The Annals of Applied Statistics
2021, Vol. 15, No. 2, 1017–1036
https://doi.org/10.1214/20-AOAS1420
© Institute of Mathematical Statistics, 2021

TENSOR-VARIATE FINITE MIXTURE MODELING FOR THE ANALYSIS OF UNIVERSITY PROFESSOR REMUNERATION

BY SHUCHISMITA SARKAR[1], VOLODYMYR MELNYKOV[2,*] AND XUWEN ZHU[2,†]

[1]*Department of Applied Statistics and Operations Research, Bowling Green State University, ssarkar@bgsu.edu*
[2]*Department of Information Systems, Statistics, and Management Science, University of Alabama,* [*]*vmelnykov@ua.edu;* [†]*xzhu20@cba.ua.edu*

There has been a long-standing interest in the analysis of university professor salary data. The vast majority of the publications on the topic employ linear regression models in an attempt to predict individual salaries. Indeed, the administration of any academic institution is interested in adequately compensating the faculty to attract and keep the best specialists available on the market. However, higher administration and legislators are not concerned with the matter of individual compensation and need to have a bigger picture for developing university strategies and policies. This paper is the first attempt to model university compensation data at the institutional level. The analysis of university salary patterns is a challenging problem due to the heterogeneous, skewed, multiway and temporal nature of the data. This paper aims at addressing all the above-mentioned issues by proposing a novel tensor regression mixture model and applying it to the data set obtained from the American Association of University Professors. The utility of the developed model is illustrated on addressing several important questions related to gender equity and peer institution comparison.

1. Problem description. University faculty remuneration has been a topic of major interest in the field of higher education management. Numerous studies have been conducted to understand the behavior of professor salaries at American universities (Becker (1975), Hearn (1999), Mincer (1958)). These studies examine various aspects of financial compensation at research institutions. For example, DeLorme, Hill and Wood (1979) consider the effect of publications, departmental affiliation as well as teaching experience and effectiveness on professor salary. Fairweather (1993) studies the relative importance of teaching, research, administration and service in the process of determining the base salary level. The obtained results show the dominance of a research-oriented faculty reward structure for each institution type regardless of the professed mission. Simpson (1981) illustrates options in designing a salary structure appropriate for an institution that gives weight both to professional concerns as to merit and personal concerns as to salary improvement over individual's career. Melguizo and Strober (2007) explore the relationship between faculty salary and university prestige. The authors examine the salary-determining factors and investigate the relationship between the faculty financial reward and institutional prestige. Some other noticable studies consider factors such as faculty rank, years of experience, gender, race and academic discipline (Ashraf and Shabbir (2006), Hearn (1999), Perna (2001), Rippner and Toutkoushian (2015), Umbach (2007)). Most of these studies focus on faculty representing a specific university. A small number of papers investigate the effect of geographical region and reported its insignificant effect on the faculty remuneration (Cohn (1973), Friedman and Kuznets (1945), Mohanty, Dodder and Karman (1986)). Primary research methods employed in the above-mentioned papers include the cohort study and multiple regression analysis. Many papers recognize the presence of severe right skewness in salary data that is usually addressed by applying a log

Received November 2019; revised October 2020.
Key words and phrases. Faculty salary data, mixture models, cluster analysis, tensor data.

transformation (Ashraf and Shabbir (2006), Rippner and Toutkoushian (2015), Toutkoushian, Bellas and Moore (2007)). Although this transformation is not designed for reaching normality, its application can relax the issue with the violation of the normality assumption to some degree. The degree of skewness, however, can vary for different factor levels and hence the log transformation oftentimes lacks flexibility. Unfortunately, nearly no paper on the topic acknowledges the presence of natural heterogeneity in the population, temporal behavior and dependence of observations within the same institution of higher education. As a result, many proposed models suffer from overly simplistic and unrealistic assumptions. Moreover, the vast majority of the above-mentioned projects focus on modeling and predicting individual faculty salary, even if the original goal of the analysis was concerned with the inference at the institutional level. However, for the purpose of strategic planning and policy development, government officials, legislators and university upper administration are not interested in the analysis of individual salaries and, therefore, consider universities as observational units. Currently, there is no literature concerned with the analysis of the remuneration data at the university level, and this paper presents the first attempt to address this deficiency.

The data set analyzed in this paper is obtained from *The Chronicle of Higher Education* web site (https://data.chronicle.com). The site contains average selfreported salaries of university professors employed at American academic institutions. The scope of our analysis is restricted to universities granting doctoral and masters' degrees. The data are categorized by the following factors: gender (*male*, *female*), professor rank (*assistant*, *associate*, *full*) and academic year (2003/2004, . . . , 2015/2016). These variables are common to consider in salary analysis (Ferreira (2013), Hexter (1990)). As *The Chronicle of Higher Education* collects information on the voluntary basis, 26% of the universities were either not represented at all or had incomplete records (mostly related to no salaries reported in specific years). Also, 37 outlying schools with clearly erroneous entries have been excluded from the study. Such entries typically show unusually high drops or spikes in the average salary and are associated with a single gender-rank combination in a particular year. Such differences cannot be explained by possible changes in university compensation policies as they are specific to just one year. At the same time, they are unlikely to represent a one-time salary adjustment, as other gender-rank groups do not exhibit similar behavior. As a result of the outlined preprocessing stage, we obtain a set of data consisting of 696 universities. As universities are treated as observational units in our study, each data point exhibits average faculty salaries presented in the form of a $2 \times 3 \times 13$-dimensional tensor. Alternatively, each observation can be viewed as a time series of 2×3 matrices. In the context of salary data analysis, the tensor structure has never been considered but has major advantages as we explain below.

University salary data are expected to be highly heterogeneous, and some popular approaches to modeling subpopulations include mixtures of regressions and mixtures of experts. Due to the traditional assumption of independent observations, these models are not readily applicable unless some modifications are considered. One possibility is to introduce random effects accounting for dependence of gender, rank and year within the same institution. Such an approach would considerably increase the complexity of the model. Another major concern with the use of nontensor approaches is the risk of assigning various gender-rank-year salary combinations from the same university to different clusters. A possible way of addressing this issue is to employ semisupervised clustering methods capable of taking into account membership constraints. Indeed, this would bring another level of complexity to the model. On the other hand, we propose employing tensor-variate distributions that effectively model dependence among tensor elements and structurally tie them together. Due to the limited developments in the area of tensor distributions, especially those that can model skewness, we propose generalizing transformation ideas presented in Melnykov and Zhu (2018, 2019).

In order to address the heterogeneity in data, we develop a transformation-based tensor regression mixture model. We utilize this model to approach major long-standing questions on university classification, gender equity and peer comparison.

The goal of our study is to investigate the impact of faculty salary on partitioning the U.S. universities and to provide insight to the officials and upper administration on strategic planning and policy development. In particular, we consider the following questions that are detailed below:

1. What is the optimal grouping of the universities based on wage patterns and how well it matches the existing Carnegie classification?

Due to a wide variety of American institutions, it is clear that the underlying population is not homogeneous. However, cluster analysis of universities based on salary patterns has never been conducted before. This paper presents the first attempt to identify groups of universities with similar remuneration behavior. Various interesting aspects can be investigated based on the found partitioning. In particular, it is of high interest to compare the obtained grouping with the famous Carnegie classification (http://carnegieclassifications.iu.edu) created by the Carnegie Commission on Higher Education in 1970. This classification has been widely used in the study of higher education for recognizing and describing institutional diversity in American higher education. Carnegie classification allocates universities into various groups based on their institutional profiles. In this study we focus on institutions granting graduate degrees. Doctoral universities are defined as institutions that award at least 20 doctoral degrees per year. They are divided into three categories based on their research caliber: *R1*: *Highest research activity*, *R2*: *Higher research activity*, and *R3*: *Moderate research activity*. Master's universities are defined as those that award fewer than 20 doctoral but at least 50 master's degrees per year. These institutions are split into three groups based on their sizes: *M1*: *Larger programs*, *M2*: *Medium programs* and *M3*: *Smaller programs*:

2. Is there a difference in salaries that can be associated with the gender factor, and, if so, what is its nature?

Although many researchers investigate the impact of specific characteristics on the level of faculty compensation, gender equity is probably the most popular one in the literature (Perna (2001), Umbach (2007), Ferreira (2013)). Methodologies developed in this direction primarily rely on regression modeling. Becker and Toutkoushian (2003) aim at measuring gender bias in the salaries of tenured faculty. The corresponding variable is found significant in the multiple regression setting. Snyder, Hyer and McLaughlin (1994) consider a multiphase approach also relying on regression analysis to identify gender-based salary inequities. Unfortunately, the above-mentioned papers employ an additive effect model assumption and do not include interaction terms. In addition, the inference based on the significance of individual variables oftentimes can be misleading due to the presence of multicollinearity. Moreover, per earlier discussion, the assumption of independence for the salary observations within the same university is hardly realistic. Our proposed approach effectively takes into consideration all outlined limitations and provides easily interpretable conclusions.

3. Does a specific university fall in the same group with its presumed peers?

The assessment of the remuneration competitiveness is a challenging problem any university administration faces on a regular basis. Strategic decisions regarding faculty wages are usually made based on the level of salaries paid at peer institutions. The groups of peers and aspirational schools are typically prespecified by the administration based on the current university profile and objectives. These groups, created for internal use, vary for different universities substantially and generally are not publicly available. Therefore, we illustrate the proposed analysis on several universities based on peer groups proposed by *The Chronicle of Higher Education* web site (http://www.chronicle.com/interactives/peers-network). Such a comparative analysis is nontrivial, and there is no literature addressing this issue at this point. This paper presents the first attempt to address this important problem.

The rest of the paper is organized as follows. Section 2 briefly outlines some necessary preliminaries on finite mixture models, matrix transformations and tensor normal distribution. Section 3 is devoted to modeling tensor-valued university salary data. In Section 4, we address the three questions of interest with thorough discussions. Finally, Section 5 concludes the paper with a summary.

2. Preliminaries.

2.1. *Matrix mixture modeling and model-based clustering.* Let $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ be a sample consisting of n independent $p \times d$ random matrices identically distributed according to the probability density function (pdf) $g(\boldsymbol{Y}; \boldsymbol{\Theta})$ given by

$$g(\boldsymbol{Y}; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{Y}; \boldsymbol{\vartheta}_k). \tag{1}$$

Equation (1) is known as a matrix-variate finite mixture model. Here, K represents the number of components, also known as mixture order. $f_k(\boldsymbol{Y}; \boldsymbol{\vartheta}_k)$ is the matrix-variate pdf of the kth mixture component with parameter $\boldsymbol{\vartheta}_k$ and π_k is the corresponding mixing proportion with restriction $\sum_{k=1}^{K} \pi_k = 1$. The estimate of the entire parameter set $\boldsymbol{\Theta} = \{\pi_1, \dots, \pi_{K-1}, \boldsymbol{\vartheta}_1, \dots, \boldsymbol{\vartheta}_K\}$ is usually obtained by employing the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin (1977)), an iterative method of finding maximum likelihood estimates (MLE) that relies on the missing data formulation. In the finite mixture modeling setting, the membership labels of $\boldsymbol{Y}_i$, denoted as z_i, $i = 1, \dots, n$, are assumed to be unknown. At the expectation (E) step of the EM algorithm, the conditional expected value of the complete-data log-likelihood function, given observed data, is obtained. Such an expectation is broadly known as the Q-function. The E step in the mixture modeling context deals with updating posterior probabilities that observation $\boldsymbol{y}_i$ originated from the kth component, that is, $\tau_{ik} = P(z_i = k | \boldsymbol{y}_i, \boldsymbol{\Theta})$. At the maximization (M) step the Q-function is optimized with respect to $\boldsymbol{\Theta}$. Iterating the E and M steps until a prespecified convergence criterion is met leads to the MLE of $\boldsymbol{\Theta}$ denoted as $\hat{\boldsymbol{\Theta}}$. If the desired mixture order K is unknown, the optimal number of components is typically selected based on one of information criteria among which the Bayesian information criterion (BIC) (Schwarz (1978)) is the most popular in the finite mixture modeling framework.

