Annual Reviews in Control 000 (2018) 1–27

Contents lists available at ScienceDirect

Annual Reviews in Control

journal homepage: www.elsevier.com/locate/arcontrol

Review article

Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data

Jinlin Zhu[a,b], Zhiqiang Ge[a,*], Zhihuan Song[a], Furong Gao[b]

[a] State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China
[b] Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

ARTICLE INFO

Article history:
Received 18 June 2018
Revised 11 September 2018
Accepted 21 September 2018
Available online xxx

Keywords:
Data mining
Robustness
Process modeling
Statistical process monitoring
Big data analytics

ABSTRACT

Industrial process data are usually mixed with missing data and outliers which can greatly affect the statistical explanation abilities for traditional data-driven modeling methods. In this sense, more attention should be paid on robust data mining methods so as to investigate those stable and reliable modeling prototypes for decision-making. This paper gives a systematic review of various state-of-the-art data preprocessing tricks as well as robust principal component analysis methods for process understanding and monitoring applications. Afterwards, comprehensive robust techniques have been discussed for various circumstances with diverse process characteristics. Finally, big data perspectives on potential challenges and opportunities have been highlighted for future explorations in the community.

Contents

* Corresponding author.
E-mail address: gezhiqiang@zju.edu.cn (Z. Ge).

https://doi.org/10.1016/j.arcontrol.2018.09.003
1367-5788/

1. Introduction

Industrial processes are usually equipped with multiple connected operating flow units and distributed computer control systems, operating in a successive chemical, physical, electrical or mechanical steps so as to conduct the manufacturing of market-oriented items. Therefore, in order to ensure the orderly operations and also to maintain system performances, data modeling and monitoring act as the high-level intelligent decision-supporting modules should be indispensable (Isermann, 1984; Venkatasubramanian, Rengaswamy, Kavuri, & Yin, 2003). To manifest process running status, one has to construct analytic models so as to make the overall understanding of the object. Due to the high complexity, however, traditional first-principle based methods become no longer readily applicable, especially for those large-scale processes (Ge, 2017; Qin, 2012). As an alternative, the community in turns has put more focus on the statistically data-driven methods (Ge, 2017; Ge, Song, & Gao, 2013; Kadlec, Grbić, & Gabrys, 2011; Tidriri, Chatti, Verron, & Tiplica, 2016; Yin, Ding, Xie, & Luo, 2014). This is due to the fact that nowadays industries are stepping into the era of Internet of things, and a large amount of data have been collected from those widely equipped distributed control systems over the past few decades, which makes the intelligent data mining and automatic decision-making become tenable (Ge, Song, Ding, & Huang, 2017). Intrinsically, data mining methods based on multivariate statistical modeling and inference theories are acting as the pivots for understanding and monitoring in a wide range of complex industrial processes and systems, such as the chemical plants (Yin, Ding, Haghani, Hao, & Zhang, 2012), semiconductor manufacturing (Qin, Cherry, Good, Wang, & Harrison, 2006), mechanical systems (Gul & Catbas, 2009), food productions (Wanihsuksombat, Hongtrakul, & Suppakul, 2010), and pharmaceutical processes (Undey & Cinar, 2002).

Generally speaking, the nature of data mining methods is the extraction of useful information from data for analytic uses (Witten, Frank, Hall, & Pal, 2016). However, the successful implementation of data mining methods should largely rely on the prerequisites on underlying process data like data accuracy and data integrity. In practice, however, such assumptions can be easily violated in the presence of missing data and outliers. Process data with missing and outlying entries are usually referred to as bad data or contaminated data. Data contaminations frequently occur from various industrial processes due to errors in data collection from measurement devices, data transmission and data management, and hence is posing a big challenge for valid data mining and statistical process monitoring (Imtiaz & Shah, 2008; Kadlec, Gabrys, & Strandt, 2009). For example, in chemical processes, the bad data may be collected by sensor failures or transmission interruptions (Zhang, Gonzalez, Huang, & Ji, 2015); in pharmaceutical processes, the pollution in data can be due to the fact that the accuracy and completeness in information gathering can be hardly ensured (Boukouvala, Muzzio, & Ierapetritou, 2010). In other words, contaminated data violate the common statistical assumptions regarding both data accuracy and data integrity. Therefore, the inaccurate and incomplete information make the classic data mining methods cumbersome. In these cases, it should be difficult for analyzers to make the hypothesis and in-depth inspection on the rational of assumptions for process data, especially for those with high dimensionalities and various process characteristics (Ge, 2018). In order to make the justified analysis, one has to consider how to improve the robustness of statistical data mining methods against low quality process data with those dirt and absent sampling information.

Technically, robust data mining methods consider analyzing with robust data preprocessing and/or direct robust data modeling. The robust data preprocessing flowchart includes the removal of outliers and/or the imputation of missing entries, while robust data modeling methods deliberately improve the robustness of certain methods when coping with contaminated process data. It should be mentioned that both procedures are focused on delivering the robust statistics and parameters for characterizing industrial processes. For instance, the data preprocessing methods remove the misleading data points so that the statistics induced by classic data mining models shall describe the clean majority in the right way, while the robust modeling procedure can alleviate the negative influence with proper intrinsic robust strategies so as to obtain better empirical explanations. Several previous relevant review works are found. Daszykowski, Kaczmarek, Vander Heyden, and Walczak (2007) gave the review on robust statistics for chemometrics, with the main concentration on dealing with outliers. Filzmoser and Todorov (2011) discussed the general ideas for robust calibration and dimension reduction in high dimensional data chunks. The work by Khatibisepehr, Huang, and Khare (2013) made some discussions on the robust design of inferential sensors Xu et al. (2015) introduced intensive robust techniques on data cleaning. Recently, the work by Kodamana et al. (2018) exhibited the views on robust model identifications under the probabilistic framework. Notice that these works either give insights on specific problems or with specific solutions for specific industrial applications. There still lacks a systematic taxonomy on the state-of-the-art robust data mining techniques, as well as their application conditions for different industrial background. In literatures, there are abundant works with respect to robust statistical data mining methods, mostly developed in the last two to three decades. During this period, the robust enhancements on statistical analytics have seen great development on the outlier detection, missing data handling and robust statistical modeling. Some of these approaches have been successfully applied in various industrial pro-

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

cess scenarios. Therefore, the motivation of this survey is to provide a comprehensive overview for those state-of-the-art robust statistical data mining methods for understanding complex industrial processes with outliers and missing values.

The rest of this work is organized as follows. In Section 2, some fundamentals on data mining and analytics for missing data and outliers have been introduced and discussed in the standard form. In Section 3, the classic techniques handling robust data preprocessing issues with missing data and outliers are presented. After that, the robust statistical data mining methods have been reviewed in detail in Section 4. In Section 5, some perspectives through the viewpoint of big data and future challenging research topics have been presented. Finally, conclusions are made.

2. Concepts and overview

2.1. Outliers and missing data

Outliers usually refer to those sampling points that are distant from the major data mass and they often do not show the consistent behaviors to the rest from a statistical perspective (Barnett & Lewis, 1974). According to the causes, outliers occur in two main practical aspects. On one hand, outliers can be occasionally introduced by measurement/record error; on the other hand, some processes generate heavy-tailed data points due to large process noise. It should be mentioned that there are no strict definitions for outliers. To distinguish outliers from normal data, literatures concerning robust subspace recovery will partition data into components which are called inliers and outliers (Vaswani, Bouwmans, Javed, & Narayanamurthy, 2018). Particularly, inliers refer to data indices that lie on or near a low-dimensional space and outliers scatter in ambient fashions. According to the locations, there are also row-wise or column-wise outliers and element-wise outliers (Brahma, She, Li, Li, & Wu, 2018). Despite various classifications, it should be largely subjective exercises to make the empirical judgment of whether or not one or more data samples should be regarded as outliers. In fact, outliers appear as the frequently encountered part of data analysis procedure which could be hardly avoided from various actual sampling scenes. In many industrial processes, outliers are notorious. However, outliers are not always to be discarded as “one person’s noise could be another person’s signal” (Angiulli & Pizzuti, 2005). In some monitoring occasions like traffic flows, credit card frauds, network intrusion detection, medical analysis and video surveillance, the outliers are actually the main concerns which should be detected and extracted for further statistical analysis.

Missing data refers to cases when there are one or more incomplete data entries for the observed variables in a database. In some special situations, the multi-rate sampling of quality variables and removed outlying entries by detection can also lead to missing entries. In fact, dealing with the missing data can be actually regarded as the fairly routine practices rather than the exceptions in industries like chemical engineering. Traditional statistical inference studies for process modeling and monitoring assume no missing data. Therefore, the missing data problem reduces the representativeness of data samples and may result in uncomfortable statistical inference.

2.2. Robust statistical methods for data mining

Notwithstanding the inaccurate and incomplete process data collections, data scientists still should find some robust ways that can best represent process information with statistics or models. In other words, such robust statistics and robust statistical models should not be excessively affected in most cases by those unpredictable missing data and outliers. However, classic statistics including mean and variance are sensitive to outliers, while classic data mining models for multivariate statistical process modeling (MSPM) methods such as principal component analysis (PCA) and factor analysis (FA) actually will take the basic assumption of clean process data. Therefore, these methods usually sink into poor performances for bad quality process data. Generally speaking, robust statistical data mining models should be resistant to assumption errors and will provide innocent estimations and predictions. Take the outliers as an example, outliers have side effects for multivariate statistical analysis which may result in model misspecification. Needless to say, those incorrectly specified model could be ill-suited for further industrial process monitoring applications. As the alternatives, outlier insensitive statistics like median will keep making well description on the main population. Likewise, those outlier-resistant methods with self-contained schemes will filter or down-weight process outliers so as to make the stable and reliable estimations or predictions for process analysis. The missing data situation is alike the outlier case except for that one aims at disposing those missing elements by deletion, imputation or inference. As a consequence, we will call statistical data mining methods as robust if they can provide reasonable statistical inference when outliers/missing data exist.

2.3. Overview for this work

According to the above analysis, the main methodologies on robust data mining can be divided by two separate parts: robust mining for data preprocessing (robust preprocessing) and robust mining for statistical modeling (robust data mining). The former one takes a serious concern on dealing with outlier and missing data explicitly by cleaning as the preprocessing. In this way, the modeling can be conducted with some traditional data mining techniques (Ge et al., 2013). In this branch, those state-of-the-art data preprocessing methods will be highlighted and overviewed in an overall taxonomy including outlier detections and data normalization. Particularly, the data normalization problem will also be considered from the aspect of robust modeling. After then, attention will be fixed on the robust data mining which will directly comprehend those imperfect data with statistical techniques. For this subfield, we will start the robust mining issue in the framework of PCA, and numerous robust solutions will be discussed. Among them, the state-of-the-art robust developments of PCA will be formulated and discussed in both non-probabilistic and probabilistic frameworks. Differences and relations will be shown for both robust prototypes. Apart from the deep and systematic review on robust PCA, other widely applied robust data mining methods will also be surveyed, in light of different process characteristics. The Fig. 1 sketches the systematic overview for this article.

3. Data preprocessing

Industrial process data are often loosely controlled which contain outliers and missing data. In the meanwhile, variables often vary significantly in different scales and the normalization should be indispensable so as to adjust ranges of values to a notionally common scale. The common scale avoids the taking over issue for statistical analysis by those dominant variables with large ranges. Therefore, successful statistical analysis largely depends upon the deliberately data preprocessing. In this section, data preprocessing methods shall be discussed for dealing with missing data and outliers. It should be mentioned that the full data preprocessing should include data cleaning, normalization, time alignment, transformation, feature extraction and selection. Among them, this review will focus on data cleaning and normalization with robust perspectives.

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

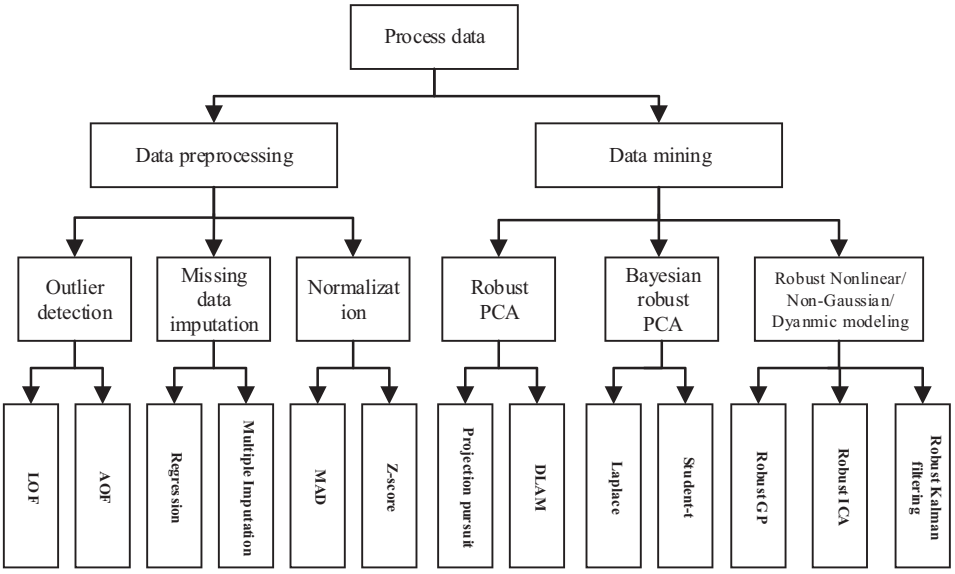


Fig. 1. Overview for this survey article. Each box is attached by 2–3 typical solutions in the bottom.

3.1. Data cleaning

In this context, univariate/multivariate outlier detection, missing data handling techniques will be focused so as to detect or correct those inaccurate/incomplete records.

3.1.1. Outlier detection

Outlier detection (or outlier mining) by statistical methods usually assumes that outliers belong to the low similarity/density regions so that outliers can be identified from the normal data clusters. In other words, the outlier detection task can be regarded as the searching and description of the normal data contours that are with high similarities or densities. As we have mentioned above, the definition of outliers largely belongs to the subjective matter and should depend on the industrial data set in practice. In literatures, although there have been a broad range of methods for outlier detection, it should be mentioned that there are no universally applicable methods. In this sense, choose of appropriate outlier detection method also relies on the specific data set and application. Without loss of generality, we will pay attention on those prevailing univariate and multivariate outlier detection methods.

3.1.1.1. Univariate outlier detection methods. Technically, univariate outlier detection methods usually rely on statistical judgments, which are based on the violation of specific assumptions on normal data distribution. Normal data are expected to be located in those regions with high density with respect to the defined density function. The most commonly used methods should be the three-sigma rule and the boxplot.

Three-sigma rule: The three-sigma rule is perhaps the most well-known one for outlier detection in data analytics. This rule is also known as the 68–95–99.7 rule to remember the percentage of data located within the 1, 2 and 3 standard deviations (sigmas) of the mean. For example, if we use the three-sigma range, and data x_n is sampled from the normal distribution $N(\mu, \sigma^2)$, then x_n can be identified as an outlier if $|x_n - \mu| > 3\sigma$. The most appealing merit for the three-sigma rule is simple and easy to be implemented. However, such rule cannot perform well for multivariate outliers as these outliers may not be well allocated in the three-sigma zone. Take the toy data in Fig. 2 as an illustration. In this example, none of the outliers (denoted as circles) can be detected by the three-sigma rule from either coordinate since all outliers are within the normal range.

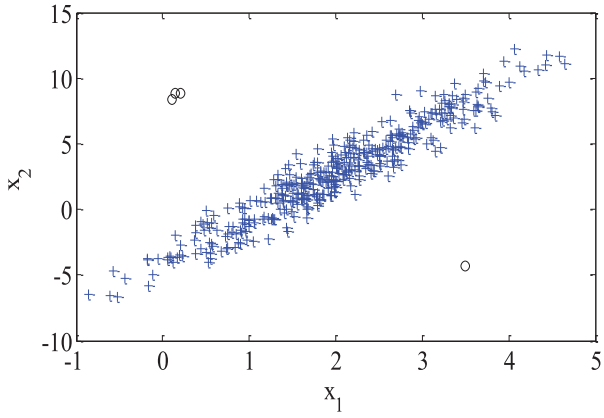


Fig. 2. Toy example with crosses as normal data and circles as outliers.

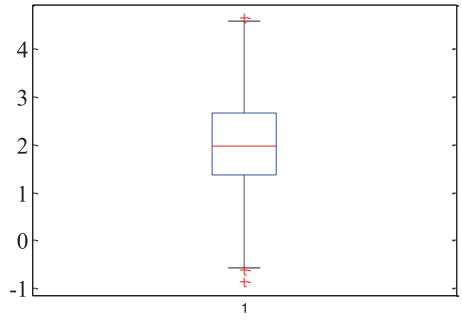


Fig. 3. Example of boxplot.

Boxplot: Boxplot is a graphical-based method for data visualization through upper/lower quartiles and extending whiskers. Those potential outliers are depicted with individual points. Take the case in Fig. 3 as an example, the up side of the box is the upper quartile Q_3 (or third quartile) splits off the highest 25% of data from the rest while the down side of the box is the lower quartile Q_1 (or first quartile) splits off the lowest 25%. The median $Q_2 = \frac{Q_1+Q_3}{2}$ (or second quartile) is the red line inside the

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

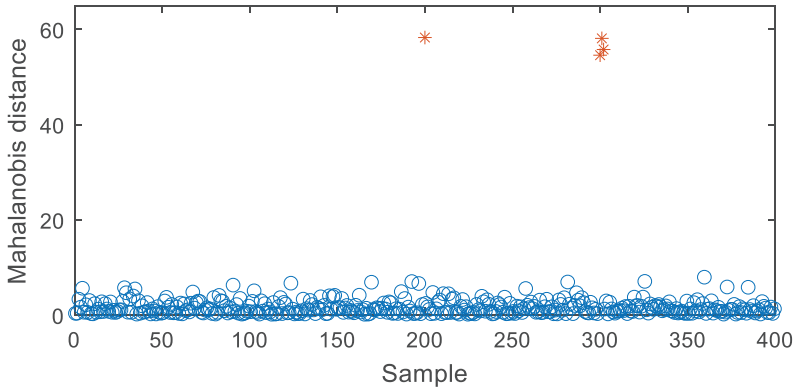


Fig. 4. Mahalanobis distances for data samples.

box. The interquartile range is then computed as $IQR = Q_3 - Q_1$. For symmetric distributions such as the Normal distribution, the whiskers of both sides will extend to the covering range of $1.5IQR$. Outliers are points located beyond the range of whiskers, namely $[Q_1 - 1.5IQR, Q_3 + 1.5IQR]$. One can speculate that the boxplot is convenient for both data visualization and outlier detection. However, like the three-sigma rule, the boxplot may not behave well for multivariate outliers.

Apart from the three-sigma rule and boxplot, other tests or criterion for univariate outlier detection can also be found, such as the Grubbs' test, the Chauvenet's criterion and the Peirce's criterion, one can refer to the relevant studies for detailed analysis (Ross, 2003).

3.1.1.2. Multivariate outlier detection methods. Due to the high complexity in nature, industrial process data are usually endowed with high dimensions, and the univariate approaches will become sluggish as outliers in multivariate conditions may be atypically scattered in the entire space, while in the same time they display normally behaviors for each single ordinate (see Fig. 4a for example). Therefore, the multivariate outlier detection methods will be considered. Those widely used approaches will include the distance-based method, density-based method and proximity-based method (Han, Pei, & Kamber, 2011).

Distance-based method. This strategy tries to convert multivariate outlier mining problems into the univariate domain by calculating Mahalanobis distances. Let $\mathbf{X} \in R^{N \times D}$ denotes the collected data matrix, N is the sample number and D is the variable number. Further let $\bar{\mathbf{x}} = (\bar{x}_1 \ \bar{x}_2 \ \cdots \ \bar{x}_D)$ represents the mean vector and $\mathbf{\Sigma}$ denotes the sample covariance, then the Mahalanobis distance for sample $\mathbf{x}_n$ can be calculated as

$$\mathrm{MD}(\mathbf{x}_n) = (\mathbf{x}_n - \bar{\mathbf{x}})\mathbf{\Sigma}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})^T \tag{1}$$

We will take the data in Fig. 2 as an illustration, the Mahalanobis distances for all data are shown in Fig. 4. One can judge that outliers (denoted with red stars) can be easily detected by using univariate techniques.

However, the Mahalonobis distance-based bound can only describe the contour of Gaussian cluster. To be worse, the distance metrics can be reckless for some nonlinear/non-Gaussian multivariate cases, such as the two-dimensional dataset scattered in Fig. 5a. This dataset is composed of two separate dense Gaussian clusters and one sparse banana-shaped cluster, outliers are spreading in the middle section with red stars. The corresponding Mahalanobis distances are shown in Fig. 5b, from which one can hardly distinguish potential outliers.

