilon_{\boldsymbol{x}} \\ \boldsymbol{y} &= \boldsymbol{f}(\boldsymbol{z}, \boldsymbol{\Theta}) + \epsilon_{\boldsymbol{y}} \end{aligned}$$

which for $\boldsymbol{x}$ is similar to the PPCA model and then the model for $\boldsymbol{y}$ is added where $\boldsymbol{f}(\boldsymbol{z}, \boldsymbol{\Theta})$ is a function that encodes n deterministic functions that depend on n different parameters, i.e., $\boldsymbol{f}(\boldsymbol{z}, \boldsymbol{\Theta}) = (f_1(\boldsymbol{z}, \theta_1), \ldots, f_n(\boldsymbol{z}, \theta_n))$. To incorporate the label information into the projections, they use information on the inter-covariance between input variables and output variables and also the intra-covariance matrix of both the inputs and the outputs. These are incorporated through the use of an EM algorithm. It is assumed in the paper that the latent distribution (for $\boldsymbol{z}$) is Gaussian. In the same work, the authors propose a semisupervised PCA (SSPCA) algorithm for the cases where only part of the data are labeled. They demonstrate how their SPPCA approach is a special case of a SSPCA algorithm as it can be seen as the case where the number of unlabeled data is 0.

4.3.2 Supervised Exponential Family PCA Using Generalized Linear Models (GLMs)

[40] used a GLM approach to model exponentially distributed features and labels to perform dimension reduction and prediction in a supervised framework. To achieve this, they used an EM approach which maximizes a weighted linear function (auxiliary function) between both (features and labels) conditional likelihoods given latent variables. Their method can be used both in regression and in classification.

Similar to this idea is the idea of [42] who used linear mixture models to achieve dimension reduction in mixed types of data. To achieve this, they had a different objective function which is the product of the conditional distribution of the labels given the features, that is, $\prod_{i=1}^{n} P(y_i|\boldsymbol{x}_i, \boldsymbol{\Theta})$. Their work mostly focused on the use

of Gaussian distribution, but the use of any distribution in the exponential family is discussed as well.

4.3.3 Supervised Exponential Family PCA via Convex Optimization

In [17], the authors extended the work by [50] by allowing the latent distribution to be any exponential family distribution using convex optimization. In this algorithm, both the labels and the data are assumed to be drawn from the latent variable based on conditional exponential family distributions. One of the most important contributions of this algorithm is the proposal of a sample-based approximation to the exponential family models which allows for the kernelization of the method and therefore allows for nonlinear feature extraction.

4.4 Other Approaches to PCA in Non-Gaussian Settings

There are many other non-Gaussian settings where PCA-like algorithms have been proposed for unsupervised feature extraction. Here, we list a sample of them.

4.4.1 Transelliptical Component Analysis

In [21] the authors proposed a high-dimensional semiparametric scale-invariant PCA type of analysis which they called transelliptical component analysis (TCA) which was applicable to the transelliptical family of distributions. The following definition was given by [21].

Definition 2 A random vector $(X_1, \dots X_p)^\mathsf{T}$ is said to follow a transelliptical distribution if and only if there exists a set of strictly monotone functions $f = \{f_j\}_{j=1}^p$ and a latent continuous elliptically distributed random vector $\boldsymbol{Z}$ with mean 0 and variance $\boldsymbol{\Sigma}$ with $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}, \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\mathsf{T}, \text{diag}(\boldsymbol{\Sigma}) = \mathbf{1}$ and $\boldsymbol{\Sigma} \succeq 0$ such that $\big(f_1(X_1), \dots, f_p(X_p)\big)^\mathsf{T} \overset{\mathcal{D}}{=} \boldsymbol{Z}$. We call such $\boldsymbol{X}$ a transelliptical random vector with parameters $(\boldsymbol{\Sigma}; f_1, \dots f_p)$.

Transelliptical distributions are considered extensions of meta-elliptical distributions (see [14]).

The idea behind their proposal is first to estimate the latent correlation matrix $\boldsymbol{\Sigma}$ with Kendall's τ correlation matrix (denoted with $\hat{\boldsymbol{R}}$) and then solving an optimization problem similar to the PCA to find the vector $\boldsymbol{u}$ which maximizes $\boldsymbol{u}^\mathsf{T}\hat{\boldsymbol{R}}\boldsymbol{u}$ under the constraint that it lies inside the unit sphere. As the authors claimed, that meant one can plug $\hat{\boldsymbol{R}}$ to any sparse PCA algorithm in the literature. Furthermore, [22] proposed a similar framework for sparse PCA in meta-elliptical distributions.