Model-based clustering usually assumes the existence of a one-to-one association between mixture components and clusters. This is an underlying assumption made in this paper as well. Under this setting the assignment of observations to clusters relies on the Bayes decision rule and is given by $\hat{z}_i = \operatorname{argmax}_k \hat{\tau}_{ik}$.

Matrix normal ($m\mathcal{N}$) distribution with the pdf given by

$$\Phi(\boldsymbol{Y}; \boldsymbol{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) = \frac{(2\pi)^{-\frac{pd}{2}}}{|\boldsymbol{\Sigma}|^{\frac{d}{2}} |\boldsymbol{\Psi}|^{\frac{p}{2}}} \exp\left\{-\frac{1}{2} \operatorname{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y} - \boldsymbol{M})\boldsymbol{\Psi}^{-1}(\boldsymbol{Y} - \boldsymbol{M})^\top\}\right\}$$

presents a convenient tool for modeling matrix-valued data. Here, $\operatorname{tr}\{\cdot\}$ represents the trace operator, $\boldsymbol{M}$ denotes the $p \times d$ mean matrix, $\boldsymbol{\Sigma}$ and $\boldsymbol{\Psi}$ are $p \times p$ and $d \times d$ covariance matrices associated with p rows and d columns, respectively. Then, the matrix normal mixture model, as proposed in Viroli (2011a, 2011b, 2012), is given by

$$g(\boldsymbol{Y};\boldsymbol{\Theta})=\sum_{k=1}^{K}\pi_k\Phi(\boldsymbol{Y};\boldsymbol{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k),$$

where $f_k(\boldsymbol{Y};\boldsymbol{\vartheta}_k)\equiv\Phi(\boldsymbol{Y};\boldsymbol{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k)$ with $\boldsymbol{\vartheta}_k\equiv\{\boldsymbol{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k\}$.

The assumption of symmetry imposed by the normal density is often unrealistic. Indeed, salary data including those considered in this paper are highly asymmetric. In the multivariate setting a variety of methods have been proposed for the analysis of skewed vector-valued data (Banfield and Raftery (1993), Lee and McLachlan (2013a, 2013b), Browne and McNicholas (2015), Franczak, Browne and McNicholas (2014), Lin, Ho and Lee (2014), O'Hagan et al. (2016)). The current literature for such matrix-variate models is limited and very recent. Two major streams of work in this direction include employing more traditional distributions in the matrix-variate framework such as matrix-variate skew-t (Gallaugher and McNicholas (2017, 2019c)), variance-gamma, normal inverse Gaussian and generalized hyperbolic distributions (Gallaugher and McNicholas (2018, 2019)) or applying the ideas of transformations to near normality, as shown in Melnykov and Zhu (2018, 2019), Sarkar et al. (2020). In this paper we also employ transformation ideas because of their convenience in generalization to tensor distributions.

Suppose $\mathcal{T}(\cdot;\lambda)$ is some transformation to normality with parameter λ in the univariate setting. If this transformation is adequately effective, the original skewed data can be modeled with the pdf $f(y;\mu,\sigma^2,\lambda)=\phi(\mathcal{T}(y;\lambda);\mu,\sigma^2)J_{\mathcal{T}}(y;\lambda)$, where $\phi(\cdot;\mu,\sigma^2)$ is the normal pdf with mean μ, variance σ^2 and $J_{\mathcal{T}}(\cdot;\lambda)$ represents the Jacobian associated with the chosen transformation. In this setting, λ can be interpreted as a skewness parameter. The choice of transformation is not unique (Box and Cox (1964), Manly (1976)), and we use the generic notation $\mathcal{T}$ to emphasize that. In the analysis of our data, however, we apply the power transformation of Yeo and Johnson (2000) given by

$$\mathcal{T}(Y;\lambda)=\begin{cases}\lambda^{-1}((Y+1)^{\lambda}-1), & \lambda\neq 0, Y\geq 0,\\ \log(Y+1), & \lambda=0, Y\geq 0,\\ (2-\lambda)^{-1}(1-(1-Y)^{2-\lambda}), & \lambda\neq 2, Y<0,\\ -\log(1-Y), & \lambda=2, Y<0.\end{cases}\tag{2}$$

that is appealing due to the provided $I\!R$ to $I\!R$ mapping and capability to model left- and right-skewed data. The Jacobian associated with this transformation has the form $J_{\mathcal{T}}(y;\lambda)=(|y|+1)^{\mathrm{sgn}(y)(\lambda-1)}$.

There is a well-known generalization of univariate transformations to vectors (Lo and Gottardo (2012), Zhu and Melnykov (2018)) assuming that an elementwise transformation can be successful in reaching joint normality. A similar assumption can be made for matrix-variate cases. In a recent paper by Melnykov and Zhu (2018), the authors proposed an additive-effect parameterization of the matrix skewness parameter $\boldsymbol{\Lambda}=(\Lambda_{jh})_{p\times d}$. Namely, $\boldsymbol{\Lambda}=\boldsymbol{\lambda}\boldsymbol{1}_d^\top+\boldsymbol{1}_p\boldsymbol{\nu}^\top$, where $\boldsymbol{\lambda}=(\lambda_1,\dots,\lambda_p)^\top$ and $\boldsymbol{\nu}=(\nu_1,\dots,\nu_d)^\top$ stand for parameters associated with rows and columns, respectively. Vectors $\boldsymbol{1}_d$ and $\boldsymbol{1}_p$ represent all-ones vectors of length d and p. In other words, $\Lambda_{jh}=\lambda_j+\nu_h$. There is a minor nonidentifiability problem due to the fact that $\boldsymbol{\Lambda}$ can be alternatively obtained as $\boldsymbol{\Lambda}=\boldsymbol{\lambda}^\star\boldsymbol{1}_d^\top+\boldsymbol{1}_p\boldsymbol{\nu}^{\star\top}$ with any $\boldsymbol{\lambda}^\star=\boldsymbol{\lambda}+a\boldsymbol{1}_p$, $\boldsymbol{\nu}^\star=\boldsymbol{\nu}-a\boldsymbol{1}_d$ and $a\in I\!R$. This issue can be effectively resolved by introducing a constraint such as $\nu_d=0$. Then, the proposed parameterization has an attractive interpretability. Vector $\boldsymbol{\lambda}$ can be seen as the parameter responsible for modeling skewness in the last column, and ν_h is the skewness adjustment to $\boldsymbol{\lambda}$ for the h^{th} column, where

$h = 1, \dots, d-1$. This leads to the matrix transformation mixture model defined in Melnykov and Zhu (2018),

$$g(\boldsymbol{Y}; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \Phi\big(\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k); \boldsymbol{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k\big) J_{\mathcal{T}}(\boldsymbol{Y}; \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k),$$

where $\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda}, \boldsymbol{\nu})$ is the matrix transformation given by

$$\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{pmatrix} \mathcal{T}(Y_{11}, \lambda_1 + \nu_1) & \dots & \mathcal{T}(Y_{1,d-1}, \lambda_1 + \nu_{d-1}) & \mathcal{T}(Y_{1d}, \lambda_1) \\ \mathcal{T}(Y_{21}, \lambda_2 + \nu_1) & \dots & \mathcal{T}(Y_{2,d-1}, \lambda_2 + \nu_{d-1}) & \mathcal{T}(Y_{2d}, \lambda_2) \\ \vdots & \ddots & \vdots & \vdots \\ \mathcal{T}(Y_{p1}, \lambda_p + \nu_1) & \dots & \mathcal{T}(Y_{p,d-1}, \lambda_p + \nu_{d-1}) & \mathcal{T}(Y_{pd}, \lambda_p) \end{pmatrix}$$

and $J_{\mathcal{T}}(\boldsymbol{Y}; \boldsymbol{\lambda}, \boldsymbol{\nu}) = |d\,\mathrm{vec}(\mathcal{T}(\boldsymbol{Y}; \boldsymbol{\lambda}, \boldsymbol{\nu}))/d\,\mathrm{vec}(\boldsymbol{Y})^{\top}|$ is the Jacobian associated with this transformation. Here, the vectorization operator vec converts a matrix into a column vector by stacking its columns on top of each other.

2.2. *Tensor normal distribution.* A generalization of the matrix normal distribution to three dimensions leads to a tensor normal distribution (Basser and Pajevic (2003), Manceur and Dutilleul (2013)) suitable for modeling tensor-valued observations. Let $\mathbb{Y} = (Y_{jht})_{p\times d\times T}$ represent a random tensor described by the following parameters: $p \times d \times T$ mean tensor $\mathbb{M}$, $p \times p$ covariance matrix $\boldsymbol{\Sigma}$, $d \times d$ covariance matrix $\boldsymbol{\Psi}$ and $T \times T$ covariance matrix $\boldsymbol{\Omega}$. The three covariance matrices measure the variability along the dimensions of the three-way tensor. It can be shown that the tensor normal ($t\mathcal{N}$), matrix normal ($m\mathcal{N}$) and multivariate normal ($\mathcal{N}$) distributions are related to each other as follows:

$$\begin{aligned} \mathbb{Y} &\sim t\mathcal{N}_{p\times d\times T}(\mathbb{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}, \boldsymbol{\Omega}) \\ &\equiv \boldsymbol{Y}^{(1)} \sim m\mathcal{N}_{p\times dT}(\boldsymbol{M}^{(1)}, \boldsymbol{\Sigma}, \boldsymbol{\Omega} \otimes \boldsymbol{\Psi}) \\ &\equiv \boldsymbol{Y}^{(2)} \sim m\mathcal{N}_{d\times pT}(\boldsymbol{M}^{(2)}, \boldsymbol{\Psi}, \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}) \\ &\equiv \boldsymbol{Y}^{(3)} \sim m\mathcal{N}_{T\times pd}(\boldsymbol{M}^{(3)}, \boldsymbol{\Omega}, \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}) \\ &\equiv \mathrm{vec}(\mathbb{Y}) \sim \mathcal{N}_{pdT}(\mathrm{vec}(\mathbb{M}), \boldsymbol{\Omega} \otimes \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}). \end{aligned} \tag{3}$$

Here, $\boldsymbol{Y}^{(1)}$, $\boldsymbol{Y}^{(2)}$, and $\boldsymbol{Y}^{(3)}$ (and, similarly, $\boldsymbol{M}^{(1)}$, $\boldsymbol{M}^{(2)}$ and $\boldsymbol{M}^{(3)}$) denote specific modes of unfolding $\mathbb{Y}$ that are defined as follows:

$$\begin{aligned} \underset{p\times dT}{\boldsymbol{Y}^{(1)}} &= [\boldsymbol{Y}_1 \quad \boldsymbol{Y}_2 \quad \dots \quad \boldsymbol{Y}_T], \\ \underset{d\times pT}{\boldsymbol{Y}^{(2)}} &= \big[\boldsymbol{Y}_1^{\top} \quad \boldsymbol{Y}_2^{\top} \quad \dots \quad \boldsymbol{Y}_T^{\top}\big], \\ \underset{T\times pd}{\boldsymbol{Y}^{(3)}} &= [\mathrm{vec}(\boldsymbol{Y}_1) \quad \mathrm{vec}(\boldsymbol{Y}_2) \quad \dots \quad \mathrm{vec}(\boldsymbol{Y}_T)]^{\top}, \end{aligned}$$

where $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_T$ are $p \times d$ matrices constituting the front slices of $\mathbb{Y}$. In equation (3), $\otimes$ denotes the Kronecker product. As we see in Section 3, each of the three different ways to unfold tensors will be helpful for separating and estimating one of the three covariance matrices.