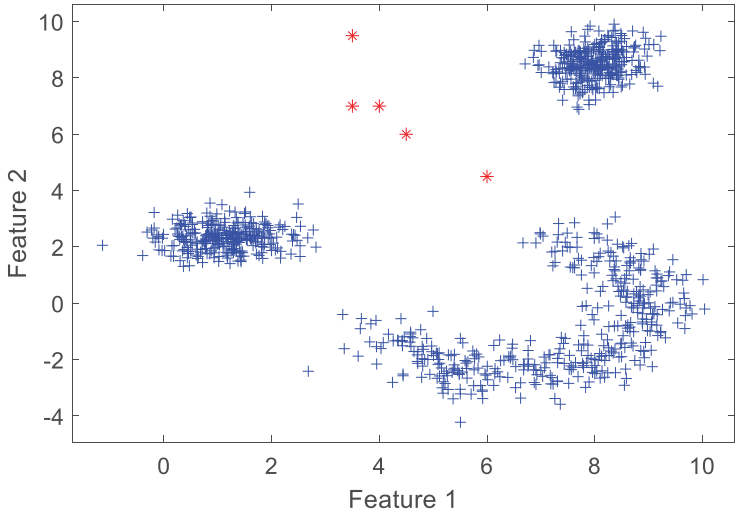
Generally speaking, in data mining literatures, multivariate outlier detection problem can be regarded as supervised or unsupervised tasks depending on whether expert knowledge can be attached for outlier labels (Hodge & Austin, 2004). However, as process data are usually with high dimensions and large volumes, those outlier labels can be hardly derived. In this sense, the outlier detection in our scope is commonly referred to as the unsupervised data mining. For high dimensional data, outlier detection is akin to finding needles in a haystack (Agyemang, Barker, & Alhajj, 2006). Generally, among various multivariate outlier mining explorations, one can find density-based methods and proximity-based methods which work well for these circumstances. For density-based methods, one first makes the proper estimation on potential data distribution and then outliers are rejected from those low-density distribution regions. The proximity-based methods consider specific metrics as the measure of similarity of data objects. Outliers are detected from observations with low similarities with the adjacent neighbors.

Density-based methods. The density-based methods can be categorized as parametric methods and non-parametric methods. For non-parametric methods, the widely applied outlier mining techniques should involve the Parzen window method and *C*-means method, while the commonly used parametric method is the Gaussian mixture model (GMM).

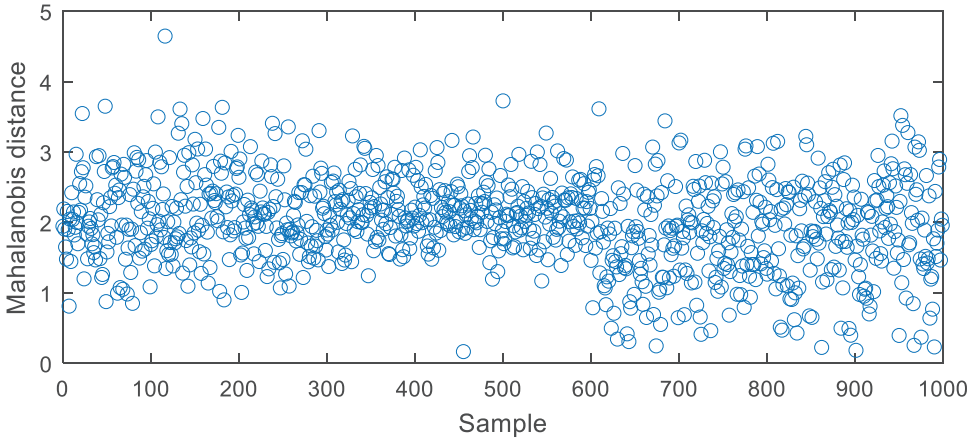
The Parzen window: The Parzen window directly estimates the distribution of data given the window function as well as the window width (Kwak & Choi, 2002). The window function defines the unit hypercube centered as the origin while the width determines the smoothness of the estimated distribution. Once the distribution has been obtained, the low-density region can be rejected as outliers. It should be mentioned that such method is very sensitive to the choice for the initial cell volume. To make the method more effective, one usually needs to adjust the window width by trial and error for a specific data. Besides, the Parzen window method requires the sufficient amount of data so as to make the appropriate statistical investigation. Despite of these defects, the Parzen window method is very simple and is one of the most commonly used approaches for data preprocessing without the sufficient a priori knowledge on data distribution (Hodge & Austin, 2004).

C-means method: Basically, the *C*-means method is used for unsupervised clustering, and hence is widely applied as the standard preprocessing method for statistical analysis. The *C*-means method uses distance metrics as the judgment of sample similarity (Duda, Hart, & Stork, 2012). The cluster centers are generated with initial sample points. The rest samples are clustered by computing the distances with all cluster centers and assigned into the nearest cluster. The cluster centers are then re-computed and repeatedly adjusted in the next new round for all samples until the location parameters converge. The cluster number of the *C*-means is a user-defined parameter that should be previously initialized. Once the clusters have been finally specified, those clusters which

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003



(a)



(b)

Fig. 5. (a) The scattered data and the outliers are plotted with red stars (b) The corresponding Mahalanobis distances. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

account for small proportions of samples can be discarded as outliers (Hodge & Austin, 2004).

Gaussian mixture model method: For multivariate data with strong non-Gaussian behaviors, one can consider the modeling with a mixture of Gaussian clusters (Hodge & Austin, 2004). Given the mixture component number as the user-defined parameter, the model parameters like component weights as well as mean and covariance can be automatically learned via the expectation maximization (EM) algorithm (Moon, 1996). Correspondingly, the component weights directly indicate the importance of local data clusters. Therefore, those local components with low weight values are low density regions which in turns can be labeled as outlier sections and should be discarded. Theoretically, the GMM is suitable for fitting any non-Gaussian continuous probability density to an arbitrary accuracy, given enough components (Härdle & Simar, 2007). Such advantage makes it flexible and appropriate for various industrial cases. However, the effectiveness of GMM is susceptive to both the user-specified component number and the initial values of parameters. From another point of view, the data fitting performances by GMM may deteriorate for those very high dimensional data.

Proximity-based methods. Different from statistical based methods which are focused on global data distribution estimations, the proximity-based methods are aiming at measuring the proximity of an instantiation to its nearest neighbors with high similarities. Outliers are defined for those samples show significant deviations by proximity measures in the data set. In previous studies, three main categories of proximity-based methods can be found, namely the k-nearest neighbor (k-NN) method, the local outlier factor (LOF) and the angle-based outlier factor (ABOF). The k-NN directly uses the local distance measure to qualify the local proximity among data instantiations to their nearest neighbors while the LOF further defines the local density for each instantiation. The ABOF is designed for high dimension data and uses angle for measuring the proximity from the perspective of vector space directions.

k-nearest neighbor method: The k-NN is known as the instance-based learning or lazy-learning method (Zhang & Zhou, 2007). Like Parzon window method, the k-NN also has

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

a user-specified parameter for the definition of nearest neighbor numbers. For each sample, one first defines the metric of distance by Mahalanobis distance or Euclidean distance with all the other exemplars to get the k nearest neighbors. There are several ways to detect outliers. A simple and intuitive way is to detect outliers by scanning the kth neighbor distance for all samples and then reject a proportion of observations with the largest distances. In this way, the multivariate outlier detection problem will be effectively converted as the univariate outlier detection problem.

Local outlier factor method: Most data mining methods take the outlier detection as a binary discrimination issue, regardless of the degree of an instantiation being as an outlier. As an improvement to such flaw, The local outlier factor (LOF) method, takes the further step by considering the outlying degree through the definition of local reachability density (Breunig, Kriegel, Ng, & Sander, 2000). By calculating the local densities with nearest neighbors, those samples with high local densities can be retained as normal data while outliers can be usually excluded by checking those with extraordinarily low local density items. The possibility of a sample being an outlier is finally reported by the local outlier factor. Data samples surrounding by dense mass are assigned with LOF values of approximately zero, while those atypical data samples are assigned with higher LOF values which also indicate the nonhomogeneous local consistencies. Benefit from the local density, the LOF mining technique can be robust to those dataset involved by both sparse and dense clusters. However, since the outlying factors for data points are essentially developed based on the nearest neighbors, the LOF performance is also affected by the determination of the number of nearest neighbors. As mentioned in the previous study, the LOF does not change monotonically according to the defined number of nearest neighbors (Breunig et al., 2000). Therefore, heuristic knowledge should be preferred so as to make the reasonable outlier isolation.

Angle-based outlier factor method: The aforementioned outlier mining methods are based upon the investigation of distances/densities in the full-dimensional Euclidean or Mahalanobis data space. Therefore, these approaches can be stuck into the 'curse of dimensionality' for high dimensional data as the relative contrast among data points converges to zeros by increasing dimensionalities (Kriegel & Zimek, 2008). As an alternative, the angle-based method can alleviate such dilemma by introducing the vector angle concept which not only uses the distance between points in the data space but also considers the directions of distance vectors (Kriegel & Zimek, 2008). Specifically, the angle-based outlier factor of an instantiation refers to the variance over the angles between one vector instantiation to all the other vectors weighted by the distances of all the involved data points. As the result, the angle-based outlier factor actually measures the relative divergence in directions of data vector objects from one to another. In this way, the utilization of vector angle can effectively alleviate the 'curse of dimensionality' by simply scanning the spectrum of angles. Accordingly, one can speculate that normal data show more broad variations while those outliers which lie outside clusters show rather small spectrums.

3.1.1.3. Remarks for multivariate outlier detection methods. Before we can make further discussions, an illustrative example will be given first and the outlier detection results for nonlinear data in Fig. 5a by all the aforementioned methods are shown in Fig. 6. Notice that the significant level threshold for all methods are set as 0.01 and normal boundaries are plotted with black solid contours.

One can speculate that in contrast with Mahalanobis distance based method, those density-based and proximity-based methods exhibit comparable performances for the complicate outlier mining situation. However, it is worthwhile to make some further remarks. Generally speaking, the non-parametric statistical methods make little assumptions on data and hence could be universally applicable on various situations. Compared with C-means approach, the GMM can be regarded as an alternative for probabilistic modeling of non-spherical clusters. However, it is unclear how GMM will perform on the general outlier detection problem. If one has insufficiently points assigned to per mixture, the estimation of the covariance matrices will become difficult, and the algorithm may tend to diverge and fail for training while one comes to the fitting solutions with infinite likelihood. There are also other concerns for the above outlier detection methods. On the one hand, as mentioned above, the main concern for most statistical methods should be how to overcome the problem of 'curse of dimensionality'. In high dimensional dataset, the data points are spreading through the entire large spatial volume with nonlinear local manifolds, which makes it harder for normal bound definition. On the other hand, most outlier detection methods are very time-consuming due to the fact that they are founded on the calculation of distances between all data records which are on the computing loading basis of $O(n^2)$. For example, the C-means, k-NN and LOF compute the k nearest neighbors should be in $O(kn)$ while the angles in angle-based method are calculated by all data pairs should be in $O(n^3)$ which should be really inefficient for large dataset. The last remark is made on the model and parameter selection. As can be inferred from Fig. 6, almost all models require the selection of intrinsic model parameters so as to achieve a better result. For example, one can also refer that in k-NN and LOF, different numbers for neighbors are required to search the reasonable boundary. From this point, one can determine that the outlier detection implementations are computational intensive mining tasks, even it is only done once.

3.1.2. Missing data preprocessing

The missing data phenomena should be another data imperfection issue which frequently appear in real experimental data. Generally, one can come across two levels for data missing: the unit level and the item level. The unit level non-response refers to the data missing due to plant shutdown or the malfunction of sensors. In that case, no data information can be recorded from the signal sources. Although such data missing can be widely encountered, the void entries in the entire unit makes up no values for information processing and empirical analysis. Therefore, this work will be focused on those statistically meaningful missing situations from the item level non-response where some variables are partly missing. Mostly, the item level missing data arise from the incomplete collection of variable entries by the intermittent data acquisition processes. In this case, the large amount of irregularly absent data collections may lead to those biased estimations on parameters and the accompanied weak model extrapolations. However, those invalid manipulations like improper case-wise deletions or wrongly replacement of missing data may also result in problematic parameter estimations. By this token, those principled methods in statistical inference should be utilized.

3.1.2.1. Three issues. To break through the limitations of unprincipled handling manners for missing data, three specific issues should be first elaborated: the missing data proportion, the missing data patterns and the missing data mechanisms.

Missing data proportion: The missing data proportion takes the first glimpse on empirical data before taking the valid measurements. Although there are no strict criterion, it is suggested that those extremely low missing rates (like only a couple of missing entries) are likely to makes no significant interference for inference, as the representative of the population can be still remained from the majority. However, the contamination covering 5%∼10% missing entries will typically result in significant biased inferences (Dong & Peng, 2013). The first sight of view on data can

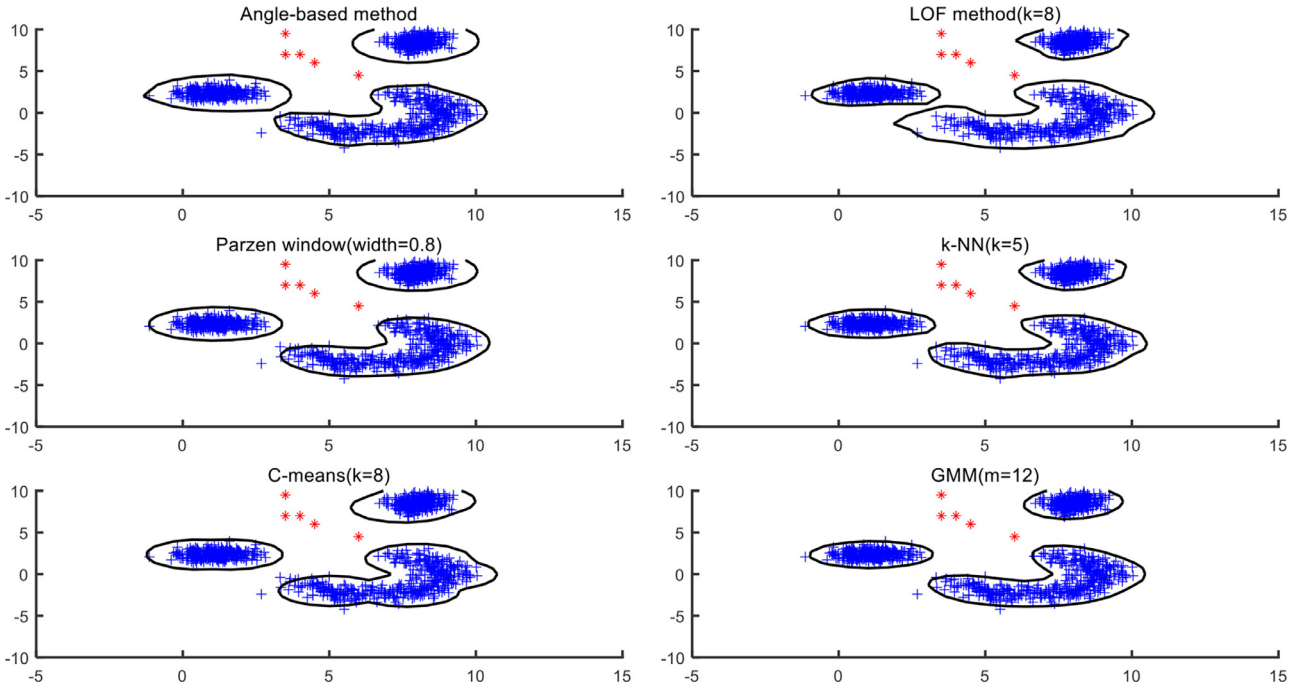


Fig. 6. Outlier detection results with various methods.

help reduce those laborious efforts by avoiding unnecessary data repairs. However, such judgment with a glance fails to consider those underlying relations regarding missing values and those observed ones. In addition of this global inspection, the missing data patterns and mechanisms make further investigations on why and how missing data may be generated, and hence have more impacts on the successful statistical inference.

Missing data patterns: There are two common missing data patterns, namely the multi-rate pattern and the general pattern. The multi-rate pattern can be frequently encountered when the data scanning of sensors are not in the same steps, or the sampling rate of a sensor will change among working phases due to different critical levels. In contrast, the general missing pattern may be stemming from the removal of outliers or the intermittent network transmissions with randomly delays on packet losses. Based on the missing data patterns, the missing data mechanisms can be further discussed.

Missing data mechanisms: The missing data mechanisms provide a probabilistic framework for elegant assumptions on missing data relations. From the practical point of view, although it is difficult to testify the integrity and correctness of such hypothesis, a fully understanding of missingness is absolutely necessary for the proper designing and application of statistical analysis methods (Graham, 2009). In literatures, three common types of missingness can be found by assuming different probabilistic relations between missing part and observed part, namely missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR) (Schafer & Graham, 2002). The MCAR mechanism assumes that the missing data should be independent of both observed part and the unobserved part. The MAR relaxes the MCAR by assuming that the missing part is only related to the observed part and is widely accepted in real application for situations such as multi-rate sampling and randomly transmission errors. The NMAR takes the weak assumption that the missing data is related to both the observed part and missing part. Since the NMAR can be hardly handled for statistical inference, we will focus on the MCAR and MAR cases. Accordingly, the missing data can be inferred from the observed part. Mention that these two cases should be widely applicable for industrial practices. In order to deal with missing entries for logic preprocessing, there are two common strategies: listwise deletion and imputation. The listwise deletion removes all data for a case that has one or more missing values. However, the deletion is always the last choice as one may suffer from significant information loss. Consider a process with 30 variables, a total of 3% missing data values are randomly missing and also spreading uniformly across the data space. Then the deletion incurs the reduction of about 60% data samples ($1–0.97^{30} = 0.5989$). In contrast, data imputations have the comparatively desirable merit with no data sacrifice. The following part will come to the state-of-the-art data imputation methods.

3.1.2.2. Data imputation. When a variable provides partially observable information, the imputation attempts to replace the missing items with plausible values from the variable itself or other variables. To impute with the partial information, there are several methods like mean substitution, hot-deck substitution, regression substitution, conditional distribution based substitution and multiple imputation.

Mean substitution: The mean substitution considers the replacement of missing values by the average of the variable. For the case of multimode process imputation, the substitution is taken by from the mean value of the within-mode distribution. The mean substitution will provide efficient and unbiased estimates for locations in those situations close to MCAR. However, the mean substitution has side effects like the distortions of variances and correlations. One can judge that such distortions will become problematic when the variance/correlation estimates are the focus of study for most statistical methods like PCA and PLS. Therefore, the mean substitution will not be recommended in most cases.

Hot-deck substitution: To preserve the distribution while imputation, the hot-deck substitution replace one missing entry at a time with the available value from a similar respondent in the same study. By doing so, it derives the better variance estimation compared with the mean imputation counterpart. In fact, the hot-deck substitution is one of the most widely used imputation methods. However, the problem will arise for this approach when sev-

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

eral missing records occur together on the data file. Another problem is that the imputed values may not be necessarily consistent for the respondent values.

Regression substitution: The regression substitution tries to plunge the missing entries with a regression estimation from other correlated auxiliary variables. Such imputation is also called conditional mean imputation. Through substitution, the imputed values are only as good as the regression model used to predict them. Therefore, this method can distort the analyses of variances and correlations, as the regression will overstate the strength of the casual relationship (the R^2 statistics are ones for all missing entries with the auxiliary variables). Another drawback is that the regression may sometimes yield improbable results which can be invalid from the reasonable domains.

Conditional distribution based substitution: The conditional distribution substitution imputes through the random draw of missing entries from the conditional distribution of uncertainties. A commonly used example is in linear regression cases when a random error will be added to the estimation followed by the normal distribution, the distribution is with zero mean and the variance is assigned by the residual mean square (Schafer & Graham, 2002). For this kind of imputation, one commonly has to define the explicit conditional distribution of missing variable given those observed variables so as to make a better substitution. Thereafter, this substitution will alleviate the problem on distortion of distributions. However, the main problem for this method is how to infer the proper distributions with unknown parameters. In some cases, the distribution can be quite complicated, which makes the method cumbersome.

Multiple imputation: The multiple imputation is a relatively modern principled imputation method for data preprocessing (Rubin, 2004). The multiple imputation handles missing data in three steps: (1) Impute those missing data multiple times so as to generate several complete data sets, the regression and MCMC method shall be used for this step; (2) Analyzes each data set using a standard statistical procedure; (3) The results are combined using simple rules to yield estimates, standard errors, and p-values that formally incorporate missing-data uncertainty. The multiple imputation may achieves better imputation than the single imputation in most cases. However, the problem is also obvious, as one has to impute several times so as to achieve the good statistical inference, which is rather computationally intensive.

Mention that the hot-deck substitution, regression substitution, conditional distribution based substitution and multiple imputation are all designed for working in MAR situations. It should also be noticed that the goal of imputation is to alleviate the impairment of missing values to statistical inference rather than the recovery the true data. One can also infer from the above analysis that the methods of imputation are less-than-ideal, as one often requires more computational loadings for the better imputations. After all, it should take some care in the trade-off selection of the imputation methods so that the efficiency can be ensured for data preprocessing while the distribution distortions can be reasonably avoided for further statistical inference.

3.2. Normalization

Normalization is a procedure used for adjusting the ranges of independent variables to a common regime, and is generally performed with proper data transformation operations. There are several such operations, namely rescaling, standardizing and normalizing. We first come to the explicit definitions.

3.2.1. Some definitions

Before the explicit techniques, the definitions are first given for distinguishing different operations. According to the definitions given by Koskela (2003), one has

Rescaling: Add or subtract a constant to a vector and then multiply or divide by a constant, the goal is to change the units of measurement. One example is to convert a temperature from Celsius to Fahrenheit.

Standardizing: Subtract a vector by a measure of location and then divide it by a measure of scale. For example, by subtracting the mean and dividing with the standard deviation, one obtains each random variable with the zero-mean and unit variance common range.

Normalizing: Divide by a norm of the vector to make the length of the vector equal to one. In literatures, the normalizing can be also regarded as the rescaling.

