Analytics, Computational Intelligence and Information Management

Reconciling business analytics with graphically initialized subspace clustering for optimal nonlinear pricing

Claire Y. T. Chen\textsuperscript{a,b}, Edward W. Sun\textsuperscript{c,d}, Wanyu Miao\textsuperscript{d}, Yi-Bing Lin\textsuperscript{c,f,g}

\textsuperscript{a} Montpellier Business School, France
\textsuperscript{b} MRM, University of Montpellier, France
\textsuperscript{c} KEDGE Business School, France
\textsuperscript{d} Olin Business School, Washington University in St. Louis, USA
\textsuperscript{e} College of AI, National Yang Ming Chiao Tung University, Taiwan
\textsuperscript{f} Miin Wu School of Computing, National Cheng Kung University, Taiwan
\textsuperscript{g} China Medical University Hospital, Taiwan

\section*{A R T I C L E   I N F O}

Article history:
Received 16 December 2022
Accepted 7 July 2023
Available online xxx

Keywords:
Analytics
Machine learning
Multivariate mixture distribution
Spectral subspace clustering
Pricing

\section*{A B S T R A C T}

The relationship between price and quantity in nonlinear pricing transcends simple proportionality, as conditional rebates and discounts may be contingent upon the quantity of goods or services purchased by consumers. This dynamic introduces significant challenges for both consumers and business operators because incomplete information arises from the inherent uncertainty of consumer behavior. In light of this, the present research elucidates the pursuit of optimal nonlinear pricing strategies by business operators through an innovative data-driven approach. Our contributions encompass two distinctive facets: a novel unsupervised spectral clustering method, termed graphically initialized subspace clustering, and a decision optimization framework. The proposed data-driven method introduces an optimization problem aimed at minimizing subspace partitioning costs, leveraging the efficient utilization of a mixture multivariate skewed t distribution to effectively capture heavy users and to characterize their parametric behavioral patterns. In addition, the decision optimization component builds upon the aforementioned method, employing a convex optimization algorithm to enable seamless modification of attributes in nonlinear pricing, while ensuring revenue consistency during pre- and post-modification. Notably, we substantiate the interpretability and practical applicability of our proposed methodology in the realm of business analytics, through empirical analysis utilizing real-world data obtained from a cellular carrier. The findings of this study confirm the efficacy and viability of our approach in enabling business operators to navigate the complexities of nonlinear pricing optimization with confidence and informed decision-making.

© 2023 Elsevier B.V. All rights reserved.

\section*{1. Introduction}

\subsection*{1.1. Motivation and problem addressed}

Cellular carriers are encountering increasing challenges to provide reasonable billing plans to heavy users (whose actual consumption exceeds the subscribed allowance) while ensuring a high-quality experience for all other consumers (see Essegaier et al., 2002; Kwon et al., 2016) under nonlinear pricing (see Chao et al., 2022; Khouja et al., 2023, for example) that contains free units (usage allowance) at a fixed (access) price and a price per unit above the allowance (overage price) (see Fibich et al., 2017). When consuming more than their allowance, customers either need to pay an overage fee or receive degraded service (Remerova et al., 2014), thus experiencing an “overage disutility” (Huang et al., 2021). Since the flat rate is not sustainable, if cellular carriers cannot charge rates that match the service (Kim & Hwang, 2009) and at the same time cellular carriers prefer to be differentiated from their competitors by providing personalized services (Sun et al., 2015), then the solution leads to a highly flexible tiered service and pricing for different markets and consumers (Gopalakrishnan et al., 2015; Julien et al., 2022).

It is difficult for cellular carriers to offer perfect billing plans if they lack an effective system that can accurately calculate rates and correctly charge for services in real time (Cooper et al., 2001).
Although tiered pricing offers broad functionality and affordability that create a less congested, bandwidth-optimized network with the flexibility to offer different types of plans for high and low bandwidth consumers, telecommunications operators still need to overcome the difficulties of gaining sufficient insight into individual subscriber preferences (Abolfathi et al., 2022b) and having the means to sell higher-end related plans to those who exceed their usage limits (Iyengar et al., 2008). In particular, fully portraying a subscriber’s consumption profile often involves constraints such as whether to collect privacy data (Tucker, 2014), whether there are scenarios that force the use of privacy data (e.g., location positioning) (Crossler & Bélanger, 2019), and whether to collect privacy data beyond the consumer’s permission (Godinho de Matos & Adjerid, 2022), thus increasing the difficulty in pricing optimization via data-driven analytics.

In view of the above-mentioned difficulties faced by cellular carriers in adopting tiered pricing, this paper addresses the following two questions.

1. How to classify a relatively reasonable price component (e.g., allowance) of nonlinear pricing based on the actual consumption extracted from real data, particularly for heavy users who consume more than their subscription?

2. How to optimize the price components (i.e., access fees, allowance, and overage surcharges) of nonlinear pricing based on the actual usage of consumers so as to assist cellular carriers’ strategic possibility to achieve their established revenue goals?

1.2. Contributions and related literature

In this study, we analyze real consumption data obtained from a cellular carrier to address two research questions. Our contribution to the field is two-fold. First, we propose a novel approach for spectral subspace clustering that is designed to handle high-dimensional data with a significant number of outliers, or extreme values. Our method employs a mixture of finite skew t distributions to cluster the data points, which are assumed to be finite, independent, and identically distributed, in the presence of arbitrary anomalous data points, such as heavy users, in a high-dimensional space. To initialize the clustering process, we employ graphical partitioning techniques. Second, based on the clustering results obtained from our proposed method using real consumption data, we develop a business analytics approach based on convex optimization. This approach aims to determine the optimal nonlinear pricing strategy that aligns with the observed consumption patterns. Overall, our study presents solutions to the research questions through the application of machine learning techniques and provides insights into the use of business analytics for pricing optimization in the context of cellular carrier data analysis.

1.2.1. A novel spectral subspace clustering method

Most popular clustering methods are centroid based (for example, the k-means algorithm), agglomeration based (or hierarchical), or graph based (see Fountoulakis et al., 2019; García Trillos & Slepčev, 2018; Peng & Wei, 2007, for example). In addition, model-based (or probabilistic) clustering is also used in a wide variety of applications such that the assignment of points to specific clusters is given as a probability (see Duan & Dunson, 2021). Although the clustering results from the centroid-based clustering method can be described and evaluated intuitively, it has two obvious drawbacks. First, it is sensitive to the kernel representing the within-cluster density particularly when there exist outliers, and second, its initialization to clusters is difficult for high-dimensional data (Srivastava et al., 2023).

Conventional spectral clustering methods treat each data point (or distribution) as a vertex of the graph and convert clustering to a graph partition problem for attributing vertices to a disjoint set (cluster). The similarity between any pair of data points (or distributions) is then viewed as the weighted edge connecting them. An intuitive two-dimensional method applies a Gaussian kernel to define the Euclidean distance and to transform it into the corresponding similarity measure. Note that here the similarity measure should be clearly defined (see Eisenach & Liu, 2019, for example). Two different ways to construct a graphic similarity measure off between data points have been usually applied: radius neighborhood graph and k-nearest neighbor graph. When the data points have different scales (i.e., distances between data points are heterogeneous) located in different regions of the space, the former encounters difficulties; while when the data points have different moments (i.e., high dispersion in same regions of the space), the latter exhibits deficiency (see Nascimento & de Carvalho, 2011, and references therein).

Several latest studies attempt to deal with these issues. For example, Rota Bulò & Pelillo (2017) provide a review of dominant sets in the clustering of edge-weighted graphs. Fountoulakis et al. (2019) show that an optimal algorithm is able to behave in a completely local manner by accessing only a small number of nodes without accessing the entire graph. Borgwardt & Happich (2019) use the polyhedral theory to characterize piece-wise linear separability of nodes, and Ah-Pine (2022) emphasize the importance of an affinity matrix for measuring edges in spectral clustering and propose an optimization algorithm to learn a doubly stochastic and nearly idempotent affinity matrix. Working on maximum (similarity) or minimum (dissimilarity) spanning trees of a graph, Caraballo et al. (2021) consider measure clustering quality with the ratio between the external separation and the internal homogeneity. Srivastava et al. (2023) proposed a robust approach of spectral clustering for high-dimensional data. They first reduce dimensions with PCA and then estimate the scaling parameter of Gaussian kernel with a predefined quantile to compute the similarity between two arbitrary points. Their method conducts clustering based on a mixture Gaussian distribution whose initial parameters are determined by k-means.

Spectral clustering performs well for clustering high-dimensional spatial data, but lacks descriptive or intuitive representations of the clustering results. In addition, Srivastava et al. (2023) point out the inefficiency of spectral clustering due to even a small number of outliers. Several studies have looked to address these shortcomings in clustering methods. Considering the complexity of data points, especially the presence of extreme values, many of the above approaches fail to address the problem in depth. Several ways to improve these clustering methods are suggested, for example, using Bayesian methods. Taking advantage of the property of pairwise differences between data points, Duan & Dunson (2021) propose a Bayesian distance clustering method that focuses on the likelihood of modeling pairwise distances to ensure robustness of the implied density.

We propose a new methodology named graphically initialized subspace clustering (GISC), specifically for capturing the behavior of heavy users, to determine the different brackets of tiered pricing based on real-time usage data. Subspace clustering is a high-dimensional technique to find valid clusters that are defined by only a subset of dimensions (i.e., a selection of multiple dimensions without necessarily having the agreement of features in all dimensions) with the possibility of overlapping both in the space of features and observations. The core of our proposed method is to cluster real-time usage data by using a multivariate mixture skew t distribution, and then to classify the tiered brackets of data plans for price differentiation with the mean of the distribution function in each cluster. The proposed GISC has two working blocks: graphic initialization to determine the centroids of overlapping clusters and EM (expecta-
tion and maximization) algorithm to parameterize the underlying distribution.

To address the computational exhaustion associated with analyzing large graphs in graphic initialization of spectral clustering, the local spectral graph clustering method is used to identify well-connected clusters around a given reference node (i.e., the seed set) without having to visit the entire graph. In our initialization process of determining the reference nodes, we propose a novel approach using an infinity norm\(^1\) for proper initialization of high-dimensional data. Applying our proposed variational formulation (i.e., Propositions 1 and 2 in Section 2.1) allows to consider only the most destructive number of non-zero points without calculating the mutual dissimilarity of all points in the entire graph. Since the variable clustering is an attempt to treat the covariance of graphically presented points with a measure of distances between points, like combinatorial optimization in general, it is NP-hard (see Ah-Pine, 2022). Therefore, Eisenach & Liu (2019) point out that fast algorithms that have been proposed to solve clustering problems are not guaranteed to produce an optimal solution to the original problem. In our proposed method, we convert the NP-hard optimization problem in a solvable way by introducing an optimization problem of reducing partitioning cost and to make the underlying clustering method explainable.

To the best of our knowledge, the proposed clustering method is the first time to combine optimized high-dimensional graphs partitioning with a multivariate mixture skew \(t\) distribution in which higher-order moments are required (see Nevo et al., 2016) in order to effectively solving the optimal nonlinear pricing problem that we shall discuss in Section 1.2.2 as follows.

### 1.2.2. Optimal nonlinear pricing

Nonlinear pricing establishes a price that is not necessarily proportional to the quantity a consumer purchases, and considers various rebates and discounts conditioned on that quantity (see Chao et al., 2022, for example). As a nonlinear pricing plan, three-part tariff (3PT) provides the underlying subscribers a number of free units (usage allowance) at a fixed (access) price and a price per unit above the allowance (overage price) in one billing cycle (see Bibich et al., 2017, for example). Therefore, the marginal price of usage is zero until the allowance is exceeded (Nevo et al., 2016). Other examples of nonlinear pricing contracts used in the telecommunications include pay-per-use (or pay-as-you-go) contracts, two-part tariffs with an access fee and a per-unit usage price, and unlimited usage contracts (Leider & Sahin, 2014). Usage-based pricing is effective at lowering usage without significantly reducing consumer welfare in comparison with unlimited plans and shifts surplus from consumers to providers (Nevo et al., 2016).

Nevo et al. (2016) point out that plan selection and information on usage (during the billing cycle) are important to identify the distribution of types. The joint distribution is highly irregular and looks obviously different from a normal distribution. High volume subscribers have substantial variation in their elasticity of demand. Consumer segments are heterogeneous in response to marketing tactics and differ in sensitivity to price changes (see Chen et al., 2020; Geng & Shulman, 2015, and reference therein). Bibich et al. (2017) compute the optimal 3PT plan as a non-smooth nested optimization problem and suggest that the carriers should consider all three levels of the tariff plan (fixed fees, unit allowances, and overage fees) to discriminate among consumer segments when consumers are heterogeneous. Their model assumes that a monopoly service provider is selling to a market of only two segments (light and heavy user) of risk-neutral consumers and implies that the optimal policy should reduce the monthly allowance and access fee, but increase overage price.