By the property of the Kronecker product, $\boldsymbol{\Omega} \otimes \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = \boldsymbol{\Omega}^{\star} \otimes \boldsymbol{\Psi}^{\star} \otimes \boldsymbol{\Sigma}^{\star}$ for any $\boldsymbol{\Sigma}^{\star} = a_1\boldsymbol{\Sigma}$, $\boldsymbol{\Psi}^{\star} = a_2\boldsymbol{\Psi}$ and $\boldsymbol{\Omega}^{\star} = a_1^{-1}a_2^{-1}\boldsymbol{\Omega}$, where $a_1, a_2 \in IR^{+}$. This gives rise to a nonidentifiability issue which can be addressed by imposing appropriate restrictions on any two of the three covariance matrices. In this paper this nonidentifiability issue is effectively dealt with by letting $|\boldsymbol{\Sigma}| = 1$ and $|\boldsymbol{\Psi}| = 1$.

3. Modeling salary data with tensor mixture model.

In the considered problem, each univeristy can be seen as a $p \times d \times T$ tensor observation with $p = 2$, $d = 3$ and $T = 13$ representing the number of levels in gender, rank and year factors, respectively.

3.1. *Tensor transformation mixture model.*

Based on the concepts introduced in Section 2, a mixture model with tensor normal components can be written as

$$g(\mathbb{Y};\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \boldsymbol{\Phi}(\mathbb{Y}; \mathbb{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k, \boldsymbol{\Omega}_k), \tag{4}$$

where $\boldsymbol{\Phi}$ is the tensor normal pdf and $\mathbb{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k, \boldsymbol{\Omega}_k$ are component-specific parameters. Based on equation (3), the mixture model in (4) can be alternatively written using one of the three unfolding modes. For instance, the third mode yields

$$g(\boldsymbol{Y}^{(3)};\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \Phi(\boldsymbol{Y}^{(3)}; \boldsymbol{M}_k^{(3)}, \boldsymbol{\Omega}_k, \boldsymbol{\Psi}_k \otimes \boldsymbol{\Sigma}_k).$$

Each mean tensor $\mathbb{M}_k$ has $pdT = 78$ unique parameters. Due to the longitudinal nature of the data, gender by rank mean salary matrices can be conveniently modeled against time using polynomial regression. Let $\boldsymbol{X} = (x_{\mathrm{tr}})_{T\times q}$ be the $T \times q$ design matrix with $x_{tr} = t^{r-1}$ for $r = 1, \dots, q$ and $\boldsymbol{B}_k^{(3)}$ be the $q \times pd$ matrix of polynomial regression coefficients. Each column of $\boldsymbol{B}_k^{(3)}$ is associated with a specific gender-rank combination. Under this specification the mode-3 mean matrix $\boldsymbol{M}_k^{(3)}$ can be written as $\boldsymbol{M}_k^{(3)} = \boldsymbol{X}\boldsymbol{B}_k^{(3)}$.

As discussed in Section 1, the normality assumption does not hold for the salary data. The application of the transformation idea in the tensor setting yields the pdf

$$g(\boldsymbol{Y}^{(3)};\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \Phi(\mathcal{T}(\boldsymbol{Y}^{(3)}; \boldsymbol{\Lambda}_k^{(3)}); \boldsymbol{X}\boldsymbol{B}_k^{(3)}, \boldsymbol{\Omega}_k, \boldsymbol{\Psi}_k \otimes \boldsymbol{\Sigma}_k) J_{\mathcal{T}}(\boldsymbol{Y}^{(3)}; \boldsymbol{\Lambda}_k^{(3)}),$$

where $\mathcal{T}(\boldsymbol{Y}^{(3)}; \boldsymbol{\Lambda}^{(3)})$ denotes the mode-3 tensor transformation. Assuming an additive effect of transformation parameters $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_p)^\top$, $\boldsymbol{\nu} = (\nu_1, \dots, \nu_d)^\top$ and $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_T)^\top$ corresponding to each tensor dimension, the transformation matrix $\boldsymbol{\Lambda}^{(3)}$ is given by $\boldsymbol{\Lambda}^{(3)} = \boldsymbol{\zeta}\boldsymbol{1}_{pd}^\top + \boldsymbol{1}_T(\boldsymbol{1}_d \otimes \boldsymbol{\lambda} + \boldsymbol{\nu} \otimes \boldsymbol{1}_p)^\top$. As the third tensor dimension in our framework represents time, it is reasonable to assume the same transformation parameter ζ at all time points, that is, $\boldsymbol{\zeta} = \zeta\boldsymbol{1}_T$. Per discussion in Section 2.1, an additional constraint, such as $\zeta_T = 0$, is needed to make the model identifiable. This leads to $\boldsymbol{\zeta} = \boldsymbol{0}_T$, and, hence, $\boldsymbol{\Lambda}^{(3)} = \boldsymbol{1}_T(\boldsymbol{1}_d \otimes \boldsymbol{\lambda} + \boldsymbol{\nu} \otimes \boldsymbol{1}_p)^\top$. As a result, the mixture model can be written as

$$g(\boldsymbol{Y}^{(3)};\boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k \Phi(\mathcal{T}(\boldsymbol{Y}^{(3)}; \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k); \boldsymbol{X}\boldsymbol{B}_k^{(3)}, \boldsymbol{\Omega}_k, \boldsymbol{\Psi}_k \otimes \boldsymbol{\Sigma}_k) J_{\mathcal{T}}(\boldsymbol{Y}^{(3)}; \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k). \tag{5}$$

Equation (5) presents the mode-3 tensor transformation mixture model that can be effectively used for the analysis of matrix time series data.

While there are relatively few parameters associated with each gender ($p(p+1)/2 = 3$) and rank ($d(d+1)/2 = 6$) covariance matrices, each $\boldsymbol{\Omega}_k$ corresponding to time has $T(T+1)/2 = 91$ unique parameters. Since there is a number of well-known covariance matrix parameterizations in the time series framework, we can explore whether the number of parameters associated with $\boldsymbol{\Omega}_k$ can be reduced without sacrificing the model fit. Most faculty at American universities have their salaries raised every year. While the amount of each raise varies among individuals on a merit basis, the current compensation level should be a strong predictor of the next-year salary. This observation suggests that an autoregressive order one

(AR(1)) time series model is a logical starting point for exploring the behavior of gender-rank salary matrices over time. The immediate explanatory analysis is not possible because the data are heterogeneous, and the partitioning as well as component-specific parameters are unavailable before the model is fitted. The justification of the proposed AR(1) structure upon fitting the model to the data can be found in Section 4.

The covariance structure corresponding to the first order autoregressive time series is given by

$$\boldsymbol{\Omega}_k = \frac{\sigma_k^2}{1-\rho_k^2}\begin{pmatrix} 1 & \rho_k & \dots & \rho_k^{T-1} \\ \rho_k & 1 & \dots & \rho_k^{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k^{T-1} & \rho_k^{T-2} & \dots & 1 \end{pmatrix},$$

where σ_k^2 and ρ_k represent the variance and correlation component-specific parameters. It is easy to show that

$$(6) \qquad |\boldsymbol{\Omega}_k| = \frac{\sigma_k^{2T}}{1-\rho_k^2} \quad \text{and} \quad \boldsymbol{\Omega}_k^{-1} = \frac{1}{\sigma_k^2}(\rho_k^2 \boldsymbol{A}_1 - \rho_k \boldsymbol{A}_2 + \boldsymbol{I}_T),$$

where $\boldsymbol{A}_1$ is a $T \times T$ diagonal matrix defined as $\boldsymbol{A}_1 = \text{diag}\{0, 1, \dots, 1, 0\}$ with $T-2$ ones on the main diagonal and $\boldsymbol{A}_2$ is the sum of the upper and lower shift matrices. The upper shift matrix is a matrix with ones on the superdiagonal and zeros elsewhere while the lower shift matrix is a similarly defined matrix with ones on the subdiagonal. In other words, $\boldsymbol{A}_2 = \text{superdiag}\{1, \dots, 1\} + \text{subdiag}\{1, \dots, 1\}$. Although there are just 13 years in the considered data set, expressions in (6) allow easy and computationally efficient calculations of the determinant and inverse of potentially high-dimensional matrix $\boldsymbol{\Omega}_k$. Also, it is worth reminding that, per discussion in Section 2.2, our model assumes $|\boldsymbol{\Sigma}_k| = 1$ and $|\boldsymbol{\Psi}_k| = 1$ for all values of k and, therefore, $|\boldsymbol{\Psi}_k \otimes \boldsymbol{\Sigma}_k| = |\boldsymbol{\Psi}_k|^p |\boldsymbol{\Sigma}_k|^d = 1$. Then, the Q-function associated with (5) under the above-listed conditions is given by

$$\begin{aligned} Q(\boldsymbol{\Theta}; \dot{\boldsymbol{\Theta}}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \ddot{\tau}_{ik}\Big(&\log \pi_k - \frac{pdT}{2}\log 2\pi - \frac{pd}{2}(T \log \sigma_k^2 - \log(1-\rho_k^2)) \\ &- \frac{1}{2\sigma_k^2}\text{tr}\{(\rho_k^2 \boldsymbol{A}_1 - \rho_k \boldsymbol{A}_2 + \boldsymbol{I}_T)(\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k) - \boldsymbol{X}\boldsymbol{B}_k^{(3)})(\boldsymbol{\Psi}_k \otimes \boldsymbol{\Sigma}_k)^{-1} \\ &\times (\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k) - \boldsymbol{X}\boldsymbol{B}_k^{(3)})^\top\} + \log J_{\mathcal{T}}(\boldsymbol{Y}_i^{(3)}; \boldsymbol{\lambda}_k, \boldsymbol{\nu}_k)\Big). \end{aligned}$$