Based on the definitions, the normalization can be conducted from several ways. If one has already conducted the irregular data cleaning, those traditional normalization methods can be readily performed; otherwise, one has to consider those robust alternatives. Notice that the robustness here particularly refer to the insensitiveness of normalization against outliers. In this survey, we will revisit some state-of-the-art normalizations and then make some discussions on robustness and efficiency.

3.2.2. Normalization methods

Specifically, the common normalization techniques are min-max scaling, decimal scaling and z-score, while for robust alternatives, one has the median absolute deviation (MAD) based scaling and tanh estimator (Jain, Nandakumar, & Ross, 2005).

Min-max scaling: The Min-max scaling is used to bring all values into the range [0, 1] and typically is done via the following transformation:

$$X_i' = \frac{X_i - \max(X_i)}{\max(X_i) - \min(X_i)} \tag{2}$$

where the minimum and maximum values are estimated from X_i. In some cases, the above range can be generalized to be between any arbitrary user-defined points a and b $(b > a)$ using the following equation:

$$X_i' = a + \frac{(X_i - \min(X_i))(b-a)}{\max(X_i) - \min(X_i)} \tag{3}$$

In that way, one will end up with smaller standard deviations. Despite the fact that the min-max can suppress the effect of outliers in some extent, one can easily speculate that the normalized outputs highly rest with the maximum range of potential outliers.

Decimal scaling: The decimal scaling makes the normalization by moving the decimal point of values from attribute which can be defined as

$$X_i' = \frac{X_i}{10^j} \tag{4}$$

where j is the smallest integer such that $\max(|X_i'|) \leq 1$. The decimal scaling can be applied for situations when the variable attribute is on the logarithmic scale. The number of decimal points moved should depend on the maximum absolute value of attribute. Identical to the min-max, such method is also sensitive to outliers.

z-score: The z-score, also called z-value, can be defined as

$$X_i' = \frac{X_i - \mu_i}{\sigma_i} \tag{5}$$

where μ_i is the mean and σ_i is standard deviation. It should be the most frequently used method in process monitoring for the sake of making transformations on variations into the unit standard normal deviation. Mention that such method is defined without assumptions of normality, which also makes it applicable for industrial processes with multiple operating modes. From the scope of

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

robust normalization, the z-score can be only performed for clean data as the population mean and standard deviation are sensitive to outliers.

MAD scaling: Analogous to z-score transformation, the MAD based scaling can be defined as

$$X_i' = \frac{X_i - median(X_i)}{MAD(X_i)} \tag{6}$$

where $median(X_i)$ estimates the median of variable and $MAD(X_i) = median(|X_i - median(X_i)|)$. Since the median and MAD statistics are both insensitive to outliers, the MAD based scaling can be regarded as the robust normalization. The MAD method is simple and usually very effective in most occasions. The only deficiency of MAD is for analyzing those large dataset cases, when the estimation on medians will become relatively inefficient.

The tanh-estimator: The tanh-estimator is also a robust normalization with scaling and transformation, which can be defined as

$$X_i' = \frac{1}{2}\left\{\tanh\left(0.01\left(\frac{X_i - \mu_{GHi}}{\sigma_{GHi}}\right)\right) + 1\right\} \tag{7}$$

where μ_{GHi} and σ_{GHi} are the mean and standard deviation for the distribution for Hampel estimators (Hampel, Ronchetti, Rousseeuw, & Stahel, 2011):

$$\psi(u) = \begin{cases} u & 0 \\ a\sin(u) & a \le |u| < b \\ a\sin(u)((c-|u|)/(c-b)) & b \le |u| < c \\ 0 & c \le |u| \end{cases} \tag{8}$$

Through the scaling with Hampel estimators, the influence of outliers at the tails of distribution will be reduced. The tanh estimator is of both high efficiency and robustness. However, a problem for this method is that the Hampel parameters should be carefully determined for definition of tails. Apparently, such procedure should be highly data dependent in practice.

As the result, one can infer that the normalization should be important and is nontrivial work. On the one hand, for those efficient methods like min-max, decimal scaling and z-score, the robustness cannot be approved. On the other hand, for robust methods, one has to consider the cost of double median estimation in MAD or making extra efforts for the delicately parameter selection in tanh-estimators. Nevertheless, the robust normalization is recommended, as it is always worth to make a good preparation for the reasonable statistical modeling.

4. Process modeling via robust data mining

One can judge from the previous analysis that the data cleaning can be arduous and inefficient. On the one hand, the procedure of outlier detection is time consuming, especially for voluminous industrial data. On the other hand, the discard of outliers will add up to the analysis difficulties of missing data. As the goals of both detection and imputation are to bootstrap the valid statistical analysis, the direct robust modeling policy can be welcomed so as to get rid of the above dilemma.

For modeling those high-dimensional industrial processes, the principal component analysis has been broadly accepted. From another point of view, most robust techniques designed for PCA can also be potentially extended into other techniques. Therefore, this section will be first focused on introducing and comparing the different robust mechanisms on PCA (RPCA). After then, we will discuss the robust modeling techniques with the further consideration of multiple industrial process characteristics so as to make the enhanced monitoring performance.

4.1. From PCA to RPCA – an overview

The PCA is indubitably the most widely used statistical modeling tool for process data analysis and dimensionality reduction. Specifically, the basic structure can be given as

$$\mathbf{X} = \mathbf{L} + \mathbf{E} \tag{9}$$

where $\mathbf{L}$ is a low rank matrix and $\mathbf{E}$ is the noise. The PCA seeks for the explanations of the variance–covariance structure contained in the data with the low lank representation by a set of orthogonal principal components or PCs through orthogonal projection $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_d]$, and the noise $\mathbf{E}$ has only been considered implicitly. The PCs can be extracted according to the data variability and the first PC contains the largest proportion of data variance. To make that, the following objective function should be maximized:

$$\begin{aligned} &\max_{\mathbf{p}_i} \ \mathbf{w}_i^T \ \mathbf{X}^T \ \mathbf{X} \ \mathbf{w}_i \\ &s.t. \ \mathbf{w}_i^T \mathbf{w}_i = 1 \end{aligned} \tag{10}$$

There are at least two general ways to determine the PCs: 1) Eigenvalue decomposition of the covariance $\mathbf{X}^T\mathbf{X} = \mathbf{W}\mathbf{\Sigma}\mathbf{W}^T$ and 2) Direct singular value decomposition (SVD) of data $\mathbf{X} = \mathbf{D}\mathbf{\Lambda}\mathbf{W}^T$, usually after z-score normalization for each attribute. Notice that the above variance maximization is equivalent to the minimization of reconstruction errors in a least-squares sense, which is

$$\begin{aligned} &\min_{\mathbf{P}} \left\|\mathbf{X} - \mathbf{X}\mathbf{W}\mathbf{W}^T\right\|_2^2 \\ &s.t. \mathbf{W}^T\mathbf{W} = I \end{aligned} \tag{11}$$

where $\|.\|_2$ is the L_2 norm. For industrial data that contain noise, the conventional PCA is only an approximation. It can be inferred from the above objective that the traditional PCA is effective in recovering the low-rank structure only when the implicit noises are Gaussian. If the noises are non-Gaussian and strong, PCA will deteriorate due to the lack of mechanism to account for noises and outliers. A single grossly corrupted entry in observations could render the estimated subspace arbitrarily far away from the expected ground truth (Candès, Li, Ma, & Wright, 2011). Besides, the traditional PCA cannot cope with missing data.

To enhance the robustness, robust PCA should be considered. Mention that many robust PCA methods were first developed and investigated for signal, image and video processing applications (Bouwmans, Aybat, & Zahzah, 2016; Bouwmans, Javed, Zhang, Lin, & Otazo, 2018; Javed et al., 2016; Mardani, Mateos, & Giannakis, 2013). Recently, some of the robust approaches have been successfully applied in industrial process monitoring. This review will first summarize the related works for various robustifications of PCA with the deterministic formulation. Following that, the robust problem will be further discussed in the probabilistic framework called Bayesian Robust PCA (BRPCA).

4.2. RPCA

4.2.1. RPCA via outlier robust methods

To make the robust statistical analysis, there are three common strategies: the PCA with robust covariance estimator, robust PCA with projection pursuit (PP) and the weighted-strategy (weighted PCA). The first class considers PCA after the robust estimations on covariance statistics. The second strategy finds projections of the data onto those directions that will maximize a certain criterion. The third way seeks for down-weighting those outliers and the goal is commonly set by minimizing the weighted squared reconstruction error.

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

4.2.1.1. Robust covariance estimator methods. A straightforward way to robustify PCA is to consider the incorporation of robust statistic estimators for covariance(or correlation) (Croux & Haesbroeck, 2000; Devlin, Gnanadesikan, & Kettenring, 1981). Traditional robust estimators are M estimators or L (i.e. trimmed) estimators which are designed for low dimensional data and may break down for analyzing high dimension processes (Campbell, 1980; Maronna, 2005). In this sense, one may consider the minimum volume ellipsoid (MVE) estimator and minimum covariance determinant (MCD) estimator. The MVE estimator looks for the minimal volume ellipsoid covering at least half the data points (Van Aelst & Rousseeuw, 2009). However, this robust estimator falls into the low efficiency of convergence. As the alternative, the minimum covariance determinant (MCD) estimator is a highly efficient robust estimator of multivariate location and scatter (Hubert & Debruyne, 2010). The objective is to find h observations (out of N) whose covariance matrix has the lowest determinant. In cases for analyzing large dataset with high-dimensionalities, a computationally efficient version of MCD has been developed called the FAST-MCD algorithm (Rousseeuw & Driessen, 1999). Based on these merits, work by Verboven and Hubert (2005) developed robust methods with the MCD based covariance estimation before performing the normal PCA. Alkan, Atakan, and Alkan (2015) compared the robust MCD-PCA, EMPCA and MI-PCA for dealing with outliers and missing data, and outliers were imputed along with the missing data. The MCD only works well for cases when the considered observation number is larger than the variable number ($h > D$). For quality monitoring/prediction cases in chemometrics, there are hundreds or thousands of variables which make the MCD methods cumbersome. To cope with this issue, Hubert, Rousseeuw, and Vanden Branden (2005) proposed the robust PCA which has combined the projection pursuit with robust scatter matrix estimation. By doing so, the projection pursuit technique was used to project the high-dimensional observations into low-dimension space and thus the robust estimator can be readily applied.

4.2.1.2. Projection pursuit methods. The projection pursuit idea for PCA was proposed by Huber (1985). To deal with the high-dimensional data, the PP searched for the low-dimensional projections that could maximize the objective function called projection index. For instance, based on optimizing the target of Eq. (10), the standard PCA-PP took the variance as the projection index. In this sense, the classic PCA can be solved from the perspective of projection pursuit. However, one should note that there are some calculation distinctions between PCA-PP and the classical PCA. Due to the utilization of the robust projective index, some optimization strategies should be considered and the set of components are usually estimated in a sequential manner. Nevertheless, due to the essential equivalence with PCA formulation, the standard PCA-PP may also yield to unreliable projection results for outlying data sets. In the light of this issue, the first robust PCA with PP was developed by Li and Chen (1985), where they used a robust measure for variance. Later on, in the work by Xie et al. (1993), the generalized simulated annealing algorithm was carried out for the global optimization. Although the robustness in both robust PP methods have been improved, the new procedures also have the drawback of more computer time. For that problem, Croux and Ruiz-Gazen introduced a fast algorithm (called CR algorithm) which was also easy to be implemented (Croux & Ruiz-Gazen, 1996). The CR algorithm was more attractive but could result in numerical accuracy issues for high dimensions. To this end, numerically stable improvements were further studied by Hubert, Rousseeuw, and Verboven (2002) and Croux and Ruiz-Gazen (2005) In another work by Croux, Filzmoser, and Oliveira (2007), a new algorithm called GRID was proposed for estimating the robust projections and the authors showed that it could outperform the CR counterpart.

4.2.1.3. The weighted methods. For the weighted strategy, the robustness will be enhanced by alleviating the statistical impacts of outliers through down-weighting. There are two types of weighted matrix. The first one is to take the weighting factor for each sample. For example, Zhan and Yin (2011) proposed the iterative weighted PCA (IWPCA). Distinct treatments have been imposed on clean data and noisy data by giving them different alignment weights. The similar way can be found in the spherical PCA (sPCA) method (Stanimirova, Daszykowski, & Walczak, 2007). In sPCA, the data were firstly centered with robust statistics and then were projected on a sphere. Through such projection, outliers far from the data majority will actually receive small weights. Based on that, the robust principal components can be derived based on the covariance of these projected data points. For missing values, the EM framework can be commonly used (Stanimirova et al., 2007). In the EM framework, the algorithm iterates between the spherical PCA construction and then the missing elements refilling with their predicted values according to the currently constructed PCA model. Despite the robustness for outliers and missing data, the sPCA has bad biased properties (Serneels & Verdonck, 2008). It should be mentioned that the weighting factors consider the sample-wise down-weighting. In other words, the weighted matrix is a diagonal one and the entire sample would be treated as outliers as long as one feature is locating outside the common range.

To deal with this issue, the second strategy makes the generalization on weighted matrix and an individual weight will be assigned to each feature in each sample. As a result, the entire sample will not be regarded as the potential outlier and only those outlier features would be assigned with less weights. Several pieces of works can be found on this topic. Delchambre (2014) designed the weighted SVD procedure which could allow one to retrieve a given number of orthogonal PCs amongst those most meaningful ones from weighted outliers/missing data entries. In another work, Skočaj et al. (2007) studied the weighted PCA algorithms for different types of weights from the spatial and temporal domains. They applied their methods for the sake of a reliable analytic system dealing with noise and occlusions for visual learning and recognition.

In should be mentioned that the weighted strategy is a simple and intuitive scheme for making robust PCA. However, it turns out that the most difficult challenge is the efficiency justification for weighted estimations on large and/or high-dimensional industrial datasets, as one has to figure out the weighting factor for each data entry. Besides, the weighting scheme should be carefully considered as it may also distort some original normal data so as to diminish those outliers' influences.

4.2.2. RPCA via decomposition into low-rank plus additive matrices (DLAM)

The second profile of approaches for robustifying PCA constructs the robust objective function to separate the outliers with a sparse corruption unit (Bouwmans & Zahzah, 2014; Bouwmans, Sobral, Javed, Jung, & Zahzah, 2017). The basic decomposition structure can be defined as

$$\mathbf{X} = \sum_{k=1}^{K} \mathbf{M}_k = \mathbf{L} + \mathbf{S} + \mathbf{E} \tag{12}$$

where $\mathbf{M}_1 = \mathbf{L} \in R^{N \times D}$ is the low-rank matrix with rank $r \leq \min(D, N)$, $\mathbf{M}_2 = \mathbf{S} \in R^{N \times D}$ is the sparse matrix and $\mathbf{M}_3 = \mathbf{E} \in R^{N \times D}$ is the noise matrix. Notice that in the case of $K = 1$, the above decomposition has actually defined the basic formulation of conventional PCA as one only need to define a low-rank matrix, while the noise term will be implicitly considered by selecting projections with largest variances. In the cases of $K = 2$ and 3, the above decompositions will explicitly consider the constraint on sparse and/or noise

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

terms. For cases when $K=2$, the class of methods can be called RPCA with principal component pursuit (PCP); while for $K=3$, robust methods are called the RPCA with stable principal component pursuit (SPCP). Furthermore, the general minimization formulation with Eq. (12) can be written as follows

$$\min_{\mathbf{M}_k} \sum_{k=1}^{K} \lambda_k f_k(\mathbf{M}_k) \quad s.t.\ C_k \tag{13}$$

where λ_k are the regularization parameters, the $f_k(.)$s are the loss functions defined for low rank, sparse and noise matrices. The constraints C_k are also defined on each term with the norms. In the following subsections, both the PCP and SPCP will be compared.

4.2.2.1. The PCP case. As the pioneer work, the PCP framework has been well defined in by Candès et al. (2011) and Chandrasekaran, Sanghavi, Parrilo, and Willsky (2011), and the optimization problem with an objective function and constraint can be given as:

$$\min_{\mathbf{L},\mathbf{S}} \ \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \quad s.t.\ \mathbf{X}-\mathbf{L}-\mathbf{S}=\mathbf{0} \tag{14}$$

where $\|\cdot\|_*$ and $\|.\|_1$ denote the nuclear norm and L_1 norm respectively. Essentially, the nuclear norm reflects the low-frequency information and the L_1 norm explores those high-frequency part containing noise, outliers and missing data. With those irregular observations, it has been proved that the low rank and sparse matrices can be exactly recovered under rather weak conditions (Candès et al., 2011). For solving the above problem, an augmented Lagrange multiplier (ALM) algorithm can be applied (Lin, Chen, & Ma, 2010). Notice that the ALM is an efficient and scalable algorithm and the computation cost is not so much higher than the classical PCA. Actually, unlike other robust methods for PCA, the most appealing advantage of RPCA-PCP lies in the fact that it will yield a polynomial-time algorithm with strong performance guarantees in various conditions (Candès et al., 2011). Due to the high effectiveness, the RPCA-PCP methods are highly focused on image analysis researches for monitoring applications like video surveillance (Bouwmans & Zahzah, 2014). Furthermore, several further improvements have been made with different optimization norms and also changed solvers with different levels of complexity. For example, there are linearized solvers (Yang & Yuan, 2013), fast solvers (Mu, Dong, Yuan, & Yan, 2011) and online solvers (Wang & Banerjee, 2013). Bouwmans et al. (2017) made a comprehensive comparison on these solvers. To study those outliers not only in the observed data matrix but also in the orthogonal complement subspace, a reinforced robust version was also proposed (Brahma et al., 2018). The results showed that the method captured semantically meaningful outliers and also found more reasonable subspaces. For robust process monitoring applications, a very recent work by Pan et al. reported that the PCP method with the low rank representation could be successfully applied into the modeling and monitoring of blast furnace simulations, where outliers were frequently mixed (Pan, Yang, An, & Sun, 2016).

4.2.2.2. The SPCP case. The PCP is limited to those cases when the low-rank term being exactly low rank and the sparse counterpart being exactly sparse. However, due to the lack of data collection control in some industrial applications, noise with stochastic or deterministic properties may affect every entry of the data matrix. Such data corruptions will disrupt both of the low-rank subspace recovery and sparse matrix definition, which in turns may result in the unstable performance of PCP. In order to make the PCP widely robust, the stable and accurate recovery in the presence of entry-wise noise must be established. In the work by Zhou et al. (2010), the stable version of PCP was proposed and the optimization problem became

$$\min_{\mathbf{L},\mathbf{S}} \ \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 \quad s.t.\ \|\mathbf{X}-\mathbf{L}-\mathbf{S}\|_F < \delta \tag{15}$$

where $\lambda = 1/\sqrt{n}$ and the bound of Frobenius norm for noise is a positive small number $\delta > 0$. In order to make the stable recovery, the above problem is converted into the dual problem (Zhou et al., 2010):

$$\min_{\mathbf{L},\mathbf{S}} \|\mathbf{L}\|_* + \lambda\|\mathbf{S}\|_1 + \frac{1}{2\mu}\|\mathbf{X}-\mathbf{L}-\mathbf{S}\|_F \tag{16}$$

And the μ is set by a small number according to the data variance in practice. To recover the low-rank and sparse terms, one could use the Accelerated Proximal Gradient (APG) algorithm proposed by Lin et al. (2009). The RPCA-SPCP extends the PCA and RPCA-PCP by considering both gross sparse errors and small entry-wise noise in the data samples. Such advantage also makes it more widely applicable to practical problems where the low-rank structure is not exact. For example, Yan, Chen, Yao, and Huang (2016) applied the RPCA-SPCP in the robust monitoring of industrial processes. Through the SPCP inducement, they recovered the robust low-rank space where the PCA-based monitoring mechanism can be readily applied. Results showed that when there were severe outliers in the training data, the performance of PCA degraded while RPCA-SPCP could still maintain its robustness and good detection efficiency.

4.3. BRPCA

The above RPCA methods have those unknown regulation parameters that have to be well tuned through empirical analysis. In addition, the optimization-based methods are deterministic and cannot describe the uncertainties for all estimated terms in concern. As the alternative, the Bayesian methods will solve these problems more flexibly and will elegantly incorporate the robust modeling issue with the probabilistic framework. Besides, probabilistic models are popular also because of their principled way to cope with overfitting problems, to compare models, to handle missing values and also to represent uncertainties (Tipping & Bishop, 1999). In this sense, several efforts have been made for probabilistic PCA so as to formulate a robust Bayesian scope. There are two general frameworks for BRPCA defined for different noise background, which will be separately introduced as follows.

