Merging anomalous data usage in wireless mobile telecommunications: Business analytics with a strategy-focused data-driven approach for sustainability

Yi-Ting Chen a,c, Edward W. Sun b,* , Yi-Bing Lin a

a College of Computer Science, National Chiao Tung University (NCTU), Hsinchu, Taiwan
b KEDGE Business School, France
c School of Business Informatics and Mathematics, University of Mannheim, Germany

A R T I C L E   I N F O

Article history:
Received 20 June 2017
Accepted 23 February 2019
Available online xxx

Keywords:
Analytics
Data mining
Decision support systems
Artificial intelligence
OR in telecommunications
Validation of OR Computations

A B S T R A C T

Mobile internet usage has exploded with the mass popularity of smartphones that offer more convenient and efficient ways of doing anything from watching movies, playing games, and streaming music. Understanding the patterns of data usage is thus essential for strategy-focused data-driven business analytics. However, data usage has several unique stylized facts (such as high dimensionality, heteroscedasticity, and sparsity) due to a great variety of user behaviour. To manage these facts, we propose a novel density-based subspace clustering approach (i.e., a three-stage iterative optimization procedure) for intelligent segmentation of consumer data usage/demand. We discuss the characteristics of the proposed method and illustrate its performance in both simulation with synthetic data and business analytics with real data. In a field experiment of wireless mobile telecommunications for data-driven strategic design and managerial implementation, we show that our method is adequate for business analytics and plausible for sustainability in search of business value.

© 2019 Published by Elsevier B.V.

1. Introduction

4G LTE, the fourth generation Long-Term Evolution (LTE) mobile telecommunications technology, standardizes high-speed wireless communications for mobile phones and data terminals, and has spread globally. Based on a recent 2016 GSA study (Global mobile Suppliers Association), 503 operators have commercially launched LTE networks in 167 countries.1 Global mobile data traffic will increase 12-fold and the number of mobile subscriptions will reach 9.3 billion by the end of 2018 (Ericsson, 2012). Such a dramatic and rapid expansion of wireless demand calls for sustainable spectrum management systemizing the managerial integration of the radio-frequency spectrum and the telecommunications infrastructure (e.g., transceiver architectures, multi-channel interference, transmission, and core networks) for efficient utilization.

On the other hand, strong usage anomalies have been observed in relation to demographic clusters, educational clusters, geographic clusters, and professional clusters. For cellular communications, Jain, Muller, and Vilcassim (1999) show that cellular service usage levels differ by customer segments (i.e., business/professional and personal). However, such segments have not been substantiated in mobile data usage. Amdocs’ 2015 State of the RAN (Radio Access Network) reported that 10 percent of mobile users has consumed 80 percent of the world’s mobile data traffic and the power users often used as much as 10 times more data than the average mobile subscriber based on 25 million voice and data connections (all with lots of smartphone usage) in major cities around the world.2 Cisco’s Visual Networking Index focused on overall global carrier trends in February 2017 reported that the top 20 percent of mobile users generate 56 percent of mobile data traffic and the top 5 percent of users consume 25 percent of mobile data traffic by September 2016.3 Both business professionals and private customers could demand high data usage with individual preference for data exhausting activities, such as videos, games, virtual reality, and augmented reality. Traditional demographic variables, education background, and career

---

1 Corresponding author.
E-mail addresses: yicheng@mail.uni-mannheim.de (Y.-T. Chen), edward.sun@kedgebs.com (E.W. Sun), linyi@nctu.edu.tw (Y.-B. Lin).

https://doi.org/10.1016/j.ejor.2019.02.046
0377-2217/© 2019 Published by Elsevier B.V.

Please cite this article as: Y.-T. Chen, E.W. Sun and Y.-B. Lin, Merging anomalous data usage in wireless mobile telecommunications: Business analytics with strategy-focused data-driven approach for sustainability, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.02.046
experience in conventional customer segmentation are no longer explicitly efficient. Therefore, the data-driven approach in business analytics for sustainable data usage is greatly desired.

Sustainable operations refer to an enterprise’s pursuit of sustainability in its corresponding systems while taking into account the economic, environmental, and social implications. In general, the goal of all businesses is to act economically, but the social aspects of sustainable operations should not be neglected (Sodhi, 2015). Jaehn (2016) highlights the “triple bottom line” — i.e., the 3Ps (profit, people and planet) ecosystem coined by Elkington (1997) — of sustainable operations on “quantitative aspects of business administration, which in addition to economic objectives aims equally at sustainability in the environmental and/or social sense”. Tang and Zhou (2012) illustrate that a 3Ps ecosystem is comprised of five core elements and various flows.

Enterprises can improve their economic sustainability (such as values and vision) by designing and producing products in an environmentally and socially responsible way. Ulhøi (1995) stresses the requirement of corporate sustainable development (CSD), which remoulds consumer preferences (wants) and steers consumption patterns towards environmentally benign activities by reducing throughput per unit of final products/services. Jenkin, Webster, and McShane (2011) discuss the green information technologies and systems that involve initiatives and programs directly or indirectly contributing to environmental sustainability. Tang and Zhou (2012) use the term “cultivate future consumers” to stress economic sustainability in creating a new market for new values (i.e., re-manufacture/ recycle, eco-friendly, ethical, socially responsible, etc.).

With dramatically increasing data traffic, mobile telecommunication operators could profit sustainably from charging consumable services. A profitable mobile tariff (i.e., billing plan) based on data/service demands is therefore vital (see Lin, Lin, Wu, & Wang (2016)). Besides generating ultimate profit, the operator should guarantee the Quality of Service (QoS) for all services provided to maintain and attract subscribers and reduce customer churn. Satisfaction service provision requires operators to enable (1) higher download and upload data rates, (2) lower packet latencies, and (3) supporting new multimedia services. Business analytics, which is an interdisciplinary context of transforming data into insight for making better decisions, accomplishing business goals, and creating values (see Mortenson, Doherty, and Robinson (2015), Royston (2013), and Vidgen, Shaw, and Grant (2017), for example). In a comprehensive review of sustainable operations, Jaehn (2016) summarizes fields of sustainable operations and classes of systems considered and discusses a number of studies in many sectors. Kunc and O’Brien (2018) address how organizations could include business analytics in their strategy processes and consider the potential role of business analytics in strategic decision support. We would like to add our study on business analytics of sustainable operations for mobile telecommunications services (i.e., Information and Communications Technology, ICT) to the OR community. We hereby attempt to present our business analytics by proposing a density-based subspace clustering method for merging anomalous data usage to reach the desirable business value.

Behaviour-based pricing is widely adopted by mobile operators in designing their tariffs: consumers are charged different prices depending on their consumption patterns (i.e., data usage) (see Lin et al., 2016; Siebert, 2015, for example). An increasing number of studies show that consumers do not always make optimal decisions (Kalayci, 2015), particularly when price obfuscation exists. In practice, we generally observe two cases: (1) consumers pay for a data plan that is above their actual demand (i.e., the specified plan is overestimated), and (2) consumers pay for a data plan that is below their actual demand (i.e., the specified plan is underestimated). In case (1), this increases resource wastage, while in case (2), consumers are required to pay an add-on, which increases future mobile operator reserves. Both cases are unsustainable, as waste and costs increase. Operators therefore need to employ precautionary measures for misestimating data usage/demand.

The reasons behind considering sustainable data plan in the mobile services are given as follows. First, a sustainable data plan is key to behaviour-based pricing, since all profitable tariff/rates are based on data usage. Second, sustainable plans could remould consumer preferences (wants) and steer consumption patterns towards environmentally benign activities. Third, sustainable plans could eliminate price obfuscation and engender efficiency. Chioveanu and Zhou (2013) point out that the “strategic choice of price presentation formats, or simply, price framing, has received relatively little attention in the economics literature in spite of its prevalence”. Fourth, sustainable data plans require sophisticated segmentation in dealing with consumer heterogeneity that interacts with consumer satisfaction. Therefore, sustainable data plans will help operators reach their profitability goals with efficient price framing and steer consumer behaviour towards environmentally benign activities in achieving sustainability.

In search of sustainable data plan, it is essential to find an intelligent segmentation method that captures (or characterizes) consumer heterogeneity of data usage/demand. Segmentation refers to distinct consumer patterns evoked by contextual differences (Schlager & Maas, 2013). Elaborating the requirements of sustainable data plans requires carefully investigating consumer heterogeneity. In this paper, we propose an intelligent segmentation based on a novel density-based subspace clustering method, i.e., a three-stage optimization procedure for data usage that illustrates (1) heterogeneity (i.e., heteroscedasticity) among clusters, and (2) intensive perturbation (i.e., anomalies) and sparsity within clusters. In our work, a penalized likelihood estimation of parameters is undertaken with a modified EM algorithm for the Multivariate Gaussian Mixture (MGM) model. Our approach can simultaneously decide the number of components to be mixed for the underlying Gaussian mixture model, the mixing weights and the parameters of the Gaussian distribution components. Our method makes two contributions. First, we apply a penalized likelihood method to identify the MGM model that combines finite multivariate Gaussian distributions for the modified EM algorithm. Second, we propose an entropy-based initialization algorithm for the modified EM algorithm. Therefore, all three stages of the modified EM algorithm (i.e., initialization, expectation, and maximization) are optimal for heuristics. Our method enhances the capabilities of the classic EM algorithm framework for a mixture model and clustering analysis due to its three characteristics: (1) robustness to perturbation and sparsity, (2) fast convergence to the global optimum, and (3) computational efficiency in heuristic searching.

We then provide a simulation study based on three conventional data sets to verify the accuracy of the proposed method when the ground truth is known and apply the proposed method with the real mobile usage data (which illustrates strong heterogeneity) to transform data into insight for business intelligence (e.g., sustainability and value creation). Our framework of business analytics for wireless mobile telecommunications can be summarized as follows:

- **Data Features**: Descriptive, MCM Clustering, Predictive
- **Mobile Data Usage**: Method Verification
- **Strategic Options**: Prescriptive

We organize this article as follows. Section 2 introduces the industry background and relevant literature focusing on current

---

4 Table in Jaehn (2016), p. 258.
challenges for mobile telecommunications in the ICT industry with respect to sustainability and our proposed method, i.e., intelligent segmentation that fosters the sustainability of behaviour-based pricing. In Section 3, we describe the proposed clustering methodology for intelligent segmentation of data. In Section 4, we present a simulation study that verifies accuracy of the proposed clustering method for synthetic data with different characteristics. In Section 5, we perform a business analytics with the primary data from a large telecom operator in Taiwan. We show that our proposed method indeed suffices the goal of business analytics in search of business value. We summarize and highlight future works in Section 6.

2. Industry background and relative studies

2.1. ICT and sustainable development

Information and Communications Technology (ICT) enables users to access, store, transmit, and utilize information by unifying sophisticated systems for telecommunications. Davenport and Prusak (1997) proposed a human-centred approach based on “Information Ecology” to design and manage information environments encompassing: (1) information strategy, (2) information politics, behaviour, and culture, (3) information staff and management processes, and (4) information architecture. Eryomin (1998) points out that the hybrid “information ecology” science could shape our lives in terms of the social and economic implications of computer and communication technologies. For example, Germany initiated an integration of cyber-physical systems, the Internet of Things, and cloud computing named “Industry 4.0” for intelligent manufacturing in the fourth industrial revolution (see Schwab, 2016).

