




Cluster-based data filtering for manufacturing big data systems

Yifu Li, Xinwei Deng, Shan Ba, William R. Myers, William A. Brenneman, Steve J. Lange, Ron Zink & Ran Jin

To cite this article: Yifu Li, Xinwei Deng, Shan Ba, William R. Myers, William A. Brenneman, Steve J. Lange, Ron Zink & Ran Jin (2021): Cluster-based data filtering for manufacturing big data systems, Journal of Quality Technology, DOI: [10.1080/00224065.2021.1889420](https://doi.org/10.1080/00224065.2021.1889420)

To link to this article: <https://doi.org/10.1080/00224065.2021.1889420>

 [View supplementary material](#) 

 Published online: 05 Mar 2021.





 [Submit your article to this journal](#) 

 [View related articles](#) 

 [View Crossmark data](#) 



Cluster-based data filtering for manufacturing big data systems

Yifu Li^a , Xinwei Deng^b , Shan Ba^c, William R. Myers^d, William A. Brenneman^d, Steve J. Lange^d ,
Ron Zink^d, and Ran Jin^a 

^aGrado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia; ^bDepartment of Statistics, Virginia Tech, Blacksburg, Virginia; ^cLinkedIn Corporation, Mountain View, California; ^dProcter and Gamble Corporation, Cincinnati, Ohio

ABSTRACT

A manufacturing system collects big and heterogeneous data for tasks such as product quality modeling and data-driven decision-making. However, as the size of data grows, timely and effective data utilization becomes challenging. We propose an unsupervised data filtering method to reduce manufacturing big data sets with multi-variate continuous variables into informative small data sets. Furthermore, to determine the appropriate proportion of data to be filtered, we propose a filtering information criterion (FIC) to balance the tradeoff between the filtered data size and the information preserved. The case study of a baby care manufacturing and a simulation study have shown the effectiveness of the proposed method.

KEYWORDS



big data; data filtering; data quality; hub clustering; smart manufacturing


1. Introduction

With the advancement of sensor and communication technologies, Industrial internet-based sensing systems can collect data over a long period of time from various manufacturing environments and machine settings. While such a sensing system is capable of collecting a large amount of data, there may contain redundant information, which significantly limits the quality and efficiency of data analysis (Kenett and Shmueli 2014). For example, a manufacturing process may generate a lot of sensor data from conformance manufacturing status, where the readings of the sensors have small variation through process modeling or control analysis, and little information can be extracted. As a result, selecting a meaningful or representative subset to reduce the overall size of the data while ensuring the quality of the subset is important to improve the efficiency of data analysis. Here the data quality is strongly related to the information richness, which can be measured by entropy (Gray 2011) and the performance of the data analysis. It is expected that a small but information-rich subset filtered from the massive raw data sets can effectively support various data analyses and reduce the time latency in computation without significantly sacrificing the performance of the analysis.

The objective of this work is to reduce the sample size of manufacturing big data while maintaining the amount of useful information for efficient data analysis. This work is motivated by a continuous baby care manufacturing system, where the data were continuously collected as multi-dimensional time-series data. The cloud-storage is typically expensive for data from such a process as there will be more than three million data points collected from a single production line over one month. Under this motivation, we focus on proposing a method to reduce the information loss (i.e., entropy loss) when data is filtered/sampled for storage. After preliminary investigation, we found out that within a time window (e.g., several hours or days), the manufacturing process produces conforming products and the data collected may contain redundant information, e.g., constant values with little information for manufacturing analysis. Big data with redundancy pose challenges to efficient data storage and timely data analysis in a continuous manufacturing process. We believe that a properly filtered data set can preserve the majority of the original data set's useful information leading to computational saving and improved data analysis performance.

Reducing the size of data while preserving the information was a major focus for data reduction under a data-rich environment (Liu et al. 2015; Xian

CONTACT Ran Jin  rran5@vt.edu  Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061. Yifu Li is now affiliated with Department of Industrial and Systems Engineering, University of Oklahoma in Norman, Oklahoma.

 Supplemental data for this article can be accessed on the publisher's website at <https://doi.org/10.1080/00224065.2021.1889420>

© 2021 American Society for Quality

et al. 2018). Along this direction, the most widely adopted approaches are probability-based filtering, which means that the probability of preserving each observation is pre-determined for filtering. Popular probability-based filtering methods include random sampling (Clarkson and Shor 1989; Liu, Sadygov, and Yates 2004) and stratified sampling (Trost 1986; Liberty, Lang, and Shmakov 2016). However, the probability-based data filtering methods often overlook inherent group structures or clustering patterns among data, and may easily preserve excessive data from large clusters, while ignoring small clusters.

In manufacturing, similar processes and equipment can produce clusters of data having similar information in terms of statistical moments (Yamaoka, Nakagawa, and Uno 1978). Filtering guided by such clustering patterns can effectively reduce the information redundancy by removing a large number of observations with similar information. A natural outcome is to utilize data clustering patterns for sampling. One recent work along this direction is Singh and Masuku (2014), which first clusters the raw data, then select a number of large data clusters to be fully preserved, while ignoring the smaller ones. However, one improvement opportunity in this work is that the clustering algorithm performed on the full data set may be over-complex, as the size of data can easily become too large to cluster. Furthermore, as the size of data is reduced through filtering, the information preserved will inevitably degrade and the resulting data analysis performance becomes worse. There is a pressing need to propose a method to optimize the selection of the filtering ratio, which is the proportion of data to be preserved after filtering, by considering the information loss.

We propose an unsupervised data filtering method along with a filtering information criterion (FIC) to automatically determine the proportion of data preserved in filtering. The proposed method aims to select representative subsets from raw data. Specifically, the unsupervised data filtering method includes two steps. The first step is to use certain index tags, such as time index tags, to segment the raw data into different segments (denoted as hubs) whenever there is a large gap between index tag values for two adjacent data observations. We assume that each hub has different characteristics compared with other hubs, as large index gaps usually indicate manufacturing events that impact the characteristics of *in situ* variables. An example of a large index gap can be caused by equipment shutdown, and the manufacturing process may run under different conditions after

the equipment is restarted. The second step is to partition each hub into clusters, extract the centroid of each cluster, and perform cluster-wise random sampling. The proposed two-step method can recover the clustering pattern from raw data and help to better preserve the information by retaining the data from each cluster with the determined filtering ratio. Furthermore, the computational speed will be significantly accelerated by performing clustering within hubs, instead of using the full data set.

To determine the best filtering ratio, we propose a filtering information criterion (FIC) to balance a tradeoff between the information preserved and the size of the filtered data set. It is worth mentioning that the proposed sampling method does not rely on any data distribution assumption while deriving the analytical form of FIC relies upon data normality assumption. A baby care manufacturing case study is used to evaluate the filtering performance based on the optimal filtering ratio selected by FIC. We further conducted simulation studies to systematically evaluate the filtering method at multiple levels of filtering ratios. The numerical results from studies show the promising performance of the proposed filtering method, compared with the benchmark methods, such as random sampling and stratified sampling.

The remainder of the paper is organized as follows. In Section 2, we introduce the proposed data filtering method and the FIC. In Section 3, we perform a case study in baby care manufacturing to test the proposed filtering method with FIC. In Section 4, we perform a simulation study inspired by the manufacturing data set for comparing the proposed data filtering method with other benchmark methods. In Section 5, we summarize the paper and discuss future work.