4.3.1. BRPCA-1

The first version of BRPCA inherits the similar low-rank and sparse diagram, but will further extend the RPCA-SPCP with probabilistic definitions. According to the work by Ding, He, and Carin (2011), the model for Bayesian RPCA (BRPCA) can be defined as

$$\mathbf{X} = \mathbf{L}+\mathbf{S}+\mathbf{E} = \mathbf{D}(\mathbf{Z}\boldsymbol{\Lambda})\mathbf{W} + \mathbf{B}\circ\mathbf{T} + \mathbf{E} \tag{17}$$

where $\mathbf{D}\in R^{N\times L_r}$, $\mathbf{Z}\in R^{L_r\times L_r}$, $\boldsymbol{\Lambda}\in R^{L_r\times L_r}$, $\mathbf{W}\in R^{L_r\times D}$, $\mathbf{B}\in R^{N\times D}$, $\mathbf{T}\in R^{N\times D}$ and $\circ$ denotes the Hadamard product. Among them, $\mathbf{Z}$ is the diagonal matrix with binary entries, $\mathbf{B}$ is the binary sparse matrix, L_r defines the largest possible rank for $\mathbf{L}$ and will be usually set as a large value. Notice that the low rank component $\mathbf{L}=\mathbf{D}(\mathbf{Z}\boldsymbol{\Lambda})\mathbf{W}$ is corresponding with the SVD decomposition, except the additional induction of $\mathbf{Z}$ which decouples the rank learning and the singular-value learning so that $r=\|\mathbf{Z}\|_0$. Through proper distribution definitions on parameters, the specific prior knowledge can be flexibly incorporated without major changes of the model structure and computations. For convenience, the probabilistic graphical model structure is illustrated in Fig. 7 and the detailed information is omitted which can be found

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

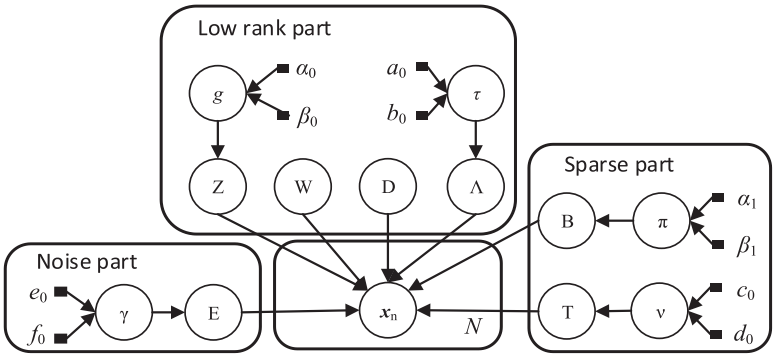


Fig. 7. Probabilistic graphical model structure for BRPCA-1.

in literature (Ding et al., 2011). Here we will put the focus for comparison between Bayesian method and the RPCA-SPCP. To see this, the objective function for optimization will be induced as the following the negative logarithm of the full posterior density function:

$$\begin{aligned} &-\log p(\mathbf{X}, H) \\ &\propto \frac{\tau}{2}\|\mathbf{\Lambda}\|_F^2 - \log\left[f_{BB}(\mathbf{Z}; H)\right] + \frac{N}{2}\sum_{l=1}^{L_r}\|\mathbf{d}_l\|_2^2 \\ &+\frac{L_r}{2}\sum_{i=1}^{D}\|\mathbf{w}_i\|_2^2 + \frac{\nu}{2}\|\mathbf{T}\|_F^2 - \log\left[f_{BB}(\mathbf{B}; H)\right] \\ &+\frac{\gamma}{2}\|\mathbf{X}-\mathbf{L}-\mathbf{S}\|_F^2 \\ &-\log\left[Ga(\tau|H)Ga(\nu|H)Ga(\gamma|H)\right] + Const. \end{aligned} \tag{18}$$

where $\mathbf{d}_l$ is the lth column from $\mathbf{D}$, $\mathbf{w}_i$ is the ith column of $\mathbf{W}$, $f_{BB}(\cdot)$ is the beta-Bernoulli prior function and H represents all hyper-parameters. One can infer from the above function that the BRPCA-1 shares several similarities with RPCA-SPCP. For example, the error $\frac{\gamma}{2}\|\mathbf{X}-\mathbf{L}-\mathbf{S}\|_F^2$ is analogous to $\frac{1}{2\mu}\|\mathbf{X}-\mathbf{L}-\mathbf{S}\|_F$, the priors are employed in $\mathbf{Z}$ and $\mathbf{\Lambda}$ to obtain the constraint for low-rank term $\mathbf{L}$, the priors are imposed on $\mathbf{B}$ and $\mathbf{T}$ so as to ensure the sparseness of $\mathbf{S}$. In fact, the distribution assignments actually impose various constraints on the statistical model, which should be alike to the non-probabilistic method. Apart from the similarities, compared with RPCA-SPCP, the main difference is that the numerical methods (variational Bayes approximation or MCMC) are used to estimate a distribution for each unknown parameter, while for optimization-based SPCP approaches one has to seek the single exact solution that minimizes the objective function. However, some unknown parameters like the noise variance can be hardly known a priori which makes the SPCP suboptimal.

It should be mentioned that like all above RPCA, the BRPCA-1 can also be applicable for analyzing those missing data situations. As one will always recover the complete low-rank space where those original missing entries has been completed after modeling and parameter learning. From this point of view, the Bayesian mechanism provides the flexible framework for robust statistical process understanding.

4.3.2. BRPCA-2

In essence, the above BRPCA-1 follows the data decomposition idea by the low-rank and sparse matrices, such strategy may distinguish those sparse outliers from small dense noise, especially for background extraction and video surveillance or traffic flow monitoring scenarios when the sparse outliers are of particular interest. For process industry monitoring, the main concerns are the low-rank terms which contain the 'recovered' process information. However, for cases when one needs to describe outliers in large dense noise, these methods may become inappropriate (Luttinen, Ilin, & Karhunen, 2012). In such cases, the true underlying distribution is heavy-tailed (has high kurtosis) due to the outliers in the data set. As the results, it is more appropriate to assume a heavy-tail distribution for observations. In literatures, two types of heavy-tail distributions can be found for tractable computations: the Student-t distribution and the Laplace distribution (or L_1 distribution) (Luttinen et al., 2012). Accordingly, the Bayesian robust PCA can be abbreviated as BRPCA-2 S and BRPCA-2 L. The generative model for both Bayesian robust PCA can be given as

$$\mathbf{x}_n = \mathbf{W}\mathbf{t}_n + \mu + \mathbf{e}_n \tag{19}$$

Notice that the generative model is derived from Eq. (12) when $K=2$ and the additional constraints will be considered on noise with heavy-tailed representations. The above definition does not contain the sparse term and the low-rank reconstruction of $\mathbf{x}_n$ is $\mathbf{W}\mathbf{t}_n + \mu$, where $\boldsymbol{\mu}$ is the offset or location vector. Like BRPCA-1, all parameters will be defined with proper distributions, the implicated details can be found in literature and are omitted here (Luttinen et al., 2012). Instead, the comparisons will be made among different robust probabilistic models. To see this, the Student-t probabilistic density function for observations is first given as

$$p(\mathbf{X}|\mathbf{W}, \mathbf{T}, \mu, \Lambda, \nu) = \prod_{n=1}^{N}\prod_{i=1}^{D} S\left(x_{ni}|\mathbf{w}_i^T\mathbf{t}_n + \mu, \tau_i^{-1}, \nu_i\right) \tag{20}$$

where τ_i is the precision, $\Lambda = \mathrm{diag}(\tau_1^{-1}, \ldots, \tau_D^{-1})$ is the variance matrix and $\nu_i > 0$ denotes the degrees of freedom. Moreover, one can infer from Eq. (20) that an individual distribution has been defined for each data entry. Such assumption is realistic for the dense noise cases when data are corrupted by intermittent outliers and missing values from data sampling sensors. The parameter ν_i regulates the thickness of the distribution tail of the ith dimension and therefore enhance its robustness against atypical observations. When ν_i tends to infinity, the Gaussian distribution will be recovered. Actually, the Student-t distribution can be interpreted hierarchically by using a Gaussian distribution with extra latent scale variable $\mathbf{U}$ as (Archambeau, Delannay, & Verleysen, 2008)

$$\begin{cases} p(\mathbf{X}|\mathbf{W}, \mathbf{T}, \mu, \Lambda, \mathbf{U}) = \prod_{n=1}^{N}\prod_{i=1}^{D} N\left(x_{ni}|\mathbf{w}_i^T\mathbf{t}_n + \mu, u_{ni}^{-1}\tau_i^{-1}\right) \\ p(\mathbf{U}) = \prod_{n=1}^{N}\prod_{i=1}^{D} Ga\left(u_{ni}|\frac{\nu_i}{2}, \frac{\nu_i}{2}\right) \end{cases} \tag{21}$$

Consequently, one can see that the Student-t scheme is akin to an infinite mixture of Gaussian distributions, and the extra scale variable controls the noise level of each individual observation. The purpose of Student-t distribution is to deploy a robust statistical model and the same goal can be achieved by another similar ap-

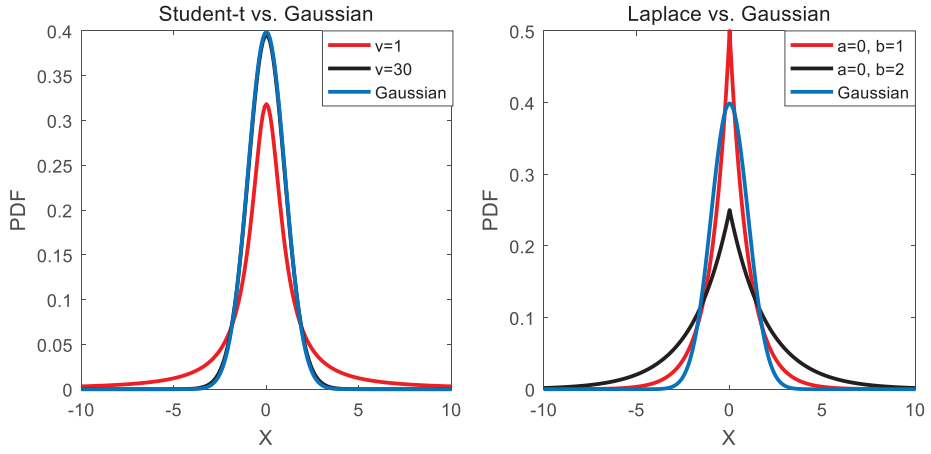


Fig. 8. Comparisons on Student-t, Gaussian and Laplace. The Student-t and Gaussian are with zero mean and the unit variance.

proach using the Laplace distribution: (Gao, 2008)

$$p(\mathbf{X}|\mathbf{W},\mathbf{T},\mu,\Lambda)=\prod_{n=1}^{N}\prod_{i=1}^{D}L\left(x_{ni}|\mathbf{w}_i^T\mathbf{t}_n+\mu,\tau_i^{-\frac{1}{2}}\right) \tag{22}$$

where $L(x|a, b)$ is Laplace probabilistic density function with location a and scale b. Similar to the Student-t distribution, the Laplace distribution can also be expanded by a superposition of an infinite number of Gaussian distributions. Likewise, the likelihood can be rewritten with the same Gaussian function in Eq. (21) by changing the prior of $\mathbf{U}$ with (Gao, 2008; Luttinen et al., 2012)

$$p(\mathbf{U})=\prod_{n=1}^{N}\prod_{i=1}^{D}IG\left(u_{ni}|1,\frac{1}{2}\right) \tag{23}$$

where $IG(u|a, b)$ is the inverse-gamma density function with shape a and scale b. As an illustration, the probability density functions for Student-t, Laplace have been compared with Gaussian in Fig. 8. From this point, one can see that the prevailing characteristic for both distributions is that they contain the extra latent variable which can regulates the heavy tail so as to accommodate those outliers. The Student-t distribution will tend to the Gaussian one as the degrees of freedom are larger than thirty (Archambeau et al., 2008; Zhu, Ge, Song, 2017c). It should be mentioned that both robust models are based on slightly different assumptions about the outliers and it still remains unclear which one is better and when. In this sense, they do not replace each other and one can choice either one for noise distribution in applications.

Based on the above robust definitions on noise, the probabilistic model can be directly constructed by further imposing robust distributions on latent variable $\mathbf{T}$ and the derived model is called robust factor analysis (RFA). If one simply assumes the isotropic variance matrix $\Lambda=\tau I$, the induced model will be robust probabilistic PCA (RPPCA). The RPPCA has been successfully applied in the robust modeling of industrial processes (Chen, Martin, & Montague, 2009). However, such definitions are not as flexible as the full Bayesian methods.

Like BRPCA-1, the full Bayesian methods will systematically impose distributions for all parameters and then make the optimization with numerical approximations. Take the BRPCA-2 S as the example, the probabilistic graphical model structure is shown in Fig. 9. And the objective function for optimization can be given as follows:

$$\begin{aligned}&-\log p(\mathbf{X},H)\\&\propto\frac{1}{2}\sum_{n=1}^{N}\sum_{i=1}^{D}u_{ni}\tau_i\left(\mathbf{w}_i^T\mathbf{t}_n+\mu_i\right)^2+\frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{L_r}\alpha_j\tau_i w_{ij}^2+\frac{1}{2}\sum_{i=1}^{D}\beta\tau_i\mu_i^2\end{aligned}$$

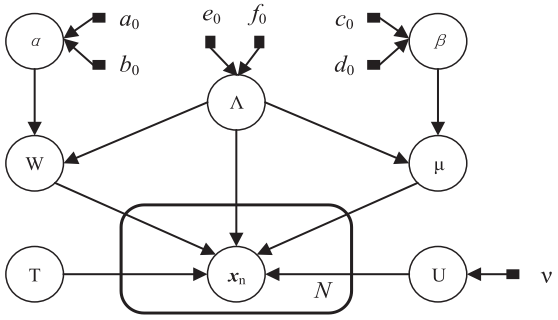


Fig. 9. Probabilistic graphical model structure for BRPCA-2S. The BRPCA-2L can be similarly obtained without the degree of freedom.

$$-\log\left[Ga(\alpha|H)Ga(\beta|H)Ga(\Lambda|H)Ga(\mathbf{U}|H)\right]+Const. \tag{24}$$

One can speculate the similar optimizer for the Laplace case which is not given in this survey. Compared with BRPCA-1, it can be inferred from the first term of above equation that the scalar has been imposed on the noise variance of each sample as a robust regulation. While the second term implies that the sparse constraint will be defined on each element of the project matrix. Like BRPCA-1, the missing entries can be completed in the low-rank subspace once the probabilistic modeling has been accomplished with variational Bayes or MCMC. In this way, the BRPCA-2 will result in a robust and sparse latent projections while the atypical entries in dense noise will be elegantly coped with the heavy-tailed accommodation. For industrial applications, some relevant works can be found. In Ge and Song (2011) the full Bayesian PCA has been used for robust monitoring of TE plants. In Liu, Pan, Sun, and Huang (2014) the variational Bayesian PCA was developed to deal with missing data in monitoring of wastewater treatment plant. A recent work by Zhu, Ge, and Song (2015d) proposed the variational Bayesian robust FA for modeling chemical process systems with outliers.

4.4. Robust data mining under different process characteristics

The PCA is the most fundamental tool for statistical process information mining and monitoring. However, the PCA may not be the best fit for all industrial modeling situations. Therefore, advanced robust data mining techniques should be employed so as to deal with different process characteristics like nonlinearity, non-Gaussianity and dynamics. This section gives the state-of-the-art review on robust statistical data mining methods for different types of industrial processes.

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

4.4.1. Nonlinear processes

For dealing with nonlinear processes, the kernel based methods like least squares support vector machine (LSSVM), extreme learning machine (ELM) and Gaussian process (GP) have been widely applied. All these methods suffer from outliers and missing data. Therefore, one has to consider the robust improvement.

4.4.1.1. Robust LSSVM. The LSSVM is based on margin-maximization performing structural risk and has excellent generalization ability. Therefore, it has been used for condition classification and product quality prediction through the regression version (Du, Lu, Li, & Deng, 2008; Ge & Song, 2010b; Thissen et al., 2004; Van Gestel et al., 2004). To deal with irregular data, several robust LSSVM methods can be found. Suykens, De Brabanter, Lukas, and Vandewalle (2002) proposed the weighted least squares support vector machine (WLSSVM) where they corrected outlying data with the weighted technique. The weights were affiliated in the cost function so that less important samples or outliers could be attached with low weights. However, such procedure should depend on LSSVM and requires the retraining of WLSSVM. To make the remedy, a heuristic weight-setting method was presented by Wen, Hao, and Yang (2008). In their work, the iterative updating algorithm was designed so as to obtain WLSSVM through a few updating steps instead of the directly retraining. These weighting schemes may be useful but also could be suboptimal solutions. For that issue, Chen, Yang, Liang, and Ye (2012) developed a recursive robust least squares support vector machine for regression applications. This method incorporated the maximum correntropy criterion and the half-quadratic optimization technique to find optimal solution. Their approach could work well for those nonzero mean and non-Gaussian noise with large outliers. However, the main difficulty lies in the selection of hyper-parameters. It is usually hard to set the proper weights of the training samples so as to make the optimal optimization. To avoid this, the robust estimator can be combined. Thereby, Yang, Tan, and He (2014) used the truncated least squares loss function for LSSVM where the influences from noise and outliers were depressed with a truncated parameter. Theoretical analysis shows that the robustness of truncated method is better than LSSVM and WLSSVM. More recently, in the work by Chen, Yan, and Li (2015), the least trimmed squares criterion was also engaged for robust estimation and then the traditional WLSSVM was combined to make the optimization. The two-stage robust estimation were reported to be better resistant to the affect of outliers.

Most of the robust LSSVMs studies are deluged in coping with outliers. However, it can be judged that different from outliers, missing values cannot be suppressed. On the other hand, the optimization with missing data can be difficult because of the non-convex nature of the objective function. To deal with missing data, Adankon, Cheriet, and Biem (2009) proposed two kinds of semi-supervised LSSVM classifier with transductive inference which can improve the generalization capacity. However, their performance should greatly depend on the well assigned hyper-parameters like the choice of kernel. As the alternatives, some imputation-based methods have been implemented. From Pelckmans, De Brabanter, Suykens, and De Moor (2005), they revealed that by defining a modified risk, the LSSVM could handle inputs contain missing entries which was alike the approach of mean substitution. Other similar missing data handling strategies can also be found in works by Yu (2012b), Wang, Li, Jiang, and Feng (2006), Zhang and Liu (2009), Aydilek and Arslan (2013) etc.

4.4.1.2. Robust ELM. The ELM approach works by fitting a single-hidden layer feed forward neural network (Huang, Zhu, & Siew, 2006). Despite the single structure, the ELM has proven to be efficient and effective learning mechanism for pattern classification and regression (Huang, Zhou, Ding, & Zhang, 2012). Therefore, it is interesting to apply the ELM framework for robust mining of industrial processes. In essence, several similar robust mechanisms for PCA and LSSVM have been adapted for improving the robustness of ELM. To see this, Horata, Chiewchanwattana, and Sunat (2013) developed three robust algorithms in their work: (1) the iteratively reweighted least squares method, (2) the multivariate least-trimmed squares, and (3) the one-step reweighted multivariate least-trimmed squares. The Extended Complete Orthogonal Decomposition (ECOD) was used in these algorithms to solve the weight computations. Results showed that compared with ELM, the robust ELMs with ECOD could be faster and also less affected by the increasing number of outliers. To account for the modeling uncertainties, the Bayesian ELM was constructed by Soria-Olivas et al. (2011), and then was extended into the robust Bayesian framework by Ning, Liu, and Dong (2015). The basic idea is similar to BRPCA-2S and BRPCA-2L where they have developed the robust nonlinear model with two heavy-tailed distributions (Laplace and Student-t) for outlier accommodation. In fact, compared with Bayeisan PCA, the basic idea for Bayesian ELM was to transform the observations into kernel space and then the generative model can be constructed similarly, expect that the ELM was to explore the data in a higher dimension space. Their results revealed that the robust Bayesian ELM could outperform the robust ELM with iteratively reweighted least squares. For dealing with missing entries, there are two mainstreams. The first way was like the work by Sovilj et al. (2016) in which they incorporated the ELM with multiple imputation for missing data estimation. Another way was to consider the semi-supervised learning framework in ELM, which has been proposed by Huang, Song, Gupta, and Wu (2014). In their work, the semi-supervised mechanism was based on the construction of graph Laplacian matrix and manifold regularization. It has been implied that the proposed SS-ELM maintained almost all the advantages of ELMs like training efficiency and direct implementation while it can also solve missing entries (Huang et al., 2014).

4.4.1.3. Robust GP. As another kind of popular nonlinear methods, the Gaussian processes have provided a principled, practical, and probabilistic approach to learning in kernel based machines (Rasmussen, 2004). The basic concept of GP is to define a distribution over functions and the inference is taking place directly in the space of functions. However, traditional GPs make the basic assumption that error terms should be i.i.d. Gaussian distributed given the function term. Nevertheless, in many industrial applications, the presence of outliers will make such assumption unaffordable, which will in turns render the quality monitoring deterioration. For this problem, Stegle, Fallert, MacKay, and Brage (2008) has proposed the Gaussian mixture clustering so as to make the explanation of the heavy-tailed noise, and then GP regression can be applied. Such framework is reasonable, but the Gaussian mixture component number should be selected with empirical knowledge. As the alternative, Jylänki, Vanhatalo, and Vehtari (2011) doped out a more flexible approach with the Student-t standpoint which was named robust GP for robust regression. The robust Expectation Propagation (EP) implementation has been established for the implementation of Student-t GP. For dealing with missing values, the GP property can be further embedded on some latent variable models (GPLVMs) like PPCA, and the derived model is known as GPLVM (Lawrence, 2005) and has been well applied in the nonlinear process modeling and monitoring by Ge and Song (2010a). In the work by Wang, Gao, Yuan, Tao, and Li (2010) the GPLVM has been further modified into the semi-supervised learning tasks under the pairwise constraints so as to improve the performance of dimensionality reduction.