Digital ecosystems deriving from ICT have been a major contributor to the evolution and growth of the global economy. The Ericsson (2013) report illustrates (1) an approximate 1% increase in GDP for every 10% increase in the broadband penetration rate, and 80 new jobs for every 1000 new broadband users, (2) doubling the broadband speed increases GDP by 0.3%, (3) broadband accessing increases household income by USD 2100 per year in OECD countries with 4 megabits per second broadband and USD 800 per year in BIC countries with 0.5 megabits per second broadband, (4) broadband speed upgrading from 0.5 megabits per second to 4 megabits per second increases household income by around USD 322 per month in OECD countries and by USD 46 per month in BIC countries. UN Secretary-General Ban Ki-moon stated, “ICTs can be an engine for achieving the Sustainable Development Goals. They can power this global undertaking.”

The mobile industry is the first sector to commit as a whole to the United Nations Sustainable Development Goals (SDGs). In September 2008, the Global System for Mobile Communications Association (GSMA) launched the Green Power for Mobile (GPM) programme in order to guide the mobile industry systematically contribute to the environmental issues. With the publication of the 2017 Mobile Industry Impact Report, GSMA highlights its increasing impact on sustainable development. Frisiani, Jubas, Lajous, and Nattermann (2017) point out that mobile operators can increase their sales to existing customers by making timely appealing offers on services or hardware after analysing the customer data. Value creation for business through ICT has been considered as a primary way of contribution to society (see Lee, 2016). Several studies have carried on how ICT can change society and how policies about ICT can influence the society and business. For example, both Green of IT and Green by IT significantly influence the national carbon emission policies and international agreements (see Jenkin et al., 2011). Corporate engagement with communal societies on ICT provides a constitution of the technology-enabled confines of corporate strategy and community influence (see Jeldstad, Snow, Miles, & Lettl, 2012). Racherla and Mandivivalla (2013) integrate a multilevel sociotechnical framework involving both micro and macro factors that are connected to the ICT infrastructure, universal access, and socioeconomics. Uratnik (2016) explain the interaction between corporate leveraging sustainable innovation and virtual community in the co-production and co-creation of value. Based on a sample of 139 countries, Gouvea, Kapelianis, and Kassiech (2018) show that ICT has a significantly positive effect on environmental sustainability.

2.2. Industrial challenges

The profitability of ICT service providers calls for continuous investments in their infrastructure and improving service quality (including matching consumption propensity, i.e., affordability). In-stalling, upgrading, and maintaining their infrastructure requires ICT service providers to pay more attention to sustainability issues not only in terms of the environmental impact, but also to efficiently reallocate scarce resources by avoiding information congestion (Anderson & De Palma, 2009). Service charge tariff or billing policies consequently aid ICT operators in these tasks. On the other hand, affordability remains a major constraint (see Crémer, Rey, & Tirole, 2000). Other tasks challenge ICT service providers when determining the level of investments and expenditure for customer acquisition and retention in competitive and dynamic markets with respect to customer affordability. Min, Zhang, Kim, and Strivastava (2016) empirically show that in wireless telecommunications markets, a firm’s acquisition cost per customer is more sensitive to market position and competition than retention cost per customer. In addressing this issue, an accurate grouping/segmentation of user data demand for an appropriate billing strategy is critical.

In analysing data demand and particularly the heterogeneity of data usage, we face several difficulties. Consumers generally have a degree of certainty about their future usage patterns. This usage uncertainty derives from hidden behaviour due to the fact that we cannot anticipate usage. As online product content changes rapidly over time, consumers are unable to recognize the effects of their current usage on their future usage. Data demand may suddenly be driven by certain unpredictable social media phenomena (e.g., surge in popularity) and users may consequently change their behaviour as well as their data consumption without prior indications. For example, the current Pokémon Go mania (a location-based augmented reality mobile game) or simply Pokémon9 has about 21 million daily active users10 with an average playing time of 33 minutes11 when the game was released in July 2016. Verizon Communications Inc., the largest wireless telecommunications provider in the US, reported that it takes around 10 megabits per hour of data usage and around 7 hours a day for 30 days straight to consume 2 gigabytes of data. Other activities caused by the surge in popularity, such as watching online high-definition videos (that usually consume as much as 350 megabits data us-

---


Please cite this article as: Y.-T. Chen, E.W. Sun and Y.-B. Lin, Merging anomalous data usage in wireless mobile telecommunications: Business analytics with strategy-focused data-driven approach for sustainability, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.02.046
age per hour), increase data demand without any prior indications. In addition, what we call hidden information implies that we cannot obtain sufficient information in our analysis due to (1) consumers’ considerable lack of ICT literacy and corresponding willingness to pay; (2) demographic variables, education background, and career experience in traditional consumer clustering analysis no longer indicate the heterogeneity of data usage; and (3) the regulator’s law/rule for computer-processed personal data protection. This leads to the extremely urgent need to find an appropriate method to analyse (e.g., characterize and forecast) consumer data usage/demand.

Another challenge in analysing data usage is overuse (see Suissa, 2015; Yan, 2015), which the World Health Organization defines a behavioural addiction/dependence syndrome irrespective of age, gender, ethnicity, career, or economic status. Such behavioural addictions clearly lead to abnormally high data demand. In some surveys, over 30% of users perceived themselves as addicted and these users consume nearly twice the data compared to non-addicted users (see Billieux, 2012). Some addicted users report data usage beyond three standard deviations from the upper hinge (see Tossell, Kortum, Shepard, Rahmatt, & Zhong, 2015).

2.3. Our method and relative studies

Commercial mobile telecommunications operators typically employ billing plans that classify customers into clusters of different amounts of mobile data consumption. In this paper, we propose a novel density-based method of demand segmentation to appropriately classify the mobile data usage clusters. In machine learning, our clustering task problem (see Santi, Aloise, & Blanchard, 2016) stems from learning a finitely valued function (i.e., the classifier) on a compact metric space. Without any prior knowledge of the data structure, clustering aims to reveal the natural data structure using similarity measures. Hence, objects that are similar should be placed in the same cluster while objects that are not should be placed in different clusters (see Meyer & Olteanu, 2013). We face the same clustering problem as Santi et al. (2016) such that all available dissimilarity matrices are used to deal with heterogeneity.

Numerous studies use the axiomatic fuzzy set (AFS) clustering methodology (see Bagirov & Yearwood, 2006; Xie, Gao, Xie, Liu, & Grant, 2016; Xu, Liu, & Chen, 2009). Kim, Lee, Lee, and Lee (2005) conduct a kernel-based classification with four clustering algorithms (i.e., K-means, Fuzzy C-means, average linkage, and mountain algorithm) and evaluate their performances for various datasets. Their results indicate that each kernel clustering algorithm evidently performs better than its conventional counterpart. Carrizosa, Mladenović, and Todosićević (2014) propose a variable neighbourhood search method in clustering networks. Bai, Dhavale, and Sarkis (2016) propose a hybrid methodology of fuzzy C-means for decision modelling in green supply chains. However, data usage does not conform to the stereotypical data investigated and entails two stylized facts: (1) features are entangled due to the heterogeneity of many dissimilarity matrices (possibly driven by hidden behaviours and hidden information), and (2) many anomalies exist due to overuse. Density-based clustering therefore constitutes an appropriate clustering method to deal with these stylized facts (see Sander, Ester, Kriegel, & Xu, 1998).

The density-based clustering method separates the feature space into high-density and low-density regions (with nonlinear separating hyper-surfaces). The connected dense regions in the feature space are defined as clusters. Such connectedness (equivalent to similarity) can be characterized by different algorithms.

Jain (2010) summarizes several methods that define connectedness in the feature space, such as (1) the Jarvis–Patrick algorithm defining the connectedness between a pair of points as the number of common neighbours they share, where neighbours are the points in the region of a predetermined radius around the point, (2) DBSCAN algorithm applies the Parzen window method in search for connected dense regions, and (3) spectral (or graph theoretic) clustering that represents the data points as nodes and weighted pairwise similarity as edges that connect the nodes. These methods rely on (1) the neighbourhood size measured by distance, and (2) the minimum number of points that a cluster can include in its neighbourhood. When heteroscedasticity14 exists among clusters, the alteration of these two parameters leads to inadmissible separation in the feature space by using only a single density threshold (see Campello, Moulavi, & Sander, 2013). As a result, probabilistic mixture models have been introduced.

The probabilistic mixture model approach assumes the existence of an underlying data generator driven by a mixture distribution function such that a cluster is defined by one or more mixture components. The EM algorithm is therefore an ideal choice to infer the parameters of these mixture models (see Lin, 2009; Ray & Ren, 2012; Yao, 2010). In addition, the Bayesian approach has been proposed to improve performance when employing mixture models and shows its superiority in terms of data analysis (see Filippone, Camastra, Masulli, & Rovetta, 2008; Jain, 2010). Alternatively, hidden Markov models (HMMs) can be applied by first mapping each data sequence into an HMM, then defining a suitable distance among HMMs, and finally proceeding to cluster the HMMs based on the distance matrix. As HMMs make explicit use of the distance, they are remarkably geared toward continuous valued time-series (possibly multi-dimensional); see Dias, Vermint & Ramos (2015) and references therein. Although HMMs allow various structures to be modelled directly, algorithms for HMMs do not estimate the number of hidden states. In order to train HMMs a (large) set of seed sequences is generally required. When the given seed sequence is long, there are many possible HMMs for it, choosing one can be difficult. On the other hand, HMMs are not conspicuously meaningful for short data sequence because a small number of hidden states still have a high probability to estimate more parameters than the number of observation, see Khreich, Granger, Miri, and Sabourin (2012) and references therein.

In the next section, we propose a three-stage optimization procedure for a density-based subspace clustering analysis in line with the classic expectation-maximization (EM) algorithm (see Bishop, 2006). In our work, a penalized likelihood estimation of parameters is carried out to modify the EM algorithm for the Multivariate Gaussian Mixture (MGM) model inference. Our approach can simultaneously decide the number of components to be mixed for the underlying Gaussian mixture model, the mixing weights, and the parameters of the Gaussian distribution components. Our framework consists in three main procedures, i.e., initialization, expectation, and maximization in the classic EM algorithm framework. In contrast to other studies (see Nguyen, Wu, & Zhang, 2014; Yang, Lai, & Lin, 2012), the main contributions of our work are two-fold. First, we consider the entropy of distances between data in the initialization stage rather than using K-means or K-means+++. Our experimental results show that this is insensitive to extreme values in the dataset. Second, we employ a decision function driven by efficiency criteria (i.e., AIC, BIC and HQC) that simultaneously consider the likelihood and penalty.

---

13 “Customers typically do not know the marginal price until after they decide to consume” (Bushnell & Mansur, 2005).

14 Heteroscedasticity (or heteroskedasticity) here refers to the circumstance in which the variability (or inconsistence) of clusters is unequal across the range of values of a feature that describes it, namely, clusters with different variability quantified by the variance or any other measure of statistical dispersion.
3. Multivariate Gaussian mixture model

This section reports the three stages of the multivariate Gaussian mixture model and shows that all are heuristically optimal.