Using a dataset of 70,510 fee-based checking accounts over 30 months, Ater & Landsman (2013) report findings of overage aversion; that is, customers’ aversion to the disutility of overage fee leads them to accept higher fixed monthly payments for the plans with larger allowances. By having well-defined plan-specific variables such as speed and prices, Nevo et al. (2016) investigate the demand for residential broadband and provide a method of demand estimation based on two types of variation: variation in prices with attributes of the subscribed plan and variation observed from the 3PT plans. Based on primary data of individual monthly billing records of cellular phones, Grubb & Osborne (2015) estimate a structural model of contract choice and usage in cellular phone services and show that forward-looking consumers are inattentive to past usage and underestimate both the mean and variance of future usage. In addition, they point out that multi-part tariffs induce marginal-price uncertainty and that higher fixed fees make consumers worse off with counterfactual simulations. Geng & Shulman (2015) investigate strategic overage pricing and indicate that using an overage fee to discourage consumption does not automatically improve profit. After theoretically investigating 3PT pricing, Bhargava & Gangwar (2018) show that 3PT is able to increase revenue and such an improvement arises from deploying an allowance for differential marginal pricing. They show that 3PT pricing confers a revenue advantage under certain demand settings in which mixture distribution and other biases (e.g., uncertainty or over-confidence) are not relevant to the allowance. From an experimental investigation, Leider & Sahin (2014) find that a very small discount of 3PT could increase revenue in comparison with only offering a pay-per-use contract. After estimating the structural preference parameters for service bundles under 3PT with a dataset from a Chinese mobile service provider, Chen et al. (2019) show that access fees have a larger impact on consumer plan choice, usage decisions, and firm revenues.

Following Nevo et al. (2016) who consider variation in mean usage across different states, our proposed method investigates the allowance determined with the mean (i.e., \(\mu\)) estimated by the parametric distributions. In addition, the multivariate mixture skew \(t\) distribution has well-defined high moments (see Adcock, 2014, for example), which satisfy the points highlighted by Nevo et al. (2016) such that higher-order moments of the usage distribution efficiently summarizes the information. In our optimization problem we incorporate the marginal-price uncertainty in Grubb & Osborne (2015), overage aversion in Ater & Landsman (2013) for overage fee modification, and consumers’ forward-looking behavior in Xu et al. (2019). Different from the existing literature, we show that increasing the allowance by actual data usage and overage fee (not exceeding 25%) while reducing the access fee slightly could be optimal for cellular carriers.

### 1.3. Outline

The rest of the paper is organized as follows. Section 2 introduces the development of our method. We begin with a graphical initialization for a similarity measure where we convert the minimization of partitioning cost into a restrict optimization problem. We then show a way to determine the initial value and the complete expectation-maximization (EM) algorithm to estimate the parameters of multivariate mixture skew \(t\) distribution. At the end, we investigate the performance of the proposed method with two

---

1 In mathematical analysis, infinity norm is also called uniform norm, supremum norm, or Chebyshev norm and assigns to real- or complex-valued bounded functions \(f\) defined on a set \(S\) the non-negative number: \(\|f\|_{\infty} = \sup\{|f(x)| : x \in S\}\). If \(f\) is a continuous function on a compact set, then it is bounded and the supremum can be replaced by the maximum. In case, the norm is also called maximum norm such that for some vector \(x\) in finite-dimensional coordinate space and \(x = (x_1, x_2, \ldots, x_n)\), it takes the form of \(\|x\|_{\infty} = \max(|x_1|, |x_2|, \ldots, |x_n|)\).
well-known datasets for clustering validation. Section 3 applies the proposed clustering methods on a real dataset of mobile data consumption for business analytics (i.e., descriptive and prescriptive analytics) on optimal nonlinear pricing of cellular carriers and shows our results based on convex optimization. Section 4 concludes.

2. The methodology

As we mention in Section 1.2.1, different from the conventional graph-based clustering methods, we start with defining a novel graphical initialization of similarity measure in Section 2.1. Data points (distributions) for clustering are characterized by a multivariate mixture skew $t$ distribution in Section 2.1.2, and the parameter estimation with EM algorithm is present in Section 2.2. We provide a summary of notations in Appendix A.

2.1. Graphic initialization for similarity measure

We use a undirected graph $G(\Theta, \varepsilon)$ to characterize the dataset, where $\Theta = \{x_i \in \mathbb{R}^M | i = 1, \ldots, N\}$ stands for the set of nodes or vertices $x_i$ (i.e., data points or distributions) and $\varepsilon = \{e_{ij} | i, j = 1, \ldots, N, \text{ and } i \neq j\}$ for the set of edges between any arbitrary nodes. In our proposed method, we treat clustering as an optimization problem of effectively searching for a graph cut of two disjoint sets in order to minimize the total weights of all edges that cross such a partition. The minimal cut paradigm is practically replaced by a normalized cut that measures both the total dissimilarity between the different clusters and the total similarity within the clusters (see Shi & Malik, 2000, for example). However, the normalized cut is NP-hard (non-deterministic polynomial-time hard) and penalizes any cluster with a small size, thus achieving more balanced clusters (see Srivastava et al., 2023, for example). To overcome these problems, we propose a new method based on infinity norm to define the similarity measure.

2.1.1. Identification of dissimilarity with minimal cost

A conventional adjacency matrix is a square matrix for representation of a finite graph. The rows and columns of the matrix are associated with the vertices (nodes) of the graph, while the entries in the matrix indicate the presence or absence of edges (connections) between pairs of vertices. In the specific context of this study, the graph being analyzed is undirected, meaning that edges are bidirectional and have no inherent directionality. As a result, the adjacency matrix for this graph is symmetric, with entries mirroring each other along the main diagonal.

**Definition 1.** Given $\varepsilon = \{e_{ij} | i, j = 1, \ldots, N, \text{ and } i \neq j\}$, let $\varepsilon\in \mathbb{R}^{N \times N}$ be an adjacency matrix, which is a symmetric matrix, representing a finite graph $G(\Theta, \varepsilon)$, such that:

$$\forall e_{ij} = 1 \quad \text{if} \quad e_{ij} \neq 0; \quad 0 \quad \text{otherwise},$$

where $\forall e_{ij}$ is an indicator function of $e_{ij}$, and $e_{ij} = \sqrt{(|x_i|_\infty - |x_j|_\infty)^2}$, which is the distance between $|x_i|_\infty$ and $|x_j|_\infty$.

When contrasted with the $l_1$-norm and $l_2$-norm, the infinity norm ($l_\infty$-norm) is a measure that specifically captures the maximum absolute value within a vector $x$. This norm is commonly employed in various optimization problems and numerical analysis methods, as it provides insight into the largest magnitude element in the vector. Studies (see Franco et al., 2018; Novak et al., 2022, for example) exemplify the utilization of the $l_\infty$-norm in the relevant literature, underscoring its significance in the field.

**Remark (1).** If $|x_i|_\infty = |x_j|_\infty$, then $e_{ij} = 0$, which is indicated by the adjacency matrix as 0; i.e., “no edge” or “no connection” between $x_i$ and $x_j$. For any $k, k \neq i, j$, we have $e_{ik} = e_{kj}$ since $\sqrt{(|x_i|_\infty - |x_k|_\infty)^2} = \sqrt{(|x_j|_\infty - |x_k|_\infty)^2}$.

Since our goal is to identify non-trivial outliers with the maximum value, we use the term “trivial” to describe values that are dominated by the maximum.

**Remark (2).** In the case where both $x_i$ and $x_j$ are at least a distance of $e_{ij}$ away from $x_k$, and $x_k$ become trivial in comparison to $x_k$ whenever $e_{ij} < e_{ik}$ and $e_{ij} < e_{kj}$. However, in rare cases where $e_{ij}$ is trivially small (i.e., approaches to zero), for some “threshold” distance $\varepsilon$, $x_i$ and $x_j$ become trivial in comparison to $x_k$ whenever $e_{ij} < e_{ik} + \varepsilon$ and $e_{ij} < e_{kj} + \varepsilon$.

In practice, it is uncommon to encounter two heavy users with exactly the same usage. Therefore, even when $|x_i|_\infty = |x_j|_\infty$, the proposed distance $e_{ij}$ can still perform well.

In numerous applications (see Melo et al., 2021; van den Brink & Rusinowska, 2022, for example), a weighted graph is employed to depict intricate relationships among nodes, wherein each edge is endowed with a non-negative integer weight or cost, signifying its associated numerical value. These weights can be directed or undirected, and enable a more nuanced representation of the graph, facilitating the consideration of varying levels of importance or significance for different edges.

**Definition 2.** Given $\varepsilon = \{e_{ij} | i, j = 1, \ldots, N, \text{ and } i \neq j\}$ and the adjacency matrix $\varepsilon$, a weight matrix $A \in \mathbb{R}^{N \times N}$ is a symmetric matrix representing the distance of the edge $e_{ij}$ between vertex $i$ and vertex $j$ in a weighted graph $G(\Theta, \varepsilon)$ and specifically defined as:

$$A_{ij} = \begin{cases} e_{ij}, & \text{if } i = j; \varepsilon_{ij} \neq 0; \\ 0, & \text{otherwise}. \end{cases}$$

**Remark (3).** $A_{ij}$ is a binary matrix with entries that are either 0 or 1, where a value of 1 indicates the presence of an edge between two vertices, and a value of 0 indicates the absence of an edge. On the other hand, $A_{ij}$ is a non-binary matrix that captures the weight (i.e., the measure of distance) between vertices in the graph.

We then define a diagonal matrix and indicator vector as follows.

**Definition 3.** Given $\varepsilon = \{e_{ij} | i, j = 1, \ldots, N, \text{ and } i \neq j\}$, $E \in \mathbb{R}^{N \times N}$ is a diagonal matrix with entries defined by:

$$E_{ij} = \begin{cases} \sum_{n=1}^{N} e_{in}, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases}$$

**Definition 4.** Given $\Theta = \{x_i \in \mathbb{R}^M | i = 1, \ldots, N\}$, an indicator vector $\eta \in \mathbb{R}^N$ facilitates the bi-partitioning of $G(\Theta, \varepsilon)$ into $D$ and $D^c$, where $D$ and $D^c$ are disjoint subsets of vertices. The entries of $\eta$ are defined as follows:

$$\eta_i = \begin{cases} 1, & i \in D; \\ -1, & i \in D^c. \end{cases}$$

In this section, we focus on recognizing dissimilarity (particularly for outliers) in $\Theta$ and estimate initial parameters of the underlying distribution by only considering $x$. In other words, identifying the outliers in $\Theta$ is equivalent to bi-partitioning $G$ into two disjoint subsets $D$ and $D^c$, where $D$ is a subset for which contains extreme values and $D^c$ is the complement set of $D$ without

---

2 See Eq. (27).
extreme values. To avoid partitioning results dominated by some edges with extreme values, we consider all edges both within a subset and across two subsets. Bi-partitioning $G$ is equivalent to minimizing the sum of edge between $D$ and $D^c$ and at the same time maximizing the sum of edge within each of them. We can define the partitioning cost$^1$ as:

$$
C(D, D^c) = \sum_{i \in D, j \in D^c} c_{ij} + \sum_{i \in D, j \in D^c} c_{ji} - \sum_{i \in D^c, j \in D^c} c_{ij} - \sum_{i \in D^c, j \in D^c} c_{ji}
$$

(5)

which considers the dissimilarity between disjoint sets and aims to avoid being dominated by the size of these sets. In this way, we need not focus more on the size balance of split sets as shown in Srivastava et al. (2023), but only on minimizing the cost of separating significantly different sets. In order to compute the partitioning cost, one needs to matricize Eq. (5).

Eq. (5) can thus be matricized with weight matrix $A$ and diagonal matrix $E$ as:

$$
C(D, D^c) = \sum_{i \in D} A_{ii} + \sum_{i \in D^c} A_{ii} - \sum_{i \in D} A_{ii} - \sum_{i \in D^c} A_{ii}
$$

(6)

In addition, let $\kappa$ be the ratio of $\sum_{j \neq 0} E_{ii}$ over $\sum_{j \in D} E_{ii}$. The $\kappa T_1 E_1$ can be further presented as:

$$
\sum_{j \neq 0} E_{ii} = \kappa \sum_{j \in D} E_{ii} = \kappa T_1 E_1.
$$

(7)

In the same way, $\sum_{j \neq 0} E_{ii}$ in Eq. (6) is then:

$$
\sum_{j \neq 0} E_{ii} = (1-\kappa) \sum_{j \in D} E_{ii} = (1-\kappa) T_1 E_1.
$$

(8)

In Eq. (6) both $\sum_{j \neq 0} A_{ij}$ and $\sum_{j \neq 0} A_{ji}$ respectively consider the sum of partial elements on the diagonal in $E$.