2. The proposed data filtering method

Denote the full manufacturing data set as $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}' \in \mathbb{R}^{n \times p}$, which contains n observations $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$, $i = 1, \dots, n$. Here we assume that there are p continuous variables of interest (e.g., signals from p sensors). Each observation \mathbf{x}_i is often associated with its index tag, denoted as t_i . For example, the index tag can be the operation time-stamp (a quantitative variable) or the current machine operational status (a categorical variable). Note that although the size of data, n , is usually very large, the corresponding manufacturing process conforms to the specifications for the majority of the time. To filter the full data \mathbf{X} , $\hat{\mathbf{X}}$ is selected from \mathbf{X} , such that $\hat{\mathbf{X}} \subset \mathbf{X}$, with $\hat{\mathbf{X}}$ contains the raw data points or summary

statistics (e.g., centroids of clusters for clustered data) from \mathbf{X} , to preserve the information in a smaller size. Here the data filtering can be formulated as,

$$\min_{\hat{\mathbf{X}}} l(\hat{\mathbf{X}}, \mathbf{X}) \text{ s.t. } \frac{|\hat{\mathbf{X}}|}{n} \leq r, \quad (2.1)$$

where $l(\hat{\mathbf{X}}, \mathbf{X})$ is a loss function, such as negative log-likelihood or entropy loss (Gray 2011), to quantify the information loss between the filtered data and the full data. The $|\cdot|$ is the cardinality or the sample size of a data set. Here $0 \leq r \leq 1$ is a tunable filtering ratio.

To filter the raw data considering the data clustering patterns, we propose to incorporate clustering into the loss function in (2.1). For example, assume that \mathbf{X} can be partitioned into k clusters, where $\mathbf{X} = \mathbf{C}_1 \cup \dots \cup \mathbf{C}_k$. The data in different clusters are heterogeneous in terms of means, variances, etc. A representative subset of such a data set after the data filtering is denoted as $\hat{\mathbf{X}} = \hat{\mathbf{C}}_1 \cup \dots \cup \hat{\mathbf{C}}_k$, where $\hat{\mathbf{C}}_i$ contains either the raw data or the summary statistic of \mathbf{C}_i . Therefore, we incorporate the clustering into (2.1) and propose cluster-based data filtering as

$$\begin{aligned} \min_{\hat{\mathbf{X}}, \mu_i, \mathbf{C}_i, \hat{\mathbf{C}}_i} l(\hat{\mathbf{X}}, \mathbf{X}) + \lambda \sum_{i=1}^k \sum_{\mathbf{x}_j \in \mathbf{C}_i} \|\mathbf{x}_j - \mu_i\|_2^2 \\ \text{ s.t. } \frac{|\hat{\mathbf{C}}_i|}{|\mathbf{C}_i|} \leq r, \text{ for } i = 1, \dots, k \end{aligned} \quad (2.2)$$

where μ_i is the centroid of each data cluster, $\|\cdot\|_2$ is the Euclidean norm, and λ is the coefficient for the clustering term, which minimizes the within-cluster distances. Here, we assume that the size of data extracted from each cluster is proportional to the sample size of each cluster. Although the loss function in (2.2) is well-defined, there are two major computational challenges when using it in filtering. The first one is that the optimization problem involves finding the optimal data partitioning and sub-sampling, which is NP-hard (Burdakov, Kanzow, and Schwartz 2015). Furthermore, solving the multi-objective problem presented in loss function in (2.2) requires the selection of λ , which can significantly increase computational cost if iterative parameter tuning procedures are adopted (e.g., 5-fold cross-validation).

To address these challenges, we propose a heuristic data filtering approach by combining the index-based data partition and the cluster-based data sampling. The method will divide the data into hubs and clusters and extract subsets of data from clusters. In the first step, we adopt the index-based data partition, which is computationally fast, to partition the full data into hubs and pave an efficient way to enable cluster-wise data filtering. In the second step, within

each hub, we perform clustering and randomly sample each cluster of data proportionally.

2.1. Cluster-based data filtering method

The index-based data partition consecutively partitions the full data along the index tags into q data hubs as $\mathbf{X} = \mathbf{H}_1 \cup \dots \cup \mathbf{H}_q$. Here, \mathbf{H}_j , where $j = 1, \dots, q$, are consecutive and non-overlapping data subsets of \mathbf{X} over time. When index tags are categorical variables, it is straight-forward to hub partition, as each unique value of a categorical variable will form one data hub. However, for continuous variables, it is not intuitive, and we propose a quantile-based method for hub partition. The data partition (segmentation) for continuous variables has many variants, such as using a likelihood criterion (Guralnik and Srivastava 1999), minimum message length approximation (Fitzgibbon, Dowe, and Allison 2002), and landmark identification (Ibragimov et al. 2014). Note that for manufacturing processes with index tags, it is not easy to assume certain probabilistic distribution properties for the index tag variable. Thus, segmentation based on likelihood and message length approximation is not applicable to our problem. Alternatively, we adopt the idea of finding landmarks (Perng et al. 2000) or perceptually important points (PIPs) (Zhang, Jiang, and Wang 2007) based on index tags for segmentation.

Note that the index tags in the manufacturing data often reflect the dynamics of a manufacturing system. A large gap between two consecutive index tags often reflects a change in the manufacturing system. For example, when the operation times are used as the index tags, a gap for several hours between two consecutive observations may indicate that there is a product change-over. The aforementioned manufacturing events can significantly vary the *in situ* conditions of manufacturing processes, and generate data in hubs. Partitioning the data into hubs will reduce the complexity of clustering, compared with the clustering in the full data set. Practitioners can always use domain knowledge to choose significant gaps among data hubs, such as identifying the process change-over, maintenance events, product receipt changes etc. However, we would like to propose a data-driven rule to identify significant gaps among data hubs when no such domain knowledge exists. Specifically, to incorporate the information on the gaps of index tags in the proposed method, we consider to use the second-order tag gap at t_i , defined as $\delta_i = (t_{i+1} - t_i) - (t_i - t_{i-1})$. When time tags of data collection are adopted

as the index tags, the second-order tag gap can identify the large index tag gaps, either in the time domain (the first-order information) or the frequency domain (the second-order information) among index tags. We consider to form the segments by partitioning the raw data \mathbf{X} at the following locations

$$\{j : \delta_j \geq q_{1-\alpha}(\delta_1, \dots, \delta_n)\}, \quad (2.3)$$

where $q_{1-\alpha}(\delta_1, \dots, \delta_n)$ is the $(1 - \alpha)$ percentile of all $\delta_1, \dots, \delta_n$. It means that the gap values at these locations are larger than the $(1 - \alpha)$ percentile of all δ_j 's.

For each identified hub, we cluster it into k_j clusters for hub \mathbf{H}_j as $\mathbf{H}_j = \mathbf{C}_1^{(j)} \cup \dots \cup \mathbf{C}_{k_j}^{(j)}$. Here we adopt the k-means clustering method (MacQueen, et al. 1967) to form clusters $\mathbf{C}_1^{(j)}, \dots, \mathbf{C}_{k_j}^{(j)}$ for each hub. Specifically, the k-means clustering method minimizes the total within-cluster sum of squares for the clusters in a hub as

$$\min \sum_{s=1}^{k_j} \sum_{\mathbf{x}_i \in \mathbf{C}_s^{(j)}} \|\mathbf{x}_i - \boldsymbol{\mu}_s^{(j)}\|_2^2, \quad (2.4)$$