Finally, Han and Liu [23] proposed the exponential component analysis (ECA) which follows a similar procedure in an effort to estimate directly the eigenvectors of the covariance matrix $\boldsymbol{\Sigma}$ and not the eigenvectors of a correlation matrix (which was the case with TCA). To achieve this, they approximated $\boldsymbol{\Sigma}$ using a multivariate Kendall's τ estimate. They also discussed sparse ECA as a method to obtain sparse features.

4.4.2 Variational Inference PCA

In [8] a variational inference approach to exponential family PCA was proposed. Variational inference is a tool used in [29] and [19] to model situations where the conditional distribution of the scores given the data becomes intractable, for example, the Poisson mixed model. In a bit more detail, the authors in [8] considered the model framework used by Bishop and Tipping (1999) in the PPCA algorithm, and the idea is that in cases where the conditional distribution of the scores given the data is intractable, then one can find a lower bound of it using a tractable product distribution. This is done using the Poisson lognormal distribution (PLN) proposed by Aitchinson and Ho (1989).

4.4.3 Exponential PCA (ePCA)

Exponential PCA (ePCA) was a recent proposal by [35] which proposed an algorithm which took a different approach. They proposed the use of a different covariance matrix estimator that is based on diagonally debiasing the sample covariance matrix. Similarly, they proposed a method for homogenization, shrinkage and heterogenization of the debiased matrix to make the procedure suitable for high-dimensional data.

4.4.4 Applying Discrete PCA in Data Analysis

In a slight different approach, [4] proposed a way to model discrete PCA and demonstrated the equivalence to discrete independent component analysis (ICA) (see [27]). Their approach specifically discussed a model for text data analysis which is our main motivation for this literature review, but at the same time, they model them using a multinomial distribution. This is because they consider that each document belongs to a topic and they consider each topic a possible outcome of a multinomial distribution.

4.4.5 PCA of Binary Data via iterated SVD

In [12] the authors combined ideas from latent structure analysis (LSA), multiple correspondence analysis (MCA) and PCA to propose a PCA method for binary data by iterated singular value decomposition (SVD).

4.4.6 Sparse Logistic PCA for Binary data

In [31] a sparse logistic PCA algorithm for binary data was proposed. They achieved this by looking at principal components from [39] perspective as was described in (1) where principal components were the linear projections that minimized the distance of the data points to their projections. They used also the likelihood approach, similar to the one proposed in [48] . To handle binary data though, they did not use the Gaussian distribution assumption as [48] (see also [44]), but instead they assumed a Bernoulli distribution imposing a logistic approach to the PCA. For a sparse solution, they use L1 penalization ([47]).

5 Sparse Exponential Family PCA

One of the biggest questions in feature extraction is whether one can obtain more meaningful and interpretable results by finding sparse features. This is the case where we try to force the loading vectors to contain a lot of 0's. Thus, we replace a lot of the small coefficients with 0, to emphasize that the variables corresponding to the zero coefficients have no real weighting in the reduced projections of the data. The most common method to achieve sparsity is by introducing some form of penalization. There has not been a lot of the literature for sparse methodology in this framework, and only the last few years this was picked up. In this section, we will discuss a few of the methods.

5.1 Sparse Probabilistic PCA

The authors in [18] introduced an L1 regularizer to the idea by [48] of probabilistic PCA. This was achieved by introducing a Laplacian prior to each element W_{ij} of the transformation matrix $\boldsymbol{W}$ since it has been shown (see [49]) that the Laplace prior is the L1 regularizer counterpart that was used in classic Sparse PCA ([53]) . They also propose two other methods for sparsity

- First, they proposed the use of an inverse Gaussian prior, since it was shown by [6] that it produces sparse models in the regression problem.
- Second they proposed the use of a Jeffreys's prior as it was shown in the literature ([15]) that it produces sparse models in regression and classification settings.

In all three ideas, the author used a two-level hierarchy model. In this model, they defined at the first level $p(W_{ij}|z_{ij}) \sim N(0, z_{ij})$ and then at the second level they put a prior distribution on z_{ij} depending on which of the three methods for imposing sparsity is used. The first case used the Laplace, the second the inverse Gaussian and the third the Jeffrey's prior.