It is advantageous to utilize Definition 4 in order to logically determine the membership of data points within a partition and to readily ascertain the size of the partitioned data points. Let $\eta^\Phi = (\eta + 1)/2$ and $\eta^\Psi = (1 - \eta)/2$. We can employ the following indicator functions compute the cost in Eq. (6):

$$
\eta^\Phi_i = \begin{cases} 1, & i \in D \\ 0, & i \in D^c \end{cases}, \quad \text{and} \quad \eta^\Psi_i = \begin{cases} 0, & i \in D \\ 1, & i \in D^c \end{cases}
$$

(9)

such that $\eta^\Phi$ keeps the indicator of all nodes in $D$ and nullifies the indicator of all nodes in $D^c$, while $\eta^\Psi$ keeps the indicator of all nodes in $D^c$ and nullifies the indicator of all nodes in $D$. Therefore, we obtain $\sum_{j \neq 0} A_{ij}$ by subtracting $\sum_{j \neq 0} A_{ij}$ from $\sum_{j \neq 0} A_{ij}$ with the following matrix form:

$$
\sum_{j \neq 0} A_{ij} = \sum_{j \neq 0} E_{ij} \quad \sum_{j \neq 0} A_{ij} = \sum_{j \neq 0} E_{ij} - \sum_{j \neq 0} A_{ij}
$$

$$
= (\eta^\Phi)^T E\eta^\Phi - (\eta^\Phi)^T A\eta^\Psi - (\eta^\Psi)^T A\eta^\Phi + (\eta^\Psi)^T E\eta^\Psi
$$

$$
= \frac{1}{4} \left( (1 + \eta)^T (E - A)(1 + \eta) \right).
$$

Similarly, $\sum_{j \neq 0} A_{ij}$ can be formulated as:

$$
\sum_{j \neq 0} A_{ij} = \sum_{j \neq 0} E_{ij} \quad \sum_{j \neq 0} A_{ij} = \sum_{j \neq 0} E_{ij} - \sum_{j \neq 0} A_{ij}
$$

$$
= (\eta^\Psi)^T E\eta^\Psi - (\eta^\Psi)^T A\eta^\Phi - (\eta^\Phi)^T A\eta^\Psi + (\eta^\Phi)^T E\eta^\Phi
$$

$$
= \frac{1}{4} \left( (1 - \eta)^T (E - A)(1 - \eta) \right).
$$

The advantage of operation with Eq. (9) is to simplify matrixing Eq. (8) since we only need partial sums of $A$ and $E$.

Based on Eqs. (7)-(10), the partitioning cost function Eq. (6) can be evaluated with $\eta$ in Eq. (4) such that:

$$
c(\eta) = \frac{1}{4} \left( (1 + \eta)^T (E - A)(1 + \eta) - (1 - \eta)^T (E - A)(1 - \eta) \right).
$$

(11)

The constant $1/4$ in Eq. (11) can be ignored when computing $c(\eta)$, and the optimal bi-partition problem can be presented as:

$$
\arg\min_\eta \frac{\eta^T (E - A) \eta}{\eta^T E \eta}
$$

(12)

which is the Rayleigh quotient and an NP-complete problem (see Zhang, 2017, and references therein). In order to find the optimal solution, we relax the range of $\eta$ from boolean in Eq. (9) to real numbers in Eq. (4).

Since Definition 4 only considers the dissimilarity information of node size, we still need dissimilarity information of node distance (measured by edges). We define it in Definition 5 and then present Lemma 2.1 accordingly as follows.

Definition 5. Given $\kappa = (\sum_{j \neq 0} E_{ii})/(\sum_{j \in D} E_{ii})$, we define $\zeta$ as follows:

$$
\zeta_i = \begin{cases} \sqrt{\frac{1}{\kappa}}, & i \in D \\ -\sqrt{\frac{1}{\kappa}}, & i \in D^c \end{cases}
$$

(13)

Lemma 2.1. With all-ones vector $\mathbf{1}$, $\zeta$ can be expressed as:

$$
\zeta = \left( \frac{1}{\sqrt{\frac{1}{\kappa} + \sqrt{\frac{1}{\kappa}}}} \right) \eta + \left( \frac{1}{\sqrt{\frac{1}{\kappa} - \sqrt{\frac{1}{\kappa}}}} \right) \mathbf{1},
$$

(14)

that satisfies two conditions: (1) $\mathbf{1}^T \zeta = \eta^T E \eta$ and (2) $\zeta^T E \mathbf{1} = 0$.

Proof. $E$ is a diagonal matrix, and then $\zeta^T E \zeta$ can be derived as:

$$
\sum_{i \in D} \zeta^T E \zeta = \sum_{i \in D} \eta^T E \eta + \left( \frac{\kappa}{1 - \kappa} \right) \sum_{i \in D} \eta^T E \eta
$$

$$
= \left( \frac{1 - \kappa}{\kappa} \right) \eta^T E \eta + \left( \frac{\kappa}{1 - \kappa} \right) (1 - \kappa) \eta^T E \eta = \eta^T E \eta,
$$

which is condition (1) and:

$$
\zeta^T E \mathbf{1} = \left( \frac{1 - \kappa}{\kappa} \right) \sum_{i \in D} E \mathbf{1} - \left( \frac{\kappa}{1 - \kappa} \right) \sum_{i \in D^c} E \mathbf{1}
$$

$$
= \left( \frac{1 - \kappa}{\kappa} \right) \mathbf{1} - \left( \frac{\kappa}{1 - \kappa} \right) (1 - \kappa) \mathbf{1} = 0,
$$

which is condition (2). □

Lemma 2.2. Let $A$ be a weight matrix and $E$ be a diagonal matrix defined by Definitions 2 and 3 for graph $G(\Theta, \mathbf{1})$, $(E - A) \mathbf{1} = 0$.

Proof. For any ith row of $(E - A)$, we have $E_{i,i} = \sum_{j} A_{i,j} = \sum_{j} E_{i,j}$; therefore, $(E - A) \mathbf{1} = 0$. □

Lemma 2.3. Given graph $G(\Theta, \mathbf{1})$ with weight matrix $A$ and diagonal matrix $E$ by Definitions 2 and 3, for any vector $\eta$ and scalars $a$ and $b$ we have $(a\eta + b\mathbf{1})^T (E - A)(a\eta + b\mathbf{1}) = a^2 \eta^T (E - A) \eta$. 

[1] Our research objective is to identify users with exceptionally high consumption by partitioning the vertices of a given graph into two disjoint sets or partitions. The cost function quantifies the connectivity between the two partitions, measuring the number or weight of edges that interconnect them. Specifically, it is calculated as the sum of edge weights between the partitions divided by the sum of edge weights within the partitions.
Proof. Since \((E-A)(a+b1) = a(E-A) \eta + b1(E-A)1 = a(E-A) \eta \), with Lemma 2.2, we obtain:

\[
(a \eta + b1)^T (E-A) (a \eta + b1) = a^2 \eta^T (E-A) \eta + ab1^T (E-A) \eta = a^2 \eta^T (E-A) \eta.
\]

From condition (1) in Lemma 2.1 with Lemma 2.3 we can derive:

\[
\frac{\xi^T (E-A) \xi}{\xi^T E \xi} = \frac{1}{4} \left( \frac{1 - \kappa}{\kappa} - \frac{\kappa}{1 - \kappa} \right)^2 \frac{\eta^T (E-A) \eta}{\eta^T E \eta} = \frac{4}{(2 - \kappa)^2} \frac{\eta^T (E-A) \eta}{\eta^T E \eta} = \frac{4 \kappa (1 - \kappa)}{2 (2 - \kappa)^2} = \frac{4 \kappa (1 - \kappa)}{2 (2 - \kappa)^2}.
\]

Since condition (2) in Lemma 2.1 holds, we generalize the partition cost on \( \xi \) as an optimization problem formally presented as follows.

**Proposition 1.** The unconstrained optimization problem of minimizing partitioning cost (in Eq. (12)) is equivalent to the following restricted optimization problem:

\[
\arg \min \frac{\xi^T (E-A) \xi}{\xi^T E \xi} \quad \text{subject to } \xi^T E1 = 0.
\]

Proof. From Eq. (15), we directly derive:

\[
\frac{\eta^T (E-A) \eta}{\eta^T E \eta} = \frac{4 \kappa (1 - \kappa)}{2 (2 - \kappa)^2} \xi^T (E-A) \xi.
\]

Since condition (2) in Lemma 2.1 holds, the unrestricted optimization problem in Eq. (12) is then expressed by restricted optimization problem Eqs. (16) and (17).

Since Proposition 1 releases the NP-hard difficulty for solving Eq. (12), then solution to the optimization problem in Propositions 1 is feasible with following proposition.

**Proposition 2.** The solution of optimal problem in Proposition 1 is \((E-A) \xi = \lambda E \xi \).

Proof. Following Proposition 1, the bi-partitioning cost becomes the minimization of the Rayleigh quotient. After using a Lagrange multiplier \( \lambda \) to Eqs. (16) and (17) we obtain the following minimization:

\[
\min \lambda (\cdot) = \xi^T (E-A) \xi - \lambda (\xi^T E \xi).
\]

The optimal solution can be derived by: \( \nabla \xi L = 2 ((E-A) - \lambda E) \xi = 0 \); and therefore we solve \((E-A) \xi = \lambda E \xi \), where the smallest eigenvalue is 0 and the second smallest eigenvector is the real valued solution to our bi-partition problem to identify outliers in \( \Theta \).

After solving the cost minimization problem in Eq. (18), we now construct \( D \) and \( D_\xi \) by \( \xi \), and \( D^\xi \) is used to estimate the initialization parameters of distributions. The distribution in our proposed method is a multivariate mixture skew t distribution, and we show the details in the following section.

2.12. The distribution

The application of symmetric elliptical distributions in data modeling and robustness studies has significantly broadened their scope, with the transformation into skew elliptical distributions effectively preserving properties and enhancing their potential for data modeling. For example, when considering heavy-tail and skewness together, the multivariate skew t distribution is a desirable choice (see Nair et al., 2022, and references therein). However, this transformation presents the challenge of having multiple methods available, as evident in the various definitions of the multivariate skew t-distribution (see Cabral et al., 2012; Lee & McLachlan, 2022, and references therein).

A stochastic representation of the multivariate skew t-distributed random variable can be expressed as:

\[
x = \mu + \xi W + ZW^\frac{1}{2},
\]

where \( \mu \) is a location vector, \( W \) is a one-dimensional random variable distributed according to an inverse gamma distribution with parameters \( (\nu/2, \nu/2) \), which includes the t-distribution as a special case, and \( Z \) is a random vector having a zero-mean multivariate normal distribution with covariance matrix \( \Sigma \) (see Demarta & McNeil, 2005; Sun et al., 2008, for example).

Since the data usage is always positive and their skewness is concentrated along the directions of the feature variables while being uncorrelated across different directions, we can utilize the following stochastic representation (see Lee & McLachlan, 2022, and reference therein):

\[
x = \mu + D |Z_0| + Z_1,
\]

where \( D = \text{diag}(\xi) \) is a matrix\(^5\) from the skewness vector \( \xi \), and \( Z_0 \) and \( Z_1 \) are \( M \times \text{dimensional normal random vectors jointly distributed and conditional on an inverse gamma distribution, that is,}

\[
\begin{bmatrix} Z_0 \\ Z_1 \end{bmatrix} \sim N_{2M} \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Gamma^{-1} \begin{bmatrix} \Omega & D \\ D & I_M \end{bmatrix} \right)
\]

where \( \Omega \) is a scale matrix, \( I_M \) denotes a \( M \times M \) identity matrix, \( \Gamma^{-1} \) is the inverse gamma distribution with parameters \( (\nu/2, \nu/2) \), and \( |Z_0|_+ \) represents a vector in which each element corresponds to the magnitude of the corresponding element in vector \( Z_0 \).