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

4.4.2. Non-Gaussian processes

Traditional statistical data mining algorithms like PCA requires the basic assumption that the extracted latent variables should follow Gaussian distributions for the viable design of control limits for detection and diagnosis (Zhu et al., 2018). However, it has been reported from many studies that the real industrial processes with non-Gaussian operating conditions rarely conform such prerequisite (Ge & Song 2013; Liu, Xie, Kruger, Littler, & Wang, 2008; Zhu, Ge, & Song, 2017b). Even though those non-Gaussian methods like independent component analysis (ICA) and finite mixture models FMMs (like the Gaussian mixture model) have been successfully applied for process understanding and monitoring (Ge & Song, 2007; Lee, Yoo, & Lee, 2004; Yu & Qin, 2008; Yu & Qin, 2009), the non-Gaussian signals are sometimes mixed with outliers and missing entries, which will be detrimental to the mining and analysis. To improve the monitoring performance, robustness should be enhanced and the reviews are given as follows.

4.4.2.1. Robust ICA. The ICA originally served as the statistical method for blind source separation and was then introduced in non-Gaussian process data mining (Comon, 1994; Lee, Yoo, & Lee, 2004). Compared with PCA which makes the analysis on at most the second-order statistics, the rational of ICA is to use the information contained in those higher order statistics of the observations so as to de-mix an additive combination of statistically independent non-Gaussian source signals based only on observations of those mixtures (Hyvärinen, Karhunen, & Oja, 2004; Lee, Girolami, & Sejnowski, 1999). It has been shown that many ICA algorithms were performed by the minimizing the KL-divergence, which is not robust in the presence of outliers (Mihoko & Eguchi, 2002). As the result, Mihoko & Eguchi (2002) proposed a robust estimator for ICA using Beta-divergence which gives smaller weights to possible outliers so that their influence was weakened. Chen, Hung, Komori, Huang, and Eguchi (2013) introduced the Gamma-divergence based ICA which was shown to be super robust against data contamination than the Beta-divergence method. There are other robust techniques as well. Cao, Murata, Amari, Cichocki, and Takeda (2003) gave a two-procedure based robust ICA. In the first procedure, the pre-whitening has been used to reduce the additive noise. Then in the second procedure, the parameterized t-distribution density model has been combined so as to deal with outliers. Similarly, Hsu, Chen, and Liu (2010) designed the outlier filter in their work so as to well describe the majority for training dataset for ICA modeling. Their results were validated with better fault detection ability on Tennessee Eastman process. More recently, Cai & Tian (2014) used the MCD estimator for the robust determination of non-Gaussian signals for modeling. In another work by Zhao et al. (2015), a reference-based negentropy based robust method was developed for ICA, which made use of the cross-statistics or cross-cumulants between estimated outputs and reference signal, and the derived ICA was demonstrated to be more robust against outliers.

Likewise, the above robust ICAs are not probabilistic models and cannot deal with missing values for training process data in the standard form. To make the principled treatment of missing data, the probabilistic ICA has been extended and a variational Bayesian ICA (VBICA) has been proposed by Chan, Lee, and Sejnowski (2003). The VBICA makes the generative model by a mixture of Gaussians on each latent direction to extract the non-Gaussian signals. In this way, the missing data problem can be handled in a flexible manner. In Salazar, Vergara, Serrano, and Igual (2010), a semi-supervised learning framework was delivered for ICA. Other similar works can be also found by Welling and Weber (1999), Peng and Zhu (2007), Rosca, Gerkmann, and Balcan (2006) etc.

4.4.2.2. Robust FMMs. While the ICA provides the implicit non-Gaussian extractions on the latent domain, the FMMs will make the relatively explicit non-Gaussian representations via the soft combination of multiple local components. For example, in those industrial processes with multiple operating modes, one or more mixture components can be assigned to each condition so as to account for the specific mode characteristics and make both the global and local diagnosis. To construct the robust finite mixture models, there are two general types, the full space model and reduced space model, in premise of whether the dimension reduction has been considered. Since both types can be unified in the probabilistic framework, reviews and discussions will be made together.

In literatures, the basic improvement on the robust mixture model is to replace the Gaussian distribution with the Student-t alternative and there have been a bunch of related works. Similar to the above robust methods, the use of t distributions will also provide a sound mathematical basis for the robust modeling for mixture estimation. The work by Peel and McLachlan (2000) proposed the t mixture model-based approach on the basis of Gaussian mixture model. After then, Archambeau et al. (2008) extended the robust mixture model in the dimensional reduction scope and the mixtures of robust probabilistic principal component analyzers were developed. Such robust model was applied for the non-Gaussian chemical process monitoring by Zhu, Ge, and Song (2014) Recently, for product quality monitoring through soft sensor predictions, the mixture of RPPCA has also been adapted with the regression formulation (RPPCR), named as mixture of robust probabilistic principal component regression (Zhu, Ge, & Song, 2015c). In another work by Lin, Lee, and Hsieh (2007), the mixture of skew t distributions has been proposed in the further treatment of heterogeneous data with asymmetric behaviors across multiple local classes. To make the account of difference variances, the work by Chatzis, Kosmopoulos, and Varvarigou (2008) built the mixture of robust factor analysis. Notice that, besides the widely used Student-t based paradigm, the L_1 mechanism has also been found for the robust mixture modeling by the work of Gao and Xu (2007). That work can be viewed as the generalization of the BRPCA-2L into the mixture fashions. Notice that benefit from their flexibilities, the above robust mixture models can be readily used for a wide range of applications with missing data. For example, the semi-supervised mixture RPPCR has been proposed by Zhu, Ge, and Song (2015b) which can simultaneously deal with outliers in measurements as well as the multi-rate sampling issue in quality variables.

4.4.3. Dynamic processes

The above review has been focused on the static robust data mining techniques which make the sample permutation invariant assumption. The sample permutation invariant refers to the requirement that modeling information should be retained by randomly permuting the time ordering (Pearson, 2002). However, such assumption cannot be hold for those dynamic process data characteristics when we have to consider the sequential relations. The dynamic characteristics should be one of the most important industrial system properties as time-wise sampling points are auto-correlated. Therefore, it is desirable to extend the statistical modeling from static to dynamic representations. In this work, the state-space representation with Bayesian inference will be commonly introduced as the preliminaries for dynamic process modeling. Then the previous works will be reviewed for robust Bayesian modeling methods.

4.4.3.1. The state-space representation and Bayesian inference. To work with dynamic processes, the discrete-time stochastic dynamical systems are employed and the state-space presentations will

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

follow the state-of-the-art definition given by Roweis and Ghahramani (1999):

$$\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \mathbf{w}_t \tag{25}$$

$$\mathbf{y}_t = g(\mathbf{x}_t) + \mathbf{v}_t \tag{26}$$

where $f(.)$ and $g(.)$ are linear/nonlinear state function and observation functions respectively, vectors $\mathbf{w}_t$ and $\mathbf{v}_t$ denote the state evolution and observation noises. Notice that sometimes for control purposes, the input terms should also be involved, which are omitted here from simplicity. A basic and widely reduced representation is the linear dynamic systems (LDSs), and the above state-space formulas can be simplified as (Barber, 2012)

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{w}_t \tag{27}$$

$$\mathbf{y}_t = \mathbf{C}\mathbf{x}_t + \mathbf{v}_t \tag{28}$$

where $\mathbf{A}$ is the state transition matrix and $\mathbf{C}$ is the observation or generative matrix. The state evolutions are of first-order Markovian dynamics and the noises are independent to states and observations. Notice that for conventional LDSs like linear Gaussian state-space models (Wen, Ge, & Song, 2012), noises are temporally white and spatially Gaussian distributed, which can be further denoted as $\mathbf{w}_t \sim N(\mathbf{0}, \mathbf{Q})$ and $\mathbf{v}_t \sim N(\mathbf{0}, \mathbf{R})$. The states are usually continuous valued with discrete sampling points. In some cases, the states can be assigned with discrete or categorical values as well. For example, in multimode processes, the operating mode indicator is in fact the discrete process status (Yu, 2012a). In some cases, for detecting abrupt changes in the states or the parameters when the system parameters like state transition matrix and observation matrix are unknown, the identification of parameter and state configurations in both continuous and discrete models can be commonly implemented with the EM algorithm in the unified Bayesian framework: the E-step is designed for state estimation of dynamic systems while the M-step will update the model parameters with the states from the current E-step, and E-step and M-step iterate until convergence (Little and Rubin, 2014; Moon, 1996).

In practical systems, one can infer that the above model given by Eqs. (27) and (28) is only an approximate to the general fashion as the state matrix and observation matrix can be nonlinear, while those noises may follow a thick-tailed, non-Gaussian distributions due to innovation and observation outliers. Therefore, various robust statistical models have been built. This review will be focused on the robust Bayesian modeling for dynamic processes. Since the robust Bayesian modeling of dynamic systems should largely rely on the robust state estimation. Consequently, the robust methods of Kalman filter and hidden Markov model will be focused, as they are the state-of-the-art dynamic process analysis techniques for processes with continuous states and discrete states respectively.

4.4.3.2. Robust Bayesian estimation with Kalman filters. Particularly, for dynamic process analysis with continuous state cases, one usually makes the latent state estimation through robust filter and/or smoother. The filter utilizes the past and current observations to predict the current state while the smoother encompasses all available data to optimally estimate the entire state sequence. Therefore, the key issue for modeling dynamic systems is to design a robust filter which is insensitive to outliers and can deal with missing data. Several robust filters have been developed.

Outlier-robust Kalman filter: As the predecessor of modern filtering, the Kalman filter (KF) is the optimal estimator for linear-Gaussian dynamical systems. Therefore, several improvements have been conducted to robustify the KF. In Masreliez and Martin (1977), the authors have introduced the heavy-tailed density term for making robust KF. In their work, the normal mixture densities were introduced where an extra normal component was weighted for the explanation of heavy tails. The normal mixture densities can be generalized into Student-t fashion. As the consequence, Agamennoni, Nieto, and Nebot (2012) has proposed the Student-t based KF so as to extend the KF into a wider range of noise distributions without the loss of computational advantages. The similar ideas can be also found in Huang, Zhang, Li, Wu, and Chambers (2017), Huang, Zhang, Xu, Wu, and Chambers (2017), and Zhu, Ge, and Song (2016b), where the t-distribution based robust KF have been applied in several state estimation and monitoring applications including cooperative localization of autonomous underwater vehicles, maneuvering target tracking and chemical plant fault detection. Apart from the Student-t robust prototypes, Gandhi and Mili (2010) proposed the robust KF by a robust estimator called generalized maximum-likelihood-type (or GM estimator) by Schweppe's proposal and the Huber function, which was enabled of high breakdown point against outliers. Actually, the GM estimator improved the KF to bounded residual and position estimations, where the former index handled the effects from observation and innovation outliers and the latter made the assessment of structural mismatches. In the work by Aravkin, Bell, Burke, and Pillonetto (2011), the Laplace distribution framework has been proposed for robust KF development and the constructed Laplace smoother was able to make a smoothed fit of dynamic process data without data removal. In another work by Nurminen, Ardeshiri, Piche, and Gustafsson (2015), the skewed t distribution was incorporated for robust inference with KF in cases with skewed measurements. More recently, Aravkin, Burke, Ljung, Lozano, and Pillonetto (2017) have discussed the problem of robust smoothing in the optimization viewpoint. They provided a comprehensive and systematic treatment framework on various constraints of state and measurement noises so as to make the outlier-resistant statistical modeling.

Outlier-robust extended KF: The robust KF can be used for the approximate inference in state-space models with heavy-tailed noises. However, such state-space models are still based on the stochastic linear dynamic processes, which could become inappropriate for nonlinear situations. As the alternatives, the extended KF (EKF) can be considered (Ljung, 1979). The key idea of EKF is to successively linearize the nonlinear model describing the current measurements, then the common KF can be applied for next state estimation. Accordingly, several robust modifications have been made for EKF. In the work by Kautz and Eskofier (2015), the authors made the claim by demonstration that an efficient way to make EKF more robust was to reject outliers based on a threshold-based detection. The boundary has been defined with the same interval as the boxplot method. Once the observation has been determined as an outlier, it will not be considered in the filtering procedure. However, such method greatly depend on the empirical knowledge. In the work by Hammes, Wolsztynski, and Zoubir (2009), the robustness of EKF has been enhanced by two strategies: the integrated robust estimator and weighted technique on residuals. Both strategies are based on the basic principle that the influence of outliers should be suppressed during the filtering. From another work by Vuković and Miljković (2015), the Student-t based probabilistic generative model has been developed to the EKF. Therefore, outliers can be naturally down-weighted during sequential data learning.

Outlier-robust unscented KF: The EKF by linearization can only behave well for certain cases when the system can be well approximated by linearization. For the general nonlinear systems, such linearization may become less accurately approximated as the higher order terms have been ignored. Another attempt at nonlinear filtering which can avoid the linearization is the unscented KF (UKF) (Julier & Uhlmann, 2004). The key idea behind UKF is

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

that it is easier to estimate a Gaussian distribution from a set of points than to approximate the arbitrary nonlinear function. The UKF starts with points that are plus/minus a certain distance (like the standard variance) away from the mean positions along each dimension, and then pipe them through the nonlinearity. Finally one fits a Gaussian to the transformed points. As a derivative free method, the UKF propagates the first two moments of the posterior distribution and is superior to EKF (Daum, 2005; Gustafsson & Hendeby, 2012). To enhance the UKF against outliers, similar robust works can be found in retrospect of the robust KF and robust EKF. For example, in Wang, Cui, and Guo (2010), the Huber-based unscented filtering was derived to modify the measurement update equations of standard UKF so as to exhibit more robustness against outlying deviations. In Chang, Hu, Chang, and Li (2012), the generalized maximum-likelihood perspective (robust estimator mechanism) has been migrated from robust KF to robust UKF. In addition, there are other robust developments with robust distributions. Accordingly, the L_1 regulation was used by Kaneda, Irizuki, and Yamakita (2013), while in Huang, Zhang, Li, and Chambers (2016), the Student-t mechanism was proposed. Compared with robust KF and robust EKF, the robust UKF should be more flexible and desirable in common practices due to the internally prevailing attributes.

KF for missing data: To deal with missing data or intermittent observations in the framework of KF, one has to consider the system parameter uncertainties and convergence criteria on expected estimation error covariance. In Sinopoli et al. (2004), the Kalman filtering problem with intermittent observations has been formalized. By defining the arrival of an observation as a Bernoulli random process, the analysis under the EM theoretical picture showed that the convergence should depend on the probability of arrival data, the eigenvalue of state transition matrix and also the structure of observation matrix. Their results have manifested that the monitoring and control of dynamic systems with uncertain parameters could be possible if the data missing uncertainty does not exceed a given threshold. Later, in works by Kluge, Reif, and Brokate (2010) and Li and Xia (2012), the stability analysis has been conducted for the EKF and UKF with unreliable observations. Both studies have statistically analyzed the system initial conditions and the upper bound on estimation uncertainties. These researches have been based on the randomly missing situations with independent packet drops. Therefore, Huang and Dey (2007) further considered the stability of Kalman filtering in a network environment with the packet losses following a Markov chain. They gave the sufficient and necessary condition for the scalar case. Based on this analysis, a technical note by Censi (2011) made a further investigation on the KF convergence property for a wider class of models when the arrival of observations were driven by a more general missing data mechanism called the semi-Markov chain. Apart from the above analysis on randomly missing cases, there are several studies on the multi-rate data fusion with KF. Basically, the multi-rate missing data pattern can be viewed as the regular missing data problem due to the incongruous sampling of sensors and measurement delays. The measurements may be contained with frequent and infrequent ones and the problem is how to make the state estimation and data fusion with the two types of measurements. In Hsiao and Pan (1996), a robust Kalman filter synthesis method was proposed by transforming the stochastic time-delay system into the uncertain stochastic system with no time delay through the proper definition of state variables. Lee, Kang, and Han (2018) studied the online optimal state estimation scheme with Kalman filter for delayed and periodic measurements. In the work by Smyth and Wu (2007), the multi-rate filtering and smoothing techniques have been considered for dynamic process monitoring. More recently, the irregular rate filtering problem has also gained interest for product quality estimation and the relevant works can be found in Guo, Zhao, and Huang (2014), Fatehi and Huang (2017) etc.

4.4.3.3. Robust Bayesian estimation with HMM. The discrete state dynamic system can be regarded as the operation on state equation by using of the winner-take-all nonlinearity (Roweis & Ghahramani, 1999). In this sense, the HMM can be viewed as the analogous case from continuous models and the robust mechanisms should also be similar. For example, in Karplus, Barrett, and Hughey (1998), the weighted HMM is proposed and the piecewise weighting function has been used for dealing with outliers. In Guha, Li, and Neuberg (2008), the rejection criterion has been incorporated to detect outliers. Apart from the soft weighting and hard detection, some special attention has been put on the more efficient Bayesian method. In the work of Chatzis, Kosmopoulos, and Varvarigou (2009), the Student-t mixtures were used for outlier tolerance and the robust HMM has been proved to an effective treatment in several robust sequential data modeling applications like bimanual gesture recognition, phonetic recognition and cognitive state decoding of brain fMRI images. Similar works can be found in other applications. In Borzeshi, Concha, Da Xu, and Piccardi (2013), the authors have established the robust extended HMM with Student-t that can consider both spatial and temporal information. The robust method has been applied into the monitoring of human actions by performing joint action segmentation and classification tasks. In Zhu, Ge, and Song (2015a), a HMM driven robust PPCA has been proposed for fault classification in dynamic processes. Benefiting from PPCA, the proposed robust method is inherently semi-supervised which can elegantly handle randomly missing data. The feasibility validation on TE pant data has shown that the robust method outperformed the traditional HMM for industrial data with irregular data samples. Recently, in the work of Ding and Shah (Ding & Shah, 2013), the Student-t based robust distribution was introduced for the hidden semi-Markov model in order to explain the state duration in dynamic systems.

4.5. Summary and remarks

4.5.1. Summary

It can be seen from above reviews all methods have the pros and cons and some are even designed for particular industrial applications of interest. Commonly, the hard detection methods like MCD can be appealed for high breakdown performances on polluted data. In this sense, it is reported to be applicable in those outlying occasions even outliers constitute nearly 50% of the entire data population (Rousseeuw & Driessen, 1999). However, the hard detection needs the empirical knowledge on process data. Besides, the detected outliers which removes outliers may also reduce the data representative ability and add up to the analysis complexity. For soft weighting methods, the outliers can be down-weighted after some rolling-backs. However, in most cases, such manipulations acts as the pre-processing and requires extra computational resources for successful statistical analysis. In contrast, the robust PCAs with DLAMs are more efficient. Such robust models provide the flexible framework for different analysis scenarios. For video monitoring analysis and robust object tracking, one actually extracts sparse outliers from the sparse term for extraction and analysis. While for industrial robust process modeling and monitoring, the interest turns to the recovered lower dimension subspace without outliers. However, to make such robust models work, some additional governing parameter should be elaborately selected. Although the parameter estimation issue can be further solved by the corresponding Bayesian formulations, as mentioned above, such type of technique may not fit for some industrial applications with dense and stochastic outliers. As the alternative, one may resort to the second type of Bayesian robust

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

PCA methods with generative models by using probabilistic interpretations for noise and information representations, such mechanism makes them flexible for modeling a wide range of heavy-tailed noise. Moreover, model and parameters can be empirically determined once the algorithm has been completed. However, the Bayesian methods rely on the EM theory which may take a couple of iterations while only get stuck at the local maxima. One should take multiple trying with different initial guesses so as to alleviate this issue.

To further explore robust mining of different process characteristics, one can judge that most of the robust schemes of PCA can be easily transferred for mining different processes. Among them, the Bayesian methods should be the most elegant and flexible ones due to the powerful representations by probabilistic models. Such merit is particularly important for the dynamic mining of sequential data when the uncertainties on model and parameters should be the vital indices for explanatory analysis. As the matter of fact, the transfer of robust techniques also attribute to some internal relations that can be induced for most of the data mining techniques. For example, the linear modeling method PCA can be analyzed as the deterministic instance of PPCA when the noise variances tend to zeros. The Bayesian ELM can be regarded as the nonlinear projection for reducing mining complexity with high dimension and the generative model is similar to PPCA. While the non-Gaussian method ICA have also be generalized by probabilistic forms with non-Gaussian latent assignations. To comprise the dynamic descriptions, static models can be further extended for the dynamic systems by filtering and smoothing. Those sophisticated relations among different data mining actually take up the bridges between linear and nonlinear methods, single and mixture methods, static and dynamic methods; under which the shared robust Bayesian modeling framework can be easily affordable. Also with such framework, the irregular entries can be elegantly handled: the missing values can be explained out during EM learning while the outliers will be tolerated by heavy tails. To this end, it is not just a coincidence as those Student-t and Laplace based robust prototypes can be readily compatible and preferred for analyzing various complex systems with nonlinear, non-Gaussian and dynamic characteristics.

4.5.2. Some remarks

It is worth to mention that in a recent overview for Robust Subspace Recovery (RSR) (Lerman & Maunu, 2018), the authors divided the robust PCA techniques into two competing camps called outlier-robust methods and sparse-corruption methods. After then they pointed out that RSR was related with the former which should attempt to recover the underlying low-dimensional subspace structure of inliers with ways like outlier-filtering, robust estimators and projection pursuit. In this sense, those robust PCA methods reviewed in Section 4.2.1 and Section 4.3.2 are related with RSR. Comparatively, the DLAM formulations discussed in Section 4.2.2 and Section 4.3.1 make the overall assumption that outliers only mix in a few indices and can be separated through the sparse-corruption component, which should belong to the sparse-corruption campaign.