3.2. *Parameter estimation.* In this section, two or one dot on top of τ_{ik} or $\boldsymbol{\Theta}$ reflects the current and previous iterations, respectively. Noting that $(\boldsymbol{\Psi}_k \otimes \boldsymbol{\Sigma}_k)^{-1} = \boldsymbol{\Psi}_k^{-1} \otimes \boldsymbol{\Sigma}_k^{-1}$, we introduce the following scatter matrices:

$$\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Sigma}} = \sum_{i=1}^{n} \ddot{\tau}_{ik}(\mathcal{T}(\boldsymbol{Y}_i^{(1)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \ddot{\boldsymbol{M}}_k^{(1)})(\dot{\boldsymbol{\Omega}}_k^{-1} \otimes \dot{\boldsymbol{\Psi}}_k^{-1})(\mathcal{T}(\boldsymbol{Y}_i^{(1)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \ddot{\boldsymbol{M}}_k^{(1)})^\top,$$

$$\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Psi}} = \sum_{i=1}^{n} \ddot{\tau}_{ik}(\mathcal{T}(\boldsymbol{Y}_i^{(2)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \ddot{\boldsymbol{M}}_k^{(2)})(\dot{\boldsymbol{\Omega}}_k^{-1} \otimes \ddot{\boldsymbol{\Sigma}}_k^{-1})(\mathcal{T}(\boldsymbol{Y}_i^{(2)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \ddot{\boldsymbol{M}}_k^{(2)})^\top,$$

$$\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Omega}} = \sum_{i=1}^{n} \ddot{\tau}_{ik}(\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \ddot{\boldsymbol{M}}_k^{(3)})(\ddot{\boldsymbol{\Psi}}_k^{-1} \otimes \ddot{\boldsymbol{\Sigma}}_k^{-1})(\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \ddot{\boldsymbol{M}}_k^{(3)})^\top,$$

where the three modes of unfolding tensor $\mathbb{M}_k$ will be useful for estimating the three covariance marices. The expectation and maximization steps of the corresponding EM algorithm are given by the following expressions:

$$\ddot{\tau}_{ik} = \frac{\dot{\pi}_k \Phi(\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \dot{\boldsymbol{\lambda}}_k, \dot{\boldsymbol{\nu}}_k); \boldsymbol{X}\dot{\boldsymbol{B}}_k^{(3)}, \dot{\boldsymbol{\Omega}}_k, \dot{\boldsymbol{\Psi}}_k \otimes \dot{\boldsymbol{\Sigma}}_k) J_{\mathcal{T}}(\boldsymbol{Y}_i^{(3)}, \dot{\boldsymbol{\lambda}}_k, \dot{\boldsymbol{\nu}}_k)}{\sum_{k'=1}^{K} \dot{\pi}_{k'} \Phi(\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \dot{\boldsymbol{\lambda}}_{k'}, \dot{\boldsymbol{\nu}}_{k'}); \boldsymbol{X}\dot{\boldsymbol{B}}_{k'}^{(3)}, \dot{\boldsymbol{\Omega}}_{k'}, \dot{\boldsymbol{\Psi}}_{k'} \otimes \dot{\boldsymbol{\Sigma}}_{k'}) J_{\mathcal{T}}(\boldsymbol{Y}_i^{(3)}, \dot{\boldsymbol{\lambda}}_{k'}, \dot{\boldsymbol{\nu}}_{k'})},$$

$$\ddot{\pi}_k = \frac{\sum_{i=1}^n \ddot{\tau}_{ik}}{n}, \qquad \{\ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k\} = \operatorname*{argmax}_{\boldsymbol{\lambda}_k, \boldsymbol{\nu}_k} Q(\boldsymbol{\Theta}; \dot{\boldsymbol{\Theta}})$$

(where parameters $\boldsymbol{\lambda}_k$ and $\boldsymbol{\nu}_k$ are estimated numerically),

$$\ddot{\boldsymbol{B}}_k^{(3)} = \left(\sum_{i=1}^n \ddot{\tau}_{ik} \boldsymbol{X}_i^\top \dot{\boldsymbol{\Omega}}_k^{-1} \boldsymbol{X}_i\right)^{-1} \sum_{i=1}^n \ddot{\tau}_{ik} \boldsymbol{X}_i^\top \dot{\boldsymbol{\Omega}}_k^{-1} \mathcal{T}(\boldsymbol{Y}_i^{(3)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k),$$

$$\ddot{\boldsymbol{\Sigma}}_k = \frac{\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Sigma}}}{|\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Sigma}}|^{1/p}}, \qquad \ddot{\boldsymbol{\Psi}}_k = \frac{\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Psi}}}{|\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Psi}}|^{1/d}}.$$

The correlation parameter ρ_k of the matrix $\boldsymbol{\Omega}_k$ can be estimated numerically by solving the cubic equation

$$\begin{aligned} &2(T-1)\operatorname{tr}\{\boldsymbol{A}_1 \ddot{\boldsymbol{W}}_k^{\boldsymbol{\Omega}}\}\rho_k^3 - (T-2)\operatorname{tr}\{\boldsymbol{A}_2 \ddot{\boldsymbol{W}}_k^{\boldsymbol{\Omega}}\}\rho_k^2 \\ &\quad - 2(T\operatorname{tr}\{\boldsymbol{A}_1 \ddot{\boldsymbol{W}}_k^{\boldsymbol{\Omega}}\} + \operatorname{tr}\{\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Omega}}\})\rho_k + T\operatorname{tr}\{\boldsymbol{A}_2 \ddot{\boldsymbol{W}}_k^{\boldsymbol{\Omega}}\} = 0, \end{aligned}$$

subject to constraint $|\rho_k| < 1$. Finally, the variance parameter of $\boldsymbol{\Omega}_k$ can be estimated by

$$\ddot{\sigma}_k^2 = \frac{\operatorname{tr}\{(\ddot{\rho}_k^2 \boldsymbol{A}_1 - \ddot{\rho}_k \boldsymbol{A}_2 + \boldsymbol{I}_T)\ddot{\boldsymbol{W}}_k^{\boldsymbol{\Omega}}\}}{pdT \sum_{i=1}^n \ddot{\tau}_{ik}}.$$

This completes the EM algorithm.

The convergence of the algorithm has been detected based on the relative change in log-likelihood values obtained from two consecutive iterations. To initialize the EM algorithm, the *emEM* procedure (Biernacki, Celeux and Govaert (2003)) is employed. At the first stage, multiple *short EM* algorithms are run starting from random points until some lax convergence criterion or a fixed number of iterations (in our case, we used five) is reached. Then, the *long EM* algorithm is run from the most promising point, as measured by the magnitude of the achieved log-likelihood value, till the final convergence is established.

Despite the application-driven character of this paper, the methodological contribution can be seen in providing an effective way of estimating mixture model parameters for multidimensional tensor data, in particular, estimating covariance matrices by means of different unfolding modes. One possible concern in the multidimensional tensor setting is the inversion of potentially high-dimensional covariance matrices $(\boldsymbol{V}_1 \otimes \cdots \otimes \boldsymbol{V}_{\dim(\mathbb{Y})-1})^{-1}$, where $\dim(\mathbb{Y})$ represents the number of tensor factors. However, such a task reduces to the calculation of the Kronecker product of several low-dimensional inverses $\boldsymbol{V}_1^{-1} \otimes \cdots \otimes \boldsymbol{V}_{\dim(\mathbb{Y})-1}^{-1}$. As a result, the developed procedure remains feasible even in multidimensional tensor framework.

4. Model utility for remuneration analysis. In this section we apply the developed model to the salary data set, verify its suitability and address the important questions discussed in Section 1.

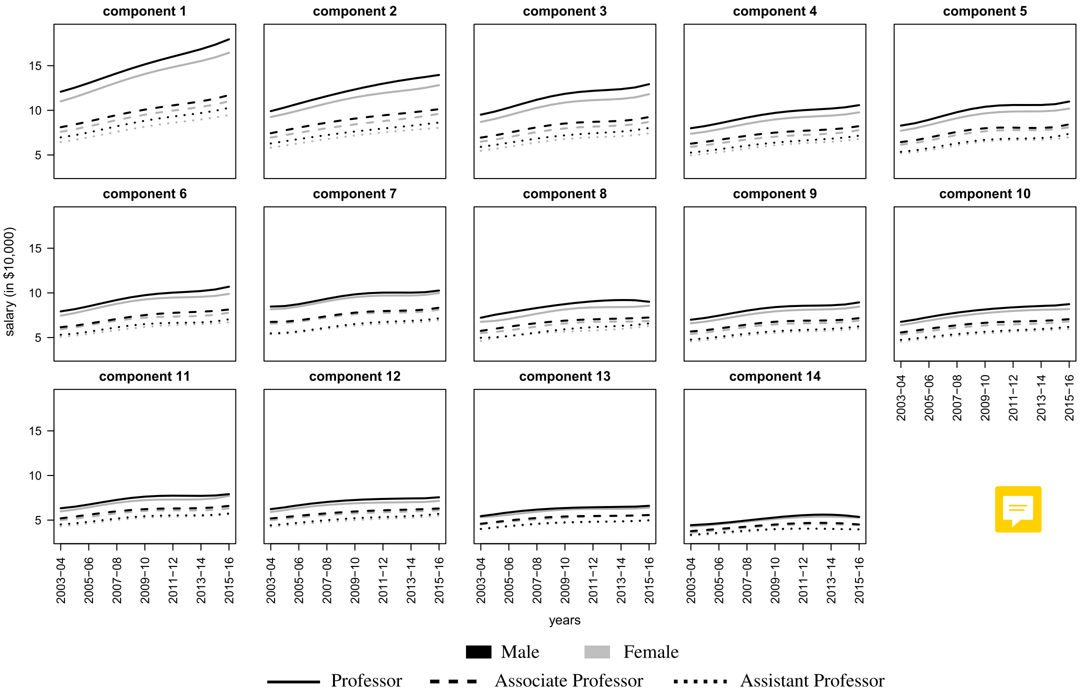

FIG. 1. *Mean salary profiles for the* 14*-cluster solution obtained based on the methodology developed in Section* 3.

4.1. *Optimal model selection and partition analysis.* The model discussed in Section 3 has been applied to the collected data. Per discussion in Section 3.1, the polynomial regression degree $q-1$ should be selected along with the number of mixture components K. We vary q from 1 to 6 (when $q=1$, only the intercept is considered, and the polynomial regression order is zero). The considered mixture order is $K=1,\dots,17$. As a result, 102 different models have been fitted and compared one to another based on BIC.

The best model producing the lowest BIC (710.3) has 14 components incorporating matrix polynomial regressions of degree 4. The AR(1) assumption for covariance matrices $\boldsymbol{\Omega}_k$ was evaluated based on the behavior of the autocorrelation function plotted against the lag for all data groups. Figure S-1 provided in the Supplement (Sarkar, Melnykov and Zhu (2021)) illustrates the close agreement between the sample autocorrelation function and relationship constructed based on the fitted model parameter, $\hat{\rho}_k$, suggesting that the choice of the AR(1) parameterization for covariance matrices $\boldsymbol{\Omega}_k$ is reasonable.