where $\boldsymbol{\mu}_s^{(j)}$ is the centroid of the cluster s from the hub \mathbf{H}_j . The number of clusters k_j for each hub is selected to maximize the average Silhouette distance of all data points in the hub (Rousseeuw 1987). Given any data point \mathbf{x}_i , the Silhouette distance calculates the difference between the average distance of \mathbf{x}_i to the other data points in the same cluster and the average distance of \mathbf{x}_i to the other data points in different clusters. The Silhouette distance is adopted because it is a state-of-the-art method at measuring how well that data points are matched to their respective clusters, hence evaluates the overall clustering performance at any given k_j (Tomasev et al. 2014). Finally, we generate samples $\hat{\mathbf{C}}_s^{(j)}$ from each cluster $\mathbf{C}_s^{(j)}$ to form the filtered data set $\hat{\mathbf{X}} = \bigcup_{j=1}^q \bigcup_{s=1}^{k_j} \hat{\mathbf{C}}_s^{(j)}$, where $\hat{\mathbf{C}}_s^{(j)}$ consist of $100r\%$ data points randomly sampled from cluster $\mathbf{C}_s^{(j)}$ plus the cluster centroid $\boldsymbol{\mu}_s^{(j)}$. Here, the centroids are important summary statistics of clusters, and adding them to the filtered dataset is crucial for the proposed method to preserve representative information of clusters. Similar to prior works such as Abramowicz et al. (2017), we consider that centroids/medoids are the representative summary statistics for clusters, and we expect that adding cluster centroids in the filtered data sets results in less information loss and better model fitting performance. In the meantime, various alternative approaches at utilizing centroids such as offering centroid a higher weight for down-stream tasks, adding more copies of centroids to the filtered data, or only preserving the centroid after filtering are beneficial future works. Further exploring the capability of centroids at balancing the

tradeoff between filtering ratio and information loss may result in better ways of performing data filtering.

2.2. Selection of the filtering ratio r

The filtering ratio r determines the degradation of data quality. For example, the entropy loss quantitatively measures the information loss between the raw and filtered data sets, which is defined as $EL = \text{trace}(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}}) - \log(\det(\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\Sigma}})) - p$, where $\text{trace}(\cdot)$ is the trace operator, $\det(\cdot)$ is the determinant operator, $\boldsymbol{\Sigma}$ is the sample covariance matrix of the raw data set, $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix of the filtered data set, and p is the number of variables in the data set.

Determining the filtering ratio r is not trivial. For example, in P&G babycare manufacturing data with a size of $n = 24915$ observations and $p = 32$ variables, the increase of loss due to the decrease of data size preserved is non-linear. In Figure 1, we compared the entropy loss acquired by random sampling under different filtering ratios. From Figure 1 we can see that the entropy loss remains stable when the filtering ratio is larger than $r = 0.2$, but increases dramatically when the filtering ratio is smaller than $r = 0.2$. As a result, $r = 0.2$ may be considered as a good choice of the filtering ratio, since it has a balanced small filtering ratio with a relatively small entropy loss. We will make use of this observation to test if our proposal can identify the optimal filtering ratio.

In practice, the exact tradeoff relationship between the filtering ratio and the entropy loss among data sets may vary significantly. The filtering ratio r can be determined by users based on their experience. However, there is a lack of statistical justification behind the *ad hoc* selection of the filtering ratio for achieving a good balance between the information preserved and the size of the filtered data $\hat{\mathbf{X}}$.

Here we propose a statistical framework on the selection of optimal value of r . Specifically, we propose FIC, as the criterion to find the optimal value of r . The FIC is motivated and modified from Akaike information criterion (AIC), a statistical model selection method (Akaike et al. 1998). Assuming that the data set \mathbf{X} has n observations and follows the independent and identically distributed normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\mu}$ as the mean and $\boldsymbol{\Sigma}$ as the covariance matrix, then the log-likelihood of the full data can be written as

$$\begin{aligned} l(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) &= -n \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \\ &= -n \frac{1}{2} \log |\boldsymbol{\Sigma}| - n \frac{1}{2} \text{trace}(\boldsymbol{\Sigma}^{-1} \mathbf{S}), \end{aligned} \quad (2.5)$$

up to some constant and where

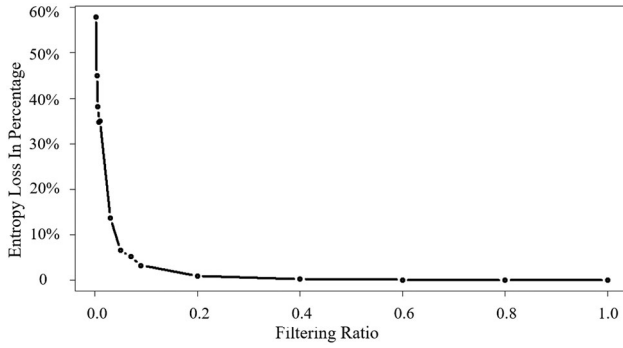


Figure 1. The increase of entropy loss as the filtering ratio decreases.

$$S = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})'$$

The original AIC is to consider the balance between the fitness of the model to the data and the model complexity, which can be expressed as

$$AIC(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -2l + 2k, \quad (2.6)$$

where l is the log likelihood function and k is the number of estimated parameters in the model.

We propose to use an AIC-like measure to find the optimal ratio r . To do this, we estimate the mean and covariance matrix from the filtered data, which are conducted by using the weighted sample mean and sample covariance matrix from the full data

$$\hat{\boldsymbol{\mu}} = \frac{1}{m} \left(\sum_{i=1}^n w_i \mathbf{x}_i + \sum_{i=1}^q w_i \boldsymbol{\mu}_i \right),$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{m} \left(\sum_{i=1}^n w_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})' + \sum_{i=1}^q v_i (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}})' \right),$$

where $w_i \in \{0, 1\}$, $v_i \in \{0, 1\}$, q is the total number of clusters identified in the full data, $\boldsymbol{\mu}_j$ is the centroid of the cluster j , $m = \sum_{i=1}^n w_i + \sum_{i=1}^q v_i$, is the total number of preserved data points, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are the mean and covariance generated from the filtered data set based on both preserved data points and cluster centroids. Then a modified AIC, denoted as filtering information criterion (FIC), for evaluating the quality of the filtered data can be written as

$$\begin{aligned} FIC(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) &= -2l(\mathbf{X}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) + 2|\hat{\mathbf{X}}| \\ &= n \log |\hat{\boldsymbol{\Sigma}}| + \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}) + 2|\hat{\mathbf{X}}| \\ &= n \left[\log |\hat{\boldsymbol{\Sigma}}| + \text{trace}(\hat{\boldsymbol{\Sigma}}^{-1} \mathbf{S}^*) + 2 \frac{|\hat{\mathbf{X}}|}{n} \right], \end{aligned} \quad (2.7)$$

where $\mathbf{S}^* = \frac{1}{n} \sum_{i=1}^n (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})'$.

Given a pre-defined set $\mathbf{r} = \{r_1, \dots, r_m\}$, the optimal filtering ratio r^* is chosen to have the smallest

corresponding FIC value resulted from the filtering. In the case study and simulation that follows, we will evaluate the performance of the proposed filtering method and the FIC.

A pseudo-code for the proposed filtering method is summarized as follows.

Algorithm 1 Pseudo code for the proposed filtering method

Step 1: Split raw data into hubs \mathbf{H}_j , where $j = 1, \dots, q$ satisfying (2.3).

Step 2: Identify k_j clusters $\mathbf{C}_s^{(j)}$, where $s = 1, \dots, k_j$, within each hub \mathbf{H}_j using k-means clustering and Silhouette distance.

Step 3: Randomly sample from each cluster $\mathbf{C}_s^{(j)}$ based on the filtering ratio r :

for each hub \mathbf{H}_j , where $j = 1, \dots, q$, **do**

for each cluster $\mathbf{C}_s^{(j)}$, where $s = 1, \dots, k_j$, **do**

Randomly sample $100r\%$ data points and the cluster centroid $\boldsymbol{\mu}_s^{(j)}$ from cluster $\mathbf{C}_s^{(j)}$.