3.1. Initialization: the first stage optimization

For \( 1 \leq i \leq M \) and \( 1 \leq n \leq N \), let \( \Theta = \{x_1, \ldots, x_n, \ldots, x_N\} \) be a set of measured data usages from \( N \) users, where \( x_n = [x_{n1}, \ldots, x_{nM}] \) is a vector and \( x_{nm} \) is the data usage of the \( m \)th user quantified by the \( m \)th measure.\(^{15}\) We assume that the measured data usages follow a \( K \) multivariate Gaussian mixture model:

\[
f(x_n) = \sum_{k=1}^{K} w_k f_k(x_n),
\]

\[
= \sum_{k=1}^{K} \frac{\exp \left( -\frac{1}{2} (x_n - \mu_k)\Sigma_k^{-1}(x_n - \mu_k) \right)}{\sqrt{2\pi}^{M}|\Sigma_k|},
\]

where \( w_k \) and \( f_k(x_n) \) denote the weight and probability of the \( k \)th Gaussian distribution. Let \( \mu_k \) and \( \Sigma_k \) be the mean vector and the covariance matrix of the \( k \)th Gaussian component, where \( 1 \leq k \leq K \). Given \( \Theta \), the log-likelihood \( L(\Theta) \) for the \( K \) multivariate Gaussian mixture model can be expressed as:

\[
L(\Theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \ln w_k f_k(x_n)
\]

\[
= \sum_{n=1}^{N} \sum_{k=1}^{K} \left( \ln w_k - \frac{M}{2} \ln (2\pi) - \frac{1}{2} \ln |\Sigma_k| - \frac{(x_n - \mu_k)^T \Sigma_k^{-1} (x_n - \mu_k)}{2} \right).
\]

then the Expectation Maximization (EM) algorithm can be derived for the above model-based clustering (elaborated in Section 3.2).

The EM algorithm is easy to implement and numerically stable under certain conditions that suffice for reliable global convergence. However, when multiple maxima exist, it may converge slowly without convergence bounded to the global maximum. In addition, the EM algorithm is sensitive to the initial values of \( w_k \), \( \mu_k \) and \( \Sigma_k \) that start the search on the likelihood surface (Bishop, 2006). The main approach to initialize the EM algorithm is our choice of the EM algorithm is \( K \)-means, where each cluster is represented by the center of the cluster gravity, but \( K \)-means is probably dominated by extreme values in the dataset and therefore results in a singular problem in the EM algorithm. To resolve this issue, we propose a stochastic initialization such that the multiple starting points are chosen and evaluated by their entropy with respect to their ambient values.

Anomalies are those data with extreme values or abnormal distances from most data in the dataset and are not necessarily related to or clustered with each other. Our experience in commercial mobile operations indicates that anomalies do not occupy the main proportion of the dataset. We use the Euclidean distance \( d_{ij} \) to be the measurement between any two data in the dataset. For \( 1 \leq i \neq j \leq N \), we have

\[
d_{ij} = \sqrt{\|x_i - x_j\|^2}.
\]

With Eq. (3), we identify anomalies in the dataset by discrete entropy. Specifically, we use the histogram approach to approximate the probability density function of the distances from one data to all other data in the dataset and estimate the discrete entropy. The histogram partitions the distances from one data to all other data in the dataset into \( B \) disjoint bins, and we therefore have a sequence of distance intervals \( d_0, d_1, \ldots, d_B \). For every \( x_n \), we construct a distance vector \( d_n = [d_{n1}, \ldots, d_{nB}] \), and then partition these distances into histogram \( p_b = [p_{b1}, \ldots, p_{bB}] \), where \( 0 \leq p_{b} \leq 1 \) and \( \sum_{b=1}^{B} p_{b} = 1 \) for \( 1 \leq b \leq B \). The discrete entropy of \( d_n \) is defined as:

\[
H(d_n) = \sum_{b=1}^{B} p_b \log(p_b).
\]

Eq. (4) is used in this paper to estimate the relative positions of data in the dataset.

If \( x_n \) is in the dense area, the distances from \( x_n \) to those data belonging to the same cluster or in the adjacent areas contribute to shorter distances. Those data not in the adjacent areas, i.e., anomalies, result in longer distances. The distribution of \( p_b \) diversifies and consequently \( H(d_n) \) is relatively large. On the other hand, if \( x_n \) is an anomaly, the distances from \( x_n \) to all others are only distributed among some specific ranges with large values and \( H(d_n) \) is relatively smaller. If we use order statistics to sort the entropy values and directly select data with the top \( K \) entropy values as initial values for \( \mu_k \), data from the same cluster is very likely to be chosen, as data with similar entropy values are usually clustered in neighbouring regions. Therefore, we use the weighted random selection approach to initialize \( \mu_k \) and give higher priorities to those data with larger entropy values. Imagine a queue with unequal-sized slots. The size of the slot for each data \( s_n \) to occupy is proportional to the entropy value of the data with respect to the total entropy values of the dataset \( \sum_{n=1}^{N} H(d_n) \), and each slot is numbered in the corresponding index. In every iteration, we generate a random number \( r \in [0, 1) \), find out which slot \( r \) falls into, and the data with the index will eventually be selected as \( \mu_k \).

Without choosing anomalies as initial parameters, we speed up the convergence of the EM algorithm to fit the Gaussian mixtures. A good initial estimate for the covariance matrix and the weight is the within-cluster covariance matrix and the fractions of the number of the data allocated to each cluster. For \( 1 \leq k \leq K \), we define \( K \) indicator functions \( \delta_k \), apply these to \( \Theta \), and define \( \delta_{k,n} = \delta_k(x_n) \) which takes the value \( 1 \) when \( x_n \) is the closest to \( \mu_k \) among the \( K \) clusters, and \( 0 \) otherwise. Let \( \Theta = \bigcup_{k=1}^{K} [x_n] \) be a subset of data which are assigned to the \( k \)th cluster, where \( \mu_k \) yields the smallest distance. Then, the initial values for \( C_k \) and \( w_k \) can be expressed as follows:

\[
w_k = \frac{\delta_k}{N},
\]

\[
C_k = \frac{1}{|\Theta|} \sum_{x_n \in \Theta} (x_n - \mu_k)(x_n - \mu_k)^T.
\]

Using Eqs. (3)–(6), the initialization process is described in Algorithm 1 in Appendix 2.

3.2. The EM algorithm

After initialization, we obtain the initial parameters for each cluster, i.e., \( \mu_k \), \( C_k \) and \( w_k \) for \( 1 \leq k \leq K \). Starting from the initial parameters, the EM algorithm iteratively updates the parameters until it yields the largest likelihood given the data, i.e., the convergence is reached. In the EM algorithm, each iteration includes two steps E and M. The E step computes the membership weights \( w_{kn} \) of data \( x_n \) in the \( k \)th cluster, which reflect the certainty of \( x_n \) belonging to the \( k \)th cluster. For \( 1 \leq k \leq K \) and \( 1 \leq n \leq N \),

\[
w_{kn} = \frac{w_k f_k(x_n)}{\sum_{l=1}^{K} w_l f_l(x_n)}.
\]

\[^{15}\text{Such measure is a feature used to characterize data usage, for example, maximum, minimum, or average usage, depending on the service provider.}\]
For $1 \leq k \leq K$, the $M$ step uses the new membership weights to update the parameters in the following order:

$$
W_k = \left( \frac{1}{N} \right) \sum_{n=1}^{N} W_{k,n},
$$

$$
\mu_k = \frac{\sum_{n=1}^{N} W_{k,n} x_n}{\sum_{n=1}^{N} W_{k,n}},
$$

and

$$
C_k = \frac{\sum_{n=1}^{N} W_{k,n} (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^{N} W_{k,n}}.
$$

Eqs. (7)-(10) are iteratively executed until convergence. Convergence is guaranteed because the likelihood is proven to increase monotonically in each iteration and is bounded. When the difference of likelihoods in the consecutive two iterations is smaller than $\epsilon (\epsilon = 10^{-10}$ in our implementation), the convergence is reached and we obtain parameters $W_k$, $\mu_k$, and $C_k$. The procedure can be summarized in Algorithm 2 in Appendix 2.

3.3. Penalized likelihood

To determine the number of the MGM components to best approximate the true distribution of the dataset, we first specify the range $[K_{\text{min}}, K_{\text{max}}]$ for the number of components in the mixture model, construct a candidate pool for all possible numbers $K$ of the MGM components where $K_{\text{min}} \leq K \leq K_{\text{max}}$, and use the model selection technique to evaluate the performance of all candidate models. This section considers some frequently used information criteria that trade off accuracy and complexity in the model construction. The likelihood of the candidate model expresses the accuracy of the fitting model, and the higher the likelihood, the better the performance fit. The total number of free parameters in the candidate model defines the complexity. At the same likelihood level, a simpler candidate model with better performance should be selected.

Akaike (1974) proposed the Akaike information criterion (AIC) that is generally considered the first model criterion. The AIC is defined as:

$$
\text{AIC} = (K - 1) + K \left[ M + \left( \frac{1}{2} \right) M(M + 1) \right] - 2 \log (L(\Theta))
$$

and some authors prefer using AIC in practice. Schwarz (1978) proposed the Bayesian information criterion (BIC) that is a popular measure to determine the number of mixture components:

$$
\text{BIC} = \left( (K - 1) + K \left[ M + \left( \frac{1}{2} \right) M(M + 1) \right] \right) \times \log (N) - 2 \log (L(\Theta)).
$$

Claeskens and Hjort (2008) suggested that in practice the criterion should be much more important than the number of parameters. However, in the BIC criterion, the term $\log(N)$ may enlarge the penalty term depending on the value of $N$. Hannan and Quinn (1979) modified BIC and proposed the Hannan-Quinn information criterion (HQC) that reduces the penalty from $\log(N)$ to $\log(\log(N))$, namely:

$$
\text{HQC} = \left( (K - 1) + K \left[ M + \left( \frac{1}{2} \right) M(M + 1) \right] \right) \times \log (\log (N)) - 2 \log (L(\Theta)).
$$

We consider AIC, BIC and HQC simultaneously and select the model with the best performance from all candidate models.

In addition to using the information criterion to evaluate the performance of candidate models, we also use the F-fold cross validation technique to examine the stability of the candidate models, which randomly partitions the original dataset $\Theta$ into $F$ equal-sized subsets $\Phi_f$, for $1 \leq f \leq F$. That is $\Theta = \bigcup_{f=1}^{F} \Phi_f$. For every possible $K$ in MGM, the cross validation process is repeatedly executed in $F$ rounds, where every data is tested once in the process as all data are equally important. Every round of cross validation involves training with $F - 1$ subsets $\Phi_f$ and validation with the remaining subset $\Phi_V$, and $\Theta = \Phi_f \cup \Phi_V$. In the training stage, we use $\Phi_f$ to estimate the appropriate $K$ pair parameters, including $W_k$, $\mu_k$ and $C_k$ for $1 \leq k \leq K$. Then in the validation stage, we use the aforementioned criteria to evaluate the fitness of the $K$-Gaussian mixture model on $\Phi_V$. If we obtain similar scores during the $F$ rounds, then we can be fairly confident about the number of mixture components. On the other hand, if the scores vary widely in the $F$ rounds, then we should be sceptical about the mixture model’s fit and clustering performance. After all candidates have been iterated, the model that minimizes all testing criteria in $F$ rounds will be selected eventually. This process is described in Algorithm 3 in Appendix 2.