Notice that for those long-time duration data sequences, the robust subspace learning topic also covers another area called Robust Subspace Tracking (RST) (Vaswani & Narayanamurthy, 2018). The RST allows the subspace to change gradually over time and the goal is to track such time-varying subspace in the presence of sparse outliers. For video tracking, there already have some successful RST applications, see related review work for details (Vaswani et al., 2018). As for the process industry monitoring, however, we have not conducted the review in this article. This is due to the fact that the safety monitoring is in some extent different with the common video tracking conditions. For most cyber physical systems, the online intelligent monitoring modules are designed to automatically uncover critical faulty system behaviors and then alarm faults upon necessity (Zhu, Ge, & Song, 2016a). Such modules often work well by following the offline modeling and then online monitoring procedures. Unfortunately, to realize the rational online version of robust tracking, adapting and monitoring, it is still not an easy job for the system to make automatic discriminations among outliers and various abnormal events in practice. From this sense, it still remains an open question as on how to make the reasonable deployment for robust subspace tracking and monitoring.

The third remark will come to the outlier tolerance for various robust data mining approaches. For most outlier-robust estimators, one could resort to the concept of breakdown points for the quantification of robustness. The breakdown point of an algorithm is defined as the largest percentage of outliers that the algorithm could withstand. In the work by Xu et al. (2015), there are complete lists of breakdown points for those widely used estimators. As for other robust subspace learning methods, assumptions and theoretical guarantees are often required rather than breakdown points. In fact, the theoretical guarantee is often used as the sufficient condition before practicing so as to know if the method is guaranteed to be correct under certain assumptions. For more performance details on robust PCA, one can refer to the work by Vaswani et al. (2018) and Lerman & Maunu (2018), where the authors compared theoretical guarantees for different robust modeling techniques.

For the last remark, it still should be emphasized here that most of the current robust techniques cannot take a general form to deal with both outliers and missing data, as different outlier treatments show different application fields while different missing data patterns impose different effects on uncertainties in parameter learning. The robust semi-supervised leaning strategy offers with such possibility on problem-solving, despite only a few related work can be found in this review. To deal with both data imperfections with such framework, one also has to consider two basic issues. On the one hand, the efficient robust state estimation mechanism should be established for outliers and missing data; on the other hand, one has to investigate the convergence property of estimation errors brought by outliers and various missing data patterns. On this basic framework, it can be anticipated that more studies should be explored and efforts are still required.

5. Recent challenges in the mining of big data: a robust perspective

Due to the great impacts for data mining, the researches on robust methods have made rapid progress over the past few decades, and are still receiving continuous focus from both academia and practical engineering. In the current age of Industry 4.0, the ubiquitous sensors and the pervasive storing of information have made the large amount of data available for statistical analysis (Slavakis, Giannakis, & Mateos, 2014). According to a well-accepted definition by Chen and Zhang (2014), “a data set can be called Big Data if it is formidable to perform capture, curation, analysis and visualization on it at the current technologies.” In this sense, the industrial process data can be regarded as a kind of big data, which has large volumes and also low-density values. The process data are invaluable for making production plans and also enhancing the competitions in markets. Hence many industrial companies are embracing the big data era (Qin, 2014). However, the increasing volumes of data also lead to more uncertainties as various data resources will have different data quality controls. Big data sets are often incomplete and inevitably contain corrupted measurements and communication errors (Chen & Zhang, 2014). Robust mining from these voluminous data will bring significant science and en-

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

gineering advances for industrial process performance monitoring and also production improvement. In this opinion, this section will bring several perspectives of recent research challenges on robust data mining issues in concern of nowadays industrial big data applications.

5.1. Robust parallel modeling

To reveal statistical insights into big data sets with data mining, those traditional analytic tools with centralized storage and one-shot processing have become no longer affordable. Instead, the whole learning mechanism could be modified in the distributed and parallelized framework, especially for those plant-wide systems with multiple units and large scales. Such framework follows the divide and conquer scheme, where multiple agents will work in parallel on disjoint data subsets and then the results will be gathered and combined for the final analysis over the entire task. For example, a recent research by Zhu, Ge, and Song (2017a) proposed the decentralized analysis of PCA on Hadoop and MapReduce, and the distributed and parallel method has been applied on the large-scale chemical plant with voluminous data. There are also alternatives, like the distributed PPCA (Elgamal, Yabandeh, Aboulnaga, Mustafa, & Hefeeda, 2015), distributed SVM (Alham, Li, Liu, & Hammoud, 2011) and distributed GMMs (Zhu, Yao, Li, & Gao, 2018). Recently, the distributed and parallel framework was also reported in quality prediction applications (Yao and Ge, 2018a,b). However, these methods are adapted from classic schemes and will fail if the data contains outliers and missing data, which are unfortunately the common cases in the big data age. In fact, for those large data sets, one can speculate two ways for parallel robust modeling. One method is to propose the parallel data cleaning techniques. The other way is to make efforts for robust parallel statistical analysis.

For parallel data cleaning, most univariate and multivariate algorithms can be migrated with the Hadoop MapReduce framework. The feasibility on statistical estimation is based on the fact that those basic statistics like mean, correlation, variance and median can be easily parallelized in calculations. One can speculate that such merits also benefit for the development of those parallel robust normalizations. While for multivariate dimensions, there also have been several attempts in recent years like the parallel kNN (Zhang, Li, & Jestes, 2012) and parallel LOF (Lozano & Acufia, 2005). Besides parallel data cleaning for robust analysis, the direct parallel robust statistical modeling is another important solution. To achieve that, one has to consider two basic issues for successful investigations in future works. On the one hand, those robust PCA methods should be considered for adaptation into those parallel scenarios for the distributed model identification and data interpretations. In this way, one can denoise the outlier contaminated entries and also impute those missing ones. The challenge here lies in how to make the numerical analysis with optimization amenable to those decentralized or parallel implementations (Slavakis et al., 2014). On the other hand, one has to further combine those specific data characteristics for robust descriptions of nonlinear, non-Gaussian and dynamic system behaviors. In this sense, large efforts are still needed on handling huge amounts of process data in the robust and parallel manner.

5.2. Robust data selection and dependence analysis

Although the big data volumes can be solved with some potent parallel-driven computing techniques, there are still issues for both academic and engineering practices. First, the benefit-to-cost ratio of such mechanism remains for further investigations. Especially for plant-wide systems with frequently update on changed product recipes and market requirements. Second, one cannot always ensure that more data will always add up values to the interpretation ability of the engaged statistical model (Li et al., 2015). Actually, the role of data mining is to extract the process knowledge to find representations on the common behaviors, from which the process knowledge can be represented and analyzed with both qualitative and quantitative domains (Zhu, Ge, Song, Zhou, & Chen, 2018). Therefore, one major problem for most parallel methods is that they are designed with implicit data mining schemes which cannot explicitly provide the qualitative domain knowledge like the compact dependence relations. Unlike the quantitative knowledge, the compact dependence relations form the basic principles in process and could be induced from the fundamental working mechanisms of most industrial productions. Moreover, the principles will normally keep unchanged once the working flowchart has been given, and hence will have little to do with the inconsistent product recipes or the enormous volume of generated data collections. Therefore, the challenging issues for big data analytics can be two-fold: selecting the most relevant data and also building a parsimonious model which can explore the underlying dependence structures in the robust manner. The following context will make the distinguished analysis on these two sides.

5.2.1. Robust data selection

In literatures, the relevant data selection can be conducted from two dimensions: feature-wise selection and time-wise sample selection (Blum & Langley, 1997; Guyon & Elisseeff, 2003; Zhu & Gao, 2018b). The data selection is in accordance with the basic requirements and attributes enrolled with industrial systems, as the redundancies exist in both features and samples. For one thing, the redundant sensor networks are always deployed so as to improve the reliability of a system and increase the accuracy for surrounding perceiving, especially for life-critical apparatus (Ali & Narasimhan, 1995; Kim, Lee, & Lee, 2010; Latif-Shabgahi, Bass, & Bennett, 2004; Sen, Narasimhan, & Deb, 1998). For another thing, due to the repetitive or continuous production in industrial processes, the characteristics of data samples may show strong consistencies over a very long period of time. For data analysis, the feature selection is to filter out or suppress those irrelevant features while the time-wise data selection is to deal with those redundant samples pertaining to the current analysis. Notice that feature selection is different from feature extraction as the latter is designed with some meaningful transformations (like the PCA) on the original space. In data science literatures, the data selection should be one of the most important hotspot and the roles can be found from at least five points: reduce samples and variables; shorter training time; simplify model structure; avoid the curse of dimensionality and improve the ability of generalization. Previous reviews and comparisons on classic feature selection and sample selection can be found such as Blum and Langley (1997), Chandrashekar and Sahin (2014), Wang, He, and Wang (2015), Xue, Zhang, Browne, and Yao (2016), and Sheikhpour, Sarram, Gharaghani, and Chahooki (2017), etc. One can judge that seldom works report on robust data selection. As nowadays data mining aims to address those larger and more complex tasks, a robust data selection diagram should indeed become more important for a better process understanding of the overwhelming industrial process data. The challenge for robust data selection is that both weighting and searching schemes will suffer from high computational cost, and are sensitive to outliers and missing values. Another practical challenge is one should further consider the data labeling problem during data selection. Since only partially labels can be available in industrial applications like fault classifications and soft sensor development. The data labeling issue determines the important data samples that should be marked so as to render the appropriate modeling and inference. However, such labeling proce-

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

dure can be hardly done in practice. For example, in soft sensor applications, those quality variables should normally undergo lab analyses. Therefore, one may only have a very limited set of labels from the response variables in order to make clear the most responsible variables for soft sensing development. For this purpose, and also as a promising research target in the future, the robust semi-supervised data selection mechanism can be investigated, from both labeled and unlabeled data items.

5.2.2. Robust parsimonious modeling via dependence analysis

Once the relevant data and features have been selected, the parsimonious modeling should be considered. Several state-of-the-art methods can be found for such parsimonious modeling like bond graph (BG) (Ould-Bouamama, El Harabi, Abdelkrim, & Gayed, 2012), signed digraph (SDG) (Lu¨& Wang, 2008), Bayesian network (BN) (Yu & Rashid, 2013) and Granger causality analysis (GCA) (Yuan & Qin, 2014). These approaches try to describe the large-scale industry system by qualitative network topologies with nodes and arcs (Yang & Xiao, 2012). The nodes usually represent variables, elements or cause events, while the arcs reveal the exchange influences or propagated effects. In this sense, the causal events may be with unidirectional or bidirectional relations. For example, the BG describes the exchange of physical energy by bonds where nodes are assigned by power variables and the edges represent the exchanged power. On the contrast, the Bayesian network usually makes the nodes as process variables and the acyclic graph will be designed for the entire system. The networks can be constructed from precise/partial process knowledge and even data-driven ways. For example, the BG and SDG are commonly built based on the detailed quantitative analysis from deep physical understanding while the BN and GCA can be suitable when the precise mathematical models are missing and only qualitative relations can be recognized (Yang & Xiao, 2012). By constructing such networks, the propagation pathways of fault evolution in the complicated industrial system can be explicitly explored which will save time and costs for troubleshooting. For those black-box processes with less qualitative knowledge, there are also data-driven inference engines for network structure constructions for methods like the BN and GCA (Chen, Anantha, & Wang, 2006; Zou & Feng, 2009). Such studies are often oriented with moderate industrial process with small data size. For high throughput large-scale dimension datasets with complicated relations and poor data qualities, the current techniques can be no longer amenable for dependence analysis, network construction as well as parameter learning. Some recent studies on machine learning have reported the parallel structure and parameter learning of Bayesian networks (Madsen, Jensen, Salmerón, Langseth, & Nielsen, 2017; Villa & Rossetti, 2014; Yue, Fang, Wang, Li, & Liu, 2015). However, seldom researches have been conducted in the field of robust process mining scope. To this end, in order to make the effective and efficient parsimonious modeling in the era of big data, both robust and parallel mechanisms should be further considered so as to cast lights on the implicit process structures, and also make the reasonable mining and inference on causal effects for process monitoring and diagnosis.

5.3. Robust data mining with noisy data labels

In literatures, real-world noises for supervised data mining can be distinguished into two types: the feature noises and label noises (Bouveyron & Girard, 2009; Frénay & Verleysen, 2014). From this aspect, the above review has been mainly dedicated to robust data mining for the feature noises. However, the comprehensive robust data mining philosophy should also cover robust techniques for label noises. The label noises will alter the correct settings into incorrect and misleading instances, and hence are more harmful than the feature noises. As mentioned in a recent survey, there are several sources of label noise like the subjective expert labeling, insufficient knowledge for labeling and also data encoding or communication problems (Frénay & Verleysen, 2014). In industrial processes with big data, one can judge that human supervisions on the labeling of process status, faults and product qualities can be either imprecise, difficult or expensive. Eventually, one has to confront the robust big data mining challenge against label noises.

Compared to the feature noises which has been widely studied, mining form the noisy labels have received rather less attention. Most of the existing methods give full confidence in the labels. Therefore, models will be misguided by the wrong labels and provide disappointing outcomes for predictions (Bootkrajang & Kabán, 2012). Despite the performance deterioration, the presence of noisy labels also increase the required number of training data samples as well as the model complexity. To avoid such perturbs, it is better to first characterize noise generation. Like the missing data, generation of label noises can be classified by three probabilistic models[235]: Noisy Completely at Random model (NCR), Noisy at Random model (NAR) and Noisy Not at Random model (NNR). Among them, the NAR label noise is the most commonly studied case. According to Frénay and Verleysen (2014), three general techniques have been summed up to address label noises: label noise cleaning, label noise-robust and label noise-tolerant methods. Specifically, the cleaning will cleanse training data by heuristically detecting mislabeled instances (like using the nearest neighborhood methods); the label noise-robust methods use regularizations to achieve better label noise-robustness (like using robust cost functions); while the label noise-tolerant approaches modify learning algorithms to reduce the influence of label noise (like using robust Bayesian methods). Recently, there was also a two-stage study which combined robust strategies of cleaning and regularization for better robust classification performances (Sabzevari, Martínez-Muñoz, & Suárez, 2018). Nevertheless, one can judge that the basic philosophy for dealing with noisy labels are very similar to the feature noise cases, except that they are dealing with the issues of label inconsistencies. Currently, the researches and robust methodologies regarding label noise are mainly designed for areas like medical analysis and image recognition (De Lusignan et al., 2010; García-Zattera, Mutsvari, Jara, Declerck, & Lesaffre, 2010; Liu & Tao, 2016; Xiao, Xia, Yang, Huang, & Wang, 2015). And it still remains an open problem in industrial processes when the process data will contain mixed outliers and also missing data entries in both explanatory and response variables. With this end in view, there are at least two promising directions which should call for further explorations. On the one hand, one should consider how to identify those feature and label outliers and then make the reasonable tolerance, correctness or recovery for both sides. On the other hand, it is also interesting on developing the dual noise-robust (i.e. both feature and label noise-robust) data mining mechanisms, preferably also in the semi-supervised framework so as to incorporate missing components.

5.4. Robust tensor analysis and batch process big data monitoring

In some industrial applications, data sets are indexed by the multidimensional arrays (or multi-way structure) which give rise to the concept of tensor (Lu, Plataniotis, & Venetsanopoulos, 2011; Zare, Ozdemir, Iwen, & Aviyente, 2018). Each dimension of a tensor is called a mode and the number of all dimensions (or ways) is the tensor order (also degree or rank). Particularly, the scalar is a zero-order tensor, the vector is a one-order tensor and the matrix is a two-order tensor. Frequently, tensors with order three or more can be called high-order tensors. For instance, in the field of industrial batch processes, data chunks are organized with the three-way ar-

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

rays or the three-order tensor as Batch × Measurements × Time. Besides process control, high-order tensors can be found in a wide range of industrial backgrounds. To see this, a recent interdisciplinary survey is recommended in which the authors summarized a bunch of tensor-based anomaly detection examples (Fanaee-T & Gama, 2016).

The mining of massive tensor data conforms the big data challenge as tensors are very high dimensional, high volume, along with a large amount of redundancy. Traditional unfolding methods will break the original correlation structure and also incur a large number of estimation parameters for analyzing the large flattened matrix. Instead, tensor theories with multilinear algebras can preserve the high-order structure while extracting the underlying interactions in each dimension (De Lathauwer, De Moor, & Vandewalle, 2000). Among them, the most well-known tensor-based subspace decomposition methods should be the Tucker decomposition and the canonical decomposition (CANDECOMP). The latter is also known as parallel factors (PARAFAC) decomposition (Lu, Plataniotis, & Venetsanopoulos, 2013). These two decompositions bear some equivalence and both have been widely used for batch process understanding and monitoring (Andersson & Bro, 2000; Faber, Bro, & Hopke, 2003; Meng et al., 2003; Smilde, 2001). On the basis of them, some higher-order singular value decomposition (HOSVD) solutions have been further developed for multilinear analysis of tensors (Lu et al., 2013). For example, one can find multilinear PCA and multilinear PLS for tensor-based monitoring and quality predictions (Guo, Li, Wang, & Zeng, 2010; Luo, Bao, Mao, & Tang, 2016). For those nonlinear and dynamic systems, developments have been introduced by employing multilinear graph embedding and the dynamic tensor embedding (Hu & Yuan, 2009; Xiaoqiang & Tao, 2017). In those reports, one can see that tensor analysis have already achieved some success in those industrial applications. From the robust viewpoint, however, most of the current tensor analysis methods will become cumbersome when the original data are partially observed and also corrupted with outliers. To overcome such limitations, the robust tensor analysis is necessary. In fact, from recent data mining literatures, one can find several related works on tensor completion (Romera-Paredes & Pontil, 2013; Tan et al., 2013), robust tensor factorization (Zhao, Zhou, Zhang, Cichocki, & Amari, 2016) and robust low-rank tensor recovery (Yang, Feng, & Suykens, 2016; Javed et al., 2015). Some of them also seem to be very promising. For example, in the work by Zhao et al. (2016) the Bayesian robust tensor factorization offered a unified probabilistic manner for outlier and missing data treatment and has claimed a wide application ranges covering anomaly detection, video background modeling and facial image denoising. Nevertheless, robust tensor analysis only appear in data mining literatures and seldom demonstrations can be found in industrial batch processes. For industrial batch process applications, the robust tensor analysis framework have to be built based on the comprehensive examination of irregular, asymmetric and unsteady information from different tensor modes. The irregular information refers to those outliers and missing data from the measurement mode as well as the uneven batch length from the time mode. Besides the irregular data problem, asymmetry tensor information exists and different modes provide different aspects for process understandings. For dimension reductions and feature extractions, one has to handle the repetitiveness in batch mode direction and also the variable colinearity in measurement mode. While for two-dimensional modeling and monitoring, dynamics should be captured simultaneously from both batch mode and time mode (Lu, Yao, Gao, & Wang, 2005). Finally, the unsteady issue points to the various nonlinear relationships owing to the successive working phases of each batch running (Zhu & Gao, 2018a). Indeed, from the data mining scope, the advances from robust matrix analysis to robust tensor analysis supply a good chance to problem solving, which serve as the important part towards intelligent batch process manufacturing, and in some extent will set the pace of research field achievement during the big data age.

5.5. Robust information fusion and multi-view monitoring

With the rapid development of smart devices and sensing technologies, there are more varieties on data formats generated with audios, images and even video streams. These sensors and devices have been widely installed in industrial systems as both the quantity and richness of industrial data will offer the potential to improve the range and quality of data mining techniques for multi-view statistical process monitoring. For example, as an emerging research field, the audio signal from embedded microphones has been widely used for the noninvasive condition monitoring in machines (Henriquez, Alonso, Ferrer, & Travieso, 2014). By extracting, mining and discriminating audio signals, one can diagnose the current status of machine degradation. The audio-based diagnosis rules are designed for monitoring with 'listening' while those image and video based methods provide the subsidiary action with 'watching'. As can be seen, images and videos are playing more and more important roles in industrial systems as they can offer the persistent and stable vision-based monitoring performances where humans can be inaccessible. Some of the vision-based monitoring applications are blast furnace (Birk, Marklund, & Medvedev, 2002), copper scraps smelting process (Zhang, Ge, Ye, & Song, 2015), polymer composites testing (Yao, Sfarra, Ibarra-Castanedo, You, & Maldague, 2017), visual servoing systems of robotics (Van et al., 2016) and railway systems (Karakose, Gencoglu, Karakose, Aydin, & Akin, 2017), etc. In some cases, information from multi-sensors could be gathered together for intelligent monitoring. For example, the vision sensors and touching sensors of temperature and pressure should work collaboratively so as to make the multi-view monitoring of blast furnace. While in polymer composites testing, the thermograph and ultrasonic signals might be combined together for the better defect monitoring in products. As the signals from different sensors will encounter imperfections like ambiguities, conflicts, incompleteness, outliers and disorders (Khaleghi, Khamis, Karray, & Razavi, 2013). Therefore, the critical problem will be how to make the robust fusion from multisensory data sources. In machine learning literatures, there are some relevant works on robust fusion with the incorporation of uncertainty theories like probabilistic theory, fuzzy set theory and rough theory (Elouedi, Mellouli, & Smets, 2004). However, there is still no common and well-established robust platform on dealing with the concurrent multi-view data imperfections in industrial processes. Therefore, as the multi-view monitoring facilities are gradually becoming the commonplace in various industrial applications, tremendous explorations will be required on the robust handling perspectives.