A random variable \( x \in \mathbb{R}^M \) follows a multivariate student's t distribution with the location vector \( \mu \in \mathbb{R}^M \), a scale matrix \( \Omega \in \mathbb{R}^{M \times M} \), and the scalar of degree of freedom \( \nu \) whose probability density function (pdf) \( f_t(x; \mu, \Omega, \nu) \) is expressed as:

\[
f_t(x; \mu, \Omega, \nu) = \frac{\Gamma \left( \frac{\nu+M}{2} \right) \Gamma \left( \frac{\nu}{2} \right)}{\Gamma \left( \frac{\nu}{2} \right) (2\pi)^{\frac{M}{2}} |\Omega|^{\frac{\nu}{2}} |\nu-M|^{\frac{\nu+M}{2}}} \left( 1 + \frac{(x-\mu)^T \Omega^{-1} (x-\mu)}{\nu} \right)^{-\frac{\nu+M}{2}}.
\]

Based on Eq. (20), a M-dimensional skew t-distribution \( f_{st}(x; \mu, \xi, c, \nu, v) \) is then characterized as:

\[
f_{st}(x; \mu, \xi, c, \nu, v) = 2^{M/2} f_t(x; \mu, \Omega, \nu) \times T \left( \delta \left( \frac{v+M}{v+\nu} \right) ; 0, A, A, M + v \right),
\]

where \( T(\cdot; 0, A, M + v) \) is the (cumulative) distribution function of student's t with a zero location vector, \( A \) for scale matrix, degrees of freedom \( (M + v) \), \( c = \Omega + DD^\xi \), \( \delta = DC^{-1}(x-\mu) \), \( L = \)

\(^4\) In the literature, both “multivariate skewed t distribution” and “multivariate skew t distribution” are used interchangeably to refer to the same concept.

\(^5\) Here, \text{diag}(\xi) returns a square diagonal matrix with the elements of vector \( \xi \) on the main diagonal.
Given random variables \( y \in \mathbb{R}^M \) and \( u \in \mathbb{R} \) with the following distributions:

\[
x \mid y, u \sim N'(\mu + Dy, u^{-1}C);
\]

\[
y \mid u \sim N'\left(0, u^{-1}\eta\right); \quad u \sim \Gamma(v/2, v/2).
\]

where \( N'(\cdot) \) denotes normal distribution, \( N''(\cdot) \) denotes half-normal distribution, and \( \Gamma(\cdot) \) denotes gamma distribution. The joint density function of multivariate skew t-distribution \( f(x; y, u) \) in Eq. (23) can be explicitly presented as:

\[
f(x; y, u) = \left((v/2)^{v/2}\left|\Omega\right|^{-1/2}\pi^{M}\Gamma(v/2)\right)^{-1}\frac{1}{\Gamma(v/2)}\frac{1}{\left|\Omega\right|^{1/2}}f_{\lambda}(x; \eta, \kappa)
\]

\[
\times \exp \left\{ -\frac{1}{2} \left((y - \delta)^T \Lambda^{-1} (y - \delta) + (x - \mu)^T \Omega^{-1} (x - \mu) + v \right) \right\}.
\]

The multivariate finite (i.e., \( K \in \mathbb{N} \)) mixture skew t distribution is then:

\[
f(x; \Psi) = \sum_{k=1}^{K} w_k f(x; \mu_k, \Sigma_k, \xi_k, \nu_k).
\]

where \( \Psi \) is a set of the parameters in the mixture model, \( \Psi = \{\mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K, \xi_1, \ldots, \xi_K, \nu_1, \ldots, \nu_K\} \), and \( w_k \) is the mixing weight for each component skew t distribution with \( w_k \geq 0 \) for \( k = 1, \ldots, K \) and \( \sum_{k=1}^{K} w_k = 1 \).

Given \( \Theta \) as all data points (e.g., total consumption in our case), the log-likelihood function of Eq. (26) is then:

\[
\ln L(\Psi|\Theta) = \sum_{k=1}^{K} \sum_{i=1}^{n} \left( \ln w_k + \frac{v_k}{2} \ln \frac{v_k}{2} - \frac{1}{2} \ln |\Omega_k| - \ln \Gamma\left(\frac{v_k}{2}\right) + \left(\frac{v_k}{2} + M - 1\right) \ln u_i + \frac{1}{2} \left((y_i - \delta)^T \Lambda^{-1} (y_i - \delta) \right) + \left((x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) + v\right) \right)
\]

where \( \ln |\Theta| \) denotes the number of observations in \( \Theta \), and the constant \( c \) is irrelevant to \( \Psi \).

2.1.3. Determination of initial value

In addition to the location vector \( \mu \), we decide the initial values for the scale matrix, degree of freedom, and skewness vector for every component skew t distribution. Given the initial \( \mu_1, \ldots, \mu_K \), we use an indicator \( z_i \) and denote \( z_{ik} = 1 \) for all \( x_i \) when \( x_i \) is the closest to \( \mu_k \) among all components in the mixture model, and it is zero otherwise. Let \( \Theta_{ik} = \cup_{z_{ik}=1} \{x_i\} \) be a subset of data assigned to the \( k \)th cluster, where \( \mu_k \) yields the smallest distance. The initial values for \( \Psi \) are then determined sequentially as follows:

\[
w_k = \|\Theta_k\|^{-1} |\Theta_k|.
\]

\[
\delta_{i,m} = \sqrt{\|\Theta_k\|^2 \left( \sum_{i=1}^{n} \delta_{i,m} \delta_{i,m} \right) ^2}
\]

\[
D_m = \frac{\pi}{\pi - 2} L_k.
\]

\[
\Omega_k = S_k - \left(\frac{1}{\pi}\right) \text{diag}(S_k M),
\]

\[
C_k = \Omega_k + D_k A_k,
\]

\[
\delta_k = \left|\Theta_k\right|^{-1} \left( \sum_{i=1}^{n} D_i A_i (x_i - \mu_k) \right).
\]

where \( S_k \) is the covariance of the \( k \)th component, \( S_k = E[(x_i - \mu_k)(x_i - \mu_k)^T] \) for \( x_i \in \Theta_k \), and \( \delta_{i,m} \) is the \( m \)th element skewness value in the \( k \)th component. In addition, \( \nu \) is jointly determined by \( \|\Theta_k\| \) and \( M \).

2.2. Expectation-maximization (EM) algorithm

After determining the initial values of \( \Psi \), we get the likelihood value based on \( \Psi \) from Eq. (27). We estimate and refine the most appropriate parameters \( \Psi \) by executing the EM algorithm iteratively until the log-likelihood converges. Each iteration includes expectation and maximization steps.

Let \( n \) denote the iteration of the EM algorithm, whereby we compute the posterior membership \( w_{ik}^n \) of each observation \( x_i \) for the \( k \)th component in the mixture model as:

\[
w_{ik}^n = \frac{w_k f(x_i; \mu_k; \Sigma_k, \xi_k, \nu_k)}{\sum_{k=1}^{K} w_k f(x_i; \mu_k; \Sigma_k, \xi_k, \nu_k)}
\]

\[
\text{for } i = 1, \ldots, N \text{ and } k = 1, \ldots, K.
\]

Different from other mixture distributions such as multivariate Gaussian distribution or student t distribution, the proposed the multivariate mixture of shifted t distribution has a complex form with the product of PDF and CDF shown in Eq. (23) that requires a special treatment shown as follows.

Since the likelihood in Eq. (27) consists of latent variables \( y \) and \( u \), in addition to computing the expectation value of the log-likelihood with the current estimation of parameters in the expectation step, we also need to compute the conditional expectations \( E(\ln w_k | x; z_{ik}^n = 1) \), \( E(w_k | x; z_{ik}^n = 1) \), \( E(w_k y_i^T | x; z_{ik}^n = 1) \), and \( E(w_k y_i y_i^T | x; z_{ik}^n = 1) \) on the \( n \)th iteration of EM algorithm.

Let \( \tilde{q}_{ik}^n \), \( q_{ik}^n \), \( \tilde{q}_{ik} \), and \( q_{ik} \) denote \( E(\ln w_k | x; z_{ik}) \), \( E(w_k | x; z_{ik}) \), \( E(w_k y_i y_i^T | x; z_{ik}) \), and \( E(w_k y_i y_i^T | x; z_{ik}) \) respectively. Then we have:

\[
\tilde{q}_{ik}^n = E(\ln w_k | x; z_{ik}^n = 1) = E(\ln w_k | x; z_{ik} = 1) = E\left(\frac{v_k + M}{v_k + \eta_k^n} \right) - \ln \left( \frac{v_k + \eta_k^n}{2} \right) - \psi\left( \frac{v_k + M}{2} \right). \tag{35}
\]

and

\[
q_{ik}^n = E(w_k | x; z_{ik}^n = 1) = E(w_k | x; z_{ik} = 1) = E\left(\frac{v_k + M + v_k^n}{v_k + \eta_k^n + v_k^n} \right) - \ln \left( \frac{v_k + \eta_k^n + v_k^n}{2} \right) - \psi\left( \frac{v_k + M + v_k^n}{2} \right). \tag{36}
\]

Let \( s_i \in \mathbb{R}^M \) follow a truncated t-distribution to the positive hyperplane, such that:

\[
f(s_i) \sim f_t\left(\delta_n \left(\frac{v_k^n + \eta\frac{v_k^n + M}{2}}{v_k^n + M + 2}\right) \Lambda_k^n, v_k^n + M + 2\right).
\]
We then have:
\[
q_{i,k}^n = E(w_i y_i | x_i; z_{i,k}^n = 1) = \left( \frac{1}{n^2} + M \right) \frac{T}{T} \left( D_k^n (\Omega_k^n)^{-1} (x_i - \mu_k^n); 0, \Lambda_k^n, M + v_k^n + 2 \right) s_i
\]
\[
= q_{i,k} E(s_i), \tag{37}
\]
and
\[
Q_{i,k}^n = E(w_i y_i y_i^T | x_i; z_{i,k}^n = 1) = \left( \frac{1}{n^2} + M \right) \frac{T}{T} \left( D_k^n (\Omega_k^n)^{-1} (x_i - \mu_k^n); 0, \Lambda_k^n, M + v_k^n + 2 \right) s_i s_i^T
\]
\[
= q_{i,k} E(s_i s_i^T). \tag{38}
\]

In maximization (M-step) of the nth iteration, we update the parameters with posterior \(w_i^n\), derived in the previous expectation (E-step) such that for \(k = 1, \ldots, K:\)
\[
w_k^n = ||\Theta||^{-1} \sum_{i=1}^{||\Theta||} w_i^n. \tag{39}
\]

We then update \(\mu_k^n, \xi_k^n, \) and \(\Omega_k^n\) with \(\hat{q}_{i,k}^n, q_{i,k}^n, \) and \(Q_{i,k}^n\) as follows:
\[
\mu_k^n = \frac{\sum_{i=1}^{||\Theta||} w_i^n (q_{i,k}^n x_i - \xi_i^n q_{i,k}^n)}{\sum_{i=1}^{||\Theta||} w_i^n q_{i,k}^n}, \tag{40}
\]
\[
\xi_k^n = \left( (\Omega_k^n)^{-1} \odot \sum_{i=1}^{||\Theta||} w_i^n Q_{i,k}^n \right)^{-1}
\]
\[
\times \text{Diag} \left( (\Omega_k^n)^{-1} \sum_{i=1}^{||\Theta||} w_i^n (q_{i,k}^n x_i - \mu_k^n)(q_{i,k}^n)^T \right). \tag{41}
\]
\[
\Omega_k^n = \sum_{i=1}^{||\Theta||} w_i^n \left( \sum_{i=1}^{||\Theta||} (w_i^n D_i^n (Q_{i,k}^n)^{-1} (x_i - \mu_k^n))(q_{i,k}^n)^T D_i^n - D_k^n q_{i,k}^n (x_i - \mu_k^n)^T + w_i^n (x_i - \mu_k^n)(x_i - \mu_k^n)^T \right). \tag{42}
\]

where \(\odot\) denotes the Hadamard product and executes the element-wise matrix product (see Zhang, 2017, for example).

To derive \(q_{i,k}^n\), we solve the following equation:
\[
\log \left( \frac{1}{n^2} \right) - \psi \left( \frac{1}{n^2} \right) = \frac{\sum_{i=1}^{||\Theta||} w_i^n (q_{i,k}^n - \hat{q}_{i,k}^n - 1)}{\sum_{i=1}^{||\Theta||} w_i^n}, \tag{43}
\]
where \(\psi(\cdot)\) is a Digamma function \(\psi(\cdot) = \Gamma'(\cdot) / \Gamma(\cdot)\). We execute E-step and M-step alternatively until the log-likelihood converges; that is, when there is no more improvement between two iterations, we obtain all parameters of the underlying multivariate mixture skew t distribution.

2.3. Validation of the proposed method

Since clustering interprets the placement of objects in high-dimensional spaces, the optimal solution can be extremely costly to discover and sometimes even unreachable or non-existent. The trade-off between accuracy and computability leads to suboptimal solutions (Iglesias et al., 2020). Therefore, validation of clustering is mandatory to refine algorithms and ensure meaningful solutions. In our validation, we first choose \(k, k = 1, \ldots, K\) well-defined and labeled classes (as the grand truth \(\Theta_k\)) with sufficiently large data points \(\Theta\), and then apply the clustering method to group \(\Theta\) into \(K\) clusters \(\Theta_k\), i.e., \(\Theta = \bigcup_{k=1}^{K} \Theta_k\). If the clustering method perfectly works, then all data points should be correctly classified into the clusters they should belong; that is, \(\Theta_k = \hat{\Theta}_k\). We use Eq. (44) to calculate this empirical accuracy:
\[
\text{Accuracy} = ||\Theta||^{-1} \left( \sum_{k=1}^{K} 1_{\hat{\Theta}_k = \Theta_k} \right), \tag{44}
\]
where \(||\Theta||\) denotes the total number of data points.