4. Accuracy evaluation

This section compares MGM with other previously proposed methods, including $K$-means, $K$-means++, Fuzzy C-means (FCM), and normalized cuts (N-Cuts) (Bai et al., 2016; Santi et al., 2016; Shi & Malik, 2000). We use three datasets to examine the clustering accuracy under different cluster characteristics. Since these datasets are generated from the given mixture model information, including the data and the labels that mark the true cluster of each data, it is easy to evaluate the accuracy of our proposed method by checking whether the clustered labels are consistent with the true labels. The first dataset is the stereotype Gaussian mixture model composed of two multivariate Gaussian distributions as per Fränti, Virmajoki, and Hautamäki (2006) (Dataset-I). The second dataset (Dataset-II) used for clustering analysis follows Yang et al. (2012). The third dataset (Dataset-III) consists of a Gamma mixture model.

4.1. Synthetic Dataset-I

As Jain (2010) and Zimek, Schubert, and Kriegel (2012) pointed out, clustering high-dimensional data is a big challenge due to the distance concentration effect, such that irrelevant features conceal relevant information when data dimension increases. Moreover, the vast majority of real world clustering datasets have the overlapping characteristic. In Dataset-I, we check the clustering accuracy of our method under the effects of (1) varying dimensions, and (2) overlapping clusters. The Gaussian distribution dimension varies from 2, 4 to 16 (see Fränti et al., 2006). Taking the bivariate Gaussian mixture model as an example, there are 10 different test cases as shown in Figure 1, where $\mu_1$ and $\mu_2$ are always located in the same positions but the value of the covariance matrix gradually increases in range from 10 to 100 along with the index increasing from 1 to 10. The overlapping region of two clusters also increases. For each test case, there are 2048 paired data including the data values and the true labels. With 3 different Gaussian distribution dimensions, there are 30 test cases in Dataset-I.

As the dataset has been labelled for the underlying true clusters, it is easy to identify errors. We report the clustering error rates with different methods in Table 1 for Dataset-I indicating that our method has the lowest error rates for the three experiments (i.e., with 2-dimensional, 4-dimensional, and 16-dimensional data). We observe two interesting facts. First, $K$-means, $K$-means++, and FCM show almost identical error rates. Second, when the dimension increases, the errors decrease for $K$-means, $K$-means++, FCM and MGM. The Dataset-I benchmark data shows strong distance-dependent characteristics, particularly
for higher dimensions, namely, by increasing the dimension, the $\mu_k$ values depart further from each other. N-Cuts do not depend on distance to separate clusters. On the other hand, the K-means, K-means++ and FCM are distance-based clustering methods, and therefore benefit from $\mu_k$ departure in clustering. MGM is not influenced by such departure of cluster centroids as it searches the density of features.

4.2. Dataset-II

The accuracy of clustering methods is easily dominated by the anomalies in the dataset, particularly for the distance-based methods. For Dataset-II, we follow the simulation method used by Yang et al. (2012) to test the clustering performance under the effects of (1) different level of perturbation (anomalies), and (2) overlapping clusters. To examine the robustness of the proposed method, we generate the parameters of the bivariate Gaussian mixture model with different $K$. We set $K_{\text{min}} = 3$ and $K_{\text{max}} = 5$ and therefore have three test cases (i.e., $K = 3, 4, 5$) in Dataset-II. We first assign $K = 3$ and apply Algorithm 1 to derive the corresponding parameters, including the mean vectors:

$$\mu_1 = (0.5 \ 0.5)^T, \quad \mu_2 = (6 \ 6)^T, \quad \mu_3 = (10 \ 10)^T,$$

and the covariance matrices:

$$C_1 = \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 7.5 & 4 \\ 4 & 7.5 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 7.5 & 4 \\ 4 & 7.5 \end{bmatrix}.$$

To model the distribution of the actual mobile data usages with four Gaussian models, we obtain the mean vectors:

$$\mu_1 = (0.5 \ 0.5)^T, \quad \mu_2 = (3 \ 3)^T, \quad \mu_3 = (10 \ 10)^T, \quad \mu_4 = (25 \ 25)^T,$$

and the covariance matrices:

$$C_1 = \begin{bmatrix} 0.4 & 0.1 \\ 0.1 & 0.4 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 16 & 10 \\ 10 & 16 \end{bmatrix}, \quad C_4 = \begin{bmatrix} 64 & 36 \\ 36 & 64 \end{bmatrix}.$$  

If we use a five-Gaussian mixture model to approximate the distribution of the actual mobile data usages, we derive the mean vectors as:

$$\mu_1 = (0.5 \ 0.5)^T, \quad \mu_2 = (2.5 \ 2.5)^T, \quad \mu_3 = (6 \ 6)^T, \quad \mu_4 = (16 \ 16)^T, \quad \mu_5 = (40 \ 40)^T,$$

and the covariance matrices:

$$C_1 = \begin{bmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 7.5 & 4 \\ 4 & 7.5 \end{bmatrix}, \quad C_4 = \begin{bmatrix} 36 & 20 \\ 20 & 36 \end{bmatrix}, \quad C_5 = \begin{bmatrix} 100 & 64 \\ 64 & 100 \end{bmatrix}.$$  

For each $K$, we sample 2500 paired data, including the data and the true labels. In each test case, we consider the perturbation of anomalies, since extremely large data usages indeed exist in our measured data of the commercial mobile operation dataset. Anomalies are sampled from a uniform distribution in the range $[50, 200]$. We gradually increase the occupied ratio of anomalies from 1% to 25% in each test case resulting from our observation of actual consumers mobile data usages where anomalies did not occupy the main proportion of the dataset. Each ratio of anomalies is run 200 times and we evaluate the clustering accuracy under the predetermined perturbations.

We compare our method with other algorithms by mixing 3, 4 and 5 bivariate Gaussian distributions with Dataset-II. For $K = 3$, the true clusters in the data are 3. Table 2 reports the results with different relative perturbation rates ranged from 1% to 25%. With relatively low perturbation rates, the anomalies occupied from 1% to 5% of the total 2500 samples for each experiment with different $K$. To compare MGM with other methods, we use the formula $(\text{Error}_{\text{MGM}} - \text{Error}_{\text{others}})/\text{Error}_{\text{others}}$ in the following comparisons. Compared with K-means, the MGM reduces the error rates by 82.7631%, 86.0226%, 87.7115%, 83.4223%, and 82.7116% respectively when $K = 3$ and the perturbation increases from 1% to 5%,
and therefore the MGM reduces the error rate by 83.5262% on average. In comparison with N-Cut, when K = 3 and perturbation is 1%, N-Cut performs better than MGM by 5.6756%. However, MGM reduces the errors by 15.0659%, 13.2058%, 22.2265%, and 23.5339% respectively, as the perturbation increases from 2% to 5%, and therefore MGM reduces the error rate by 13.6693% on average. In general, when K = 3 and the perturbation is relatively low, the average error rates the MGM reduced are 83.5262%, 91.4121%, 77.8743%, and 13.6693% compared with K-means, K-means++, FCM, and N-Cut, respectively. Similarly, we see an improvement in accuracy on average of 72.8607%, 91.0197%, 81.5388%, and 7.8851% for K = 4 and 82.7115%, 90.9156%, 84.0753%, and 13.2058% for K = 5 when comparing the MGM with the other four methods, respectively.

Table 2 also indicates that MGM outperforms other methods when the perturbation rates are high, i.e., the occupation rate ranges from 7.5% to 25% of the total observations. Compared with the K-means, when K = 3 and the perturbation gradually increases from 7.5% to 25%, the MGM reduces the error rates by 74.6281%, 79.2354%, 80.2249%, 79.9498% and 81.4575%, respectively, and the MGM reduces the error rate by 79.9911% on average. For K = 3, average accuracy improves with the MGM by 79.9931%, 91.4096%, 83.1264%, and 37.6314% compared with K-means, K-means++, FCM, and N-Cuts, respectively. Similarly, we can see an average improvement in accuracy of 90.15%, 94.5191%, 91.274%, and 55.9284% for K = 4 and 89.1595%, 95.3613%, 92.8644%, and 62.1131% for K = 5 when comparing our method with the other four methods, respectively.

We find that when the perturbation crosses over 5%, e.g., 7.5%, the advantage of our proposed method becomes very significant over the N-Cut. Along with the increase in perturbation, our method shows consistently superior performance.

4.3. Dataset-III

For Dataset-III, we consider the clustering performance under the effects of (1) the Gamma mixture model, (2) different level of perturbation (anomalies), and (3) overlapping clusters. The most important property of the Gamma distribution is its skewness and kurtosis, which allows a skew center in one cluster and relaxes the constraint that distribution in one cluster should always be symmetric. The shape parameter $\alpha$ and the scale parameter $\beta$ determine the Gamma distribution, where the mean is $\alpha\beta$ and the variance is $\alpha\beta^2$.

Like Dataset-II, we determine all parameters of the Gamma mixture model with different $K$ values. When $K = 3$, we obtain the shape parameters $\alpha_1 = 0.5$, $\alpha_2 = 4.8$, $\alpha_3 = 40/3$, and the scale parameters $\beta_1 = 1$, $\beta_2 = 1.25$, $\beta_3 = 0.75$. When $K = 4$, the shape parameters are $\alpha_1 = 0.5$, $\alpha_2 = 3$, $\alpha_3 = 6.25$, $\alpha_4 = 9.75$, and the scale parameters are $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1.6$, $\beta_4 = 2.56$. When $K = 5$, the shape parameters are $\alpha_1 = 0.5$, $\alpha_2 = 25/12$, $\alpha_3 = 4.8$, $\alpha_4 = 64/9$, $\alpha_5 = 16$, and the scale parameters are $\beta_1 = 1$, $\beta_2 = 1.2$, $\beta_3 = 1.25$, $\beta_4 = 2.25$, $\beta_5 = 2.5$, respectively.

Like Dataset-II, we consider the perturbation of anomalies sampled from the uniform distribution over the range [50, 200]. We gradually increase the occupied ratio of anomalies from 1% to 25% in each test case and check whether the proposed method is robust against perturbations. Our method is compared with other algorithms using 3, 4, and 5 Gamma mixture distributions with Dataset-III. Table 3 summarizes the results with different perturbation rates ranged from 1% to 25%.

In comparison with K-means, when K = 3 and the perturbation increases from 1% to 5%, MGM reduces the error rates by 82.2718%, 80.2254%, 80.2771%, 82.8381%, and 84.2248%; therefore, MGM reduces the error rate by 81.9675% on average. Compared with N-Cut, MGM reduces the error rates by 17.5153%, 14.4056%, 8.9411%, 21.6312%, and 18.9445% when K = 3 and the perturbation increases from 1% to 5%, and the MGM therefore reduces the error rate by 16.2875% on average.

Overall, when K = 3 and the perturbation is relatively low, the error rates reduced by MGM are 81.9675%, 90.9060%, 80.6375%, and 16.2875% compared to K-means, K-means++, FCM, and N-Cuts, respectively. Similarly, we can see an average improvement of accuracy of 89.4245%, 93.5029%, 85.8516%, and 10.9469% for K = 4 and 89.2690%, 95.1978%, 91.8338%, and 15.8988% for K = 5 when comparing our method with the other four methods, respectively.