6. Conclusion

This paper presented a review on state-of-the-art principals and ways for robust data mining techniques so as to deal with process outliers and missing data in various industrial application backgrounds. Detailed descriptions and comparisons have been illustrated on data preprocessing approaches including data cleaning and normalizations. Based on that, the robust statistical process modeling techniques have been introduced and discussed on the core basis of PCA. Afterwards, several robust perspectives have been carried out on statistical data mining methods through different process characteristics including nonlinear, non-Gaussian and dynamics. At last, recent challenges have been presented in the ongoing prevalent scope regarding robust mining from industrial big data.

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

Based on this exposition, it is clear that the traditional statistical modeling methods can be elegantly extended for mining various robust process data scenarios with different properties. The robust data-driven methods have been and will continue to act as the driving force with the ever-increasing interest on big data analytics. It is our hope for this paper to serve as a taxonomy and also a tutorial of advances elucidated from a multitude of works on robust data mining, and to provide the process system engineering community with a picture of the contemporary state and matters for future endeavors.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (NSFC)(61722310), the Natural Science Foundation of Zhejiang Province (LR18F030001), the Fundamental Research Funds for the Central Universities2018XZZX002-09. Zhu and Gao also acknowledge the fund support by the Hong Kong Research Grant Council under project number 16233316, and the National Natural Science Foundation of China (no. 61433005).

References

Adankon, M. M., Cheriet, M., & Biem, A. (2009). Semisupervised least squares support vector machine. *IEEE Transactions on Neural Networks, 20*(12), 1858–1870.
Agamennoni, G., Nieto, J. I. &, & Nebot, E. M. (2012). Approximate inference in state-space models with heavy-tailed noise. *Signal Processing, IEEE Transactions on, 60*(10), 5024–5037.
Agyemang, M., Barker, K., & Alhajj, R. (2006). A comprehensive survey of numeric and symbolic outlier mining techniques. *Intelligent Data Analysis, 10*(6), 521–538.
Alham, N. K., Li, M., Liu, Y., & Hammoud, S. (2011). A MapReduce-based distributed SVM algorithm for automatic image annotation. *Computers & Mathematics with Applications, 62*(7), 2801–2811.
Ali, Y., & Narasimhan, S. (1995). Redundant sensor network design for linear processes. *AIChE Journal, 41*(10), 2237–2249.
Alkan, B. B., Atakan, C., & Alkan, N. (2015). A comparison of different procedures for principal component analysis in the presence of outliers. *Journal of Applied Statistics, 42*(8), 1716–1722.
Andersson, C. A., & Bro, R. (2000). The N-way toolbox for MATLAB. *Chemometrics and Intelligent Laboratory Systems, 52*(1), 1–4.
Angiulli, F., & Pizzuti, C. (2005). Outlier mining in large high-dimensional data sets. *IEEE Transactions on Knowledge and Data Engineering, 17*(2), 203–215.
Aravkin, A., Burke, J. V., Ljung, L., Lozano, A., & Pillonetto, G. (2017). Generalized Kalman smoothing: Modeling and algorithms. *Automatica, 86*, 63–86.
Aravkin, A. Y., Bell, B. M., Burke, J. V., & Pillonetto, G. (2011). An L1-Laplace Robust Kalman Smoother. *IEEE Transactions on Automatic Control, 56*(12), 2898–2911.
Archambeau, C., Delannay, N., & Verleysen, M. (2008). Mixtures of robust probabilistic principal component analyzers. *Neurocomputing, 71*(7), 1274–1282.
Aydilek, I. B., & Arslan, A. (2013). A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. *Information Sciences, 233*, 25–35.
Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press.
Barnett, V., & Lewis, T. (1974). *Outliers in statistical data*. Wiley.
Birk, W., Marklund, O., & Medvedev, A. (2002). Video monitoring of pulverized coal injection in the blast furnace. *IEEE Transactions on Industry Applications, 38*(2), 571–576.
Blum, A. L., & Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence, 97*(1-2), 245–271.
Bootkrajang, J., & Kabán, A. (2012). In Label-noise robust logistic regression and its applications. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 143–158). Springer: 2012.
Borzeshi, E. Z., Concha, O. P., Da Xu, R. Y., & Piccardi, M. (2013). Joint action segmentation and classification by an extended hidden Markov model. *IEEE Signal Processing Letters, 20*(12), 1207–1210.
Boukouvala, F., Muzzio, F. J., & Ierapetritou, M. G. (2010). Predictive modeling of pharmaceutical processes with missing and noisy data. *AIChE Journal, 56*(11), 2860–2872.
Bouveyron, C., & Girard, S. (2009). Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition, 42*(11), 2649–2658.
Bouwmans, T., Aybat, N. S., & Zahzah, E.-H. (2016). *Handbook of robust low-rank and sparse matrix decomposition: Applications in image and video processing*. Chapman and Hall/CRC.
Bouwmans, T., Javed, S., Zhang, H., Lin, Z., & Otazo, R. (2018). On the applications of Robust PCA in image and video processing. *Proceedings of the IEEE, 106*(8), 1427–1457.
Bouwmans, T., Sobral, A., Javed, S., Jung, S. K., & Zahzah, E.-H. (2017). Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset. *Computer Science Review, 23*, 1–71.
Bouwmans, T., & Zahzah, E. H. (2014). Robust PCA via principal component pursuit: A review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding, 122*, 22–34.
Brahma, P. P., She, Y., Li, S., Li, J., & Wu, D. (2018). Reinforced robust principal component pursuit. *IEEE Transactions on Neural Networks and Learning Systems, 29*(5), 1525–1538.
Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). In LOF: Identifying density-based local outliers. In *ACM Sigmod Record* (pp. 93–104). ACM.
Cai, L., & Tian, X. (2014). A new fault detection method for non-Gaussian process based on robust independent component analysis. *Process Safety and Environmental Protection, 92*(6), 645–658.
Campbell, N. A. (1980). Robust procedures in multivariate analysis I: Robust covariance estimation. *Applied Statistics*, 231–237.
Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM), 58*(3), 11.
Cao, J., Murata, N., Amari, S.-i., Cichocki, A., & Takeda, T. (2003). A robust approach to independent component analysis of signals with high-level noise measurements. *IEEE Transactions on Neural Networks, 14*(3), 631–645.
Censi, A. (2011). Kalman filtering with intermittent observations: Convergence for semi-Markov chains and an intrinsic performance measure. *IEEE Transactions on Automatic Control, 56*(2), 376–381.
Chan, K., Lee, T.-W., & Sejnowski, T. J. (2003). Variational Bayesian learning of ICA with missing data. *Neural Computation, 15*(8), 1991–2011.
Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., & Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization, 21*(2), 572–596.
Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16–28.
Chang, L., Hu, B., Chang, G., & Li, A. (2012). Multiple outliers suppression derivative-free filter based on unscented transformation. *Journal of Guidance, Control, and Dynamics, 35*(6), 1902–1906.
Chatzis, S. P., Kosmopoulos, D. I., & Varvarigou, T. A. (2008). Signal modeling and classification using a robust latent space model based on $ t $ distributions. *IEEE Transactions on Signal Processing, 56*(3), 949–963.
Chatzis, S. P., Kosmopoulos, D. I., & Varvarigou, T. A. (2009). Robust sequential data modeling using an outlier tolerant hidden Markov model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 31*(9), 1657–1669.
Chen, C., Yan, C., & Li, Y. (2015). A robust weighted least squares support vector regression based on least trimmed squares. *Neurocomputing, 168*, 941–946.
Chen, C. P., & Zhang, C.-Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences, 275*, 314–347.
Chen, P., Hung, H., Komori, O., Huang, S.-Y., & Eguchi, S. (2013). Robust independent component analysis via minimum $\gamma $-divergence estimation. *IEEE Journal of Selected Topics in Signal Processing, 7*(4), 614–624.
Chen, T., Martin, E., & Montague, G. (2009). Robust probabilistic PCA with missing data and contribution analysis for outlier detection. *Computational Statistics & Data Analysis, 53*(10), 3706–3716.
Chen, X., Yang, J., Liang, J., & Ye, Q. (2012). Recursive robust least squares support vector regression based on maximum correntropy criterion. *Neurocomputing, 97*, 63–73.
Chen, X.-W., Anantha, G., & Wang, X. (2006). An effective structure learning method for constructing gene networks. *Bioinformatics, 22*(11), 1367–1374.
Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*(3), 287–314.
Croux, C., Filzmoser, P., & Oliveira, M. R. (2007). Algorithms for projection–pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems, 87*(2), 218–225.
Croux, C., & Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *Biometrika, 87*(3), 603–618.
Croux, C., & Ruiz-Gazen, A. (1996). In A fast algorithm for robust principal components based on projection pursuit. *Compstat*, 211–216.
Croux, C., & Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: The projection-pursuit approach revisited. *Journal of Multivariate Analysis, 95*(1), 206–226.
Daszykowski, M., Kaczmarek, K., Vander Heyden, Y., & Walczak, B. (2007). Robust statistics in data analysis – A review: Basic concepts. *Chemometrics and Intelligent Laboratory Systems, 85*(2), 203–219.
Daum, F. (2005). Nonlinear filters: Beyond the Kalman filter. *IEEE Aerospace and Electronic Systems Magazine, 20*(8), 57–69.
De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000). A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications, 21*(4), 1253–1278.
De Lusignan, S., Khunti, K., Belsey, J., Hattersley, A., Van Vlymen, J., Gallagher, H., et al. (2010). A method of identifying and correcting miscoding, misclassification and misdiagnosis in diabetes: A pilot and validation study of routinely collected data. *Diabetic Medicine, 27*(2), 203–209.
Delchambre, L. (2014). Weighted principal component analysis: A weighted covariance eigendecomposition approach. *Monthly Notices of the Royal Astronomical Society, 446*(4), 3545–3555.
Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1981). Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association, 76*(374), 354–362.
Ding, J., & Shah, S. (2013). A robust hidden semi-Markov model with application

to aCGH data processing. *International Journal of Data Mining and Bioinformatics, 8*(4), 427–442.
Ding, X., He, L., & Carin, L. (2011). Bayesian robust principal component analysis. *IEEE Transactions on Image Processing, 20*(12), 3419–3430.
Dong, Y., & Peng, C.-Y. J. (2013). Principled missing data methods for researchers. *SpringerPlus, 2*(1), 1–17.
Du, Y., Lu, J.-p., Li, Q., & Deng, Y. (2008). Short-term wind speed forecasting of wind farm based on least square-support vector machine. *Power System Technology, 32*(15), 62–66.
Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. New York: John Wiley & Sons.
Elgamal, T., Yabandeh, M., Aboulnaga, A., Mustafa, W., & Hefeeda, M. (2015). In spca: Scalable principal component analysis for big data on distributed platforms. In *Proceedings of the 2015 ACM SIGMOD international conference on management of data* (pp. 79–91). ACM: 2015.
Elouedi, Z., Mellouli, K., & Smets, P. (2004). Assessing sensor reliability for multisensor data fusion within the transferable belief model. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34*(1), 782–787.
Faber, N. K. M., Bro, R., & Hopke, P. K. (2003). Recent developments in CANDECOMP/PARAFAC algorithms: A critical review. *Chemometrics and Intelligent Laboratory Systems, 65*(1), 119–137.
Fanaee-T, H., & Gama, J. (2016). Tensor-based anomaly detection: An interdisciplinary survey. *Knowledge-Based Systems, 98*, 130–147.
Fatehi, A., & Huang, B. (2017). Kalman filtering approach to multi-rate information fusion in the presence of irregular sampling rate and variable measurement delay. *Journal of Process Control, 53*, 15–25.
Filzmoser, P., & Todorov, V. (2011). Review of robust multivariate statistical methods in high dimension. *Analytica Chimica Acta, 705*(1-2), 2–14.
Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE transactions on Neural Networks and Learning Systems, 25*(5), 845–869.
Gandhi, M. A., & Mili, L. (2010). Robust Kalman filter based on a generalized maximum-likelihood-type estimator. *IEEE Transactions on Signal Processing, 58*(5), 2509–2520.
Gao, J. (2008). Robust L1 principal component analysis and its Bayesian variational inference. *Neural Computation, 20*(2), 555–572.
Gao, J., & Xu, R. Y. (2007). In Mixture of the robust L1 distributions and its applications. In *Australasian joint conference on artificial intelligence* (pp. 26–35). Springer: 2007.
García-Zattera, M. J., Mutsvari, T., Jara, A., Declerck, D., & Lesaffre, E. (2010). Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in Medicine, 29*(30), 3103–3117.
Ge, Z. (2017). Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems, 171*(Supplement C), 16–25.
Ge, Z. (2018). Process data analytics via probabilistic latent variable models: A tutorial review. *Industrial & Engineering Chemistry Research, 57*(38), 12646–12661.
Ge, Z., & Song, Z. (2007). Process monitoring based on independent component analysis– principal component analysis (ICA– PCA) and similarity factors. *Industrial & Engineering Chemistry Research, 46*(7), 2054–2063.
Ge, Z., & Song, Z. (2010a). Nonlinear probabilistic monitoring based on the Gaussian process latent variable model. *Industrial & Engineering Chemistry Research, 49*(10), 4792–4799.
Ge, Z., & Song, Z. (2010b). A comparative study of just-in-time-learning based methods for online soft sensor modeling. *Chemometrics and Intelligent Laboratory Systems, 104*(2), 306–317.
Ge, Z., & Song, Z. (2011). Robust monitoring and fault reconstruction based on variational inference component analysis. *Journal of Process Control, 21*(4), 462–474.
Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics in the process industry: The role of machine learning. *IEEE Access, 5*, 20590–20616.
Ge, Z., Song, Z., & Gao, F. (2013a). Review of recent research on data-based process monitoring. *Industrial & Engineering Chemistry Research, 52*(10), 3543–3562.
Ge, Z., & Zhihuan, S. (2013). Non-Gaussian Process Monitoring. In *Multivariate Statistical Process Control.* (pp. 13–27). London: Springer.
Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology, 60*, 549–576.
Guha, S., Li, Y., & Neuberg, D. (2008). Bayesian hidden Markov modeling of array CGH data. *Journal of the American Statistical Association, 103*(482), 485–497.
Gul, M., & Catbas, F. N. (2009). Statistical pattern recognition for structural health monitoring using time series modeling: Theory and experimental verifications. *Mechanical Systems and Signal Processing, 23*(7), 2192–2204.
Guo, J., Li, Y., Wang, G., & Zeng, J. (2010). In Batch Process monitoring based on multilinear principal component analysis. In *Intelligent system design and engineering application (ISDEA), 2010 international conference on, 2010* (pp. 413–416). IEEE.
Guo, Y., Zhao, Y., & Huang, B. (2014). Development of soft sensor by incorporating the delayed infrequent and irregular measurements. *Journal of Process Control, 24*(11), 1733–1739.
Gustafsson, F., & Hendeby, G. (2012). Some relations between extended and unscented Kalman filters. *IEEE Transactions on Signal Processing, 60*(2), 545–555.
Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research, 3*(Mar), 1157–1182.
Hammes, U., Wolsztynski, E., & Zoubir, A. M. (2009). Robust tracking and geolocation for wireless networks in NLOS environments. *IEEE Journal of Selected Topics in Signal Processing, 3*(5), 889–901.
Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust statistics: The approach based on influence functions*: 196. New York: John Wiley & Sons.
Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Waltham, MA.: Elsevier.
Härdle, W., & Simar, L. (2007). *Applied multivariate statistical analysis*: 22007. Springer.
Henriquez, P., Alonso, J. B., Ferrer, M. A., & Travieso, C. M. (2014). Review of automatic fault diagnosis systems using audio and vibration signals. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 44*(5), 642–652.
Hodge, V. J., & Austin, J. (2004b). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*(2), 85–126.
Horata, P., Chiewchanwattana, S., & Sunat, K. (2013). Robust extreme learning machine. *Neurocomputing, 102*, 31–44.
Hsiao, F.-H., & Pan, S.-T. (1996). Robust Kalman filter synthesis for uncertain multiple time-delay stochastic systems. *Journal of Dynamic Systems, Measurement, and Control, 118*(4), 803–808.
Hsu, C.-C., Chen, L.-S., & Liu, C.-H. (2010). A process monitoring scheme based on independent component analysis and adjusted outliers. *International Journal of Production Research, 48*(6), 1727–1743.
Hu, K., & Yuan, J. (2009). Batch process monitoring with tensor factorization. *Journal of Process Control, 19*(2), 288–296.
Huang, G., Song, S., Gupta, J. N., & Wu, C. (2014). Semi-supervised and unsupervised extreme learning machines. *IEEE Transactions on Cybernetics, 44*(12), 2405–2417.
Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42*(2), 513–529.
Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2006). Extreme learning machine: Theory and applications. *Neurocomputing, 70*(1-3), 489–501.
Huang, M., & Dey, S. (2007). Stability of Kalman filtering with Markovian packet losses. *Automatica, 43*(4), 598–607.
Huang, Y., Zhang, Y., Li, N., & Chambers, J. (2016). Robust Student'st based nonlinear filter and smoother. *IEEE Transactions on Aerospace and Electronic Systems, 52*(5), 2586–2596.
Huang, Y., Zhang, Y., Li, N., Wu, Z., & Chambers, J. A. (2017a). A novel robust student's t-based Kalman filter. *IEEE Transactions on Aerospace and Electronic Systems, 53*(3), 1545–1554.
Huang, Y., Zhang, Y., Xu, B., Wu, Z., & Chambers, J. (2017b). A new outlier-robust student's t based Gaussian approximate filter for cooperative localization. *IEEE/ASME Transactions on Mechatronics, 22*(5), 2380–2386.
Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, 435–475.
Hubert, M., & Debruyne, M. (2010). Minimum covariance determinant. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(1), 36–43.
Hubert, M., Rousseeuw, P. J., & Vanden Branden, K. (2005). ROBPCA: A new approach to robust principal component analysis. *Technometrics, 47*(1), 64–79.
Hubert, M., Rousseeuw, P. J., & Verboven, S. (2002). A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems, 60*(1-2), 101–111.
Hyvärinen, A., Karhunen, J., & Oja, E. (2004). *Independent component analysis*: 46. John Wiley & Sons.
Imtiaz, S., & Shah, S. (2008). Treatment of missing values in process data analysis. *The Canadian Journal of Chemical Engineering, 86*(5), 838–858.
Isermann, R. (1984). Process fault detection based on modeling and estimation methods – A survey. *Automatica, 20*(4), 387–404.
Jain, A., Nandakumar, K., & Ross, A. (2005). Score normalization in multimodal biometric systems. *Pattern Recognition, 38*(12), 2270–2285.
Javed, S., Bouwmans, T., & Jung, S. K. (2015). In Stochastic decomposition into low rank and sparse tensor for robust background subtraction. *6th International conference on imaging for crime prevention and detection (ICDP-15)*.
Javed, S., Jung, S. K., Mahmood, A., & Bouwmans, T. (2016). In Motion-aware graph regularized RPCA for background modeling of complex scenes. In *Pattern recognition (ICPR), 2016 23rd international conference on* (pp. 120–125). IEEE: 2016.
Julier, S. J., & Uhlmann, J. K. (2004). Unscented filtering and nonlinear estimation. *Proceedings of the IEEE, 92*(3), 401–422.
Jylänki, P., Vanhatalo, J., & Vehtari, A. (2011). Robust Gaussian process regression with a Student-t likelihood. *Journal of Machine Learning Research, 12*(Nov), 3227–3257.
Kadlec, P., Gabrys, B., & Strandt, S. (2009). Data-driven soft sensors in the process industry. *Computers & Chemical Engineering, 33*(4), 795–814.
Kadlec, P., Grbić, R., & Gabrys, B. (2011). Review of adaptation mechanisms for data–driven soft sensors. *Computers & Chemical Engineering, 35*(1), 1–24.
Kaneda, Y., Irizuki, Y., & Yamakita, M. (2013). In Robust unscented Kalman filter via l 1 regression and design method of its parameters. In *Control conference (ASCC), 2013 9th Asian* (pp. 1–6). IEEE: 2013.
Karakose, E., Gencoglu, M. T., Karakose, M., Aydin, I., & Akin, E. (2017). A new experimental approach using image processing-based tracking for an efficient fault diagnosis in pantograph–catenary systems. *IEEE Transactions on Industrial Informatics, 13*(2), 635–643.
Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics, 14*(10), 846–856.
Kautz, T., & Eskofier, B. M. (2015). A robust Kalman framework with resampling and optimal smoothing. *Sensors, 15*(3), 4975–4995.
Khaleghi, B., Khamis, A., Karray, F. O., & Razavi, S. N. (2013). Multisensor data fusion: A review of the state-of-the-art. *Information Fusion, 14*(1), 28–44.
Khatibisepehr, S., Huang, B., & Khare, S. (2013). Design of inferential sensors in the process industry: A review of Bayesian methods. *Journal of Process Control, 23*(10), 1575–1596.