Following Srivastava et al. (2023), we evaluate the performance and accuracy of the proposed method by using two open datasets MNIST⁶ and USPS⁷ that are the data of handwritten digits composed of ten different digit images and their corresponding labels. The MNIST data are composed of 70,000 images where each image has a 784-dimension vector. The USPS data includes 11,000 images, where each image is 16 × 16 pixels, which we flatten into a 256-dimension vector. In order to deal with high dimensionality of each image, we apply the t-distributed stochastic neighbor embedding (t-SNE) algorithm to reduce the high-dimensional vector into a heavy-tailed distributed lower-dimensional (i.e., 3D) vector specified by t-SNE.

In this work we first randomly choose the subset of different digits from all ten classes of digits (labeled from 0 to 9). Among these 10 different classes of digits, we randomly choose 3, 4, 5, and 6 different classes and investigate the performance of clustering methods for all combinations with \(C^{10}_{3} = 120, C^{10}_{4} = 210, C^{10}_{5} = 252, \) and \(C^{10}_{6} = 210\) times of validation, respectively. In this study each digit has 1000 data points and when choosing 3, 4, 5, and 6 different classes, there are a total of 3000 data points for three classes, 4000 data points for four classes, and so on. We treat each class as a well-defined and labeled cluster.

We show one 3D graphic sample in Figs. 1 and 2, and these results serve to cluster with the grand truth (\(\Theta_k\)) for validating the clustering results generated by the underlying methods. For example, when we choose five clusters, there is a total of 252 combinations of five different digit classes from ten classes, such as \(\{1, 2, 3, 4, 5\}\) or \(\{3, 5, 7, 8, 9\}\). We can perform the validation in total 252 times with 5000 data points each time and then compare the clustering results with the grand truth. If the clustering method perfectly works, then the accuracy should be one in Eq. (44).

Based on the established grand-truth clusters in Figs. 1 and 2, we calculate the accuracy by performing the proposed graphically initialized subspace clustering methods with three different multivariate mixture distributions: multivariate mixture Gaussian distribution (MMG), multivariate mixture t distribution (MMT), and multivariate mixture skew t distribution (MMST). We also consider three other methods, K-mean, K-mean++, and support vector machine (SVM),⁸ in comparison with the spectral clustering methods.

Table 1 shows the average accuracy in percentage with corresponding standard deviation. We see that the proposed spectral clustering with MMST distribution illustrates the highest value of mean and the lowest value of standard deviation for all four different cases. This indicates a better performance than all other methods investigated in this study.

3. Business analytics

In this section we apply the proposed method in Section 2 for data-driven business analytics. The decision problem is to deter-
Fig. 1. Clusters in 3D of MNIST data after dimensionality reduction and each cluster contains 1000 data points.

Table 1
Comparing mean and standard deviation (in parentheses) of classification accuracy in percentage for six methods, K-means, K-means++, support vector machine, multivariate mixture Gaussian distribution (MMG), multivariate mixture $t$ distribution (MMT), and multivariate mixture skew $t$ distribution (MMST), based on the MNIST and USPS dataset.

<table>
<thead>
<tr>
<th>Clusters</th>
<th>Runs</th>
<th>K-Means</th>
<th>K-Means++</th>
<th>SVM</th>
<th>MMG</th>
<th>MMT</th>
<th>MMST</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>3</td>
<td>(10)</td>
<td>75.2402</td>
<td>78.5225</td>
<td>94.0399</td>
<td>95.1134</td>
<td>95.1825</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>(10)</td>
<td>74.0251</td>
<td>74.3653</td>
<td>86.7405</td>
<td>91.2293</td>
<td>94.9049</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>(10)</td>
<td>64.7820</td>
<td>65.8081</td>
<td>78.6434</td>
<td>86.4655</td>
<td>87.3369</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(12.4113)</td>
<td>(11.9410)</td>
<td>(8.4050)</td>
<td>(6.8475)</td>
<td>(2.8347)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>(10)</td>
<td>56.6009</td>
<td>60.0198</td>
<td>70.2131</td>
<td>81.8596</td>
<td>84.7340</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(10.7835)</td>
<td>(10.0545)</td>
<td>(8.5794)</td>
<td>(7.1286)</td>
<td>(4.1196)</td>
</tr>
<tr>
<td>USPS</td>
<td>3</td>
<td>(10)</td>
<td>71.2853</td>
<td>76.0943</td>
<td>90.4968</td>
<td>94.3202</td>
<td>94.3443</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(17.0069)</td>
<td>(16.8548)</td>
<td>(10.9575)</td>
<td>(8.3669)</td>
<td>(7.2880)</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>(10)</td>
<td>69.5552</td>
<td>69.6500</td>
<td>84.1095</td>
<td>91.1290</td>
<td>94.8543</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(12.3301)</td>
<td>(12.2132)</td>
<td>(9.6588)</td>
<td>(6.9163)</td>
<td>(2.4414)</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>(10)</td>
<td>64.6975</td>
<td>65.4468</td>
<td>76.3304</td>
<td>87.4700</td>
<td>88.9587</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(10.3870)</td>
<td>(10.2784)</td>
<td>(9.0563)</td>
<td>(6.6159)</td>
<td>(2.8502)</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>(10)</td>
<td>60.2326</td>
<td>61.7328</td>
<td>68.4766</td>
<td>84.0807</td>
<td>86.8843</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(9.2266)</td>
<td>(8.7373)</td>
<td>(9.1725)</td>
<td>(6.3190)</td>
<td>(3.2903)</td>
</tr>
</tbody>
</table>
mine data allowance under a 3PT pricing regime by suggesting an optimal tiered tariff scheme with different allowances that ultimately match the actual usage as directly described by the data. In our study, cellular operators are faced with two types of consumers described in the literature. They have to propose tiered pricing with attributed allowance on a combination of the actual usage of these two types of consumers. The optimization referred to here has two main meanings: first, no reduction in the total revenue of the cellular operator; second, minimization of price changes so that the number of subscribers affected by price changes is minimal.

We first conduct descriptive analytics in Section 3.1, which employ real usage data provided by a cellular carrier for its consumers who have mixture subscriptions of flat rate plan and tiered rate plan under the privacy-preserving principle of strict anonymity that allows only using features of the real usage data. We performed the proposed clustering methods in Section 2 on the actual data usage. We then work on prescriptive analytics in Section 3.2 to demonstrate how to optimize a new tiered pricing scheme that considers two types of consumers. We provide a summary of notations in Appendix B.

3.1. Descriptive analytics

In this study, we use the aforementioned method in Section 2 to perform cluster analysis according to the data usage of the consumers in the sample. All consumers are considered to exhibit a mixed distribution (i.e., multivariate mixture skew $t$ distribution) and each cluster can be characterized by a component (i.e., multivariate skew $t$) distribution according to the usage characteristics of that cluster. In this way, we can estimate the parameters of the corresponding distribution based on actual usage, and thus provide indicative information (such as the moment of distribution) on the usage for subsequent decision making.

3.1.1. The data

In this section we work on the real usage data provided by a cellular carrier\(^9\) to exemplify our business analytics. Under the EU’s General Data Protection Regulation (GDPR) reinforced on 25 May 2018, we use 10,000 anonymous consumers’ mobile data usages of twelve months randomly sampled for the contractual period from

---

\(^9\) The complete dataset and code will be released as supplementary materials on the author's webpage.
Table 2

Descriptive statistics of the data usage (in gigabyte) of all consumers.

<table>
<thead>
<tr>
<th></th>
<th>M</th>
<th>Mean</th>
<th>Variance</th>
<th>Minimum</th>
<th>25-Percentile</th>
<th>Median</th>
<th>75-Percentile</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.8633</td>
<td>15.3132</td>
<td>0.0000</td>
<td>0.8619</td>
<td>2.9386</td>
<td>7.1644</td>
<td>204.0747</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7.5264</td>
<td>15.5830</td>
<td>0.0000</td>
<td>1.1056</td>
<td>3.0752</td>
<td>8.2296</td>
<td>279.3722</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>7.5120</td>
<td>14.4450</td>
<td>0.0000</td>
<td>1.1709</td>
<td>3.4922</td>
<td>8.3780</td>
<td>301.0661</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>8.4706</td>
<td>18.9625</td>
<td>0.0000</td>
<td>1.0859</td>
<td>3.4581</td>
<td>9.0128</td>
<td>479.4200</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>8.4079</td>
<td>15.8987</td>
<td>0.0001</td>
<td>1.2964</td>
<td>3.9015</td>
<td>10.2266</td>
<td>349.3561</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>8.1985</td>
<td>15.9642</td>
<td>0.0000</td>
<td>1.3797</td>
<td>3.8732</td>
<td>9.3858</td>
<td>332.6890</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>9.1501</td>
<td>18.4520</td>
<td>0.0000</td>
<td>1.5182</td>
<td>4.2607</td>
<td>10.0815</td>
<td>309.3104</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>8.7416</td>
<td>17.1331</td>
<td>0.0006</td>
<td>1.4647</td>
<td>4.2151</td>
<td>10.4528</td>
<td>291.7135</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9.1364</td>
<td>16.7733</td>
<td>0.0000</td>
<td>1.5200</td>
<td>4.4592</td>
<td>10.8817</td>
<td>305.1616</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9.9379</td>
<td>17.7749</td>
<td>0.0001</td>
<td>1.5033</td>
<td>4.2574</td>
<td>11.4681</td>
<td>195.0865</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>9.0019</td>
<td>17.2656</td>
<td>0.0009</td>
<td>1.4925</td>
<td>4.2589</td>
<td>10.2147</td>
<td>260.1111</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>9.9166</td>
<td>19.5891</td>
<td>0.0000</td>
<td>1.4206</td>
<td>4.4417</td>
<td>11.1860</td>
<td>236.4082</td>
<td></td>
</tr>
</tbody>
</table>

Table 3

The mean± and variance (in parentheses) of the goodness-of-fit measured by AIC, BIC, and MDL for three different K-multivariate models composed of Gaussian (MMG), student’s t (MMT), and skewed t (MMST).

<table>
<thead>
<tr>
<th>K</th>
<th>MMG AIC</th>
<th>MMG BIC</th>
<th>MMG MDL</th>
<th>MMT AIC</th>
<th>MMT BIC</th>
<th>MMT MDL</th>
<th>MMST AIC</th>
<th>MMST BIC</th>
<th>MMST MDL</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>(1.0002)</td>
<td>(1.0002)</td>
<td>(1.0002)</td>
<td>(1.4518)</td>
<td>(1.4518)</td>
<td>(1.4518)</td>
<td>(1.0526)</td>
<td>(1.0526)</td>
<td>(1.0526)</td>
</tr>
<tr>
<td>5</td>
<td>(4.1868)</td>
<td>(4.9313)</td>
<td>(8.9877)</td>
<td>(2.7553)</td>
<td>(9.5771)</td>
<td>(7.5562)</td>
<td>(1.6720)</td>
<td>(5.7945)</td>
<td>(7.5090)</td>
</tr>
<tr>
<td>6</td>
<td>(5.0007)</td>
<td>(5.0007)</td>
<td>(4.0006)</td>
<td>(5.0007)</td>
<td>(5.0007)</td>
<td>(5.0007)</td>
<td>(3.9500)</td>
<td>(3.9500)</td>
<td>(3.9500)</td>
</tr>
</tbody>
</table>

*: α = 0.05.

April 2014 to December 2021 from the TCO-NCTU database that is established based on the telecommunication operator’s regulation and representing a wide range of mobile services without considering any specific features of demographics, psychographics, or sociographics. We aggregate the actual usage to monthly frequency and treat it as a feature of the data consumption. Table 2 provides descriptive statistics of all consumers in our dataset.

3.1.2. The clustering results

Since the non-parametric methods (i.e., K-Means, K-Means++, and SVM) used in validation in Section 2.3 are not suitable due to their dimensional constraints, we only show the results based on three parametric methods. Table 3 shows the clustering results with multivariate mixture distributions of Gaussian (MMG), t (MMT), and skewed t (MMST). We use three criteria (i.e., AIC, BIC, and MDL) to evaluate (i.e., the smaller the better) how good their clustering performance is in terms of goodness-of-fit when matching the data. The results in Table 3 indicate there are actually four clusters of total usage as all three methods lead to the same conclusion.

Since all three methods lead to the same number of clusters, we consider four component distributions to estimate the parameters of the multivariate mixture distributions and then divide all consumers into four clusters accordingly. Each consumer has 12 data usage observations, and therefore for i consumers in one cluster, there is a total of 12 × i data usage observations in that cluster. Table 4 provides descriptive statistics for each of the four clusters partitioned by these four multivariate mixture distributions.