When the perturbation is relatively high, i.e., the occupation rate ranges from 7.5% to 25% of the total observations, the average accuracy has improved with the MGM by 80.8424%, 89.8340%, 86.6415%, and 32.8704% compared with K-means, K-means++, FCM, and N-Cuts respectively. Similarly, when K = 3, comparing our method with the aforementioned four methods, the average improvements of accuracies are 89.4897%, 92.6798%, 91.7409%, and 39.8774% respectively. Similarly, for K = 5, the accuracy improvements are 90.3343%, 94.6990%, 93.4242%, and 54.3619%. We find that when the perturbation crosses over 5%, e.g., 7.5%, the advantage of MGM becomes very significant over N-Cut.

As previously mentioned, K-means, K-means++ and FCM are distance-based clustering methods, and they are beneficial when the data belonging to different clusters depart further from each other. On the other hand, if there are some anomalies in the dataset, the performance is easily influenced by anomalies.
Compared with distance-based clustering methods, N-Cut is much closer to spectral clustering methods. The N-Cuts clusters data are not directly based on distances but on measuring the total dissimilarity between different clusters and the total similarity within the clusters. When there are some elongated clusters in the dataset, N-Cuts becomes sensitive to anomalies. Along with the increase in perturbation, our method shows consistently superior performance.

4.4. Accuracy evaluation

Clustering is fundamental when dealing with high-dimensional data, however, there is distressingly little general theory on it for application to a particular data. When we conduct clustering for big data, difficulties arise due to (1) undeterminable number of clusters, (2) perturbation (or anomalies), and (3) non-spherical and overlapping in a dataset. In practice, we usually observe that when perturbation increases the data points are either elongated in certain directions (i.e., the data shape is non-spherical) or entangled (i.e., clusters are overlapping). In our simulation, we assume that the ground-truth number of clusters is determinable. Each clustering algorithm makes specific structural assumptions that need to be considered about the dataset, that is, the shape of the clusters. Table 4 summarizes the attributes of each dataset investigated in our study.

Dataset I serves as a benchmark for checking the impact of dimensionality. In Table 1, we can see that the proposed MGM outperforms other methods. For Datasets II and III, we evaluate all methods with respect to the impact of perturbation and clusters overlapping. We shift the perturbation rate from 1% to 25% against the total data size to verify the sensitivity of each method. Figs. 2 and 3 illustrate the performance of all methods. It is clear that N-Cuts and MGM outperform others. N-cuts is an unbiased measure of disassociation between subgroups by evaluating simultaneously the total dissimilarity between groups and the total similarity within groups. Comparing N-Cuts and MGM, we can see that when increasing the perturbation and number of clusters, MGM is significantly better. The advantage of MGM is to use variations in data density to define clusters.

K-means and K-means++ are sensitive to perturbation as they assume that each cluster has roughly equal numbers of observations and clusters are spherical. FCM is also distance based, such that it assigns membership to objects corresponding to each cluster center determined by the distance between the centroid and the object. Objects on the boundaries between clusters are assigned membership degrees between 0 and 1 indicating their partial association. It is sensitive to perturbation and assigns outliers low (or even no) membership degree. The drawback of N-Cuts is from the minimum cut criteria applied because it occasionally supports cutting isolated objects due to the small values achieved by partitioning them.

We investigate the Gaussian mixture distribution and Gamma mixture distribution in our simulation. They are members of the exponential family distributions. The Gaussian distribution is fundamental and the Gamma model illustrates rich characteristics with different combinations of shape and scale parameters. Other models can be used to evaluate the performance. It should be noted that the exponential distribution, Erlang distribution, and chi-squared distribution are all special cases of the Gamma distribution.

We do not consider some synthetic cluster shapes such as rings or spirals because they are not realistic to describe the human behaviour. It should be cautious when duplicating the proposed method on the data with those pretentious cluster shapes.

Table 3

<table>
<thead>
<tr>
<th>K</th>
<th>Method</th>
<th>Perturbations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1%</td>
</tr>
<tr>
<td>3</td>
<td>K-means</td>
<td>33.82</td>
</tr>
<tr>
<td></td>
<td>FCM</td>
<td>18.79</td>
</tr>
<tr>
<td>4</td>
<td>K-means</td>
<td>36.66</td>
</tr>
<tr>
<td></td>
<td>K-means++</td>
<td>75.00</td>
</tr>
<tr>
<td></td>
<td>FCM</td>
<td>16.38</td>
</tr>
<tr>
<td>5</td>
<td>K-means</td>
<td>36.57</td>
</tr>
<tr>
<td></td>
<td>K-means++</td>
<td>80.00</td>
</tr>
<tr>
<td></td>
<td>FCM</td>
<td>22.40</td>
</tr>
<tr>
<td></td>
<td>N-Cuts</td>
<td>4.96</td>
</tr>
<tr>
<td></td>
<td>GMM</td>
<td>3.03</td>
</tr>
</tbody>
</table>

Table 4

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Dataset-I</th>
<th>Dataset-II</th>
<th>Dataset-III</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perturbation</td>
<td>0%</td>
<td>1%-25%</td>
<td>1%-25%</td>
</tr>
<tr>
<td>Clusters overlapping</td>
<td>Moderate</td>
<td>Increased</td>
<td>Increased</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>2</td>
<td>3, 4, 5</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td>Dimensionality</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Probability density</td>
<td>Gaussian mixture</td>
<td>Gaussian mixture</td>
<td>Gamma mixture</td>
</tr>
</tbody>
</table>
management for sustainability. Following the 3Ps in Jaehn (2016) we hereby interpret the goals of sustainability as follows: stimulating the operator’s continuous investments (profit), fostering the consumer’s personalized adoption to avoid smartphone addiction (person), and mitigating inefficiency caused by reckless decisions (planet). Laffont and Tirole (1999) explain the economic theory of pricing scheme applied by the telecommunications industry. Our research is built on the nonlinear pricing literature and we list some related works in Table 5. We first exemplify a descriptive analytics in Section 5.1 based on the method we proposed in Section 3 with the real usage data provided by a telecom company (TEL). Based on the outcome of descriptive analytics, we perform a predictive analytics in Section 5.2 with machine learning and simulation for understanding the current business situation and formulating strategic options of revenue management. After that we present our prescriptive analytics in Section 5.3 that attempts to advise possible outcomes and actions.

5.1. Descriptive analytics

Descriptive analytics identifies patterns and trends in data and categorises, characterises, and classifies them into useful information in order to understand past and current business performance, see Kunc and O’Brien (2018) and references therein. Mobile operators analyze their users’ behaviour on the product/service and time frequency in order to guide business growth. Behavioural description of mobile data builds a fine, complete user’s portrait through the study of usage data.

In this section, we apply descriptive analytics on 1000 anonymous users’ mobile data usages of twelve months randomly

Fig. 2. Comparison of the error rates of five methods performed on the Dataset-II (Gaussian mixture) that contains different perturbation rates ranged from 1% to 25%.

Fig. 3. Comparison of the error rates performed by five methods on the Dataset-III (Gamma mixture) that contains different perturbation rates ranged from 1% to 25%.

Table 5
Overview of related studies on pricing methods of communications in the literature.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Object</th>
<th>Pricing scheme</th>
<th>Network congestion</th>
<th>Linearity</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bagh and Bhargava (2013)</td>
<td>Bundle service</td>
<td>Two and three parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Brito, P., and Vereda (2010)</td>
<td>Not specified</td>
<td>Two parts</td>
<td>No</td>
<td>Linear</td>
<td>–</td>
</tr>
<tr>
<td>Chen and Huang (2016)</td>
<td>Data usage</td>
<td>Two parts</td>
<td>Yes</td>
<td>Nonlinear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Ferrer, Moa, and Olivares (2010)</td>
<td>Voice usage</td>
<td>Three parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>–</td>
</tr>
<tr>
<td>Fibich, Klein, Koeigsberg, and Muller (2017)</td>
<td>Voice usage</td>
<td>Three parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>–</td>
</tr>
<tr>
<td>Iyengar, Ansari, and Gupta (2007)</td>
<td>Voice usage</td>
<td>Three parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>Empirical</td>
</tr>
<tr>
<td>Lee, Mo, Jin, and Park (2012)</td>
<td>Data usage</td>
<td>Flat and two parts</td>
<td>Yes</td>
<td>Nonlinear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Ma, Deng, Xue, Shen, and Lan (2017)</td>
<td>Data usage</td>
<td>Three parts</td>
<td>Yes</td>
<td>Nonlinear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Masuda and Whang (2006)</td>
<td>Voice usage</td>
<td>Two parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>–</td>
</tr>
<tr>
<td>Sumantam, Chakraborty, and Sharma (2015)</td>
<td>Voice usage</td>
<td>Two parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Wang, Ma, and Xu (2017)</td>
<td>Cloud service</td>
<td>Flat and two parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>Numerical</td>
</tr>
<tr>
<td>Yang and Ng (2010)</td>
<td>Data usage</td>
<td>Two parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>Empirical</td>
</tr>
<tr>
<td>This study</td>
<td>Data usage</td>
<td>Two parts</td>
<td>No</td>
<td>Nonlinear</td>
<td>Empirical</td>
</tr>
</tbody>
</table>

Please cite this article as: Y.-T. Chen, E.W. Sun and Y.-B. Lin/European Journal of Operational Research xxx (xxxx) xxx
sampled from the contractual period from April 2014 to June 2016. Following the Computer-Processed Personal Data Protection Law in Taiwan\textsuperscript{17}, they are randomly picked up by TEL from its primary databank and regarded as a representative of the new 4G service without considering any specific features of demographics, psychographics, or sociographics. We refer to it as the TEL data in this study and treat each monthly usage as a specific feature of usage. In addition, the 12 dimensional features used to characterize the cumulative data consumption are synchronized. Due to the legislation in force, we are not clarified for any labels of these 12 features (e.g., concrete months of the contractual period). Our experiment is therefore a double-blind test. From practitioner’s viewpoint, it is a misconception that many organizations think they need “perfect” data to proceed with analytics. James Guszczka, the Deloitte Consulting chief data scientist pointed out that there’s no absolute standard that describes what data is sufficient vs. insufficient and limited data set can help make valuable classifications or predictions.\textsuperscript{18}

In order to improve the statistical significance, we apply the bootstrapping method of Sun and Meini\textsuperscript{(2012)} to enlarge the sample to 10,000 observations with the original 12 features in our back-testing. We show the complete procedure in Algorithm 4 in Appendix 2. We refer to this bootstrapped dataset as the augmented TEL data in our study. The advantage of bootstrapping, particularly its contribution to both natural and social sciences, has been addressed in several studies (see the Statistical Science 2003 special issue, vol. 18, no. 2). In addition, bootstrapping techniques can be used for an efficiency measurement (Fallah-Fini, Triantis, & Seaver, 2012) and to obtain a good estimation of the statistical properties of the original population, see Cerquet, Falbo, and Pelizzari (2017) and references therein.

The descriptive statistics of the TEL data and augmented TEL data are shown in Table 6. We recognize that there exists heterogeneity in the usage data. In order to well-characterize them, we apply the density-based subspace clustering proposed in Section 3. We specify the range of $K$ between 3 and 6, as most operators generally provide customers with more than 3 different data plans\textsuperscript{19} and TEL offers six data plans for 4G services. By running the proposed method, we obtain the optimal mixture model with four ($K = 4$) multivariate Gaussian distributions, i.e., four distinguishable clusters. Table 8 reports the fitting measures (i.e., AIC, BIC, and HQC) showing that grouping consumers into 4 clusters leads to the best fit solution such that all these criteria are steadily minimized in F-rates. In order to graphically illustrate the clustering results, we apply the $t$-Distributed Stochastic Neighbour Embedding (t-SNE) proposed by van der Maaten and Hinton (2008) for dimensionality reduction by employing the Barnes-Hut approximation. Fig. 4 illustrates four different clustering results i.e., $K = 3, 4, 5$ and 6, respectively.