An adequate visual representation of the selected model and corresponding clustering solution is not trivial due to the data dimensionality and relatively large sample size. In this section we study the mean profiles of the 14 components detected and investigate the behavior of skewness parameters. The mean profiles associated with the 14 components are illustrated in Figure 1 by means of panel plots. Within each plot a specific component is shown with six gender-rank mean profiles, assistant, associate and full professors are reflected with dotted, dashed and solid curves, respectively. Two different colors are used to distinguish females (gray) and males (black). The horizontal axis reflects the period of years under consideration while the vertical axis shows the salary size given in $10,000 units. The specific partition of universities corresponding to Figure 1 can be found in Table S-1 in the Supplementary Material (Sarkar, Melnykov and Zhu (2021)). As we can see, although there are some roughly linear or quadratic trends, many components do exhibit curvatures that require polynomials of higher degrees.

TABLE 1

University allocation to clusters by the Carnegie classification. M1, M2 and M3 labels represent master's universities with large, medium and small programs, respectively. R1, R2 and R3 labels denote doctoral universities with the highest, higher and moderate research activity, respectively

	Private institutions						Public institutions						
k	M3	M2	M1	R3	R2	R1	M3	M2	M1	R3	R2	R1	Total
1	0	0	2	1	5	19	0	0	0	0	0	6	33
2	1	1	6	9	4	9	3	1	1	0	10	4	49
3	1	0	3	5	6	1	1	1	4	1	14	55	92
4	1	4	17	0	3	0	3	4	16	4	26	1	79
5	0	0	12	3	1	1	0	1	7	1	2	2	30
6	0	6	19	2	2	1	1	2	5	5	3	1	47
7	0	0	1	0	0	0	1	2	25	2	2	0	33
8	0	3	3	1	0	0	1	1	2	1	1	0	13
9	2	3	11	2	0	0	3	10	41	4	9	1	86
10	5	7	25	3	0	0	4	10	20	8	4	0	86
11	2	4	22	1	0	0	2	5	7	1	0	1	45
12	2	4	10	3	0	0	5	6	18	1	1	0	50
13	5	8	4	0	0	0	6	6	3	0	1	0	33
14	3	7	5	4	0	0	0	1	0	0	0	0	20
Total	22	47	140	34	21	31	30	50	149	28	73	71	696

Among some interesting trends that can be highlighted, males seem to be paid higher than females at all ranks. Moreover, this trend is consistent for all 14 components. This observation and the nature of the salary difference will be further investigated in Section 4.2. Overall, there is a clear trend to increase salaries over time, even though a particular size of raises varies from cluster to cluster dramatically. Moreover, there has been a very limited growth among the lowest paid universities over the past five years. Another interesting observation can be made with regard to a "hump" observed for many components around the years 2009–2012. It can be well explained by salary raises after the economy bounced back after the real estate bubble crisis. Another general remark can be made regarding the rank salary differences: the higher the rank, the higher the salary for all 14 components.

Table 1 presents the obtained partition summarized based on the Carnegie classification as well as the public or private nature of institutions. The highest mean salaries across all categories are observed in universities from cluster 1. Table 1 suggests that among the majority of schools there are private doctoral with high or very high research activity. Also, there are six public institutions with very high research activity. Some representative schools from this group are Princeton University, Stanford University, Rice University, University of Pennsylvania and University of California at Berkeley. The only two master's level institutions in this cluster are Bentley University and Santa Clara University; Clusters 2 and 3 are also characterized by rather high mean salary profiles. While cluster 2 is broadly spread among all research groups of private and public universities, the biggest share of cluster 3 can be claimed by public schools with high or very high research activity. Cluster 3 includes such institutions as Michigan State University, North Carolina State University, Ohio State University, University of Alabama and University of Louisville. Groups 4, 5, 6 and 7 generally present lower mean salary profiles than those in the first three clusters. Cluster 4 primarily consists of large master's schools, such as Drake University, Samford University and University of Minnesota at Duluth as well as public doctoral universities with high research activity, for example, Bowling Green State University, Illinois State University and Michigan Technological University Clusters 5, 6 and 11 can be characterized by a high proportion of large master's institutions,

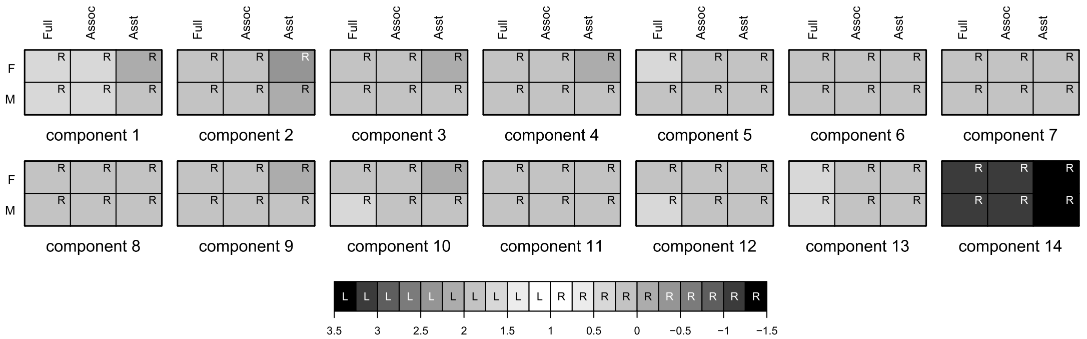

FIG. 2. *Heat map of the estimated skewness parameters associated with gender and faculty rank factors for the* 14*-component solution. For each presented matrix, rows correspond to gender levels* (*female and male*) *and columns represent faculty ranks* (*full professor, associate professor and assistant professor*)*. Color hues annotated with letters "L" and "R" correspond to skewness parameters associated with the left and right skewness, respectively.*

especially from the private sector. Universities from the latter cluster, however, pay considerably lower salaries than those in the former two groups. Clusters 7, 9 and 12 also consist of large master's schools, primarily from the public sector though. It can be also noticed that more than 75% of schools from cluster 7 are from the states of California and Pennsylvania and cluster 9 consists of a large number of universities from the state of New York. Cluster 13 contains master's institutions, especially those with small and medium sizes. The lowest mean salary profile is associated with a rather small cluster 14 that includes just 20 universities. It can be noticed that 40% of the schools in this cluster are located in Puerto Rico. We can conclude that the selected model is reasonable, and the obtained partition corresponds well to the Carnegie classification and public or private university profiles.

Another interesting aspect of the analysis includes the study of skewness parameters. Figure 2 presents the heat map of the estimated skewness parameters associated with gender and faculty rank for each of the 14 clusters. The legend relates gray color hues to the skewness parameter values ranging from -1.5 to 3.5. The different intensity of color hues reflects the degree of skewness. Due to the form of the transformation presented in equation (2), the symmetry is attained at $\lambda = 1$. The hues marked with letter "R" are related to $\lambda < 1$ and represent the cases of right skewness, with more substantial skewness observed for smaller values of λ. Similarly, the hues annotated with "L" correspond to $\lambda > 1$ and are associated with left skewness. In this case, the higher λ values, the larger left skewness. As we can see, all components exhibit the behavior consistent with right skewness. Indeed, this finding is well expected for the salary data. Several interesting remarks can be made based on the presented heat map. Component 14 displays the darkest shades implying the severe skewness in the data. This component contains many universities from the outside of the United States main land, and the corresponding salary data exhibit high variability and skewness. Another interesting result is that salaries of assistant professors are generally more skewed than those of senior faculty. In addition, within the *assistant professor* category, the skewness related to the *female* category is typically higher than that for *male*. This interesting finding is particularly noticeable for components 1, 2, 3, 4, 9 and 10 corresponding to major universities with high research activity and large master's schools. It suggests that there are more cases of extreme salaries among assistant professors, especially females. However, this effect shades out along with the progress to the higher ranks.

4.2. *Gender equity analysis.* As we can see from the profiles in Figure 1, there seems to be a difference in mean salaries that can be associated with the gender factor. In all clusters

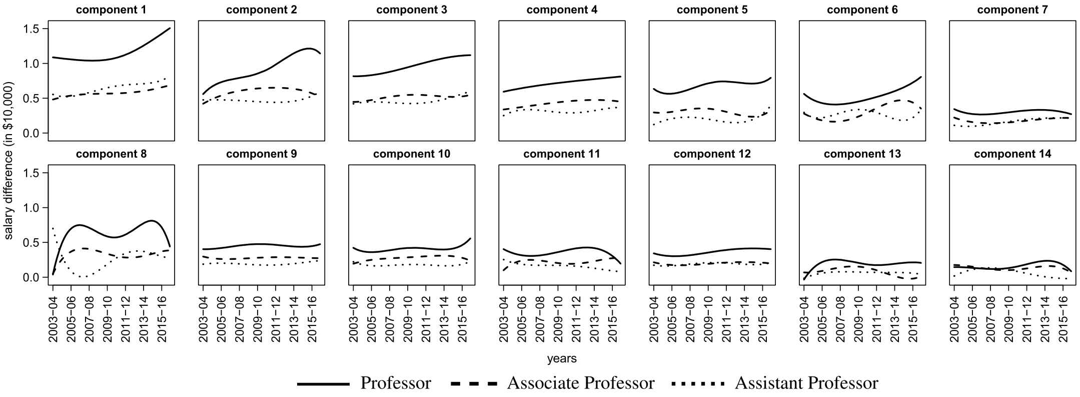

FIG. 3. *Difference in mean salaries (male–female) for the* 14-*component solution.*

and at all ranks, mean salaries of male professors are somewhat higher than those of female professors. In general, the higher the salary of a group is, the larger the difference between male and female professors is observed. Although no cause-and-effect conclusion can be drawn in this observational study, it is still of interest to check whether the gender gap exists or the observed difference is simply the result of data variation. Moreover, Figure 1 suggests that salaries observed for both genders exhibit similar patterns, even though they occur at different levels.

This leads us to two propositions that we aim to explore in this section: (1) the difference between male and female mean salaries remains unchanged over time and (2) there is no difference in mean salaries that can be explained by the gender factor at all. To formulate these propositions mathematically, we consider alternative ways to parameterize the matrix of regression coefficients $\boldsymbol{B}_k^{(3)}$.

Figure 3 presents the differences in mean salaries between the genders for full, associate and assistant professors. As before, the solid, dashed and dotted lines serve to distinguish the three ranks. Overall, we can see that the salary difference remains roughly at the same level for assistant and associate professors across the vast majority of clusters. For full professors some components indicate that the difference may slightly change over time. However, this change is rather minor relative to the data variability and observed just for few high research activity clusters. The differences are almost always positive. This discussion provides additional motivation to study the first proposition.

Using the methodology provided in the Appendix, we find that the BIC value corresponding to the first proposition is -161.6. The observed improvement in BIC is rather dramatic, as the original model has the BIC value of 710.3. Thus, we can conclude that there is no support for the inclusion of different slopes for males and females into the model. In the second proposition we check whether there is support for the inclusion of different intercepts. The value of BIC associated with the model under this proposition is 158.2. The increase in BIC clearly suggests that the role of unequal intercepts in the model is substantial. The conducted analysis supports the claim that there is a gender gap observed in faculty salaries. The nature of the gap is such that starting salaries are unequal, but wage changes over time are essentially the same for both genders. As the final remark of this section, we would like to mention that the clustering solution under the model with equal slopes is very similar to the one presented in Section 4.1. There are just 17 universities that change their membership while the other 679 institutions keep their assignments.