Denote these observations for each cluster as $\hat{\mathbf{C}}_s^{(j)}$.

end forend for

Step 4: Combine all cluster $\hat{\mathbf{C}}_s^{(j)}$ and form the filtered data $\hat{\mathbf{X}} = \bigcup_{j=1}^q \bigcup_{s=1}^{k_j} \hat{\mathbf{C}}_s^{(j)}$.

We would like to remark that although the proposed filtering method does not rely on distributional assumption of the underlying data, the derivation of FIC does rely on the Multivariate Normality assumption of data. In the later studies, we tested the performance of both FIC and proposed filtering method on various data sets with different distributions. The result shows that FIC can make reasonable filtering ratio suggestions even when the data do not follow the Multivariate Normal assumption. However, we believe that relaxing the multivariate Normality assumption to work with data having a variety of and mixed distribution can become a valuable future work. Note that the formulation in (2.7) can be viewed as an information discrepancy between $\hat{\boldsymbol{\Sigma}}$ and \mathbf{S}^* . Following this observation, it is possible to use FIC as a general guideline to select the optimal filtering ratio for data with proper mean and covariance matrix. The analysis of the case study provides further evidence on the use of FIC.

3. Case study

In collaboration with the P&G baby care manufacturing sector, we evaluate the performance of the proposed method. A data collected from the P&G production line has a size of $n = 24915$ observations and $p = 32$ variables including the time tag. Although the physical meanings of variables are not disclosed

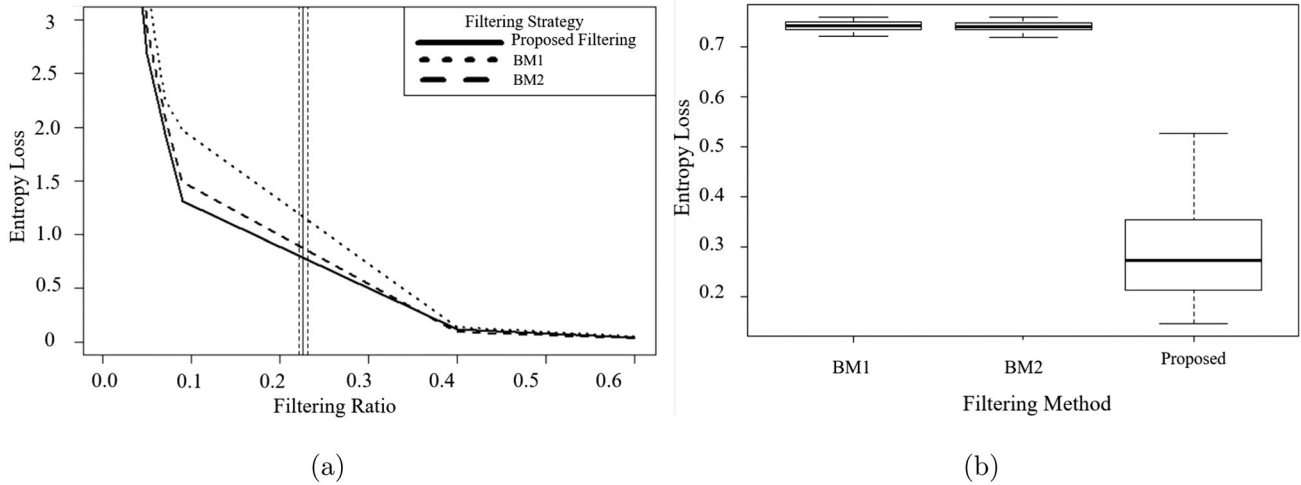


Figure 2. (a) The mean and standard errors (in vertical solid and dotted lines) of the optimal filtering ratio over 50 replications using the proposed filtering method versus the mean entropy loss at different filtering ratios for all methods over 50 replications. (b) The comparison in a boxplot of entropy loss for all methods using the fixed filtering ratio over 50 replications.

here due to the data non-disclosure agreement, we summarized the summary statistics, the histogram of each centralized variable, and normality test of the data set in the [supplementary materials](#). Based on the summary, we observe that although the whole dataset doesn't follow multivariate Normality, which requires all variables to follow Normal distribution, approximately half of variables passed the univariate Normality test while several others resemble the Normality patterns. As a result, FIC's assumption is partially fulfilled to a certain degree for the case study data so that likelihood in FIC can still quantify the similarities between the raw data and the filtered data to a certain degree. However, we want to remind readers again that relaxing the multivariate Normality assumption for FIC to improve the generality of the proposed method is a valuable future work.

Typically, a manufacturing big data system can have millions of observations for analysis. This selected data set is to provide a representative performance evaluation in a big data environment for two reasons. The first reason is that some engineering-driven partition can divide a big data set into sub-data with smaller sample sizes for filtering. The second reason is that production engineers often avoid waiting until the collected data for analysis becomes big sizes, which can cause a significant delay in decision-making.

The objective of this case study is two-fold. First, we demonstrate that an effective filtering ratio can be selected by FIC to achieve a good balance between the size of data preserved and the data quality. Second, we evaluate the performance of the

proposed filtering method in comparison with several benchmark methods.

We used the entropy loss (Gray 2011) to evaluate the entropy loss of the filtered data sets compared with the raw data efficiently. Two benchmark methods, denoted as BM_1 and BM_2 were included for comparison. The BM_1 is random sampling (Liu, Sadygov, and Yates 2004) which extracts observations randomly from the full data set, and the BM_2 is stratified sampling (Liberty, Lang, and Shmakov 2016) which extracts one observation within every equally spaced and non-overlapping time window. For example, given a filtering ratio r and n data points, we stratified the raw data into approximately nr consecutive and equally sized data segments and preserved the first observation of each segment. The threshold for forming hubs by segmentation in (2.3) set $\alpha = 0.01$. To illustrate the impact of different α on gap identification, we compare the gaps above the $(1 - \alpha)$ percentile with $\alpha = 0.02, 0.01, 0.001$ versus the gaps identified in the natural log space in the appendix. To evaluate the performance of the proposed filtering method, 50 replications with 90% of randomly extracted data were performed on varying and fixed filtering ratios to ensure reproducibility. Here the time order of the randomly extracted data is preserved in each replication.

In Figure 2, we summarized the average entropy loss for each filtering method with varying (Figure 2a) and fixed (Figure 2b) filtering ratios. The mean value and the standard error of the optimal filtering ratios selected by FIC over 50 replications are symbolized as solid and dashed vertical lines respectively in Figure 2a. In Figure 2a, the entropy loss appears not to improve significantly when the filtering ratio is larger