Our descriptive analytics reveals there is likely to be ignored imperfection when matching six different data plans to four user groups. We shall conduct predictive analytics to investigate what could happen in revenue if TEL relocates these users clustered by the MGM method to its data plans.

5.2. Predictive analytics for formulating strategic options

Predictive tools (e.g., machine learning and simulation) will strengthen formulating strategic options as they provide decision maker with actionable insights from the data, (see Kunc & O’Brien (2018)). Our descriptive analytics infers the imperfection of current TEL data plans in matching the user clusters. In this section, we run a predictive analytics that related to sensitivity analysis to predict revenue changes by introducing the MGM based data plan.

The current data plans are directly obtained from TEL and listed in Table 9 where six data plans $A = \{a_{K_1}^{1}, \ldots, a_{K_2}^{1}\}$ are bracketed from 0.55 gigabytes to 16 gigabytes. In Table 9, the two-part tariff (a lump-sum fee and a per-unit charge) applied by TEL is shown. The tariff rates $G = \{g_{K_1}^{1}, \ldots, g_{K_2}^{1}\}$ indicate the basic subscription fees and two options $a_1$ and $a_2$ for data overage surcharge: NT$ 100 per 0.2 gigabytes and NT$ 250 per gigabyte, i.e. $p_1 = 100$ and $p_2 = 250$. Customers will be informed by TEL\textsuperscript{20} before their subscribed plan to be completely exhausted. Then the consumer can voluntarily choose or deny the overage plan.

In our predictive analytics, we plan to recognize the revenue changes for TEL after using the new brackets of basic data plans determined by the mean (or median) of the MGM components, see Section 5.2. The MGM method clusters the actual usage into $K$ plans after confirming $K$ 12-dimensional MGM distributions. Each distribution has a mean vector ($\mu_{K}$). Then the mean or median of each mean vector $\mu_{K}$ will be served as the bracket for the basic data plan. Revenue management is to decide the fees for basic plan and overage surcharge for each bracket. For the real data we investigated, TEL does not clarify either the basic plan initially subscribed or the overage plan the user decided. We assume a rational consumer will always choose the most beneficial plan. For example, when the subscribed data allowance is almost exhausted,}

---

\textsuperscript{17} Taiwan has adopted the Personal Information Protection Act (PIPA) since 01 October, 2012 and EU adopted the General Data Protection Regulation (GDPR) in 25 May 2018.

\textsuperscript{18} https://deloitte.wsj.com/cio/2016/02/04/you-dont-need-perfect-data-for-analytics-analytics/.

\textsuperscript{19} For example, Verizon Wireless in US offers three plans for the 4G LTE data, AT&T in US and Deutsche Telekom in Germany offer four, Hutchison 3G in UK offers five, and NTT Docomo in Japan offers six.

\textsuperscript{20} Developments in charging and billing architectures have been discussed in the literature. See, among others, Koutsopoulou, Kaloxyllos, Alonistioti, Merakos, and Kawamura (2004), de Reuver, de Konin, Bouwman, and Lemstra (2009), and Lee, Murray, and Qiao (2015),

---

TEL will remind the customer of choosing an overage plan or stoppage. With the fees in Table 9, if a consumer knows that her excessive usage will finally exhausted above 0.4GB but no more than 1 gigabyte (to be charged at least NT$300), she will choose the overage plan of 1 gigabyte at beginning that only charges NT$250. Therefore, we consider this fact for the overage and apply the minimum principle to minimize the total payment. For example, if a consumer’s usage is 5.1 gigabytes, the total charge is NT$1686 since the “billing plan iv” applies NT$1336 (with 4 gigabytes data plan) with an overcharge of NT$ 350 (i.e., NT$250+NT$100). If a consumer’s usage is 5.5 gigabytes, the total payment is then NT$1736 charged by the “billing plan iv” because the total charges of NT$1836 (i.e., NT$1336 + NT$250 × 2) by the “billing plan iv” is eliminated when applying the minimum principle. We illustrate the machine learning procedure of determining tariff for the MGM data plans by Algorithm 5 in Appendix 2.

The total revenue (Rev) of TEL data plan \( \Lambda \) with rates \( \Gamma \) (see Table 9) and estimated overage (see Table 7) can be calculated as follows: (see Algorithm 6 in Appendix 2):

\[
\text{Rev}(\Theta, \Gamma, \Lambda, a_i, p_i) = \sum_{k=1}^{6} \sum_{m=1}^{12} \left( \sum_{n=1}^{m} \gamma_k \right) \\
+ \max \left( 0, \min \left( \rho_1 \text{cell} \left( x_{n,m} - \lambda_k \over a_1 \right), \rho_2 \text{cell} \left( x_{n,m} - \lambda_k \over a_2 \right) \right) \right)
\]

where \( a_i \) is the overage plan and \( p_i \) is its rate for \( i = 1, 2, \Theta \) is the actual data usage whose characteristics are shown in Table 6, \( \theta_i \) stands for the corresponding estimated number of users for each data plan, and \( \Theta = \left\{ \theta_i \right\} \) is the set of all \( \theta_i \).

With the proposed MGM method, after identifying the number of clusters \( K \), the new data plan brackets \( \Lambda = \left\{ \lambda^i_k \right\} \) can be determined by the mean or median of \( \mu_k \). The total revenue based on the new basic plan \( \left( K, \Gamma, \Lambda^i \right) \) is calculated by

\[
\text{Rev}(K, \Theta, \Gamma^i, \Lambda^i, a_i, p_i) = \sum_{k=1}^{K} \sum_{m=1}^{12} \gamma_k^i \\
+ \max \left( 0, \min \left( \rho_1 \text{cell} \left( x_{n,m} - \lambda^i_k \over a_1 \right), \rho_2 \text{cell} \left( x_{n,m} - \lambda^i_k \over a_2 \right) \right) \right) \\
\]

where \( \theta_k \) is the number of users for each MGM based data plan and \( \Theta = \left\{ \theta_k \right\} \).

We propose new data plans by using 3, 4, 5, and 6 multivariate Gaussian distributions based on the descriptive results. In addition, we reduce the 12 dimensions to one dimension with the scalar mean or median of the underlying K univariate mixture Gaussian distributions. Table 10 illustrates the new data plans for different clusters based on the original data (10,000 users) and augmented data (10,000 users). For example, when the resulting cluster is 3 with the original 12 dimensions, we obtain two different data plans based on the MGM-mean bracket and MGM-median bracket. The mean-based bracket sets the data plan into three tiers bracketed by 0.7671 gigabytes, 4.4972 gigabytes, and
**ARTICLE IN PRESS**


Please cite this article as: Y.-T. Chen, E.W. Sun and Y.-B. Lin, Merging anomalous data usage in wireless mobile telecommunications: Business analytics with strategy-focused data-driven approach for sustainability, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.02.046

---

**Table 10**

New data plan calculated with the proposed clustering method and the corresponding revenue increase (RI). The clustering is based on the original dimension ($M=12$) and one reduced feature ($M=1$).

<table>
<thead>
<tr>
<th>$K$</th>
<th>Number of users = 1000</th>
<th>Number of users = 10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data Plan Brackets (in gigabyte)</td>
<td>R.I</td>
</tr>
<tr>
<td>$M=12$</td>
<td>3 0.7617 4.4972 17.9662</td>
<td>0.0503</td>
</tr>
<tr>
<td>4 0.8508 3.0034 7.9632 21.2971</td>
<td>0.1327</td>
<td>0.2245 7.3196 23.8056</td>
</tr>
<tr>
<td>Mean 5 0.2450 1.3972 4.5755 14.5199</td>
<td>0.1311</td>
<td>0.2827 3.5918 9.1414</td>
</tr>
<tr>
<td>6 0.2390 1.0033 2.7219 5.8098 13.6533</td>
<td>0.0840</td>
<td>0.0503 0.8025 4.5538 17.6432</td>
</tr>
<tr>
<td>$M=13$</td>
<td>3 1.4528 6.8454 20.5755</td>
<td>0.2004</td>
</tr>
<tr>
<td>4 0.7994 3.8532 12.1280</td>
<td>0.2278</td>
<td>0.3863 3.7845 21.0881</td>
</tr>
<tr>
<td>Median 5 0.2883 1.3972 4.5755 14.5199</td>
<td>0.1879</td>
<td>0.7044 2.7712 6.3432</td>
</tr>
<tr>
<td>6 0.1678 0.9956 3.0034 7.9632 13.6533</td>
<td>0.1370</td>
<td>0.0741 0.7106 2.1130</td>
</tr>
<tr>
<td>$M=1$</td>
<td>3 1.7927 6.4637 18.0513</td>
<td>0.1299</td>
</tr>
<tr>
<td>4 0.8508 3.0034 10.8944 24.4235</td>
<td>0.1370</td>
<td>1.2666 4.1088 10.3088</td>
</tr>
<tr>
<td>Mean 5 0.2883 1.3972 4.5755 14.5199</td>
<td>0.1307</td>
<td>0.9916 2.5965 5.8860</td>
</tr>
<tr>
<td>6 0.1678 0.9956 3.0034 7.9632 13.6533</td>
<td>0.1298</td>
<td>1.2666 4.1088 10.3088</td>
</tr>
<tr>
<td>Median 5 0.2883 1.3972 4.5755 14.5199</td>
<td>0.1307</td>
<td>0.9916 2.5965 5.8860</td>
</tr>
<tr>
<td>6 0.1678 0.9956 3.0034 7.9632 13.6533</td>
<td>0.1298</td>
<td>1.2666 4.1088 10.3088</td>
</tr>
</tbody>
</table>

**Fig. 4.** Empirical results of clustering the TEL data based on our method with different $K$. 
### Table 11

New data plan calculated with the proposed clustering method and the corresponding revenue increase (RI). The clustering is based on the random 6 months (M = 6) to obtain the data plan brackets. Revenue increase (RI) is based on the left 6 months. We report the mean and standard deviation (in parentheses) of 924 combinations.

<table>
<thead>
<tr>
<th>Number of users</th>
<th>K</th>
<th>Data Plan Brackets (in gigabyte)</th>
<th>R.I.</th>
<th>Number of users</th>
<th>K</th>
<th>Data Plan Brackets (in gigabyte)</th>
<th>R.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>3</td>
<td>1.4955</td>
<td>6.1195</td>
<td>0.1156</td>
<td>1.4972</td>
<td>6.1164</td>
<td>0.1166</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.7096</td>
<td>2.9312</td>
<td>0.2495</td>
<td>0.7103</td>
<td>2.9167</td>
<td>0.2314</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.4317</td>
<td>1.9111</td>
<td>0.1220</td>
<td>0.4770</td>
<td>2.1867</td>
<td>0.0893</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.3792</td>
<td>1.6731</td>
<td>0.0737</td>
<td>0.3729</td>
<td>1.7555</td>
<td>0.0644</td>
</tr>
</tbody>
</table>
| 17.9662 gigabytes, when the median-based brackets are 14 gigabytes, 6.8454 gigabytes, and 20.5755 gigabytes. Although the corresponding basic data rates are NT$ 636, NT$ 1336, and NT$ 2636 for both mean-based and median-based plans, new plans increase data allocation by 38.49%, 12.43%, and 12.29% for the mean-based plans and 164.15%, 71.14%, and 37.92% for the median-based plan, respectively. Since the tariff rates for the TEL data plans remain unchanged, the unit rates of the MGM based plans are reduced.