4.3. *Comparison with peers.* As mentioned in Section 1, one of the major questions higher education administration faces is whether the faculty and staff are adequately compensated. While cases of overpays are not so common, remuneration below the market value eventually leads to severe issues such as high faculty turnover as well as the lack of productivity and involvement in routine university life and operation. Determining an objective market value is a complex problem involving multiple socioeconomic characteristics. Therefore, it is a common practice in the American education system to rely on the comparison with peer institutions identified by the upper administration. The composition of the peer list is a stand-alone problem that is beyond the scope of this paper. However, it is worth mentioning that some universities prefer to compare with their rivals while others aim at enhancing their research or teaching profile and hence focus on aspirational peer-to-be schools. In both cases a comparison with the selected peers allows identifying the existing deficiencies and addressing them in order to match the characteristics of the selected peer institutions.

In this section we propose an approach to assess the compensation competitiveness of a certain university to its peers. By construction, the peers are specifically selected by administrators in such a way that they form a group of schools with presumably similar characteristics (including remuneration patterns) and thus required to belong to the same cluster. This leads us to the framework of so-called partially supervised or semisupervised clustering with positive equivalence constraints, that is, a situation with some observations (called a block) required to belong to the same data group. For additional information on positive constraints and related to them negative constraints, we refer the reader to the papers by Shental et al. (2003), Basu, Banerjee and Mooney (2004) and Melnykov, Melnykov and Michael (2016). Indeed, the university under consideration is not guaranteed to belong to the same group with its peers simply because the remuneration records can differ from those of its peers. The decision on whether the university under consideration belongs to the same group with its peers can be made based on the analysis of posterior probabilities obtained in the course of the EM algorithm in the semisupervised clustering setting.

Let B be the number of blocks $\mathcal{B}_1, \dots, \mathcal{B}_B$ so that, within each block, data points are connected with positive constraints and thus must be assigned to the same group. By construction the blocks are mutually exclusive and collectively exhaustive, that is, $\mathcal{B}_{b_1} \cap \mathcal{B}_{b_2} = \varnothing \text{for} b_1 \neq b_2$ and $\bigcup_{b=1}^{B} \mathcal{B}_b = \{1, \dots, n\}$. Such a setup involves a rather minor update of the EM algorithm. In the E-step the posterior probabilities τ_{ik} should be the same for all observations within the same block, that is, for all peer institutions. It can be shown that posterior probabilities for blocks need to be updated by the following expression:

$$\ddot{\tau}_{kb} = \frac{\dot{\pi}_k^{|\mathcal{B}_b|} \prod_{i \in \mathcal{B}_b} \Phi(\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \dot{\boldsymbol{\lambda}}_k, \dot{\boldsymbol{\nu}}_k); \boldsymbol{X}\boldsymbol{B}_k, \dot{\boldsymbol{\Omega}}_k, \dot{\boldsymbol{\Psi}}_k \otimes \dot{\boldsymbol{\Sigma}}_k) J_{\mathcal{T}}(\boldsymbol{Y}_i^{(3)}, \dot{\boldsymbol{\lambda}}_k, \dot{\boldsymbol{\nu}}_k)}{\sum_{k'=1}^{K} \dot{\pi}_{k'}^{|\mathcal{B}_b|} \prod_{i \in \mathcal{B}_b} \Phi(\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \dot{\boldsymbol{\lambda}}_{k'}, \dot{\boldsymbol{\nu}}_{k'}); \boldsymbol{X}\boldsymbol{B}_{k'}, \dot{\boldsymbol{\Omega}}_{k'}, \dot{\boldsymbol{\Psi}}_{k'} \otimes \dot{\boldsymbol{\Sigma}}_{k'}) J_{\mathcal{T}}(\boldsymbol{Y}_i^{(3)}, \dot{\boldsymbol{\lambda}}_{k'}, \dot{\boldsymbol{\nu}}_{k'})},$$

where $|\mathcal{B}_b|$ represents the cardinality of the block $\mathcal{B}_b$, that is, the number of peers. The modification of the corresponding M step is very minor: prior probabilities need to be updated by $\ddot{\pi}_k = n^{-1} \sum_{b=1}^{B} |\mathcal{B}_b| \ddot{\tau}_{kb}$. Other expressions undergo similar straightforward modifications.

Now, we demonstrate how the proposed procedure can be applied to specific universities. While university administration can readily use the presented approach to their schools of interest, the most challenging and interesting cases in this context are the ones with high classification uncertainty, and, thus, we will focus on them. To choose such universities, we calculated the estimated entropy value given by $\hat{\eta} = \sum_{i=1}^{n} \hat{\eta}_i$, where $\hat{\eta}_i = -\sum_{k=1}^{K} \hat{\tau}_{ik} \log \hat{\tau}_{ik}$ represents the contribution of the ith university. It is easy to see that values of $\hat{\eta}_i$ close to zero imply that the university is easy to classify. On the contrary, large $\hat{\eta}_i$ values suggest higher classification uncertainty, with the upper bound being equal to $\log K$. A concept known as a membership extent (White and Murphy (2016)) or effective number of species (Hill (1973)) is closely related to $\hat{\eta}_i$. It is defined as $\exp(\hat{\eta}_i)$ and can be used to estimate the effective number of clusters the ith data point belongs to.

Figure 4 displays entropy distributions by cluster assignments summarized in the form of box plots. Each panel contains box plots representing the six university types according to

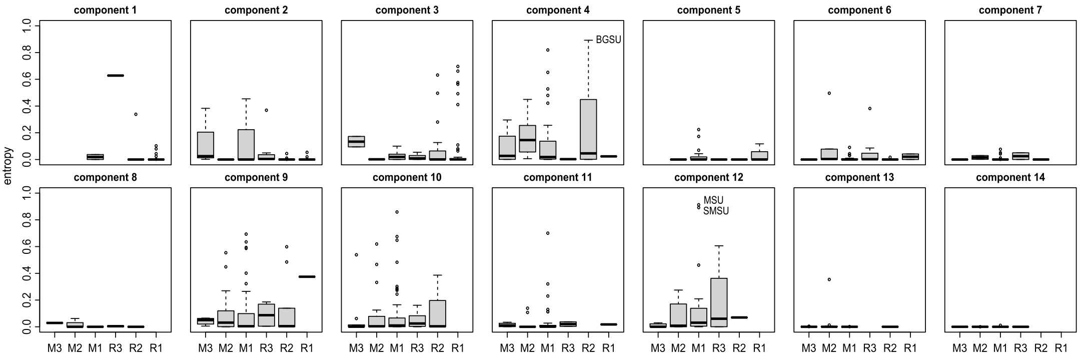

FIG. 4. *Distributions of entropy contributions of American universities arranged by cluster assignments.*

the Carnegie classification. If a box plot corresponding to a certain type is missing, it means that the cluster does not include universities from that particular Carnegie group. As we can see, the highest uncertainty is associated with universities in cluster 4. This cluster includes large master's universities and public doctoral universities with high research activity. After examining the uncertainty of assignments, we notice that many universities from cluster 4 have a high probability of assignment to cluster 3 or 9. The former consists of primarily public universities with highest research activity while the latter includes large master's schools from the public sector. Clusters 9 and 10 also exhibit a large number of universities with uncertain assignments. This uncertainty is mostly explained by the proximity of these two groups, as the largest misclassification probabilities are associated with them. Some uncertainty observed in cluster 12 is also due to the interference with clusters 9 and 10. We can observe that a number of clusters show low variability in posterior probabilities. Among them there are clusters 5, 6, 7, 8, 13 and 14. The vast majority of the institutions in these clusters are master's level.

The three universities with the largest entropy values are labeled in Figure 4. Among them, Bowling Green State University ($\hat{\eta}_i = 0.892$) is from the category *R2* (doctoral university with high research activity). The other two high entropy schools, namely, Murray State University (0.913) and Southeast Missouri State University (0.892) belong to *M1* category (large master's university). Table S-1 in the Supplementary Material (Sarkar, Melnykov and Zhu (2021)) contains the cluster assignment, with the maximum posterior probability and entropy contribution reported for all universities. While many universities have their own lists of peers (oftentimes available just for internal use), per our discussion in Section 1 we illustrate the procedure on the list of peers provided by *The Chronicle of Higher Education*.

The methodology is illustrated first on Bowling Green State University. There are 14 peers of this school: Ball State University, Binghamton University, Illinois State University, Indiana University of Pennsylvania, Miami University Oxford, Northern Arizona University, Northern Illinois University, Ohio University Athens, University of North Carolina Greensboro, University of North Texas, University of Northern Colorado, University of Southern Mississippi, University of Texas Arlington and Western Michigan University If the block of peers is the first one, then block $\mathcal{B}_1$ contains 14 peers while the remaining blocks are all singletons, that is, $|\mathcal{B}_2| = \cdots = |\mathcal{B}_{683}| = 1$. The probability that Bowling Green State University is assigned to the same cluster with its peers is given by $\sum_{k=1}^{K} \hat{\tau}_{1k}\hat{\tau}_{jk}$, where j is the number of the singleton block representing Bowling Green State University. The BIC value obtained as a result of fitting the model with positive constraints is 746.9. It is important to remark that, due to the constraints introduced in semisupervised settings, BIC value is always higher than that in unrestrained framework with every observation being treated as a separate block. The block of 14 peer institutions is assigned to a cluster with posterior probability 1.000. Bowling Green State University is assigned to the same group based on the probability 0.936. Overall, the probability of belonging to the same cluster with its peers is $\sum_{k=1}^{K} \hat{\tau}_{1k}\hat{\tau}_{jk} = 0.936$. This result suggests that Bowling Green State University provides remuneration similar to that at its peer institutions.

Southeast Missouri State University has 11 peers including Arkansas State University main campus, Eastern Kentucky University, Murray State University, Northwest Missouri State University, Southern Illinois University Carbondale, Southern Illinois University Edwardsville, University of Central Missouri, University of Missouri Columbia, University of Missouri Kansas City, University of Missouri St. Louis and Western Kentucky University. The BIC value of the fitted model is 925.03. The block of 11 peer universities is classified to a cluster with the maximum posterior probability 1.000. However, Southeast Missouri State University is assigned to another group with 0.983 chance. Overall, there is probability $\sum_{k=1}^{K} \hat{\tau}_{1k}\hat{\tau}_{jk} = 0.017$ that Southeast Missouri State University belongs to the same salary cluster with its peers. Thus, this institution's remuneration history is rather different from those of its peers.