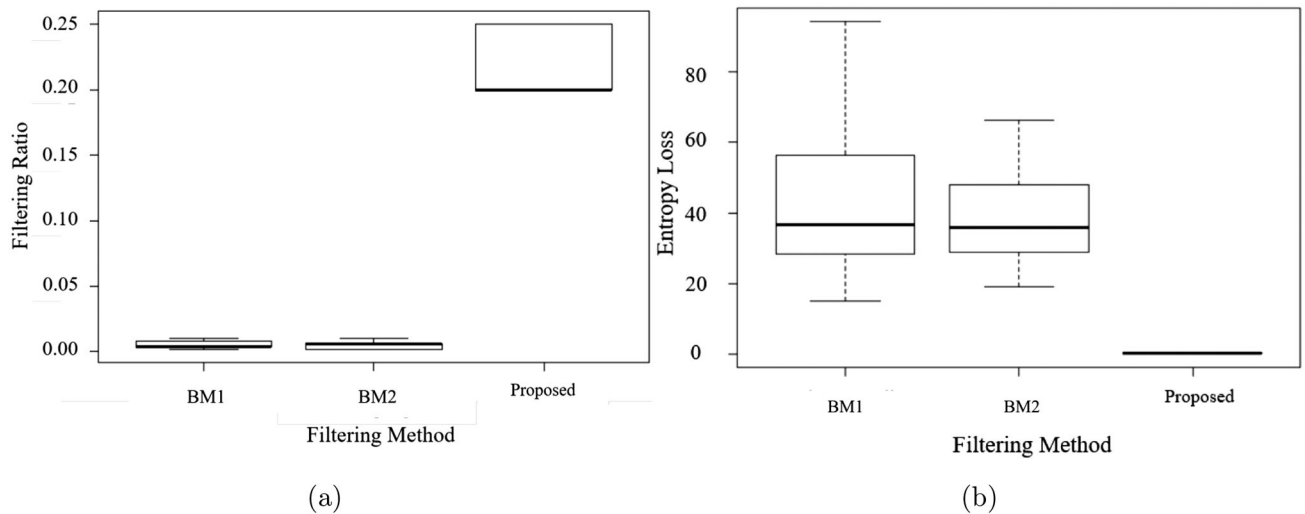


Figure 3. (a) The boxplots of the optimal filtering ratios determined over 50 replications. (b) The entropy losses obtained under the optimal filtering ratio over 50 replications.

than 0.4 and increased dramatically as the filtering ratio is smaller than 0.1. As a result, the desirable filtering ratio should be between 0.1 and 0.4. In Figure 2a, the average optimal filtering ratio was 0.235, which is within the desirable range (0.1-0.4) with very small standard errors. Furthermore, we observed that the cluster-based filtering method (the solid curve in Figure 2a) showed superior performance compared with the benchmark methods at the selected filtering ratio. We further use the same optimal filtering ratio determined by FIC for the proposed filtering method to the two benchmark methods to compare the entropy loss over 50 replications. As a summary, the boxplots in Figure 2b show that the cluster-based filtering method provides a significantly smaller entropy loss than the benchmark methods.

We also summarize the filtering performance for different methods when the filtering ratios are determined by FIC over 50 replications. Specifically, we used FIC to select the optimal filtering ratios, which yielded the lowest FIC scores for the three filtering methods, e.g., random sampling, stratified sampling, the proposed method, in each replication. The optimal filtering ratios and the corresponding entropy loss from the optimal filtering ratios are summarized in the boxplots shown in Figure 3. Figure 3a shows the boxplot of the selected filtering ratios over 50 replications. Figure 3b shows the boxplot of the entropy loss obtained under the selected filtering ratios over 50 replications. We can see that the proposed filtering method favored higher filtering ratios, which were between 0.2-0.5, while the benchmark filtering methods favored much smaller filtering ratios (0.01-0.05). As FIC jointly leverages the likelihood and the size

of data preserved, it is straightforward to see that FIC favors preserving more data points for the proposed filtering method. As a result, we conclude that the proposed filtering method can efficiently preserve the likelihood with higher filtering ratios to yield the lowest FIC. On the other hand, a much smaller filtering ratio is preferred by FIC for the benchmark filtering methods as they are less effective at preserving the likelihood where only decreasing the filtering ratio will lead to a lower FIC score. Furthermore, although the cluster-based filtering method had a relatively higher filtering ratio, it offered much smaller entropy loss. We summarize that: (1) the filtering information criterion (FIC) can successfully help to identify the optimal filtering ratio, balancing the tradeoff between the entropy loss and the size of data preserved; (2) the cluster-based filtering method can more efficiently preserve the data likelihood compared to the benchmark filtering methods; and (3) the FIC will favor saving more data when the underlying filtering methods can better preserve the likelihood.

Furthermore, to show that the proposed method can be easily applied to dataset with much larger size, we performed hub identification using $\alpha = 0.01$ on a dataset with $n = 24,915$ (the case study data) and a dataset with $n = 1,003,708$ data points from the P&G babycare manufacturing process. The result is that hub identification generated data hubs in a comparably average size of 199.902 and 196.1811 from two datasets. Such similar hub sizes indicate that the scale-up of the proposed method to new data sets from the same process can be straight-forward since hub

Table 1. Simulation setting summary.

Parameters	Low	High
Number of Clusters (NC)	30	1000
Model Sparsity (MS)	0.3	0.7
Signal-to-Noise Ratio (SNR)	3	10

The same settings were applied on Normal, t, and Uniform distribution

identification partitions raw data in significantly different sizes into similar-sized hubs for further processing.

4. The simulation study

We further evaluated the filtering performance in a simulation study. In particular, we considered varying four simulation settings: the inherent number of data clusters, the signal-to-noise ratios, the model sparsity, and the distributions that variables follow. To mimic the characteristics of the real manufacturing data set, the simulation data were generated as follows: for two settings on the inherent number of clusters, we respectively clustered the real data set into NC clusters using k-means clustering, each denoted as \hat{X}_f . Then we extracted the original time stamp for each observation in the cluster f as t_f and the means of all variables as μ_f , where $f = 1, \dots, NC$ from real data. Then we simulated each data cluster with a mixed of exponential, non-exponential, symmetrical, and skewed distributions: 1. $X_f \sim N(\mu_f, \Sigma)$, 2. $X_f \sim t(\mu_f, \Sigma, df = 10)$, 3. $X_f \sim Uniform(\mu_f - 0.5, \mu_f + 0.5)$, 4. $X_f \sim Gamma(k, \theta)$, where the shape parameter $k \sim Uniform(1, 5)$, and scale parameters $\theta \sim Uniform(1, 2)$, 5. a mixed distribution with each variable following either Student's t, Uniform, or Gamma distribution with the parameters described above (2,3,4) with 1/3 of probability. X_f has the same size as \hat{X}_f , and the regularized covariance matrix $\Sigma = [\sigma_{ij}]$ with $\sigma_{ij} = 1$ when $i = j$, and $\sigma_{ij} = 0.3$ when $i \neq j$. To investigate the different impact of covariance structure, we can investigate the results from datasets following t-distribution with the same covariance matrix and Uniform distribution, which did not rely on covariance. Finally, we aggregated all simulated clusters X_f with the time stamps t_f to produce the simulated data matrix X .

We further created simulation models to evaluate the modeling performance. To generate the response variable y , we used a linear model given by $y = \tilde{X}\beta + \varepsilon$, where $y = (y_1, \dots, y_n)'$ was the response, \tilde{X} was the simulated data matrix with all the main effect variables $(t, x_1, \dots, x_p)'$ in X and the two-factor interactions in the multiplication form $(tx_1, \dots, x_{p-1}x_p)'$. $\beta = (\beta_t, \beta_1, \dots, \beta_p, \beta_{t1}, \dots, \beta_{(p-1)p})'$ is the model parameter

vector corresponding to main effect variables and the interaction terms, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ were the residual terms with $\varepsilon_i \sim N(0, \sigma^2)$ as independent and identically distributed. To obtain the value of β , we first trained a linear model using the real data set, and extracted the model coefficients as $\hat{\beta}$. Then, we randomly set $MS\%$ of main effect parameters $(\beta_t, \beta_1, \dots, \beta_p)'$ and 10% of interaction term parameters $(\beta_{t1}, \dots, \beta_{(p-1)p})'$ in β as significant (non-zeros), with values randomly chosen from $\hat{\beta}$. Furthermore, we varied the signal-to-noise ratio (SNR) when generating the error term ε , where $SNR = \frac{var(\tilde{X}\hat{\beta})}{var(\varepsilon)}$, and $var(\cdot)$ represents the variance (Friedman, Hastie, and Tibshirani 2001). In summary, two levels of the three factors are shown in Table 1.