In the following, we will apply these parameters estimated by different distributions as key indicators of usage to our decision optimization.

3.2. Prescriptive analytics

The fundamental impetus for mobile operators to implement price adjustments predicated on users’ data usage lies in the pursuit of aligning their pricing strategy with the actual costs incurred in providing the service. Notably, data usage serves as a critical determinant of network capacity, as mobile networks possess finite capabilities in handling data traffic, and heavy data users can exert a strain on network resources. Through the strategic adaptation of prices based on usage patterns, operators can effectively incentivize customers to utilize their data resources in a more judicious manner, mitigating the risk of network overload. Consequently, data usage emerges as a pivotal factor that significantly influences the cost associated with provisioning mobile services. By leveraging usage patterns as a basis for price adjustments, operators can adeptly manage their costs and sustain profitability, thus embodying a judicious approach to pricing optimization in the mobile telecommunications industry. In this section, we conduct prescriptive analytics using the clustering result in Section 3.1 obtained with the proposed technology described in Section 2 to make optimal decisions with a data-driven approach that mobile operators could adopt when adjusting prices.

In our study, a cellular carrier plans to introduce a new tiered pricing of 3PT regime based on a sample of 10,000 consumers’ actual usage over twelve months. The current 3PT billing plan is given by Table 5. The cellular carrier should determine new allowances to match the currently existing prices without distorting...
the realized revenue from the previous allowance. For data protection, all consumers and their subscribed plans are anonymous.

We also consider price data in correspondence with mobile data subscriptions. Table 5 shows the cellular carrier's current billing plans and additional overage fees in detail.  

We carry out our study with the following steps. First, the actual usage is allocated to a new data plan in which the new tiers are suggested by the characteristics of the underlying multivariate mixture skew t (MMST) distribution: see our proposed clustering method in Section 2. Second, the current billing rate remains unchanged, and we connect the new data allowance to the basic rate by applying the nearest neighbor search (NNNS) algorithm such that the integer-valued usage is matched to the new tier of each basic plan. Third, we apply the minimum principle in search of the nearest low rate that minimizes the total costs of consumers. Finally, we calculate the revenue by adding total charges of the fixed access rate and overage for the corresponding data usage.

The notion of customer value represents a salient consideration in the context of price adjustments for mobile operators. Acknowledging that not all customers hold equal value to operators, the strategic adaptation of prices based on usage patterns allows for a tailored pricing strategy that caters to the unique needs and preferences of distinct customer segments, ultimately maximizing revenue generation from each customer. However, it is important to note that price modifications, particularly in the form of increases, may impact customer retention, potentially leading to a reduction in revenue. In light of this, we conduct an analysis on the effects of revenue reduction stemming from decreased customer retention, focusing on three multivariate probability distributions that characterize different tiers of customer segments. Through this investigation, we aim to glean insights into the intricate interplay between pricing adjustments, customer value, and revenue outcomes, thus elucidating the nuanced dynamics of pricing optimization in the mobile telecommunications industry.

3.2.1. Determination of billing plans

A three-part tariff contains an access fee (fixed price) $\beta$ with an allowance $\beta$ and a per-unit fee $\rho$ on each overage unit $\alpha$, respectively. We use $B$ for the set of all tiered access fees $\alpha$.

for the set of per-unit prices, and $\|B\|$ and $\|A\|$ for the number of them respectively. Thus, $B = \{(\beta_1, \rho_1), \ldots, (\beta_m, \rho_m)\}$ and $A = \{(\alpha_1, \rho_1), \ldots, (\alpha_N, \rho_N)\}$. The actual monthly (i.e., $m$ and $m = 1, \ldots, M$) data usage of the $i$th consumer is $x_i \in \mathbb{R}^M$, and the consumers’ total consumption is then $\theta = \sum_{i=1}^M |x_i|$. Following Nevo et al. (2016) who find (1) consumption is highly impacted by the subscribed allowance and overage reduces welfare (e.g., overage aversion in Ater & Landsman, 2013), (2) consumers’ willingness to pay is heterogeneous as their usage is in different ways, and (3) consumers are forward-looking and economically responsive to variation of usage (see Xu et al., 2019), we then consider the following two types of consumers: those who consider the allowance (i.e., the quantity given an access fee) and those who economically count for total cost (i.e., price). For the former, they have stable consumption and rarely consume overage and their decision of subscribing to a plan $\beta_0$ is to minimize their specific cost function $C_i(\cdot)$ on overage payment as follows:

$$
\beta_i = \min_{\beta_j} C_i(\beta_j | x_i) := \arg\min_{\beta_j} \left( \beta_j - \frac{1}{M} \sum_{m=1}^M x_{i,m} \right)^2,
$$

(45)

where $j \in [1, \|B\|]$, $\beta_j$ is the $j$th data allowance, and $x_{i,m}$ is the $i$th consumer’s actual data usage in the $m$th month. For the latter, consumers make their decision of subscribing to a plan $\beta$ so as to minimize their specific cost function $C_i(\cdot)$ on total payment for access and overage as follows:

$$
\beta_i = \min_{\beta_j} C_i(\beta_j | p_k) := \arg\min_{\beta_j} \left( \sum_{m=1}^M m \cdot p_k + \sum_{m=1}^M \sum_{n=1}^N n \cdot \delta_{m,n} \cdot \rho_n \right).
$$

(46)

where $p_k$ is the access fee of the subscribed plan, $\delta_{m,n}$ is the $n$th quantity of overage in $m$th month, and $\rho_n$ is $n$th unit overage price. Several studies highlight consumers’ uncertainty about their (future) usage, which cellular carriers can exploit through 3PT pricing (Fibich et al., 2017). We do not rely on the argument based on consumers’ risk aversion, but we agree that risk neutrality is appropriate here because consumers can obtain exact historical usage records provided by cellular carriers and be economically responsive to them (Xu et al., 2019).

If the usage at the $m$th month $x_{i,m}$ does not exceed the subscribed allowance $\beta_i$, then these two types of consumers do not
Algorithm 1 Calculate average of individual consumer $\delta(x, \beta_k, A)$.
Input: allowance $\beta_k$, average $A$, and individual consumption $x_i$. 
Output: the consumption matrix of average unit $\delta$ of ith consumer for $m = 1$ to $M$ do
for $n = 1$ to $|A|$ do
if $n = |A|$ then
  $\delta_{m,n} = \left[ \frac{x_m - \beta_k}{\alpha_n} \right]$, if $x_m - \beta_k > \alpha_n - 1$ and $x_m - \beta_k < \alpha_n$. 
  0. otherwise
else if $n > 1$ and $n < |A|$ then
  $\delta_{m,n} = \left[ \frac{x_m - \beta_k - \sum_{i=1}^{n-1} \alpha_i \delta_{m,i}}{\alpha_n} \right]$, if $0 < x_m - \beta_k$. 
  $- \sum_{i=1}^{n-1} \alpha_i \cdot \delta_{m,n_1} < \alpha_n - 1$. 
  0. otherwise
else
  $\delta_{m,n} = \left[ \frac{x_m - \beta_k - \sum_{i=1}^{n-1} \alpha_i \delta_{m,i}}{\alpha_n} \right]$, if $x_m - \beta_k > \alpha_n - 1$. 
  $- \sum_{i=1}^{n-1} \alpha_i \cdot \delta_{m,n_1} > 0$. 
  0. otherwise
end if
end for
end if
end for

consume average. When they necessarily consume average (i.e., $x_i \text{ exceeds } \beta_k$), they shall purchase possibly a large enough amount of average to satisfy their usage. As shown in Table 5, the prices per GB of four different-sized average are proportionally inconsistent (i.e., nonlinear) on average usage. These prices will give consumers different priority when purchasing average. For a rational consumer, the purchase of average is to utmostly satisfy the excess usage at a possibly less cost. As shown in Table 5, for large average, 3GB or 5GB seems more economical and practical for consuming average. On the other hand, if the actual usage is just slightly larger than the average, then the smaller size of average costs less. Given $M$ months of actual usage $x_i \in \mathbb{R}^M$ and the allowance subscribed $\beta_k$ by Type I consumer in Eq. (45) and Type II consumer in Eq. (46), the partial payment of average consumption $\delta(x_i, \beta_k, A) \rho_1$ can be calculated by Algorithm 1 with a vector of all unit average fee $\rho$ and $\delta \in \mathbb{R}^{M \times |A|}$.

Since the actual usage consumed by ith customer $\theta_k$ is different from the subscribed one $\beta_k$, we shall use the total actual usage $\Theta = \bigcup_{k=1}^{K} \theta_k$, to calculate the carrier’s total revenue as follows:

$$R(\Theta, A, B) = \sum_{k=1}^{K} \sum_{i=1}^{|\Theta_k|} \left[ 1_{\text{access fee}} | p_k + [\delta(x_i, \beta_k, A) \rho_1] | \right].$$

where $A$ and $B$ indicate different components, like access and average bundles, of the 3PT billing plan.

3.2.2. Strategic modification of pricing

The descriptive results in Table 4 exhibit that there exist four usage clusters from the actual usage data. The current billing plans in Table 5 indicate the cellular carrier is currently performing six access plans and four average plans. According to such tiered pricing, the possibility of mismatch between price and usage is shown to be very high (Baldacci et al., 2022). A strategic modification is therefore pragmatic for different consumers (see Geng & Shulman, 2015; Gopalakrishnan et al., 2015; Julienne et al., 2022, for example). This section presents a modification based on the analytic and empirical results in the related literature as follows. First, we need to avoid significant losses in revenue as much as possible. Given $\varepsilon = R(\Theta, A', B') - R(\Theta, A, B)$ as the difference between new revenue after modification and the existing revenue with currently billing plans, we minimize the cellular carrier’s cost function of operations $C_0(\varepsilon)$ when $\varepsilon \to 0$ with the given revenue calculated by Eq. (47). Second, when prices change, the changes in benefits to different consumers also differ. We need the number of worse-off customers to be smaller than the number of better-off customers.

We apply the proposed clustering method to estimate the mean usage from each cluster. In order to avoid the impact of the usage of a specific month in the sample, we adopt a random sampling method to drop 6 months (i.e., 50% of total data) from 12 months in the estimation (we call it in-sample and the rest is out-of-sample). The total sample size is then a permutation of $\binom{12}{6}$, 924, for both in-sample and out-of-sample usage data. Table 6 shows the random sampling results of mean usage that serves as allowance ($\beta^*$ for $B^*$). We then need to find a k-tiered optimal access fee $p$ by solving the following problem:

$$\min_p \Psi(\varepsilon),$$

s.t. $\sum_{k=1}^{K} \sum_{i=1}^{|\Theta_k|} \left( 1_{\text{access fee}} | p_k + [\delta(x_i, \beta_k, A) \rho_1] | - R(\Theta, A, B) \right)$

$$p_j \leq \frac{p_{j-1}}{\beta_{j-1}} \leq \frac{p_{j-1}}{\beta_{j-1}} \leq \frac{p_{j-1}}{\beta_{j-1}} \leq \beta_{j-1} \leq \beta_{j-1}.$$ 

where we only consider the changes for the access bundle $B^*$ (i.e., access fee and allowance). We consider all consumers are either Type I or Type II defined by Eqs. (45) and (46), respectively. In Table 6, the allowance $\beta^*$ determined by the proposed method differs from the billing plans in Table 5. The unit access fee per gigabyte of allowance in Table 5 is monotonic decreasing; i.e., the increment of the tariff is less than the increment of $\beta$. Therefore, we consider this monotonically decreasing constraint as a small discount suggested in Leider & Sahnin (2014) when applying convex optimization to solve $p^*$.