Murray State University includes 19 universities listed as its peers: Central Connecticut State University, Eastern Illinois University, Eastern Washington University, Frostburg State University, Indiana State University, Northwest Missouri State University, Oakland University, Pittsburg State University, Plymouth State University, Rhode Island College, Southeast Missouri State University, Stephen F. Austin State University, University of Central Missouri, University of Montevallo, University of Nebraska Omaha, University of Tennessee Chattanooga, University of Tennessee Martin, Western Carolina University and Western Illinois University. The BIC value of the fitted model is 921.6. The block of 19 peer universities is classified based on the maximum posterior probability 1.000. Murray State University is assigned to a different cluster based on the maximum posterior probability 0.991. The overall probability that Murray State University compensates its faculty at the same salary level with its peers is 0.009. We conclude that there is strong indication that compensation patterns at this university are different from those of its peers.

As a final comment of this section, we would like to remark that the presented analysis is based on the list of peers provided at *The Chronicle of Higher Education*, and the results can change considerably if the actual list of peers is different.

5. Discussion. This paper studies faculty remuneration at American master's and doctoral universities. Unlike most of the studies conducted in this field and focused on establishing causal relationships between individual faculty remuneration and potentially contributing factors, our objective was to focus at the institutional level and develop a model capable of taking into account the highly heterogeneous and skewed nature of the salary data. The methodological novelty of this paper is in establishing a novel framework for tensor mixture modeling and model-based clustering. Following the approach considered in this paper, tensor observations of any (not just three-way) dimensionality can be effectively modeled. The proposed methodology is useful in applications with tensor-variate observational units.

The obtained clustering result partitions the 696 universities under consideration into 14 groups by compensation patterns observed for gender-rank combinations over 13 years. It shows strong correspondence with the Carnegie classification as well as private or public university profile. The proposed model has been utilized to answer several important and thought-provoking questions. In particular, gender equity, the matter of ultimate importance in labor market, has been investigated. The obtained results are consistent with the gender-related discrepancy in remuneration patterns. Based on the conducted model comparison, a gap in starting salaries favoring male professors is detected for all faculty ranks in the 14 components. Further analysis suggests that salary raise patterns observed over time for male and female professors are indistinguishable.

Another important contribution of the paper is the proposed approach to the comparison of the salary level at a specific university with that of its peers. Nearly every university administration is interested in remunerating faculty adequately, as insufficient compensation traditionally leads to high faculty turnover, lack of involvement, low productivity and other related issues. Thus, the proposed procedure can be useful for evaluating university remuneration policies and calibrating them as needed to provide competitive compensation packages at the peer level. The developed approach is based on the notion of positive equivalence constraints in the semisupervised clustering framework that are utilized to combine peers into a single block. The procedure assesses the chances of the university under consideration to be clustered together with its presumed peers. The employed entropy measure along with the vector of posterior probabilities are used to quantify the uncertainty in the class assignment. This was effectively illustrated in the analysis provided in Section 4.3 where three schools with the highest entropy value have been discussed.

Although the data used in this paper are selfreported, the obtained results are very reasonable and can suggest more comprehensive studies to address other important questions. Some possible future work in this direction includes the analysis of official university salary data, possibly with the division by disciplines or colleges. Such analysis would involve an additional tensor dimension and could be approached using the procedure similar to the one described in this paper with just minor modifications. Another interesting development in this area was proposed by one of the anonymous reviewers involves the inclusion of geographical information into the model, as there could be some state-related aspects of salary behavior. Indeed, a state-related variable along with many other characteristics such as the proximity to metropolitan area, population size, cost of living, school quality, crime and unemployment rates are related to the remuneration process in a very complex and unobvious way. Although very valuable and important, a massive study involving these and other similar features would shift the focus of the paper from salary patterns to the degree of life quality satisfaction of university professors. This interesting project is beyond the scope of our paper and is left for the future consideration.

An important advantage of the analyzed data is their public availability. While the original data can be found at *The Chronicle of Higher Education* web site (https://data.chronicle.com), the cleaned version of the data set is made available through the R package MatTransMix (Zhu and Melnykov (2020)). The code related to this project is available from the authors upon request.

APPENDIX

Per the discussion in Section 3, matrix $\boldsymbol{B}_k^{(3)}$ has dimensions $q \times pd$. We can write $\boldsymbol{B}_k^{(3)} = \begin{bmatrix} \boldsymbol{\beta}_{k,1} & \cdots & \boldsymbol{\beta}_{k,pd} \end{bmatrix}$, where $\boldsymbol{\beta}_{k,j}$ is the q-variate vector of coefficients in the kth component, that is, specific for the jth combination of factor levels. As there are two gender ($p = 2$) and three rank ($d = 3$) levels in our context, we obtain

$$\boldsymbol{B}_k^{(3)} = [\boldsymbol{\beta}_{k1} \quad \boldsymbol{\beta}_{k2} \quad \boldsymbol{\beta}_{k3} \quad \boldsymbol{\beta}_{k4} \quad \boldsymbol{\beta}_{k5} \quad \boldsymbol{\beta}_{k6}],$$

where odd and even columns correspond to females and males, respectively. Vectors $\boldsymbol{\beta}_{k1}$ and $\boldsymbol{\beta}_{k2}$ represent the full professor rank. They are followed by two columns related to associate professors. Finally, vectors $\boldsymbol{\beta}_{k5}$ and $\boldsymbol{\beta}_{k6}$ are associated with the rank of assistant professor.

To address the first proposition, we rewrite matrix $\boldsymbol{B}_k^{(3)}$ in the following way:

$$\boldsymbol{B}_k^{(3)} = \begin{bmatrix} \beta_{k11} & \beta_{k21} & \beta_{k31} & \beta_{k41} & \beta_{k51} & \beta_{k61} \\ \tilde{\boldsymbol{\beta}}_{k1} & \tilde{\boldsymbol{\beta}}_{k2} & \tilde{\boldsymbol{\beta}}_{k3} & \tilde{\boldsymbol{\beta}}_{k4} & \tilde{\boldsymbol{\beta}}_{k5} & \tilde{\boldsymbol{\beta}}_{k6} \end{bmatrix},$$

where β_{kj1} is the intercept coefficient and $\tilde{\boldsymbol{\beta}}_{kj}$ is the part of vector $\boldsymbol{\beta}_{kj}$ that corresponds to regression slopes. The first proposition can be formulated mathematically by setting nonintercept polynomial regression coefficients equal for both genders. It suggests the following restrictions: $\tilde{\boldsymbol{\beta}}_{k1} = \tilde{\boldsymbol{\beta}}_{k2}$, $\tilde{\boldsymbol{\beta}}_{k3} = \tilde{\boldsymbol{\beta}}_{k4}$ and $\tilde{\boldsymbol{\beta}}_{k5} = \tilde{\boldsymbol{\beta}}_{k6}$. The first proposition can be incorporated into the model by replacing the unrestricted matrix $\boldsymbol{B}_k^{(3)}$ with its constrained counterpart,

$$\begin{bmatrix} \beta_{k11} & \beta_{k21} & \beta_{k31} & \beta_{k41} & \beta_{k51} & \beta_{k61} \\ \tilde{\boldsymbol{\beta}}_{k1} & \tilde{\boldsymbol{\beta}}_{k1} & \tilde{\boldsymbol{\beta}}_{k3} & \tilde{\boldsymbol{\beta}}_{k3} & \tilde{\boldsymbol{\beta}}_{k5} & \tilde{\boldsymbol{\beta}}_{k5} \end{bmatrix} \equiv \boldsymbol{c}\boldsymbol{b}_k^\top + \boldsymbol{C}_2 \boldsymbol{B}_k^{[1]} \boldsymbol{C}_1,$$

where the elements of this expression are defined as follow:

$$\boldsymbol{c} = \begin{bmatrix} 1 & \boldsymbol{0}_{q-1} \end{bmatrix}^\top, \qquad \boldsymbol{b}_k = \begin{bmatrix} \beta_{k11} & \beta_{k21} & \beta_{k31} & \beta_{k41} & \beta_{k51} & \beta_{k61} \end{bmatrix}^\top,$$

$$\underset{3\times 6}{\boldsymbol{C}_1} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \qquad \underset{q\times(q-1)}{\boldsymbol{C}_2} = \begin{bmatrix} \boldsymbol{0}_{q-1}^\top \\ \underset{(q-1)\times(q-1)}{\boldsymbol{I}} \end{bmatrix} \quad \text{and}$$

$$\underset{(q-1)\times 3}{\boldsymbol{B}_k^{[1]}} = \begin{bmatrix} \tilde{\boldsymbol{\beta}}_{k1} & \tilde{\boldsymbol{\beta}}_{k3} & \tilde{\boldsymbol{\beta}}_{k5} \end{bmatrix}.$$

The proposed changes can be incorporated into the EM algorithm. The parameters of the model under considered constraints can be estimated from the following two expressions:

$$\ddot{\boldsymbol{b}}_k = \frac{\sum_{i=1}^n \ddot{\tau}_{ik} (\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \boldsymbol{X}_i \boldsymbol{C}_2 \ddot{\boldsymbol{B}}_k^{[1]} \boldsymbol{C}_1)^\top \dot{\boldsymbol{\Omega}}_k^{-1} \boldsymbol{X}_i \boldsymbol{c}}{\sum_{i=1}^n \ddot{\tau}_{ik} \boldsymbol{c}^\top \boldsymbol{X}_i^\top \dot{\boldsymbol{\Omega}}_k^{-1} \boldsymbol{X}_i \boldsymbol{c}},$$

$$\begin{aligned} \ddot{\boldsymbol{B}}_k^{[1]} = {} & \left(\sum_{i=1}^n \ddot{\tau}_{ik} \boldsymbol{C}_2^\top \boldsymbol{X}_i^\top \dot{\boldsymbol{\Omega}}_k^{-1} \boldsymbol{X}_i \boldsymbol{C}_2 \right)^{-1} \\ & \times \left(\sum_{i=1}^n \ddot{\tau}_{ik} \boldsymbol{C}_2^\top \boldsymbol{X}_i^\top \dot{\boldsymbol{\Omega}}_k^{-1} (\mathcal{T}(\boldsymbol{Y}_i^{(3)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) - \boldsymbol{X}_i \boldsymbol{c} \ddot{\boldsymbol{b}}_k^\top)(\dot{\boldsymbol{\Psi}}_k^{-1} \otimes \dot{\boldsymbol{\Sigma}}_k^{-1}) \boldsymbol{C}_1^\top \right) \\ & \times (\boldsymbol{C}_1 (\dot{\boldsymbol{\Psi}}_k^{-1} \otimes \dot{\boldsymbol{\Sigma}}_k^{-1}) \boldsymbol{C}_1^\top)^{-1}. \end{aligned}$$

The rest of the EM algorithm considered in Section 3 remains unchanged.