We evaluated the performance of the proposed filtering method with 90% of data randomly extracted from the raw data over 50 replications to ensure reproducibility. Same as the case study, We normalized all the variables in this study. We evaluated the performance measures with both fixed filtering ratios (0.1, 0.01, 0.005) and the optimal filtering ratio (marked as *) determined by FIC and the proposed filtering method. The α value for identifying the significant second-order index gap in (2.3) is set to be $\alpha = 0.01$. Then the filtered data sets were used to estimate a linear model for the evaluation goodness-of-fit.

Five performance measures were used including entropy loss (Gray 2011), R^2 , adjusted- R^2 , the CPU time for the data filtering step, and the CPU time for the modeling step. R^2 quantifies the goodness-of-fit for the model on the filtered training data while the adjusted- R^2 (Gelman and Pardoe 2006) simultaneously evaluates the goodness-of-fit and the complexity of the model. Besides the previously adopted random sampling and stratified sampling methods, the third benchmark method was to directly use the full data set for modeling. Using the full data yielded zero entropy loss but is not necessarily the best model performance. The simulation was performed on a workstation with CPU Xeon Processor E5-2687W, 3.10 GHz, 64 GB RAM.

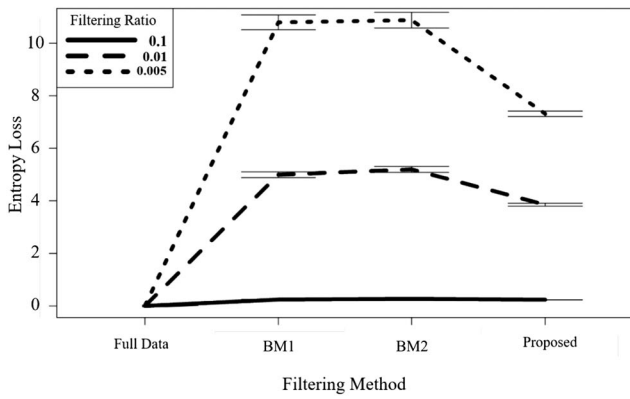
For the simplicity of presentation, we only include the tables of the results with varying inherent number of clusters in the manuscript (Tables 2 and 3), while presenting the rest of tables corresponding to the other scenarios in the [supplementary materials](#). Both tables show that the proposed method significantly outperformed the two benchmark methods on entropy loss as the filtering ratio decreased. Furthermore, the proposed method outperformed all methods in comparison on R^2 and adjusted- R^2 . As the inherent

Table 2. Means and standard errors (within parenthesis) of five performance measures based on 50 simulation replications for Normal distribution with $NC = 30$, $MS = 0.3$, and $SNR = 3$.

Filtering ratio	Filtering Method	Entropy Loss	R^2	Adjusted- R^2	Filtering Time (in seconds)	Modeling Time (in seconds)
0.1	Full Data	-	0.748	0.748	-	39.664
		-	(0.000)	(0.000)	-	(0.220)
	BM_1	0.270	0.750	0.750	0.106	10.073
		(0.005)	(0.001)	(0.001)	(0.008)	(0.146)
0.01	BM_2	0.247	0.752	0.752	0.123	10.292
		(0.004)	(0.001)	(0.001)	(0.009)	(0.223)
	Proposed	0.237	0.753	0.753	9.496	10.359
		(0.003)	(0.001)	(0.001)	(0.178)	(0.154)
0.005	Full Data	-	0.748	0.748	-	39.664
		-	(0.000)	(0.000)	-	(0.220)
	BM_1	5.194	0.764	0.762	0.052	0.725
		(0.111)	(0.005)	(0.005)	(0.004)	(0.013)
0.104*	BM_2	4.987	0.768	0.765	0.051	0.745
		(0.108)	(0.005)	(0.005)	(0.004)	(0.013)
	Proposed	3.849	0.789	0.787	9.622	0.717
		(0.056)	(0.006)	(0.006)	(0.179)	(0.017)
0.005	Full Data	-	0.748	0.748	-	39.664
		-	(0.000)	(0.000)	-	(0.220)
	BM_1	10.875	0.774	0.769	0.049	0.457
		(0.298)	(0.009)	(0.010)	(0.003)	(0.011)
0.104*	BM_2	10.790	0.791	0.786	0.068	0.484
		(0.281)	(0.009)	(0.009)	(0.007)	(0.010)
	Proposed	7.312	0.833	0.829	9.595	0.403
		(0.101)	(0.011)	(0.011)	(0.176)	(0.012)
0.104*	Full Data	-	0.748	0.748	-	39.664
		-	(0.000)	(0.000)	-	(0.220)
	BM_1	0.249	0.754	0.753	0.059	11.036
		(0.008)	(0.001)	(0.001)	(0.006)	(0.244)
0.104*	BM_2	0.233	0.752	0.752	0.044	11.213
		(0.006)	(0.001)	(0.001)	(0.003)	(0.252)
	Proposed	0.225	0.752	0.752	34.714	10.078
		(0.006)	(0.001)	(0.001)	(0.316)	(0.162)

Table 3. Means and standard errors (within parenthesis) of five performance measures based on 50 simulation replications for Normal distribution with $NC = 1000$, $MS = 0.3$, and $SNR = 3$.

Filtering ratio	Filtering Method	Entropy Loss	R^2	Adjusted- R^2	Filtering Time (in seconds)	Modeling Time (in seconds)
0.1	Full Data	-	0.750	0.750	-	39.821
		-	(0.000)	(0.000)	-	(0.169)
	BM_1	0.319	0.753	0.753	0.104	7.680
		(0.013)	(0.001)	(0.001)	(0.007)	(0.104)
0.01	BM_2	0.261	0.752	0.752	0.129	7.589
		(0.006)	(0.001)	(0.001)	(0.010)	(0.110)
	Proposed	0.302	0.754	0.754	10.019	7.705
		(0.013)	(0.001)	(0.001)	(0.337)	(0.096)
0.01	Full Data	-	0.750	0.750	-	39.821
		-	(0.000)	(0.000)	-	(0.169)
	BM_1	3.836	0.785	0.782	0.050	0.721
		(0.254)	(0.005)	(0.005)	(0.003)	(0.012)
0.005	BM_2	3.992	0.772	0.769	0.062	0.743
		(0.245)	(0.005)	(0.005)	(0.007)	(0.012)
	Proposed	3.203	0.792	0.789	9.994	0.649
		(0.176)	(0.009)	(0.009)	(0.334)	(0.009)
0.005	Full Data	-	0.750	0.750	-	39.821
		-	(0.000)	(0.000)	-	(0.169)
	BM_1	6.812	0.791	0.786	0.054	0.431
		(0.478)	(0.008)	(0.008)	(0.005)	(0.010)
0.115*	BM_2	8.652	0.788	0.782	0.058	0.505
		(0.740)	(0.010)	(0.010)	(0.005)	(0.008)
	Proposed	5.480	0.838	0.833	9.997	0.371
		(0.363)	(0.013)	(0.013)	(0.330)	(0.009)
0.115*	Full Data	-	0.750	0.750	-	39.821
		-	(0.000)	(0.000)	-	(0.169)
	BM_1	0.275	0.754	0.754	0.043	9.333
		(0.014)	(0.001)	(0.001)	(0.002)	(0.257)
0.115*	BM_2	0.215	0.753	0.752	0.041	8.993
		(0.012)	(0.001)	(0.001)	(0.001)	(0.244)
	Proposed	0.269	0.755	0.754	35.238	8.411
		(0.017)	(0.001)	(0.001)	(0.452)	(0.246)



(a) Comparison of entropy loss.