We define percentage benefit $\frac{\Psi(p^*) - \Psi(p)}{\Psi(p)}$ to 100% to evaluate the size of $\varepsilon$ and show the results in Table 7. The results show that access bundle adjustments lower overall revenue. When we consider the attractiveness of a billing plan with the monotonic decreasing constraint on unit access fee for a given allowance, it results in a revenue loss of 1.0380% to 9.9692% in our study, which is not negligible for cellular carriers. The monotonic decreasing constraint is equivalent to lowering the unit access fee, which is not

Table 6

<table>
<thead>
<tr>
<th>µ</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
<th>Cluster 3</th>
<th>Cluster 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>Std</td>
<td>Mean</td>
<td>Std</td>
<td>Mean</td>
</tr>
<tr>
<td>MMG</td>
<td>1.0693</td>
<td>(1.1771)</td>
<td>6.3723</td>
<td>(3.9258)</td>
</tr>
<tr>
<td>MMT</td>
<td>0.8827</td>
<td>(0.3048)</td>
<td>3.6308</td>
<td>(1.1824)</td>
</tr>
<tr>
<td>MMST</td>
<td>0.5690</td>
<td>(0.2735)</td>
<td>3.0148</td>
<td>(1.0167)</td>
</tr>
</tbody>
</table>

Table 7

<table>
<thead>
<tr>
<th>Percentage Benefit</th>
<th>MMG</th>
<th>MMT</th>
<th>MMST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type I</td>
<td>Mean</td>
<td>-9.9692</td>
<td>-5.1027</td>
</tr>
<tr>
<td>Consumer</td>
<td>Std</td>
<td>(1.8175)</td>
<td>(2.1419)</td>
</tr>
<tr>
<td>Type II</td>
<td>Mean</td>
<td>-8.1778</td>
<td>-3.1140</td>
</tr>
<tr>
<td>Consumer</td>
<td>Std</td>
<td>(1.8228)</td>
<td>(2.1077)</td>
</tr>
</tbody>
</table>
optimal given the $\beta^*$. Our results are consistent with Grubb & Osborne (2015) in that a higher access fee makes consumers worse off, while a lower access fee makes cellular carriers worse off. On the other hand, the results differ from the experimental study of Leider & Sahin (2014) who report a small discount in 3PT could increase revenue.

### 3.2.3. Convex optimization for price bundle in 3PT

The price bundle contains an access fee and unit price of overage in 3PT pricing. In Section 3.2.2, we show that only modifying the access bundle does not necessarily increase revenue if such a change does not reflect the actual consumption features. In order to derive an eligibility billing plan, we introduce an affordable way based on the strategy given in the previous section by simultaneously increasing overage by $\xi$, $\xi > 0$, which controls the adjustment altitude (i.e., the lower and upper bounds) of overage prices. Therefore, the new revenue is attributed by access component $B^*$ and overage component $A^*$ of the price bundle in 3PT. The convex optimization problem in Eq. (46) is then given as follows:

$$\min_{\mathbf{p}, \mathbf{p}^*} C_0(\epsilon).$$

s.t. $\sum_{k=1}^{n} \left( 1 - \epsilon \right) \cdot p^*_k + \beta_1(\mathbf{x}, \mathbf{\beta}^*, A^*, \mathbf{p}^*) - R(\Theta, A, B);$ $p^*_j \leq \frac{\beta_j}{\beta_j} \cdot p^*_{j-1} \quad j = 1, \ldots, |A^*|$ and $\beta_{j-1} \leq \beta_j \leq \beta_j;$ $0 < (1 - \xi) \rho \leq \rho^*_{\xi} \leq (1 + \xi) \rho, \quad \xi > 0, n = 1, \ldots, |A^*|.$

(49)

To solve the optimization problem in Eq. (49) following the idea of Bubeck (2015), we set $\xi = 0.05, 0.1, 0.15, 0.2, 0.25$, and 0.3 by attempting to incorporate the simulation results in Grubb & Osborne (2015) under overage aversion (Ater & Landsman, 2013). When the total payment of $i$th consumer is lower than before the price adjustment, we define it as better-off; otherwise, it is defined as worse-off. We show the number of beneficiary (i.e., better-off) consumers in Table 8. We can see whether it is for the Type I or the Type II consumers that the new billing plans clustered by the multivariate mixture skew $t$ distribution will benefit more consumers in general in comparison with the current billing plan in Table 5. In addition, Fig. 3 graphically shows the details of better-off consumers in comparison with their current billing plans and the distribution of beneficiary consumers.

We now compare the welfare changes after price modification and show the in-sample and out-of-sample comparisons in Figs. 4 and 5, respectively. We see when the total revenue of the cellular carrier remain unchanged (i.e., $C(\epsilon)$ approaches to zero) that heavy users will be penalized with a price to subsidize those who use their data within the allowance.

We next use the performance dummy $\xi^*(i)$ for evaluating the robustness of the results in Fig. 3 (i.e., the beneficiary consumers with billing plans clustered by the underlying multivariate mixture distributions) with the following variance components model:

$$Y_{i,m} = a_i + \sum_{j=1}^{6} \xi^*(i) b_j X_{i,m} + \epsilon_{i,m}.$$

where $Y$ stands for the number of beneficiary consumers attributed by one multivariate mixture distribution and for the number of beneficiary consumers attributed by comparing the multivariate mixture distribution. Here, $\xi^*(i) = 1$ for the $i$th consumer who impacts $j$th change of $\xi$ in Table 8 and $\xi^*(i) = 0$ otherwise. We report the results in Table 10 and see that the number of beneficiary consumers given by the MMST distribution is larger than that of the MMG and MMT distributions when increasing the overage fee from 5% to 30%. We thus conclude that when simultaneously modifying the access and overage components (i.e., decreasing access fee in $B^*$ and increasing overage fee in $A^*$) of the pricing bundle in 3PT, the allowance $\beta$ generated by the proposed method with MMST distribution will financially benefit more consumers while keeping the carrier’s revenue unchanged.

### 3.2.4. Implementation for customer retention

Retention is a measure of the long-term loyalty and interest of users towards a product or service. As opposed to switch or churn (see Gattermann-Itscher & Thonemann, 2021; Höppner et al., 2020; Maldonado et al., 2020, for example), it is centered on capturing users’ longevity of persistence and is often measured by the duration of their continuous relationship with a service provider. The literature identifies variety-seeking behavior as a major factor influencing customer retention, and consumers with low product involvement tend to exhibit higher levels of variety-seeking behavior, meaning that products with low perceived involvement are less likely to retain their consumers. From this perspective, those heavy users are likely to retain a high degree of loyalty (see Gu et al., 2022, for example). Therefore, the impact of price changes on their retention may greatly affect the changes in the operator’s revenue. Therefore, it is necessary to analyze the change in revenue brought about by the change in customer retention.
Fig. 3. Proportion of beneficial customers from price optimization compared by the previous billing plans.
Fig. 4. Welfare analysis of impacted customers (in sample) from price modification.
Fig. 5. Welfare analysis of impacted customers (out of sample) from price modification.
Table 9
Comparison of percentage revenue decrease under four price tiers, calculated using multivariate mixture distributions, for different levels of retention. Retention is determined by the data consumption of users whose expenses fall below the 90th, 95th, and 99th percentiles for both Type I and Type II customers.

<table>
<thead>
<tr>
<th>Retention</th>
<th>Price tier</th>
<th>Type I consumer</th>
<th>Type II consumer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MMG</td>
<td>MMT</td>
</tr>
<tr>
<td>90-Percentile</td>
<td>1</td>
<td>0.3426</td>
<td>0.2239</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>3.3213</td>
<td>2.5610</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8.5569</td>
<td>5.0050</td>
</tr>
<tr>
<td>(Mean)</td>
<td>1</td>
<td>6.2402</td>
<td>5.1023</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>1.8294</td>
<td>1.4357</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>4.7358</td>
<td>2.9654</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>8.6766</td>
<td>8.3627</td>
</tr>
<tr>
<td>(Mean)</td>
<td>1</td>
<td>3.8583</td>
<td>3.2352</td>
</tr>
<tr>
<td>99-Percentile</td>
<td>1</td>
<td>0.0509</td>
<td>0.0471</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.4594</td>
<td>0.4021</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>1.2731</td>
<td>0.7877</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>3.0204</td>
<td>2.8526</td>
</tr>
<tr>
<td>(Mean)</td>
<td>1</td>
<td>1.2009</td>
<td>1.0224</td>
</tr>
</tbody>
</table>

Table 10
Comparing the number of beneficiary customers who are impacted by average changes under different multivariate mixture distributions. When the underlying value of $\xi$ is greater than 1, the distribution of $Y$ creates a significant number of beneficiary customers of the distribution of $X$.

<table>
<thead>
<tr>
<th>Type I consumer</th>
<th>Type II consumer</th>
</tr>
</thead>
<tbody>
<tr>
<td>In sample</td>
<td>$a$ $\xi$</td>
</tr>
<tr>
<td>----------------</td>
<td>-----------------</td>
</tr>
<tr>
<td>$X_{\text{MMG}}$</td>
<td>$b$ 0.8225</td>
</tr>
<tr>
<td>$X_{\text{MMT}}$</td>
<td>$SE^*$ 1.3427</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 0.6131</td>
</tr>
<tr>
<td>$Y_{\text{MMG}}$</td>
<td>$b$ 0.8014</td>
</tr>
<tr>
<td>$Y_{\text{MMT}}$</td>
<td>$SE^*$ 6.8997</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.1616</td>
</tr>
<tr>
<td>$X_{\text{MGT}}$</td>
<td>$b$ 0.8100</td>
</tr>
<tr>
<td>$X_{\text{MST}}$</td>
<td>$SE^*$ 6.3613</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.2733</td>
</tr>
<tr>
<td>$Y_{\text{MGT}}$</td>
<td>$b$ 0.7744</td>
</tr>
<tr>
<td>$Y_{\text{MST}}$</td>
<td>$SE^*$ 6.8852</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.1584</td>
</tr>
<tr>
<td>$X_{\text{MGT}}$</td>
<td>$b$ 0.8264</td>
</tr>
<tr>
<td>$X_{\text{MST}}$</td>
<td>$SE^*$ 7.2253</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.1438</td>
</tr>
<tr>
<td>$Y_{\text{MGT}}$</td>
<td>$b$ 0.8776</td>
</tr>
<tr>
<td>$Y_{\text{MST}}$</td>
<td>$SE^*$ 6.124</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.4287</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Out of sample</th>
<th>$a$ $\xi$</th>
<th>$a$ $\xi$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_{\text{MMG}}$</td>
<td>$b$ 0.7744</td>
<td>$b$ 0.9856</td>
</tr>
<tr>
<td>$X_{\text{MMT}}$</td>
<td>$SE^*$ 6.8852</td>
<td>$SE^*$ 5.2182</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.1584</td>
<td>T-stat$^t$ 5.1579</td>
</tr>
<tr>
<td>$Y_{\text{MMG}}$</td>
<td>$b$ 0.8264</td>
<td>$b$ 1.9591</td>
</tr>
<tr>
<td>$Y_{\text{MMT}}$</td>
<td>$SE^*$ 7.2253</td>
<td>$SE^*$ 7.0076</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.1438</td>
<td>T-stat$^t$ 7.0076</td>
</tr>
<tr>
<td>$X_{\text{MGT}}$</td>
<td>$b$ 0.8776</td>
<td>$b$ 1.4903</td>
</tr>
<tr>
<td>$X_{\text{MST}}$</td>
<td>$SE^*$ 6.124</td>
<td>$SE^*$ 1.4903</td>
</tr>
<tr>
<td></td>
<td>T-stat$^t$ 1.4287</td>
<td>T-stat$^t$ 1.4903</td>
</tr>
</tbody>
</table>

Under the new price tiers, the total revenue is calculated by:

$$R(\Theta, \mathbf{A}^*, B^*) = \sum_{i \in \Theta} \sum_{k \in \mathbb{P}} \left[ 1_{M_k} \left| \mathbf{p} - |\mathbf{d}(x, \mathbf{b}^*_k, \mathbf{A}^*) - \mathbf{p}^*| \right| \right].$$

We assume no switching cost (e.g., with number portability) for mobile subscribers to replace or terminate the contract, although there exists a correlation between customer switching costs and economic performance of firms in the literature (see Abolfathi et al., 2022a, for example). The previous optimal price adjustment in Section 3.2.3 is based on two types of users with different preferences, and it is for searching for an optimization plan that affected the least number of people under the premise of constant revenue. Based on the research results in Maldonado et al. (2020), we further assume that type I and type II users of each price tier are equally affected by the price modification and will churn only when their total consumption (including add-on) exceeds a certain cut-off point of the corresponding expenses in that price tier. From Eq. (50), for the ith consumer who has $x_i \in \Theta_i$, the total expense is calculated as:

$$C(x_i, \mathbf{p}^*) = \sum_{k \in \mathbb{P}} \mathbf{d}(x, \mathbf{b}^*_k, \mathbf{A}) \cdot \mathbf{p}^*.$$
tion probability based on Eq. (50). To consider the varying profitability of different customers, Lemmens & Gupta (2020) propose a profit-based loss function that weights more heavily those customers who have the greatest (positive or negative) impact on the profit. Therefore, when we analyze revenue reduction due to customer retention decrease after price modification, we rank customers’ cut-off point of their expenses, see Eq. (51), by percentile in each tier to illustrate their incremental impact of retention on the total revenue.