5.2 Sparse Exponential Family PCA

The work presented in [36] is the first to propose a general theory for sparse exponential family PCA algorithms. They combined the idea of [9] with the regularizer in [44] to propose a general framework for dimension reduction in the exponential family distribution. More specifically, they propose the solution of the following optimization problem:

$$\min_{\boldsymbol{Z}:\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}=\boldsymbol{I}} \min_{\boldsymbol{W},\boldsymbol{b}} \sum_{n} \boldsymbol{A}(\boldsymbol{W}\boldsymbol{z}_n + \boldsymbol{b}) - \mathrm{tr}((\boldsymbol{Z}\boldsymbol{W}^{\mathsf{T}} + \boldsymbol{1}\boldsymbol{b}^{\mathsf{T}})\boldsymbol{X}^{\mathsf{T}}) + P(\boldsymbol{W}, \boldsymbol{b})$$

where $\boldsymbol{W}$ is the $p \times q$ principal loading matrix, $\boldsymbol{Z}$ is the $n \times q$ principal component score matrix with each row being $\boldsymbol{z}_n$, $\boldsymbol{A}$ is a semiorthonormal matrix and $P(\cdot)$ is the

penalty term which the author suggests to be equal to

$$\lambda_0 \|\boldsymbol{Z}\boldsymbol{W}^{\mathsf{T}} + \mathbf{1}\boldsymbol{b}^{\mathsf{T}}\|^2 + \sum_{l=1}^{q} \lambda_q |\boldsymbol{W}_q|.$$

where $\boldsymbol{W}_q$ is the loadings of the q^{th} principal component. The authors explained that the l2-norm ensures a stable reconstruction of principal components when $\boldsymbol{X}$ is not a full rank matrix and the value of λ_q controls the sparsity of loading vectors.

5.3 Sparse Generalized PCA

In [45] a method for sparse exponential family PCA was derived, using the generalized PCA idea of Landgraf and Lee (2015) and combining it with the $L1$ and SCAD penalties. In particular, they minimized

$$L(\boldsymbol{U}, \boldsymbol{\mu}) = D(\boldsymbol{U}, \boldsymbol{\mu}) + \lambda_1 |\boldsymbol{U}| + \lambda_2 \text{SCAD}(\boldsymbol{U}; \alpha, \lambda) \tag{4}$$

where $D(\boldsymbol{U}, \boldsymbol{\mu})$ is given in (3), λ_i $(i = 1, 2)$ are tuning parameters which improves the number of zeroes in the model, and $\text{SCAD}(\boldsymbol{U}; \alpha, \lambda)$ is the SCAD penalty which is usually defined by its derivative :

$$\text{SCAD}'(\boldsymbol{U}; \alpha, \lambda) = \lambda I(\boldsymbol{U} \leq \lambda) + \frac{\alpha\lambda - \boldsymbol{U}}{\alpha - 1} I(\boldsymbol{U} < \lambda)$$

where λ and α are tuning parameters.

In order to solve the optimization problem, which is nonconvex, they used a majorization–maximization (MM) algorithm. In a simulation study, they showed that all considered penalties (the $L1$ penalty, the SCAD penalty, and a linear combination of the two) had similar performance and so suggest using the $L1$ for its simplicity. They focused more on PCA for Poisson distributed data due to their application in analyzing text data.

5.4 Sparse Simple Poisson PCA

In [46] putthe authors used an adaptive $L0$ penalty due to [16] on the simple exponential PCA algorithm. They place the penalty on $\boldsymbol{W}$, the analogue of the loadings matrix. The adaptive version of the $L0$ is chosen for its differentiability everywhere, unlike the original version of the penalty, which allows for an efficient gradient-based optimization method. Their alteration yields an optimization problem which needs to be solved iteratively, as the adaptive penalty is calculated using values from the previous iteration. To illustrate their variant of simple exponential PCA, they gave the Poisson case and used examples drawn from text data.

5.5 Choice of Regularization

To close this section, we will briefly discuss the choice of regularization. There is a long discussion in the statistics community which regularization to choose between $L0$ and $L1$. This is mostly due to the fact that $L0$ regularization seems the most logical as it penalizes the number of nonzero coefficients, but at the same time a solution to the $L0$ regularization is NP-hard and therefore is computationally much more expensive than the $L1$ regularization which is based on a convex objective function. In [7] previous discussion on this topic by [2] and [24] was extended. They suggest that the choice of regularization should be based on 3 aspects:

- The problem/dataset: The authors suggest that $L0$ should be preferred when the cardinality of the model parameters is constrained and raise concerns over the use of l_1 regularization as a surrogate to the $L0$ objective function. This warning is due to the tendency of $L1$ to favor zeroes, that is to overshrink the parameter estimates.
- Optimization aspect: Computationally $L1$ is convex and therefore easy to find a solution. For the $L0$, a solution can be found through alternative approximations appeared in the literature, but these algorithms offer no guarantee of finding an optimal solution.
- Statistical aspect: When we are looking for an optimal solution based on stability and replicability, we can make a different choice of regularizer based on the data generating process as different data may prefer different results.