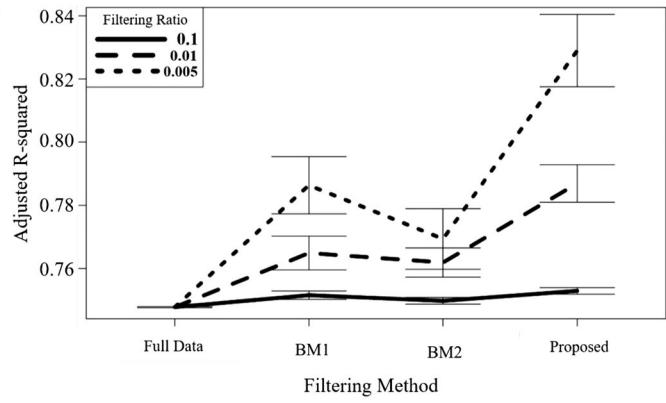
(b) Comparison of adjusted R^2 .

Figure 4. The means and standard errors (in black error bars) of two performance measures based on 50 simulation replications.

number of clusters increased from 30 (Table 2) to 1000 (Table 3), each cluster contained less data and information. As a result, the clustering pattern is becoming less significant in data composition and the modeling performance is less affected by such a pattern. However, the proposed method always outperformed the benchmark filtering methods for the majority of performance measures in both scenarios. Although the proposed filtering method took a longer time at the filtering step, it costs less than 10 seconds on average and was deemed as a time-efficient method by production engineers. The rest of the results, corresponding to the Normal, t, Uniform, Gamma, and mixed distribution, show that the proposed methods outperformed benchmark methods for the majority of scenarios at reducing entropy loss and achieving better goodness-of-fit. One thing drew our attention is that the proposed method yielded higher entropy loss in some runs for Gamma distribution (skewed distribution). One reason is that the proposed method focuses on preserving major/large clusters at low filtering ratio so that it will likely ignore the data points from small clusters. As a result, the proposed method is not as good as random and stratified sampling on preserving information from those small data clusters representing the tail of highly skewed variables.

Additionally, we generated the Figure 4 based on the information from Table 2. It is seen that there was a significant reduction of entropy loss achieved by the proposed filtering method over the two benchmark filtering methods. The proposed filtering method outperformed the benchmark methods for the adjusted R^2 as well. Such improvements became more significant as the filtering ratio decreased.

5. Discussion

As sensor and communication technologies advance, Industrial Internet-based sensing systems are capable of collecting massive data for process modeling and control. However, such systems also generate a lot of redundant information which significantly limits the quality and efficiency of the data analysis. As a result, extracting representative and high-quality data subsets is important to improve the efficiency of the data analysis. In this work, we proposed a filtering method and new criteria (FIC) to facilitate data analysis by selecting a small but effective data subset. Specifically, the proposed method partitions raw data into clusters and proportionally extracts a data subset from each cluster. The proposed cluster-based data filtering method outperformed the benchmark filtering methods on performance measures, such as information loss, the modeling goodness-of-fit in the case study and the simulation. Furthermore, the new filtering ratio selection criteria (FIC) has shown its effectiveness in terms of balancing the tradeoffs between the size of data preserved in filtering and the quality of the filtered data.

Another challenge at storing and analyzing big data is when the number of variables/dimensions of data grows high. To mitigate these issues, one can adopt variable screening or dimension reduction methods to reduce the data dimension before applying filtering and direct analysis. State-of-the-art methods on dimension reduction include but are not limited to principle component analysis (Dunteman 1989), sure independence screening (Fan and Lv 2008), manifold embedding (Nie et al. 2010), and kernel dimension reduction (Wang et al. 2010). This paper leads to a few future research directions. Data analysis under big

data environments have become prevalent and even sometimes necessary because of the computational performance requirement. However, efficiently utilizing big data through analysis that is both time-efficient and accurate is still a challenging problem. The proposed method can also facilitate internet-of-things (IoT)-based data collection, communication, storage, and analysis. For example, as the inline data de-duplication (real-time data filtering for continuous data streaming), has become more popular in recent years (Zhou, Liu, and Li 2013), future work will focus on the conversion of the current offline data filtering to online. Specifically, we need to investigate when to update the filtering ratio due to process changes, product changes, etc. Furthermore, an integer programming heuristic or relaxation technique (Schrijver 1998) can be derived to help directly solve the data filtering problem shown in the objective function (2.1) so that the filtering performance may be further improved. Lastly, we will investigate filtering for functional data (Sun, Huang, and Jin 2017) and imaging data (Li et al. 2019) as *in situ* data measurement for manufacturing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research is supported by a research collaboration program between Procter & Gamble Co. and Virginia Tech, and the National Science Foundation grant CMMI-1634867.

About the authors

Yifu Li is an assistant professor of Industrial & Systems Engineering at the University of Oklahoma. He received his Ph.D. and B.S. degree in Industrial and Systems Engineering from Virginia Tech. He previously worked as Research Assistant at Grado Department of Industrial and Systems Engineering of Virginia Tech and Feinberg School of Medicine of Northwestern University. His research interest is mainly on studying the interface between manufacturing data-driven modeling and data quality. He is a member of IEEE, INFORMS and IISE.

Xinwei Deng is an associate professor in the Department of Statistics at Virginia Tech. He received his Bachelor's degree in mathematics from Nanjing University, and PhD degree in industrial engineering from Georgia Tech. His research interests focus on statistical modeling and data analysis, including high-dimensional classification, graphical model estimation, and the interface between experimental design

and machine learning. He is an elected member of ISI, and a member of INFORMS and ASA.

Shan Ba is a data science applied researcher at LinkedIn. He had previously worked as a group data scientist at the Procter & Gamble Company and an assistant professor of statistics at the Fariborz Maseeh Department of Mathematics and Statistics, Portland State University. He received his Ph.D. in Industrial Engineering from Georgia Institute of Technology.

William R. Myers is a Principal Statistician at the Procter & Gamble Company. He received his Bachelor's degree in statistics from Radford University and a Ph.D. in biostatistics from Virginia Commonwealth University. His expertise are in the design and analysis of both physical experiments and computer experiments and statistical modeling.

William A. Brenneman is a Research Fellow and the Global Statistics Discipline Leader at Procter & Gamble in the Data and Modeling Sciences Department and an Adjunct Professor of Practice at Georgia Tech in the Stewart School of Industrial and Systems Engineering. Since joining P&G, he has worked on a wide range of projects that deal with statistics applications in his areas of expertise: design and analysis of experiments, robust parameter design, reliability engineering, statistical process control, computer experiments, machine learning and statistical thinking. He was also instrumental in the development of an in-house statistics curriculum. He received a Ph.D. in Statistics from the University of Michigan, an MS in Mathematics from the University of Iowa and a BA in Mathematics and Secondary Education from Tabor College. William is a Fellow in both the American Statistical Association (ASA) and the American Society for Quality (ASQ). He has served as ASQ Statistics Division Chair, ASA Quality and Productivity Section Chair and as Associate Editor for *Technometrics*. William also has seven years of experience as an educator at the high school and college level.

Steve J. Lange is Managing Member of ProcessDev, LLC, a manufacturing process consultancy, and retired as a Research Fellow from the Procter & Gamble Company, where he had a 35-year career in Research and Development, developing processes and materials for new products and product improvements.