We then compare the operator’s revenue increases after applying the new billing plans by (Revenue)MGM – (Revenue)TEL. We summarize the results in Table 10. We find that when K = 4 the MGM revenue increases most and the median-based data plan constituted of four brackets leads to the highest revenue increase (i.e., 22.78% for 1000 users and 23.69% for 10,000 users). It is precisely the best result we obtained from the MGM clustering (i.e., K = 4) in terms of revenue increase.

5.3. Prescriptive analytics

Prescriptive analytics in our study focuses on improving prediction accuracy and prescribing better decision options that bring new “values”. The values here could be the company’s future expected revenues and profits or intangible assets, from intellectual capital, and revealed opportunity to future growth potential and such a process can be considered to find, measure, create, and protect value across functional areas. Heyns (2015) shows that analytics enables organizations to understand and embrace emerging opportunities and align products and services with changing customer needs creating additional value for stakeholders in the process and effectively grow, optimise and protect value. Ransbotham and Kiron (2017) highlight that companies experienced in analytics are increasingly gaining competitive advantage and analytics for innovations lead to new products, services, and processes or improve existing ones.

In Section 5.2, we obtained the new brackets of data plans by setting M = 1 and M = 12. When M = 12, the classification is well cultivated by covering all observable patterns (or variety) that the machine learning algorithm can identify. It is unfortunately not accessible in real business as the data plans are determined before collecting all actual usage. On the other hand, when M = 1 the bias in data will increase which consequently leads to predictive inaccuracy. Therefore, we divide the data set into training data (used to determine the data plan brackets) and test data (used to evaluate the revenue of these plans). We randomly pick up six months from training and the rest six months for testing. There are 924 combinations when order does not matter and replacements are not allowed. We apply Algorithm 6 on 924 different training/testing settings and present the results in Table 11. Unsurprisingly, we recognize the results from Table 10 that when K = 4 the proposed MGM leads to the highest revenue increase. Figs. 5 and 6 plot the revenue of all 924 combinations based on the MGM data plans (i.e., K = 3, 4, 5 and 6) in comparison with the TEL revenue.

Laffont and Tirole (1999) consider the two-part tariffs as two complementary services (connection and consumption). Many studies have shown that these two prices should be coordinated (see Bagh and Bhargava, 2015, for example). It is worth losing some revenue on connections to boost variable consumption, and conversely. Even though the revenue from the charge of basic plans at beginning decreases, the revenue from the average surcharge increases and the total revenue finally increases. The more the data are charged for, the more the total revenue increases. On the other hand, the revenue of basic plan can increase after assigning more users who have large data consumption with higher rate. The proposed MGM method separates adjacent components efficiently as we have shown in Section 4. For members in a close cluster, the mean vector describes the average consumption and the covariance matrix of MGM allows discrepancy of consumption. The covariance structure of MGM accommodates discrepant membership, see Fig. 4. Components adhered to discrepancy are completely separated by the underlying Gaussian distributions to different clusters. We illustrate the distributions of users clustered by the TEL billing plans and MGM method in Table 12.

In our predictive analytics, we assume other things are constant and simply adopt the current TEL tariff. We match the MGM based data plan to the current TEL tariff (see Algorithm 5 in Appendix 2). Decision makers can determine the fees for basic and overage plans in different ways. Based on the proposed MGM method, we draw advantages from the mixture of multivariate normality by increasing the mean and accommodating itself to skewness, which can be used to preserve the revenue from the basic plan and meantime increase overage charge.

Once the rational consumers recognize when they remain in the regime of low rate basic plan their overage surcharge of the new data plan in fact increase, they will adjust their behaviour by either subscribing an adequate basic data plan at the beginning.

---

21 An extra charge against a consumer who alters the initial contractual subscription will apply in practice. For example, TEL provides a time-varying regime of sur-
or foster their consumption to avoid excessive usage. Both choices led to efficient data usage, and congestion of data traffic could be weakened. The operators can efficiently allocate data flows, therefore, 3Ps triplet of sustainability is tenable.

5.4. Discussion and limitations

Frisiani et al. (2017) point out that mobile operators can increase their sales to existing customers by making timely appealing offers on services or hardware after analysing the customer data. We show that the operator’s total revenue can increase benefited from analysing the user data. In our study, the users who choose to consume the extremely large volume of add-on should pay more since this volume is not previously subscribed in the initial contract. We should note a fact that TEL requires compensation for early termination that are calculated differently from person to person with (received monthly discount + terminal equipment subsidy) \times (number of remaining days of contract/total length of contract)\(^2\). It is very hard to confirm if all sampled users have sufficient ICT literacy to well estimate their usage. In addition, some consumers continuously pay for the add-on may be due to some price inelasticity effects, e.g., appropriated telecommunication subsidies paid by the employer. On the other hand, many consumers benefit from the new plan as they could move to a lower tariff class for basic plan. If the net change between the revenue decrease caused by down-shifting billing plans and revenue increase caused by charging add-ons remains positive, the operator can still benefit.

\(^2\) See Julien, Rey, and Sand-Zantman (2013) for example.
Besides the ICT industry, multipart tariff schemes have been generally adopted for behaviour-based pricing by service providers such as for insurance, car rental company, and membership organization. Consumers are required to pay a variable per-unit fee for usage beyond a fixed allowance. We propose a method for business analytics that uses density-based subspace clustering method to assess the contribution of these pricing schemes to revenue management for heterogeneous consumers with anomalous usage. An important aspect of our method is that it allows for inference with entangled (or mixed) behaviour measured by multiple dimensions. It can be applied not only by using the log data but also with other types of data at different frequencies and metrics such as big data. We have highlighted several technical features of our method in Section 2.3 comparing with other established methods. With this empirical study, we show its applicability and rigidity particularly for clustering anomalous behaviours.

Nonetheless, our study does not come without limitations. Changing of consumption will impact the total revenue. We take the billing plans for granted and work with it in a static way without considering any sequential interaction between the operator and consumer. Obviously, the operator can formulate best billing plans by altering the current rates after applying a comprehensive market research but we cannot clarify if the sustainability is still achievable under another price regime without careful calculation. The method is data-driven and relies on the distribution of samples. Applying optimal sampling method (see Thompson, 2012, for example) might obtain the most information-rich representatives that suffice for robust parameter estimates. However, randomization inference were overly concise of obscuring important points for coming to a small number of samples. As to this aspect, it is nontrivial to untiringly search for a large sample. In addition, researcher should attempt to make further exploration of the consumers’ log data in accordance with the legislation of privacy protection.

6. Conclusion

The methods used to cluster heterogeneous demand differ in the choice of objective function, the underlying probabilistic assumptions, the data generating models, and heuristics. This paper provides a tool based on the density-based subspace clustering approach (i.e., the three-stage iterative optimization procedure) for intelligent segmentation of heterogeneous consumer demand with anomalies. This approach is particularly practical as it (1) mitigates dependency on theoretical assumptions, (2) relies on easily recognized features and benefits from the recent development in big data technologies, (3) improves computational efficiency. In this paper, we discuss how this method was developed taking into account the recognized data features with a focus on the performance of the method under business engineering sustainability scenarios. In addition, as a decision-supporting system for business analytics, the algorithmic optimization ensures identifying the best combination of pricing strategies that jointly achieves managerial goals and sustainability. This study demonstrates not only the proposed method’s ability as an optimization tool in supporting managerial decision-making, but the analytic technique can also be implemented for sustainable management in various industries involving a high variety of consumer demand.

In this study, our method attempts to mitigate the effect of anomalies in clustering heterogeneous demand. We illustrate the critical components of the method and several innovative techniques to bridge existing studies on heterogeneous demand with anomalies and evaluating sustainable business engineering operations. We first calibrate and validate our method based on the simulated stylized data representing different behavioural patterns of demand heterogeneity. We then carry out business analytics by implementing the proposed method with the real data to formulate strategic options in search of business value. The results highlight the superior performance of our method in comparison with several alternative methods and analytics performance shows that our method increases revenue and strengthens the sustainable goal compared to its conventional counterpart.

The preliminary results from this study pave the way for subsequent investigations using both simulation and machine learning techniques for business analytics (e.g., descriptive, predictive, and prescriptive analytics). Future study could consider the billing plans with sequential interaction between the operator and consumer. In addition, billing plans altering the current rates shall be
investigated for their achievability. A particular case of analysing how the proposed method can be achieved with practical constraints (such as the telecommunications operator cost structure) can be considered together with a cost-and-benefit analysis.

Acknowledgement

The authors would like to thank the three anonymous reviewers and the guest editors for providing valuable comments. This work was financially supported 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 (MOE) in Taiwan. Chen was supported by the research project funded by InfoTech Frankfurt am Main, Germany under USt-IdNr. DE320245686. Lin was supported in part by the Ministry of Science and Technology (MOST) under Grant 106-2221-E-009-006 and Grant 106-2221-E-009-049-MY2, in part by the “Aiming for the Top University Program” of National Chiao Tung University and the Ministry of Education, Taiwan, and in part by Academia Sinica AS-105-TP-A07 and Ministry of Economic Affairs (MOEA) 106-EC-17-A-24-0619.

Appendix A. Symbols and notations

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>number of users</td>
</tr>
<tr>
<td>M</td>
<td>number of features that are used to characterized data usages</td>
</tr>
<tr>
<td>K</td>
<td>number of clusters of users</td>
</tr>
<tr>
<td>x_n</td>
<td>data usages of the n-th user measured in the observation period where x_n is a row vector with M elements and n = 1, …, N</td>
</tr>
<tr>
<td>Θ</td>
<td>the set of measured data usages where ( \sum_{n=1}^{N}</td>
</tr>
<tr>
<td>μ_k</td>
<td>mean vector of the k-th Gaussian distribution where ( \mu_k ) is a row vector with M elements and k = 1, …, K</td>
</tr>
<tr>
<td>C_k</td>
<td>covariance matrix of the k-th Gaussian distribution, where ( k = 1, \ldots, K )</td>
</tr>
<tr>
<td>L(Θ)</td>
<td>probability density function of ( x_n ) from the k-th Gaussian distribution with parameters ( \mu_k ) and ( C_k ) where n = 1, …, N and k = 1, …, K</td>
</tr>
<tr>
<td>J_k(x_n)</td>
<td>the indicator function for the k-th cluster which takes the value 1 when ( x_n ) is the closest to ( \mu_k ) among K clusters and 0 otherwise</td>
</tr>
<tr>
<td>s_k</td>
<td>probability of ( x_n ) to be selected as the initial ( \mu_k ) of the cluster where ( 0 \leq s_k \leq 1 ) and ( 1 \leq n \leq N )</td>
</tr>
<tr>
<td>d_i,j</td>
<td>Euclidean distance between ( x_i ) and ( x_j ) where ( x_i, x_j \in \Theta ) and ( 1 \leq i \neq j \leq N )</td>
</tr>
<tr>
<td>B</td>
<td>number of disjoint bins in the histogram to approximate the probability density function of the distances from ( x_n ) to all other data in ( \Theta )</td>
</tr>
<tr>
<td>( a_b )</td>
<td>a distance interval partitioned by the B-bin histogram where ( 0 \leq b &lt; B )</td>
</tr>
<tr>
<td>( p_{b,j} )</td>
<td>a value between 0 and 1 representing the proportion of data in ( \Theta ), whose distances ( d_{b,j} ) lie in the range of ( a_{b-1} ) and ( a_b ) where ( 1 \leq n \neq j \leq N ), ( 1 \leq b \leq B ), ( 0 \leq p_{b,j} \leq 1 ) and ( \sum_{b=1}^{B} p_{b,j} = 1 )</td>
</tr>
<tr>
<td>( p_{b} )</td>
<td>vector of ( p_{b,j} ) where ( 1 \leq n \leq N )</td>
</tr>
<tr>
<td>( H(d_{b}) )</td>
<td>number of folds in the cross validation</td>
</tr>
<tr>
<td>( Φ_T )</td>
<td>training set used in the cross validation that is a subset in ( Θ )</td>
</tr>
<tr>
<td>( Φ_V )</td>
<td>validation subset used in the cross validation that is a subset in ( Θ )</td>
</tr>
<tr>
<td>( Λ )</td>
<td>basic data plan given by TEL</td>
</tr>
<tr>
<td>( Γ )</td>
<td>tariff rate for ( Λ )</td>
</tr>
<tr>
<td>( a_{l}, p_l )</td>
<td>TEL fees</td>
</tr>
<tr>
<td>( A' )</td>
<td>data plan derived by MGM</td>
</tr>
<tr>
<td>( Γ' )</td>
<td>tariff rate for ( A' )</td>
</tr>
</tbody>
</table>

Appendix B. Algorithms

Algorithm 1 Initialization algorithm.