In the probabilistic framework we discuss in this paper, the choice of regularization is driven by the choice of prior distribution. As was mentioned earlier in [18], it was discussed for example that a Laplace prior is equivalent to having an $L1$ regularization in the probabilistic framework. They also suggested the use of inverse Gaussian as a prior which can be adjusted (with a specific set of parameters) to approximate $L2$ regularization and noninformative Jeffrey's prior which approximates $L0$ regularization.

6 Discussion

As we can see from the methods we reviewed in this manuscript, there are a number of exponential family PCA algorithms available in the literature, as well as more specific algorithms focusing on specific distributions, i.e., Gaussian, Poisson, binomial etc. Each of these methods was discussed in a different framework and has been applied to a variety of dataset. Although we do not claim that the list of methods in this literature review is exhaustive, we believe that we have touched upon a number of different methods and ideas, to give an overview of the existing methodology and to motivate further research in this interesting topic.

6.1 Implementation tools

Throughout this literature review on the exponential family PCA in this manuscript, we refer a lot to some methods that are used for the implementation of the algorithms.

In this section, we will briefly discuss some of these tools, which are very well-known tools in the statistical literature.

6.1.1 Expectation–Maximization (EM) Algorithm

First, we present the EM algorithm which was around in a number of formats before being formally introduced by [13] who laid out all the theoretical foundations of the algorithm. The algorithm is mainly used in cases where we need to maximize a complicated likelihood function which cannot be solved analytically. In many cases, in addition to the parameters that are not known and need to be estimated there exist unobserved latent variables $\boldsymbol{Z}$ which need to be approximated. The two steps are as follows:

- *Expectation* Use the current estimate of the parameters, let's denote them with $\boldsymbol{\theta}$ (at the first step this is an initial value for each parameter) and calculate the expectation of the log likelihood with respect to the conditional distribution $\boldsymbol{Z}|\boldsymbol{X}$ and $\boldsymbol{\theta}$, that is $E_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}}(\log L(\boldsymbol{\theta}; \boldsymbol{z}, \boldsymbol{x}))$, where $\boldsymbol{X}$ are the known data, $\boldsymbol{Z}$ is the latent variables, and $L(\cdot)$ is the likelihood
- *Maximization* Find the new value of the parameters $\boldsymbol{\theta}^N$ that maximizes the expected value in the previous step.

These two steps are repeated until we have a relatively stable estimate of the parameters $\boldsymbol{\theta}$.

The EM algorithm has been used extensively in a number of statistical methods in the literature. A lot of the likelihood-based methods presented in this review are implemented using EM algorithm.

6.1.2 Majorize/Maximize (MM) Algorithm

MM algorithm may stand for any combination of the words majorize/minorize–maximize/minimize depending on whether we are trying to maximize or minimize an objective function. Although as an idea it existed for some time, it was formally conceptualized and better studied by [26] who applied it on quantile regression. The main idea of the MM algorithm is to exploit the convexity of a function by using a surrogate function in order to majorize (minorize) the objective function. By maximizing/minimizing the surrogate function, one tries to maximize/minimize the objective function.

6.1.3 Maximum a posteriori (MAP) Estimation

As was discussed throughout the literature review, the methodology discussed in the exponential family PCA framework is based on likelihood estimation with latent variables. This leads naturally to the use of a Bayesian approach to estimation. MAP estimation is used to estimate essentially a quantity that represents the mode of the posterior distribution we give to our parameter of interest (which we denote with $\boldsymbol{\theta}$). If we give a uniform prior distribution to $\boldsymbol{\theta}$, then the MAP estimator coincides with the maximum likelihood estimator (MLE). It is important to note that in general the MAP estimator does not coincide with the Bayes estimator.

6.1.4 Automatic Relevance Determination (ARD)

Automatic relevance determination is a way to reduce the number of features in high-dimensional problems. It effectively prunes away redundant features by using a parameterized data-dependent prior distribution. ARD (see [37]) has been extended in a number of direction, and currently, it is considered a collection of algorithms that achieve dimension reduction in different settings. For example, the sparse Bayesian learning (SBL) method is considered a special case of the ARD algorithm. One of the properties of ARD is that it usually gives better estimators than a MAP estimator in selecting an optimal feature set.