Ron Zink is a Principal Engineer at the Procter & Gamble Company with over 20 years of Research & Development experience in the consumer products industry. He is a named inventor for over 40 US patents for consumer products, manufacturing equipment and processes. Ron received a BS in Chemical Engineering from the University of Florida.

Ran Jin is an Associate Professor and the Director of Laboratory of Data Science and Visualization at the Grado Department of Industrial and Systems Engineering at Virginia Tech. He received his Ph.D. degree in Industrial Engineering from Georgia Tech, Atlanta, his Master's degrees in Industrial Engineering, and in Statistics, both from the University of Michigan, Ann Arbor, and his bachelor's degree in Electronic Engineering from Tsinghua

University, Beijing. He has been working with leading manufacturing companies in the aerospace, semiconductor, personal care, optical fiber industries. His research focuses on machine learning in manufacturing, manufacturing computation services and cognitive-based interactive visualization.

ORCID

Yifu Li  <http://orcid.org/0000-0003-0602-8429>

Xinwei Deng  <http://orcid.org/0000-0002-1560-2405>

Steve J. Lange  <http://orcid.org/0000-0002-8542-5969>

Ran Jin  <http://orcid.org/0000-0003-3847-4538>

References

- Abramowicz, K., P. Arnqvist, P. Secchi, S. S., De Luna, S. Vantini, and V. Vitelli. 2017. Clustering misaligned dependent curves applied to varved lake sediment for climate reconstruction. *Stochastic Environmental Research and Risk Assessment* 31 (1):71–85. doi: [10.1007/s00477-016-1287-6](https://doi.org/10.1007/s00477-016-1287-6).
- Akaike, H., E. Parzen, K. Tanabe, and G. Kitagawa. 1998. *Selected papers of Hirotugu Akaike*. New York: Springer Science & Business Media.
- Burdakov, O., C. Kanzow, and A. Schwartz. 2015. On a reformulation of mathematical programs with cardinality constraints. In *Advances in global optimization*, ed. D. Gao, N. Ruan, and W. Xing, 3–14. Cham, Switzerland: Springer International Publishing.
- Clarkson, K. L., and P. W. Shor. 1989. Applications of random sampling in computational geometry, II. *Discrete & Computational Geometry* 4 (5):387–421. doi: [10.1007/BF02187740](https://doi.org/10.1007/BF02187740).
- Dunteman, G. H. 1989. *Principal components analysis*. Newbury Park: Sage Publications.
- Fan, J., and J. Lv. 2008. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70 (5):849–911. doi: [10.1111/j.1467-9868.2008.00674.x](https://doi.org/10.1111/j.1467-9868.2008.00674.x).
- Fitzgibbon, L. J., D. L. Dowe, and L. Allison. 2002. Change-point estimation using new minimum message length approximations. In *Proceedings of Pacific Rim International Conference on Artificial Intelligence*, 244–54.
- Friedman, J., T. Hastie, and R. Tibshirani. 2001. *The elements of statistical learning*. New York: Springer.
- Gelman, A., and I. Pardoe. 2006. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics* 48 (2):241–51. doi: [10.1198/004017005000000517](https://doi.org/10.1198/004017005000000517).
- Gray, R. M. 2011. *Entropy and information theory*. Heidelberg: Springer Science & Business Media.
- Guralnik, V., and J. Srivastava. 1999. Event detection from time series data. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 33–42. doi: [10.1145/312129.312190](https://doi.org/10.1145/312129.312190).
- Ibragimov, B., B. Likar, F. Pernuš, and T. Vrtovec. 2014. Shape representation for efficient landmark-based segmentation in 3-D. *IEEE Transactions on Medical Imaging* 33 (4):861–74. doi: [10.1109/TMI.2013.2296976](https://doi.org/10.1109/TMI.2013.2296976).
- Kenett, R. S., and G. Shmueli. 2014. On information quality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 177 (1):3–38. doi: [10.1111/rssa.12007](https://doi.org/10.1111/rssa.12007).
- Li, Y., H. Sun, X. Deng, C. Zhang, H. P. Wang, and R. Jin. 2020. Manufacturing quality prediction using smooth spatial variable selection estimator with applications in aerosol jet® printed electronics manufacturing. *IIEE Transactions* 52 (3):321–333.
- Liberty, E., K. Lang, and K. Shmakov. 2016. Stratified sampling meets machine learning. In *Proceedings of 33rd International Conference on Machine Learning*, 2320–9.
- Liu, H., R. G. Sadygov, and J. R. Yates. 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Analytical Chemistry* 76 (14):4193–201. doi: [10.1021/ac0498563](https://doi.org/10.1021/ac0498563).
- Liu, K., Y. Mei, and J. Shi. 2015. An adaptive sampling strategy for online high-dimensional process monitoring. *Technometrics* 57 (3):305–319. doi: [10.1080/00401706.2014.947005](https://doi.org/10.1080/00401706.2014.947005).
- MacQueen, J., et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281–97.
- Nie, F., D. Xu, I. W. H. Tsang, and C. Zhang. 2010. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions on Image Processing* 19 (7):1921–1932. doi: [10.1109/TIP.2010.2044958](https://doi.org/10.1109/TIP.2010.2044958).
- Perng, C., H. Wang, S. Zhang, and D. S. Parker. 2000. Landmark: A new technique for similarity-based pattern querying in time series databases. In *Proceedings of International Conference on Data Engineering*, 1–17.
- Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20:53–65. doi: [10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- Schrijver, A. 1998. *Theory of linear and integer programming*. New York: John Wiley & Sons.
- Singh, A. S., and M. B. Masuku. 2014. Sampling techniques & determination of sample size in applied statistics research: An overview. *International Journal of Economics, Commerce and Management* 2:1–22.
- Sun, H., S. Huang, and R. Jin. 2017. Functional graphical models for manufacturing process modeling. *IEEE Transactions on Automation Science and Engineering* 14 (4):1612–21. doi: [10.1109/TASE.2017.2693398](https://doi.org/10.1109/TASE.2017.2693398).
- Tomasev, N., M. Radovanovic, D. Mladenec, and M. Ivanovic. 2014. The role of hubness in clustering high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering* 26 (3):739–51. doi: [10.1109/TKDE.2013.25](https://doi.org/10.1109/TKDE.2013.25).
- Trost, J. E. 1986. Statistically nonrepresentative stratified sampling: A sampling technique for qualitative studies. *Qualitative Sociology* 9 (1):54–7. doi: [10.1007/BF00988249](https://doi.org/10.1007/BF00988249).
- Wang, M., F. Sha, and M. Jordan. 2010. Unsupervised kernel dimension reduction. *Advances in Neural Information Processing Systems* 23:2379–2387.
- Xian, X., A. Wang, and K. Liu. 2018. A nonparametric adaptive sampling strategy for online monitoring of big data streams. *Technometrics* 60 (1):14–25. doi: [10.1080/00401706.2017.1317291](https://doi.org/10.1080/00401706.2017.1317291).

- Yamaoka, K., T. Nakagawa, and T. Uno. 1978. Statistical moments in pharmacokinetics. *Journal of Pharmacokinetics and Biopharmaceutics* 6 (6):547–58. doi: [10.1007/BF01062109](https://doi.org/10.1007/BF01062109).
- Zhang, Z., J. Jiang, and H. Wang. 2007. A new segmentation algorithm to stock time series based on PIP approach. In *Proceedings of 2017 Wireless Communications, Networking and Mobile Computing*, 5609–12.
- Zhou, R., M. Liu, and T. Li. 2013. Characterizing the efficiency of data deduplication for big data storage management. In *Proceedings of 2013 Workload Characterization, IEEE International Symposium*, 98–108.