<table>
<thead>
<tr>
<th>Input:</th>
<th>the dataset ( \Theta ); the number of clusters ( K ); the number of histogram bins ( B )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>initial parameters for Gaussian mixture model, including ( \mu_k, \Sigma_k ) and ( w_k ) for ( 1 \leq k \leq K )</td>
</tr>
<tr>
<td>1:</td>
<td>for ( i = 1 ) to ( N ) do</td>
</tr>
<tr>
<td>2:</td>
<td>for ( j = 1 ) to ( N ) do</td>
</tr>
<tr>
<td>3:</td>
<td>( d_{i,j} = \sqrt{</td>
</tr>
<tr>
<td>4:</td>
<td>end for</td>
</tr>
<tr>
<td>5:</td>
<td>for ( b = 1 ) to ( B ) do</td>
</tr>
<tr>
<td>6:</td>
<td>( p_{b,k} = \left( \frac{1}{N} \right) \sum_{j=1}^{N} I(d_{i,j} \leq a_b) )</td>
</tr>
<tr>
<td>7:</td>
<td>end for</td>
</tr>
<tr>
<td>8:</td>
<td>( H(d_{i}) = \sum_{b=1}^{B} p_{b,k} \log(p_{b,k}) )</td>
</tr>
<tr>
<td>9:</td>
<td>end for</td>
</tr>
<tr>
<td>10:</td>
<td>for ( n = 1 ) to ( N ) do</td>
</tr>
<tr>
<td>11:</td>
<td>( s_n = \left( \frac{H(d_n)}{\sum_{i=1}^{N} H(d_i)} \right) )</td>
</tr>
<tr>
<td>12:</td>
<td>end for</td>
</tr>
<tr>
<td>13:</td>
<td>for ( k = 1 ) to ( K ) do</td>
</tr>
<tr>
<td>14:</td>
<td>( r = \text{Random}(0, 1) )</td>
</tr>
<tr>
<td>15:</td>
<td>( i^* = \text{argmin}</td>
</tr>
<tr>
<td>16:</td>
<td>( \mu_k \leftarrow x_n )</td>
</tr>
<tr>
<td>17:</td>
<td>end for</td>
</tr>
<tr>
<td>18:</td>
<td>for ( k = 1 ) to ( K ) do</td>
</tr>
<tr>
<td>19:</td>
<td>( \theta_k = [x :</td>
</tr>
<tr>
<td>20:</td>
<td>( w_k = \frac{</td>
</tr>
<tr>
<td>21:</td>
<td>( C_k = \left( \frac{1}{</td>
</tr>
<tr>
<td>22:</td>
<td>end for</td>
</tr>
</tbody>
</table>

Algorithm 2 EM algorithm.

<table>
<thead>
<tr>
<th>Input:</th>
<th>the dataset ( \Phi \in \Theta ); the initial parameters derived from Algorithm 1, including ( \mu_k, \Sigma_k ) and ( w_k ) for ( 1 \leq k \leq K )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Output:</td>
<td>the refined parameters for the mixture Gaussian model: ( \mu_k, C_k, w_k ); and ( \theta_k ) for ( 1 \leq k \leq K )</td>
</tr>
<tr>
<td>1:</td>
<td>repeat</td>
</tr>
<tr>
<td>2:</td>
<td>Compute ( w_{k,m} ) for ( 1 \leq k \leq K ) and ( 1 \leq m \leq</td>
</tr>
<tr>
<td>3:</td>
<td>Update ( w_{k,m} ), ( \mu_k ) and ( C_k ) for ( 1 \leq k \leq K ) with Eqs. (8)–(10)</td>
</tr>
<tr>
<td>4:</td>
<td>Compute ( L(\Phi) ) with Eq. (2)</td>
</tr>
<tr>
<td>5:</td>
<td>until ( L(\Phi) ) is converged</td>
</tr>
<tr>
<td>6:</td>
<td>for ( m = 1 ) to (</td>
</tr>
<tr>
<td>7:</td>
<td>( \delta_{m} = \text{argmax} w_{k,m} ) with Eq. (7)</td>
</tr>
<tr>
<td>8:</td>
<td>end for</td>
</tr>
<tr>
<td>9:</td>
<td>for ( k = 1 ) to ( K ) do</td>
</tr>
<tr>
<td>10:</td>
<td>for ( m = 1 ) to (</td>
</tr>
<tr>
<td>11:</td>
<td>( \theta_k \leftarrow \bigcup_{m=1}^{N}</td>
</tr>
<tr>
<td>12:</td>
<td>end for</td>
</tr>
<tr>
<td>13:</td>
<td>end for</td>
</tr>
</tbody>
</table>

---

Please cite this article as: Y.-T. Chen, E.W. Sun and Y.-B. Lin, Merging anomalous data usage in wireless mobile telecommunications: Business analytics with strategy-focused data-driven approach for sustainability, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.02.046
Algorithm 3 Estimating the optimal number of components in MGM and their parameters.

**Input:** the dataset $\Theta$; the number of partitions in cross validation $F$; the range of candidate components for MGM $[K_{\text{min}}, K_{\text{max}}]$

**Output:** the parameters for Gaussian mixture models, including the number of components $K$, $\mu_k$, $C_k$, and $w_k$ for $1 \leq k \leq K$

1. Randomly partition $\Theta$ into $F$ disjoint subsets $\Phi_f$ for $1 \leq f \leq F$
2. For $k = K_{\text{min}}$ to $K_{\text{max}}$
3. For $f = 1$ to $F$
4. $\Phi_f = \Theta \setminus \Phi_f$
5. $\Phi_f = \Phi_f$
6. Using $\Phi_f$ to initialize parameters $\mu_k$, $C_k$, and $w_k$ with Algorithm 1.
7. Using $\Phi_f$ to obtain parameters $\mu_k$, $C_k$, and $w_k$ with Algorithm 2.
8. $AIC_{f,k} = \Phi_f$ to evaluate AIC, see Eq. (11)
9. $BIC_{f,k} = \Phi_f$ to evaluate BIC, see Eq. (12)
10. $HQC_{f,k} = \Phi_f$ to evaluate HQC, see Eq. (13)
11. End for
12. $K \leftarrow \arg \min_{k_{\text{max}} \leq K \leq k_{\text{min}}} AIC_{f,k}$
13. $\mu_k \leftarrow \mu_k$ for $1 \leq k \leq K$
14. $C_k \leftarrow C_k$ for $1 \leq k \leq K$
15. $w_k \leftarrow w_k$ for $1 \leq k \leq K$

Algorithm 4 Bootstrapping algorithm.

**Input:** the dataset $\Theta$; the number of bootstraps $N_B$; the sampling threshold $\epsilon$

**Output:** the set of bootstrapped data $\Theta_B$

1. Initialization: $\Theta_B = \emptyset$
2. While $|\Theta_B| < N_B$
3. For $n = 1$ to $N_B$
4. Generating uniformly distributed random number: $r = \text{Random}(0, 1)$
5. If $r \geq \epsilon$ and $|\Theta_B| < N_B$ then
6. $\Theta_B = \Theta_B \cup \{\xi_n\}$
7. End if
8. End for
9. End while

Algorithm 5 Determination of the tariff for MGM

**Input:** the number of clusters $K$; the TEL data allowance $\Lambda = \bigcup_{1 \leq s \leq S} \{A_s\}$; the TEL tariff $\Gamma = \bigcup_{1 \leq s \leq S} \{Y_s\}$; the data allowance determined by MGM $\Lambda' = \bigcup_{1 \leq s \leq S} \{A'_s\}$

**Output:** the tariff for MGM $\Gamma' = \bigcup_{1 \leq s \leq S} \{Y'_s\}$

1. For $i = 1 \to K$
2. $j \leftarrow \text{arg \ min}_{1 \leq s \leq S} |Y'_s - \gamma_j|$
3. $\gamma'_i \leftarrow \gamma_j$
4. End for

Algorithm 6 Total revenue calculation.

**Input:** the dataset $\Theta$; the TEL data allowance $\Lambda$; the TEL tariff $\Gamma$; the TEL coverage plan $\alpha_i$ and its rate $p_i$ for $i = 1, 2$

**Output:** the total revenue $\text{Rev}(\Theta, \Lambda, \alpha, p)$

1. For $n = 1 \to N$
2. $\delta_n \leftarrow \text{arg \ min}_{1 \leq k \leq K} \gamma_k + \sum_{m=1}^{12} \lambda_k + V(X_{nm}, \lambda_k)$
3. End for
4. For $k = 1 \to 6$
5. For $n = 1 \to N$
6. $\theta_k = \sum_{i=1}^{N} \lambda_{\Phi_{ki}}$
7. End for
8. End for
9. $\text{Rev}(\Theta, \Lambda, \alpha) = \sum_{i=1}^{6} \sum_{k=1}^{12} \lambda_k + V(X_{nm}, \lambda_k)$

10. Function $V(x, \lambda)$
11. $n_1 = 0$ Initializing the number of $a_1$ (0.2 gigabytes)
12. $n_2 = 0$ Initializing the number of $a_2$ (1 gigabyte)
13. If $x > \lambda$ then
14. $n_1 = |x - \lambda|$
15. Else $a_1$
16. $n_2 = |x - \lambda|$
17. End if
18. End if
19. $V = p_1 n_1 + p_2 n_2$
20. End function

References


Please cite this article as: Y.-T. Chen, E.W. Sun and Y.-B. Lin, Merging anomalous data usage in wireless mobile telecommunications: Business analytics with strategy-focused data-driven approach for sustainability, European Journal of Operational Research, https://doi.org/10.1016/j.ejor.2019.02.046


Heyns, H. (2015). Becoming an analytics-driven organisation to create value. Ernst & Young LLP UK.