6.2 Open Problems

There are a number of open topics that are obviously missing from the literature in this based on how it presented in this review. We discuss below some of the more interesting questions one may try to address:

1. One of the things we have not discussed extensively is nonlinear feature extraction (or nonlinear PCA) in the exponential family distribution framework. With a number of recent developments in the kernel methods, including the development in kernel PCA and the extensive list of applications kernel PCA has been used one would have expected that there will be an equivalent list of the literature in nonlinear PCA for exponential family distributions. To the best of our knowledge, there is very little that has been done in this direction. It will also be interesting if there was a unified framework.
2. Some of the exponential family distribution PCA algorithms presented here have not been studied extensively and especially the more recent ones. For example, methods like ePCA by [35] or the copula PCA by [20] and the transelliptical component analysis by [21] have not been extended to include sparsity.
3. Other dimension reduction algorithms like canonical correlation or independent component analysis have not been studied as extensively in the exponential family framework. Although this methodology is not necessarily part of this literature review, a quick search on the available literature for these reveals that there is much less effort to extend those in the different directions on the exponential family distribution framework than the effort the research community has put in doing this for the PCA.
4. Finally, most of these methods have been presented in the literature and have not been extensively used in practical problems, mainly due to the lack of a single source of code that can help practitioners by giving them the option to run multiple of these methods easily to compare their outputs. Therefore, we believe that one can start writing a package in R/Python to implement as many of these methods in a unified framework, which will allow practitioners to easily apply them in real data.

Acknowledgements The authors would like to thank the AE and the reviewers for their helpful and constructive comments who improved this literature review presentation.

References

1. Aitchison J, Ho C (1989) The multivariate poisson-log normal distribution. Biometrika 76:643–653
2. Bertsimas D, Pauphilet J, van Parys B (2020) Sparse regression: scalable algorithms and empirical performance. Stat Sci 35:555–578
3. Bishop CM (1998) Bayesian PCA. In: Proceedings of the 1998 annual conference on advances in neural information processing systems (NIPS)
4. Buntine W, Jakulin A (2004) Applying discrete PCA in data analysis. In: Proceedings of the 20th conference on uncertainty in artificial intelligence, pp 59–66
5. Cai TT, Ma Z, Wu Y (2013) Sparse PCA: optimal rates and adaptive estimation. Ann Stat 41:3074–3110
6. Caron F, Doucet A (2008). Sparse bayesian nonparametric regression. In: International conference on machine learning, pp 88–95
7. Chen Y, Taeb A, Buhlmann P (2020) A look at robustness and stability of l_1- versus l_0-regularization: discussion of papers by Bertsimas et al and Hastie et al. Stat Sci 35:614–622
8. Chiquet J, Mariadassou M, Robin S (2017) Variational inference for probabilistic poisson PCA. https://arxiv.org/pdf/1703.06633.pdf
9. Collins M, Dasgupta S, Schapire R. E (2001). A generalization of principal components to the exponential family. In: Proceedings of the 14th annual conference of neural information processing systems (NIPS)
10. Cook RD (2007) Fisher lecture: dimension reduction in regression. Stat Sci 22:1–40
11. d'Aspremont A, El-Ghaoui L, Jordan MI, Lanckriet GRG (2007) A direct formulation for sparse PCA using semidefinite programming. SIAM Rev 49:434–448
12. de Leeuw J (2006) Principal component analysis of binary data by iterated singular value decomposition. Comput Stat Data Anal 50:21–39
13. Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38
14. Fang HB, Fang KT, Kotz S (2012) The meta-elliptical distributions with given marginals. J Multivar Anal 82:1–16
15. Figueiredo MAT (2001) Adaptive sparseness using Jeffreys prior. NIPS, pp 679–704
16. Frommlet F, Nuel G (2016) An adaptive ridge procedure for L0 regularization. PLoS ONE 11:1–23
17. Guo Y (2009) Supervised exponential family principal component analysis via convex optimization. In: Proceedings of the 21st international conference on neural information processing systems, pp 569–576
18. Guan Y, Dy JG (2009) Sparse probabilistic principal component analysis. J Mach Learn Res 5:185–192
19. Hall P, Ormerod JT, Wand MP (2011) Theory of gaussian variational approximation for a Poisson mixed model. Statistica Sinica 369–389
20. Han F, Liu H (2012a) Semiparametric principal component analysis. In: Proceedings of the 25th annual conference on neural information processing systems, pp 171–179
21. Han F, Liu H (2012b). Transelliptical component analysis. In: Proceedings of the 25th annual conference on neural information processing systems, pp 368–376
22. Han F, Liu H (2014) Scale-invariant sparse PCA on high-dimensional meta-elliptical data. J Am Stat Assoc 109:275–287
23. Han F, Liu H (2018) ECA: high-dimensional elliptical component analysis in non-Gaussian distributions. J Am Stat Assoc 113:252–268
24. Hastie T, Tibshirani R, Tibshirani R (2020) Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Stat Sci 35:579–592
25. Hotelling H (1933) Analysis of a complex of statistical variables into principal components. J Educ Psychol 24:417–441