In the context of the optimal price plan derived from three distinct multivariate mixture distributions, the allocation of subscribing consumers to different price tiers is achieved through clustering based on their respective usage patterns. Consequently, the number of users assigned to each tier may not be uniform. In light of this, the utilization of percentiles, as opposed to percentages, to rank heavy users who are at risk of churning confers several advantages. First, it ensures that customers within each price tier are assigned an equivalent probability of churning, rather than a consistent headcount ratio. Second, as the data usage within each price tier may not exhibit symmetrical distribution, percentiles effectively capture the degree of skewness from the central tendency of the probability distribution. Lastly, percentiles offer interpretability in managing non-integer calculations, particularly when rounding off numbers beyond the decimal point, which is a task that may pose challenges when using percentages.

In our study we choose the 90th, 95th, and 99th percentiles as the threshold values to define churn in each price tier, based on the multivariate mixture distributions. Specifically, only users whose expenses exceed the respective percentile cut-off point for their price tier will be considered as churned users. By employing this approach, we can accurately quantify the change in total revenue (i.e., the reduction in total revenue attributed to churn), which can be calculated as follows:

\[
\text{Revenue Decrease} = \frac{\text{Revenue without Churn} - \text{Revenue with Churn}}{\text{Revenue without Churn}} \times 100\%.
\]

The findings of our study are presented in Table 9, revealing several important observations. First, as consumer retention decreases, the average total revenue exhibits a corresponding decline. For instance, when the tail percentile of churn increases from 1 to 10, indicating a decrease in consumer retention, the average revenue for Type I consumers decreases from 1.2006% to 6.2402% under MMG tiered pricing, from 1.0224% to 5.1023% under MMST tiered pricing, and from 0.9109% to 4.6906% under MMST tiered pricing. Similarly, the average revenue for Type II consumers decreases from 1.2011% to 6.2443% under MMG tiered pricing, from 1.0550% to 5.0804% under MMST tiered pricing, and from 1.0020% to 4.6153% under MMST tiered pricing.

Second, our analysis reveals that MMST tiered pricing results in the least reduction in revenue with decreasing consumer retention for both Type I and Type II consumers, while MMG tiered pricing leads to the highest revenue decrease with the same magnitude of consumer retention decrease.

Third, our findings indicate that consumers who subscribe to higher-priced services have a greater impact on revenue reduction with the same magnitude of consumer retention decrease. These results are consistent with the empirical findings reported in the literature; see Lemmens & Gupta (2020); Reme et al. (2022) and references therein.

This study solely reflects changes in consumer retention by the total data consumption, denoted as \( \Theta \) in Eq. (50). This implies that when calculating revenue, the data usage of consumers who churn is excluded from the entire contract period \( M \), assuming a consistent usage pattern throughout the contract period. However, the method proposed in Section 2.1 for determining the price tier aims to capture the heterogeneity in data usage patterns among different consumers. This does not lead to a paradoxical result, but rather demonstrates that employing a model of skewness can effectively capture individual differences in data usage patterns, making it a valuable tool for optimal price decision-making.

The results obtained therefore indicate that the MMST model enables optimized price adjustments without altering consumer retention, while minimizing the reduction in total revenue, even in scenarios where consumer retention is impacted due to price modifications. This finding underscores the effectiveness of the MMST model in achieving optimal price adjustments that balance consumer retention and revenue optimization.

3.2.5. Discussion and future works

In the literature of nonlinear pricing, empirical investigations with a static comparative study focus on adjusting either the access fee (see Ater & Landsman, 2013, for example) or overage fee (see Geng & Shulman, 2015, for example) to ensure cellular carriers’ profitability. The theoretic model given by Fibich et al. (2017) suggests an optimal policy should reduce the monthly allowance and access fee, but increase overage price. Our work considers all three components of the tariff plan (fixed fees, unit allowances, and overage fees) for tiered nonlinear pricing and discriminates consumer segments for heterogeneous consumers who are data-driven. The optimal decision in our study shows the possibility of keeping revenue unchanged while the cellular carrier can increase allowance and overage fee and at the same time reduce access fee. In other words, on the premise of increasing the allowance, cellular carriers can optimize revenue with any customized pricing of consumers’ usage, which is suggested by Juilen et al. (2022).

Similar to Fibich et al. (2017) who consider only two segments (light and heavy users) of risk-neutral consumers, we focus on actual data usage and consider two consumer segments: inattentive in price (Type I) and inattentive in allowance (Type II) based on the findings of Grubb & Osborne (2015) about marginal-price uncertainty. Type I consumers satisfy the findings in Ater & Landsman (2013) whereby customers adopt non-cost-minimizing plans with a large allowance and high fixed fee, while Type II consumers suffer the overage aversion in Ater & Landsman (2013) as well. Our results confirm the finding in Geng & Shulman (2015) that an overage fee strategy could be more profitable when it effectively steers consumers away from costly overage.

Many studies point out the demand uncertainty when consumers choose a service plan caused by the temporal separation between their subscription choice decision and usage decision. When there is higher demand uncertainty, a higher monthly access fee and usage allowance are preferred by consumers (see Chen et al., 2019). Type I consumers in our optimization problems make firm pricing decisions (without counting for demand uncertainty) become obviously suboptimal. Temporal separation and subsequent demand uncertainty lead to distinctive cost implications for Type II consumers who exhibit a strong characteristic of forward-looking in Xu et al. (2019). In a future work, an optimization problem shall be investigated for a pooled sampling of Type I and Type II consumers as Xu et al. (2019) show that not all users are forward-looking and 40% of consumers are indeed forward-looking in their data.

Price modification can impact consumer retention, and in our research, we assume that there are no switching costs associated with consumers changing service providers. Due to limitations in identifying individual consumer termination situations, such as early termination within the contract term or non-renewal after contract expiration, we did not consider the dynamics of consumer termination during the contract period in our study. However, in future research, with proper permissions from data providers, in-
corporating dynamic changes in consumer behavior during the contract period can be explored, as it is of practical interest.

In our study we assume that changes in customer retention are uniform for both types of customers. This is an area that could be explored further in future research, especially considering the potential differences in retention patterns between the two types of consumers at different price tiers.

4. Conclusion

In this paper we propose a novel clustering method that attempts to integrate the spectral approach (with graphic initialization) and Bayesian approach (through EM algorithm) to reduce the impact of extreme values. Inspired by the idea of spectral clustering, we introduce a graphic initialization to distinguish outliers in the global density, instead of applying the conventional spectral approach to directly assign the original data points into balanced clusters. This not only reduces the impact of the subsequent estimation process using the probability model, but also reduces the limitation of the lack of data descriptiveness in the spectral approach. For estimation, we continue with the centroid-based probabilistic approach and use multivariate mixture skew $t$ distribution estimated by the likelihood of the Bayesian approach to better characterize the extreme points when describing each cluster. We then apply the clustering method in business analytics to find optimal nonlinear pricing. By solving a convex optimization problem, we show that increasing the allowance by actual data usage and overage fee (no more than 25%) with a smaller access fee reduction could be optimal for cellular carriers without jeopardizing realized revenue.

Some empirical studies on nonlinear (or 3PT) pricing for telecommunication services consider consumer uncertainty that leads consumers to pay more for nonlinear plans and leave large unused allowances in a way that seems irrational. One argument is based on consumers’ risk aversion, stating that consumers prefer deterministic payments under a nonlinear pricing regime. In our study, we do not interpret our results according to consumers’ risk preferences, since cellular operators provide consumers with instantaneous usage status, or consumers can access this information at any time. Such information transparency reduces consumers’ uncertainty about their usage, so risk attitudes become neutral. For risk-neutral consumers, overconfidence and biased beliefs about their usage distribution, especially at higher moments of their usage distribution, may not necessarily be favorable to cellular carriers offering nonlinear prices. While companies can still leverage consumers’ uncertainty about usage well with nonlinear prices, their profitability still depends on the operational costs of implementing such a pricing scheme.

To our knowledge, our analysis provides the first data-driven optimization study of how price bundling combinations in nonlinear pricing can help cellular carriers modify their fixed allowances, even if the modifications are not solely for the purpose of increasing revenue. We show that nonlinear pricing (i.e., 3PT) may be more profitable even when access prices are reduced and overage charges are increased to the detriment of a small number of consumers served by the cellular carrier. We also note another new aspect of our work. Subsidizing consumers with lighter usage by allowing some heavy users to be penalized by overage charges, thus allowing them to maintain the overall fixed allowance at a higher level that would satisfy the actual usage of the majority of consumers.

Acknowledgments

This work was supported in part by the Center for Open Intelligent Connectivity from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education, National Science and Technology Council (NSTC 110-2622-8-A49-022, 112-2221-E-A49-049), and NCKU Miin Wu School of Computing in Taiwan. We would like to extend our sincere gratitude to Professor W. Brent Lindquist for organizing the Mathematics & Statistics seminars at Texas Tech University, USA. Additionally, we express our immense appreciation to all the participants for their generous contribution of valuable time, thoughtful reviews, constructive feedback, and insightful comments. Their significant contributions have greatly enriched our research.

Appendix A. Summary of notations in Section 2

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td></td>
</tr>
<tr>
<td>$</td>
<td></td>
</tr>
<tr>
<td>$\Theta$</td>
<td>Zero vector, where a subscript indicates its dimension</td>
</tr>
<tr>
<td>$\Gamma$</td>
<td>Vector of all ones, where a subscript indicates its dimension</td>
</tr>
<tr>
<td>$K$</td>
<td>Actual consumption of the $i$th consumer in the $m$th period</td>
</tr>
<tr>
<td>$\mu_j$</td>
<td>Vector of actual consumption in M dimension for the $i$th consumer</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>All data points, e.g., total data consumption</td>
</tr>
<tr>
<td>$\Psi$</td>
<td>Set of all parameters of the mixture multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$M$</td>
<td>Total number of consumption periods in $\Theta$</td>
</tr>
<tr>
<td>$K$</td>
<td>Total number of components for the mixture multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$z_i$</td>
<td>Indicator vector for the membership of the $i$th consumer in each cluster</td>
</tr>
<tr>
<td>$w_k$</td>
<td>Mixture weight for the $k$th component multivariate skew $t$ distribution, $\sum_k w_k = 1$</td>
</tr>
<tr>
<td>$v_k$</td>
<td>Degrees of freedom in the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$\varepsilon_k$</td>
<td>Skewness vector for the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$\mu_k$</td>
<td>Location vector of the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$\Delta_k$</td>
<td>Diagonal matrix of $\varepsilon_k$ of the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$\delta_j$</td>
<td>Matrix of $\Delta_j C_j^{-1} (x - \mu_j)$ of the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$C_j$</td>
<td>Matrix of $\Omega_j + \Delta_j \Delta_j^T$ of the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$\eta_j$</td>
<td>Scalar of $(x - \mu_j)^T C_j^{-1} (x - \mu_j)$ for the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$\Theta_k$</td>
<td>Matrix of the $k$th component multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$A_k$</td>
<td>Matrix of the $k$th multivariate skew $t$ distribution</td>
</tr>
<tr>
<td>$e_{i,j}$</td>
<td>Edge between $x_i$ and $x_j$, where $i, j \in |\Theta|$ and $i \neq j$, $\bigcup_j e_{i,j} = \mathcal{E}$</td>
</tr>
<tr>
<td>$\mathcal{A}$</td>
<td>Adjacency matrix of $\mathcal{E}$ and its element</td>
</tr>
<tr>
<td>$A$</td>
<td>Adjacency matrix of $\mathcal{E}$ and its element</td>
</tr>
<tr>
<td>$\Xi$</td>
<td>Diagonal matrix of $\varepsilon$ and its element</td>
</tr>
<tr>
<td>$G(\Theta, \mathcal{E})$</td>
<td>Graph of $\Theta$ with edges $\mathcal{E}$</td>
</tr>
<tr>
<td>$D(\mathcal{D}, \mathcal{D}')$</td>
<td>Disjoint sets and $\mathcal{D} = \mathcal{D} \cup \mathcal{D}'$</td>
</tr>
<tr>
<td>$\epsilon(\cdot)$</td>
<td>Cost function of partitioning</td>
</tr>
</tbody>
</table>

Appendix B. Summary of notations in Section 3

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_j, \alpha$</td>
<td>Overage plan in gigabyte generated by the cellular carriers, $\bigcup_j \alpha_j = \alpha$</td>
</tr>
<tr>
<td>$\beta_i, \beta$</td>
<td>Allowance in gigabyte, $\bigcup_j \beta_i = \beta$</td>
</tr>
<tr>
<td>$\rho_j, \rho$</td>
<td>Tariffs in USD given by the cellular carrier (see Table 5), $\bigcup_j \rho_j = \rho$</td>
</tr>
<tr>
<td>$p, p'$</td>
<td>Tariffs in USD derived with Eq. (49); see Algorithm 1</td>
</tr>
</tbody>
</table>

Greek Symbols

<table>
<thead>
<tr>
<th>Letter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Probability density function or distribution function</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Probability density function or distribution function</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Probability density function or distribution function</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Probability density function or distribution function</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Probability density function or distribution function</td>
</tr>
</tbody>
</table>
References