Under the second proposition, restrictions $\boldsymbol{\beta}_{k1} = \boldsymbol{\beta}_{k2}$, $\boldsymbol{\beta}_{k3} = \boldsymbol{\beta}_{k4}$ and $\boldsymbol{\beta}_{k5} = \boldsymbol{\beta}_{k6}$ ensure that the gender factor is not incorporated in the model at all. Then, matrix $\boldsymbol{B}_k^{(3)}$ from the unrestricted model should be replaced with its constrained version,

$$[\boldsymbol{\beta}_{k1} \quad \boldsymbol{\beta}_{k1} \quad \boldsymbol{\beta}_{k3} \quad \boldsymbol{\beta}_{k3} \quad \boldsymbol{\beta}_{k5} \quad \boldsymbol{\beta}_{k5}] \equiv \boldsymbol{B}_k^{[2]} \boldsymbol{C}_1,$$

where

$$\underset{q\times 3}{\boldsymbol{B}_k^{[2]}} = [\boldsymbol{\beta}_{k1} \quad \boldsymbol{\beta}_{k3} \quad \boldsymbol{\beta}_{k5}].$$

It can be shown that, under this setting, the matrix of coefficients $\boldsymbol{B}_k^{[2]}$ can be estimated in the course of the EM algorithm based on the expression

$$\begin{aligned} \ddot{\boldsymbol{B}}_k^{[2]} = {} & \left(\sum_{i=1}^n \ddot{\tau}_{ik} \boldsymbol{X}_i^\top \dot{\boldsymbol{\Omega}}_k^{-1} \boldsymbol{X}_i \right)^{-1} \left(\sum_{i=1}^n \ddot{\tau}_{ik} \boldsymbol{X}_i^\top \dot{\boldsymbol{\Omega}}_k^{-1} \mathcal{T}(\boldsymbol{Y}_i^{(3)}; \ddot{\boldsymbol{\lambda}}_k, \ddot{\boldsymbol{\nu}}_k) \right) \\ & \times (\dot{\boldsymbol{\Psi}}_k^{-1} \otimes \dot{\boldsymbol{\Sigma}}_k^{-1}) \boldsymbol{C}_1^\top (\boldsymbol{C}_1 (\dot{\boldsymbol{\Psi}}_k^{-1} \otimes \dot{\boldsymbol{\Sigma}}_k^{-1}) \boldsymbol{C}_1^\top)^{-1}. \end{aligned}$$

SUPPLEMENTARY MATERIAL

Supplement to "Tensor-variate finite mixture modeling for the analysis of university professor remuneration" (DOI: 10.1214/20-AOAS1420SUPP; .pdf). Justification of AR(1) covariance structure and partitioning obtained for the 14-component solution discussed in Section 4.1.

REFERENCES

ASHRAF, J. and SHABBIR, T. (2006). Are there racial differences in faculty salaries? *J. Econ. Finance* **30** 306–316.

BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. MR1243494 https://doi.org/10.2307/2532201

BASSER, P. J. and PAJEVIC, S. (2003). A normal distribution for tensor-valued random variables: Applications to diffusion tensor MRI. *IEEE Trans. Med. Imag.* **22** 785–794.

BASU, S., BANERJEE, A. and MOONEY, R. J. (2004). Active semi-supervision for pairwise constrained clustering. In *Proceedings of the Fourth SIAM International Conference on Data Mining* 333–344. SIAM, Philadelphia, PA. MR2388453

BECKER, G. S. (1975). Front matter, human capital: A theoretical and empirical analysis, with special reference to education. In *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education* 1–22 2nd ed. NBER.

BECKER, W. E. and TOUTKOUSHIAN, R. K. (2003). Measuring gender bias in the salaries of tenured faculty members. *New Directions for Institutional Research* **117** 5–20.

BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Comput. Statist. Data Anal.* **413** 561–575. MR1968069 https://doi.org/10.1016/S0167-9473(02)00163-9

BOX, G. E. P. and COX, D. R. (1964). An analysis of transformations. (With discussion). *J. Roy. Statist. Soc. Ser. B* **26** 211–252. MR0192611

BROWNE, R. P. and MCNICHOLAS, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canad. J. Statist.* **43** 176–198. MR3353379 https://doi.org/10.1002/cjs.11246

COHN, E. (1973). Factors affecting variations in faculty salaries and compensation in institutions of higher education. *The Journal of Higher Education* **44** 124–136.

DELORME, C. D. J., HILL, R. C. and WOOD, N. J. (1979). Analysis of a quantitative method of determining faculty salaries. *Journal of Economic Education* **11** 20–25.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38. MR0501537

FAIRWEATHER, J. S. (1993). Faculty reward structures: Toward institutional and professional homogenization. *Res. High. Educ.* **34** 603–623.

FERREIRA, A. P. (2013). Are all faculty members being compensated equally? A multi-method approach to investigating faculty salary. Open Access Dissertation.

FRANCZAK, B. C., BROWNE, R. P. and MCNICHOLAS, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Trans. Pattern Anal. Mach. Intell.* **36** 1149–1157.

FRIEDMAN, M. and KUZNETS, S. (1945). *Income from Independent Professional*. National Bureau of Economic Research, New York.

GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2017). A matrix variate skew-*t* distribution. *Stat* **6** 160–170. MR3653050 https://doi.org/10.1002/sta4.143

GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognit.* **80** 83–93.

GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2019). Package MatSkew: Matrix skew-*t* parameter estimation.

GALLAUGHER, M. P. B. and MCNICHOLAS, P. D. (2019c). Three skewed matrix variate distributions. *Statist. Probab. Lett.* **145** 103–109. MR3873895 https://doi.org/10.1016/j.spl.2018.08.012

HEARN, J. C. (1999). Pay and performance in the university: An examination of faculty salaries. *The Review of Higher Education* **22** 391–410.

HEXTER, H. (1990). Faculty salaries in perspective. *Research Briefs* **1**.

HILL, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology* **54** 427–432.

LEE, S. X. and MCLACHLAN, G. J. (2013a). Model-based clustering and classification with non-normal mixture distributions. *Stat. Methods Appl.* **22** 427–454. MR3127088 https://doi.org/10.1007/s10260-013-0237-4

LEE, S. X. and MCLACHLAN, G. J. (2013b). On mixtures of skew normal and skew t-distributions. *Adv. Data Anal. Classif.* **7** 241–266. MR3103965 https://doi.org/10.1007/s11634-013-0132-8

LIN, T.-I., HO, H. J. and LEE, C.-R. (2014). Flexible mixture modelling using the multivariate skew-t-normal distribution. *Stat. Comput.* **24** 531–546. MR3223539 https://doi.org/10.1007/s11222-013-9386-4

LO, K. and GOTTARDO, R. (2012). Flexible mixture modeling via the multivariate t distribution with the Box–Cox transformation: An alternative to the skew-t distribution. *Stat. Comput.* **22** 33–52. MR2865054 https://doi.org/10.1007/s11222-010-9204-1

MANCEUR, A. M. and DUTILLEUL, P. (2013). Maximum likelihood estimation for the tensor normal distribution: Algorithm, minimum sample size, and empirical bias and dispersion. *J. Comput. Appl. Math.* **239** 37–49. MR2991957 https://doi.org/10.1016/j.cam.2012.09.017

MANLY, B. F. J. (1976). Exponential data transformations. *Biometrics Unit* **25** 37–42.

MELGUIZO, T. and STROBER, M. H. (2007). Faculty salaries and the maximization of prestige. *Res. High. Educ.* **48** 633–668.

MELNYKOV, V., MELNYKOV, I. and MICHAEL, S. (2016). Semi-supervised model-based clustering with positive and negative constraints. *Adv. Data Anal. Classif.* **10** 327–349. MR3541239 https://doi.org/10.1007/s11634-015-0200-3

MELNYKOV, V. and ZHU, X. (2018). On model-based clustering of skewed matrix data. *J. Multivariate Anal.* **167** 181–194. MR3830641 https://doi.org/10.1016/j.jmva.2018.04.007

MELNYKOV, V. and ZHU, X. (2019). Studying crime trends in the USA over the years 2000–2012. *Adv. Data Anal. Classif.* **13** 325–341. MR3935201 https://doi.org/10.1007/s11634-018-0326-1

MINCER, J. (1958). Investment in human capital and personal income distribution. *J. Polit. Econ.* **66** 281–302.

MOHANTY, D., DODDER, R. and KARMAN, T. (1986). Faculty salary analysis by region, rank, and discipline from 1977–1978 to 1983–1984. *Res. High. Educ.* **24** 304–317.

O'HAGAN, A., MURPHY, T. B., GORMLEY, I. C., MCNICHOLAS, P. D. and KARLIS, D. (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Comput. Statist. Data Anal.* **93** 18–30. MR3406193 https://doi.org/10.1016/j.csda.2014.09.006

PERNA, L. W. (2001). Sex differences in faculty salaries: A cohort analysis. *The Review of Higher Education* **24** 283–307.

RIPPNER, J. A. and TOUTKOUSHIAN, R. K. (2015). The 'Big Bang' in public and private faculty salaries. *Journal of Education Finance* **41** 103–123.

SARKAR, S., MELNYKOV, V. and ZHU, X. (2021). Supplement to "Tensor-variate finite mixture modeling for the analysis of university professor remuneration." https://doi.org/10.1214/20-AOAS1420SUPP

SARKAR, S., ZHU, X., MELNYKOV, V. and INGRASSIA, S. (2020). On parsimonious models for modeling matrix data. *Comput. Statist. Data Anal.* **142** 106822, 26. MR3992446 https://doi.org/10.1016/j.csda.2019.106822

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHENTAL, N., BAR-HILLEL, A., HERTZ, T. and WEINSHALL, D. (2003). Computing Gaussian mixture models with EM using equivalence constraints. In *Advances in NIPS* **15**.

SIMPSON, W. B. (1981). Faculty salary structure for a college or university. *The Journal of Higher Education* **52** 219–236.

SNYDER, J. K., HYER, P. B. and MCLAUGHLIN, G. W. (1994). Faculty salary equity: Issues and options. *Res. High. Educ.* **35** 1–19.

TOUTKOUSHIAN, R. K., BELLAS, M. L. and MOORE, J. V. (2007). The interaction effects of gender, race, and marital status on faculty salaries. *The Journal of Higher Education* **78** 572–601.

UMBACH, P. D. (2007). Gender equity in the academic labor market: An analysis of academic disciplines. *Res. High. Educ.* **48** 169–192.

VIROLI, C. (2011a). Finite mixtures of matrix normal distributions for classifying three-way data. *Stat. Comput.* **21** 511–522. MR2826689 https://doi.org/10.1007/s11222-010-9188-x

VIROLI, C. (2011b). Model based clustering for three-way data structures. *Bayesian Anal.* **6** 573–602. MR2869958 https://doi.org/10.1214/11-BA622

VIROLI, C. (2012). On matrix-variate regression analysis. *J. Multivariate Anal.* **111** 296–309. MR2944423 https://doi.org/10.1016/j.jmva.2012.04.005

WHITE, A. and MURPHY, T. B. (2016). Exponential family mixed membership models for soft clustering of multivariate data. *Adv. Data Anal. Classif.* **10** 521–540. MR3575730 https://doi.org/10.1007/s11634-016-0267-5

YEO, I.-K. and JOHNSON, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika* **87** 954–959. MR1813988 https://doi.org/10.1093/biomet/87.4.954

ZHU, X. and MELNYKOV, V. (2018). Manly transformation in finite mixture modeling. *Comput. Statist. Data Anal.* **121** 190–208. MR3759207 https://doi.org/10.1016/j.csda.2016.01.015

ZHU, X. and MELNYKOV, V. (2020). MatTransMix: An R package for clustering matrices. R package version 0.1.9.