26. Hunter DR, Lange K (2000) Quantile regression via an MM algorithm. J Comput Graph Stat 9:60–77
27. Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley
28. Joliffe IT (2002) Principal component analysis. Springer
29. Karlis D (2005) EM algorithm for mixed Poisson and other discrete distributions. Astin Bull 35:3–24
30. Landgraf AJ, Lee Y (2015) Generalized principal component analysis: projection of saturated model parameters, Technical Report 892, Ohio State University Statistics Department
31. Lee S, Huang JZ, Hu J (2010) Sparse logistic principal component analysis for binary data. Ann Appl Stat 4:1579–1601
32. Li J, Tao D (2010) Simple exponential family PCA. In: Proceedings of the 13th international conference on artificial intelligence and statistics 9:453–460
33. Lindsay BG (1983) The geometry of mixture likelihoods: a general theory. Ann Stat 11:86–104
34. Liu K, Li Q, Wang H, Tang G (2019) Spherical principal component analysis. In: Proceedings of the 2019 SIAM international conference on data mining, pp 387–395
35. Liu LT, Dobriban E, Singer A (2018) ePCA: high dimensional exponential family PCA. Ann Appl Stat 12:2121–2150
36. Lu M, Huang JZ, Qian X (2016) Sparse exponential family principal component analysis. Pattern Recogn 60:681–691
37. MacKay DJC (1995) Probable networks and plausible predictions–a review of practical Bayesian methods for supervised neural networks. Network Comput Neural Syst 6(3):469–505
38. Mohammed S, Heller K, Ghahramani Z (2009) Bayesian exponential family PCA. In: Proceedings of the 2008 conference on advances in neural information processing systems
39. Pearson K (1901) On lines and planes of closest fit to system of points in space. Philos Mag Ser 6(2):559–572
40. Rish I, Grabarnilk G, Cecchi G, Pereira F, Gordon G (2008) Closed-form supervised dimensionality reduction with generalized linear models. In: Proceedings of international conference on machine learning (ICML), Helsinki, Finland
41. Sajama, Orlitsky A (2004) Semiparametric exponential family PCA. In: Proceedings of the 2004 conference on advances in neural information processing systems, pp 1177–1184
42. Sajama, Orlitsky A (2005). Supervised dimension reduction using mixture models. In: Proceedings of the international conference on machine learning (ICML)
43. Schölkopf B, Smola A, Müller K-R (1997) Kernel principal component analysis. In: Proceeding of the international conference on artificial neural networks, pp 583–588
44. Shen H, Huang JZ (2008) Sparse principal component analysis via regularized low rank matrix approximation. J Multivar Anal 99:1015–1034
45. Smallman L, Artemiou A, Morgan J (2018) Sparse generalised principal component analysis. Pattern Recogn 83:443–455
46. Smallman L, Underwood W, Artemiou A (2019) Simple Poisson PCA: an algorithm for (sparse) feature extraction with simultaneous dimension determination. Comput Stat 35:559–577
47. Tibshirani RJ (1996) Regression shrinkage and selection via the lasso. J R Stat Soc B 58:267–288
48. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. J R Stat Soc B 61:611–622
49. Williams P (1995) Bayesian regularization and pruning using a Laplace prior. Neural Comput 7:117–143
50. Yu S, Yu K, Tresp V, Kriegel H-P, Wu M (2006) Supervised probabilistic principal component analysis. In: Proceedings of 12th ACM SIGKDD international conference on KDD
51. Zass R, Shashua A (2007) Nonnegative sparse PCA. Adv Neural Inf Process Syst 19:1561–1567
52. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc B 67:301–320
53. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15:265–286