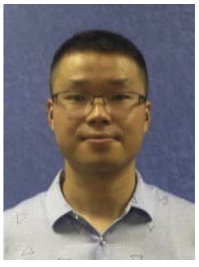
Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

Kim, M. H., Lee, S., & Lee, K. C. (2010). Kalman predictive redundancy system for fault tolerance of safety-critical systems. *IEEE Transactions on Industrial Informatics, 6*(1), 46–53.
Kluge, S., Reif, K., & Brokate, M. (2010). Stochastic stability of the extended Kalman filter with intermittent observations. *IEEE Transactions on Automatic Control, 55*(2), 514–518.
Kodamana, H., Huang, B., Ranjan, R., Zhao, Y., Tan, R., & Sammaknejad, N. (2018). Approaches to robust process identification: A review and tutorial of probabilistic methods. *Journal of Process Control, 66*, 68–83.
Koskela, T. (2003). Neural network methods in analysing and modelling time varying processes. *Helsinki University of Technology*.
Kriegel, H.-P., & Zimek, A. (2008). In Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 444–452). ACM: 2008.
Kwak, N., & Choi, C.-H. (2002). Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(12), 1667–1671.
Latif-Shabgahi, G., Bass, J. M., & Bennett, S. (2004). A taxonomy for software voting algorithms used in safety-critical systems. *IEEE Transactions on Reliability, 53*(3), 319–328.
Lawrence, N. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *Journal of Machine Learning Research, 6*(Nov), 1783–1816.
Lee, H., Kang, S., & Han, S. (2018). Real-time optimal state estimation scheme with delayed and periodic measurements. *IEEE Transactions on Industrial Electronics, 65*(7), 5970–5978.
Lee, J.-M., Yoo, C., & Lee, I.-B. (2004a). Statistical process monitoring with independent component analysis. *Journal of Process Control, 14*(5), 467–485.
Lee, T.-W., Girolami, M., & Sejnowski, T. J. (1999). Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural Computation, 11*(2), 417–441.
Lee, Y., Yoo, C., & Lee, I.-B. (2004b). Statistical process monitoring with independent component analysis. *Journal of Process Control, 14*, 467–485.
Lerman, G.; Maunu, T. (2018)An overview of robust subspace recovery. *arXiv preprint*.
Li, G., & Chen, Z. (1985). Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *Journal of the American Statistical Association, 80*(391), 759–766.
Li, L., Su, X., Wang, Y., Lin, Y., Li, Z., & Li, Y. (2015). Robust causal dependence mining in big data network and its application to traffic flow predictions. *Transportation Research Part C: Emerging Technologies, 58*, 292–307.
Li, L., & Xia, Y. (2012). Stochastic stability of the unscented Kalman filter with intermittent observations. *Automatica, 48*(5), 978–981.
Lin, T. I., Lee, J. C., & Hsieh, W. J. (2007). Robust mixture modeling using the skew t distribution. *Statistics and Computing, 17*(2), 81–92.
Lin, Z.; Chen, M.; Ma, Y. (2010)The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint.
Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M., & Ma, Y. (2009). *Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix* Coordinated Science Laboratory Report no. UILU-ENG-09-2214.
Little, R. J., & Rubin, D. B. (2014). *Statistical analysis with missing data*. John Wiley & Sons.
Liu, T., & Tao, D. (2016). Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 38*(3), 447–461.
Liu, X., Xie, L., Kruger, U., Littler, T., & Wang, S. (2008). Statistical-based monitoring of multivariate non-Gaussian systems. *AIChE Journal, 54*(9), 2379–2391.
Liu, Y., Pan, Y., Sun, Z., & Huang, D. (2014). Statistical monitoring of wastewater treatment plants using variational Bayesian PCA. *Industrial & Engineering Chemistry Research, 53*(8), 3272–3282.
Ljung, L. (1979). Asymptotic behavior of the extended Kalman filter as a parameter estimator for linear systems. *IEEE Transactions on Automatic Control, 24*(1), 36–50.
Lozano, E., & Acufia, E. (2005). In Parallel algorithms for distance-based and density-based outliers. In *Fifth IEEE International Conference on Data Mining (ICDM'05)* (pp. 729–732). 2005.
Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. (2013). *Multilinear subspace learning: Dimensionality reduction of multidimensional data*. Chapman and Hall/CRC.
Lu, H., Plataniotis, K. N., & Venetsanopoulos, A. N. (2011). A survey of multilinear subspace learning for tensor data. *Pattern Recognition, 44*(7), 1540–1551.
Lu, N., & Wang, X. (2008). Fault diagnosis based on signed digraph combined with dynamic kernel PLS and SVR. *Industrial & Engineering Chemistry Research, 47*(23), 9447–9456.
Lu, N., Yao, Y., Gao, F., & Wang, F. (2005). Two-dimensional dynamic PCA for batch process monitoring. *AIChE Journal, 51*(12), 3300–3304.
Luo, L., Bao, S., Mao, J., & Tang, D. (2016). Quality prediction and quality-relevant monitoring with multilinear PLS for batch processes. *Chemometrics and Intelligent Laboratory Systems, 150*, 9–22.
Luttinen, J., Ilin, A., & Karhunen, J. (2012). Bayesian robust PCA of incomplete data. *Neural Processing Letters, 36*(2), 189–202.
Madsen, A. L., Jensen, F., Salmerón, A., Langseth, H., & Nielsen, T. D. (2017). A parallel algorithm for Bayesian network structure learning from large data sets. *Knowledge-Based Systems, 117*, 46–55.
Mardani, M., Mateos, G., & Giannakis, G. B. (2013). Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalies. *IEEE Transactions on Information Theory, 59*(8), 5186–5205.
Maronna, R. (2005). Principal components and orthogonal regression based on robust scales. *Technometrics, 47*(3), 264–273.
Masreliez, C., & Martin, R. (1977). Robust Bayesian estimation for the linear model and robustifying the Kalman filter. *IEEE transactions on Automatic Control, 22*(3), 361–371.
Meng, X., Morris, A., & Martin, E. (2003). On-line monitoring of batch processes using a PARAFAC representation. *Journal of Chemometrics, 17*(1), 65–81.
Mihoko, M., & Eguchi, S. (2002). Robust blind source separation by beta divergence. *Neural Computation, 14*(8), 1859–1886.
Moon, T. K. (1996). The expectation-maximization algorithm. *IEEE Signal Processing Magazine, 13*(6), 47–60.
Mu, Y., Dong, J., Yuan, X., & Yan, S. (2011). In Accelerated low-rank visual recovery by random projection. In *Computer vision and pattern recognition (CVPR), IEEE conference on* (pp. 2609–2616). IEEE.
Ning, K., Liu, M., & Dong, M. (2015). A new robust ELM method based on a Bayesian framework with heavy-tailed distribution and weighted likelihood function. *Neurocomputing, 149*, 891–903.
Nurminen, H., Ardeshiri, T., Piche, R., & Gustafsson, F. (2015). Robust inference for state-space models with skewed measurement noise. *IEEE Signal Processing Letters, 22*(11), 1898–1902.
Ould-Bouamama, B., El Harabi, R., Abdelkrim, M. N., & Gayed, M. B. (2012). Bond graphs for the diagnosis of chemical processes. *Computers & Chemical Engineering, 36*, 301–324.
Pan, Y., Yang, C., An, R., & Sun, Y. (2016). Fault detection with improved principal component pursuit method. *Chemometrics and Intelligent Laboratory Systems, 157*, 111–119.
Pearson, R. K. (2002). Outliers in process modeling and identification. *IEEE Transactions on Control Systems Technology, 10*(1), 55–63.
Peel, D., & McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing, 10*(4), 339–348.
Pelckmans, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2005). Handling missing values in support vector machine classifiers. *Neural Networks, 18*(5-6), 684–692.
Peng, H., & Zhu, S. (2007). Handling of incomplete data sets using ICA and SOM in data mining. *Neural Computing and Applications, 16*(2), 167–172.
Qin, S. J. (2012). Survey on data-driven industrial process monitoring and diagnosis. *Annual Reviews in Control, 36*(2), 220–234.
Qin, S. J. (2014). Process data analytics in the era of big data. *AIChE Journal, 60*(9), 3092–3100.
Qin, S. J., Cherry, G., Good, R., Wang, J., & Harrison, C. A. (2006). Semiconductor manufacturing process control and monitoring: A fab-wide framework. *Journal of Process Control, 16*(3), 179–191.
Rasmussen, C. E. (2004). Gaussian processes in machine learning. In *Advanced lectures on machine learning* (pp. 63–71). Springer.
Romera-Paredes, B., & Pontil, M. (2013). A new convex relaxation for tensor completion. In *Advances in neural information processing systems* (pp. 2967–2975). 2013.
Rosca, J., Gerkmann, T., & Balcan, D.-C. (2006). In Statistical inference of missing speech data in the ICA domain. In *Proceedings of the ICASSP, 2006* (pp. 617–620).
Ross, S. M. (2003). Peirce's criterion for the elimination of suspect experimental data. *Journal of Engineering Technology, 20*(2), 38–41.
Rousseeuw, P. J., & Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics, 41*(3), 212–223.
Roweis, S., & Ghahramani, Z. (1999). A unifying review of linear Gaussian models. *Neural Computation, 11*(2), 305–345.
Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*: 81. New York: John Wiley & Sons.
Sabzevari, M., Martínez-Muñoz, G., & Suárez, A. (2018). A two-stage ensemble method for the detection of class-label noise. *Neurocomputing, 275*, 2374–2383.
Salazar, A., Vergara, L., Serrano, A., & Igual, J. (2010). A general procedure for learning mixtures of independent component analyzers. *Pattern Recognition, 43*(1), 69–85.
Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.
Sen, S., Narasimhan, S., & Deb, K. (1998). Sensor network design of linear processes using genetic algorithms. *Computers & Chemical Engineering, 22*(3), 385–390.
Serneels, S., & Verdonck, T. (2008). Principal component analysis for data containing outliers and missing elements. *Computational Statistics & Data Analysis, 52*(3), 1712–1727.
Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. Z. (2017). A survey on semi-supervised feature selection methods. *Pattern Recognition, 64*, 141–158.
Sinopoli, B., Schenato, L., Franceschetti, M., Poolla, K., Jordan, M. I., & Sastry, S. S. (2004). Kalman filtering with intermittent observations. *IEEE transactions on Automatic Control, 49*(9), 1453–1464.
Skočaj, D., Leonardis, A., & Bischof, H. (2007). Weighted and robust learning of subspace representations. *Pattern Recognition, 40*(5), 1556–1569.
Slavakis, K., Giannakis, G. B., & Mateos, G. (2014). Modeling and optimization for big data analytics: (Statistical) learning tools for our era of data deluge. *IEEE Signal Processing Magazine, 31*(5), 18–31.
Smilde, A. K. (2001). Comments on three-way analyses used for batch process data. *Journal of Chemometrics, 15*(1), 19–27.
Smyth, A., & Wu, M. (2007). Multi-rate Kalman filtering for the data fusion of displacement and acceleration response measurements in dynamic system monitoring. *Mechanical Systems and Signal Processing, 21*(2), 706–723.
Soria-Olivas, E., Gomez-Sanchis, J., Martín, J. D., Vila-Francés, J., Martinez, M., & Mag-

dalena, J. R. (2011). BELM: Bayesian extreme learning machine. *IEEE Transactions on Neural Networks, 22*(3), 505–509.
Sovilj, D., Eirola, E., Miche, Y., Björk, K.-M., Nian, R., et al. (2016). Extreme learning machine for missing data using multiple imputations. *Neurocomputing, 174*, 220–231.
Stanimirova, I., Daszykowski, M., & Walczak, B. (2007). Dealing with missing values and outliers in principal component analysis. *Talanta, 72*(1), 172–178.
Stegle, O., Fallert, S. V., MacKay, D. J., & Brage, S. (2008). Gaussian process robust regression for noisy heart rate data. *IEEE Transactions on Biomedical Engineering, 55*(9), 2143–2151.
Suykens, J. A., De Brabanter, J., Lukas, L., & Vandewalle, J. (2002). Weighted least squares support vector machines: Robustness and sparse approximation. *Neurocomputing, 48*(1-4), 85–105.
Tan, H., Feng, G., Feng, J., Wang, W., Zhang, Y.-J., & Li, F. (2013). A tensor-based method for missing traffic data completion. *Transportation Research Part C: Emerging Technologies, 28*, 15–27.
Thissen, U., Üstün, B., Melssen, W. J., & Buydens, L. M. (2004). Multivariate calibration with least-squares support vector machines. *Analytical Chemistry, 76*(11), 3099–3105.
Tidriri, K., Chatti, N., Verron, S., & Tiplica, T. (2016). Bridging data-driven and model-based approaches for process fault diagnosis and health monitoring: A review of researches and future challenges. *Annual Reviews in Control, 42*, 63–81.
Tipping, M. E., & Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61*(3), 611–622.
Undey, C., & Cinar, A. (2002). Statistical monitoring of multistage, multiphase batch processes. *IEEE Control Systems, 22*(5), 40–52.
Van Aelst, S., & Rousseeuw, P. (2009). Minimum volume ellipsoid. *Wiley Interdisciplinary Reviews: Computational Statistics, 1*(1), 71–82.
Van Gestel, T., Suykens, J. A., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G., et al. (2004). Benchmarking least squares support vector machine classifiers. *Machine Learning, 54*(1), 5–32.
Van, M., Wu, D., Ge, S. S., & Ren, H. (2016). Fault diagnosis in image-based visual servoing with eye-in-hand configurations using kalman filter. *IEEE Transactions on Industrial Informatics, 12*(6), 1998–2007.
Vaswani, N., Bouwmans, T., Javed, S., & Narayanamurthy, P. (2018). Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery. *IEEE Signal Processing Magazine, 35*(4), 32–55.
Vaswani, N., & Narayanamurthy, P. (2018). Static and dynamic robust PCA and matrix completion: A review. In *Proceedings of the IEEE: 106* (pp. 1359–1379).
Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., & Yin, K. (2003). A review of process fault detection and diagnosis: Part III: Process history based methods. *Computers & Chemical Engineering, 27*(3), 327–346.
Verboven, S., & Hubert, M. (2005). LIBRA: A MATLAB library for robust analysis. *Chemometrics and Intelligent Laboratory Systems, 75*(2), 127–136.
Villa, S., & Rossetti, M. (2014). Learning continuous time bayesian network classifiers using mapreduce. *Journal of Statistical Software, 62*(3), 1–25.
Vuković, N., & Miljković, Z. (2015). Robust sequential learning of feedforward neural networks in the presence of heavy-tailed noise. *Neural Networks, 63*, 31–47.
Wang, H., & Banerjee, A. (2013). Online alternating direction method (longer version). *(arXiv preprint*.
Wang, X., Cui, N., & Guo, J. (2010a). Huber-based unscented filtering and its application to vision-based relative navigation. *IET Radar, Sonar & Navigation, 4*(1), 134–141.
Wang, X., Gao, X., Yuan, Y., Tao, D., & Li, J. (2010b). Semi-supervised Gaussian process latent variable model with pairwise constraints. *Neurocomputing, 73*(10-12), 2186–2195.
Wang, X., Li, A., Jiang, Z., & Feng, H. (2006). Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. *BMC Bioinformatics, 7*(1), 32.
Wang, Z. X., He, Q. P., & Wang, J. (2015). Comparison of variable selection methods for PLS-based soft sensor modeling. *Journal of Process Control, 26*, 56–72.
Wanihsuksombat, C., Hongtrakul, V., & Suppakul, P. (2010). Development and characterization of a prototype of a lactic acid–based time–temperature indicator for monitoring food product quality. *Journal of Food Engineering, 100*(3), 427–434.
Welling, M., & Weber, M. (1999). In Independent component analysis of incomplete data. In *1999 6th Joint symposium on neural compuatation proceedings, 1999* (pp. 162–168).
Wen, Q., Ge, Z., & Song, Z. (2012). Data-based linear Gaussian state - space model for dynamic process monitoring. *AIChE Journal, 58*(12), 3763–3776.
Wen, W., Hao, Z., & Yang, X. (2008). A heuristic weight-setting strategy and iteratively updating algorithm for weighted least-squares support vector regression. *Neurocomputing, 71*(16-18), 3096–3103.
Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.
Xiao, T., Xia, T., Yang, Y., Huang, C., & Wang, X. (2015). In Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2691–2699). 2015.
Xiaoqiang, Z., & Tao, W. (2017). Tensor dynamic neighborhood preserving embedding algorithm for fault diagnosis of batch process. *Chemometrics and Intelligent Laboratory Systems, 162*, 94–103.
Xie, Y. L., Wang, J. H., Liang, Y. Z., Sun, L. X., Song, X. H., & Yu, R. Q. (1993). Robust principal component analysis by projection pursuit. *Journal of Chemometrics, 7*(6), 527–541.
Xu, S., Lu, B., Baldea, M., Edgar, T. F., Wojsznis, W., Blevins, T., et al. (2015). Data cleaning in the process industries. *Reviews in Chemical Engineering, 31*(5), 453–490.
Xue, B., Zhang, M., Browne, W. N., & Yao, X. (2016). A survey on evolutionary computation approaches to feature selection. *IEEE Transactions on Evolutionary Computation, 20*(4), 606–626.
Yan, Z., Chen, C.-Y., Yao, Y., & Huang, C.-C. (2016). Robust multivariate statistical process monitoring via stable principal component pursuit. *Industrial & Engineering Chemistry Research, 55*(14), 4011–4021.
Yang, F., & Xiao, D. (2012). Progress in root cause and fault propagation analysis of large-scale industrial processes. *Journal of Control Science and Engineering, 2012*, 1–10.
Yang, J., & Yuan, X. (2013). Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization. *Mathematics of Computation, 82*(281), 301–329.
Yang, X., Tan, L., & He, L. (2014). A robust least squares support vector machine for regression and classification with noise. *Neurocomputing, 140*, 41–52.
Yang, Y., Feng, Y., & Suykens, J. A. (2016). Robust low-rank tensor recovery with regularized redescending M-estimator. *IEEE Transactions on Neural Networks and Learning Systems, 27*(9), 1933–1946.
Yao, L., & Ge, Z. (2018a). Scalable Semi-supervised GMM for Big Data quality prediction in multimode processes. *IEEE Transactions on Industrial Electronics*.
Yao, L., & Ge, Z. (2018b). Big data quality prediction in the process industry: A distributed parallel modeling framework. *Journal of Process Control, 68*, 1–13.
Yao, Y., Sfarra, S., Ibarra-Castanedo, C., You, R., & Maldague, X. P. V. (2017). The multi-dimensional ensemble empirical mode decomposition (MEEMD). *Journal of Thermal Analysis and Calorimetry, 128*(3), 1841–1858.
Yin, S., Ding, S. X., Haghani, A., Hao, H., & Zhang, P. (2012). A comparison study of basic data-driven fault diagnosis and process monitoring methods on the benchmark Tennessee Eastman process. *Journal of Process Control, 22*(9), 1567–1581.
Yin, S., Ding, S. X., Xie, X., & Luo, H. (2014). A review on basic data-driven approaches for industrial process monitoring. *IEEE Transactions on Industrial Electronics, 61*(11), 6418–6428.
Yu, J. (2012a). Multiway discrete hidden Markov model-based approach for dynamic batch process monitoring and fault classification. *AIChE Journal, 58*(9), 2714–2725.
Yu, J. (2012b). A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Computers & Chemical Engineering, 41*, 134–144.
Yu, J., & Qin, S. J. (2008). Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models. *AIChE Journal, 54*(7), 1811–1829.
Yu, J., & Qin, S. J. (2009). Multiway Gaussian mixture model based multiphase batch process monitoring. *Industrial & Engineering Chemistry Research, 48*(18), 8585–8594.
Yu, J., & Rashid, M. M. (2013). A novel dynamic bayesian network-based networked process monitoring approach for fault detection, propagation identification, and root cause diagnosis. *AIChE Journal, 59*(7), 2348–2365.
Yuan, T., & Qin, S. J. (2014). Root cause diagnosis of plant-wide oscillations using Granger causality. *Journal of Process Control, 24*(2), 450–459.
Yue, K., Fang, Q., Wang, X., Li, J., & Liu, W. (2015). A parallel and incremental approach for data-intensive learning of Bayesian networks. *IEEE Transactions on Cybernetics, 45*(12), 2890–2904.
Zare, A., Ozdemir, A., Iwen, M. A., & Aviyente, S. (2018). Extension of PCA to higher order data structures: An Introduction to Tensors, Tensor Decompositions, and Tensor PCA. *Proceedings of the IEEE*, (99).
Zhan, Y., & Yin, J. (2011). Robust local tangent space alignment via iterative weighted PCA. *Neurocomputing, 74*(11), 1985–1993.
Zhang, C., Li, F., & Jestes, J. (2012). In Efficient parallel kNN joins for large data in MapReduce. In *Proceedings of the 15th international conference on extending database technology* (pp. 38–49). ACM.
Zhang, H., Ge, Z., Ye, L., & Song, Z. (2015a). Vision-based fan speed control system in the copper scraps smelting process. *Asian Journal of Control, 17*(5), 1742–1755.
Zhang, K., Gonzalez, R., Huang, B., & Ji, G. (2015b). Expectation–maximization approach to fault diagnosis with missing data. *IEEE Transactions on Industrial Electronics, 62*(2), 1231–1240.
Zhang, M.-L., & Zhou, Z.-H. (2007). ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition, 40*(7), 2038–2048.
Zhang, Y., & Liu, Y. (2009). Data imputation using least squares support vector machines in urban arterial streets. *IEEE Signal Processing Letters, 16*(5), 414–417.
Zhao, Q., Zhou, G., Zhang, L., Cichocki, A., & Amari, S.-I. (2016). Bayesian robust tensor factorization for incomplete multiway data. *IEEE transactions on Neural Networks and Learning Systems, 27*(4), 736–748.
Zhao, W., Shen, Y., Yuan, Z., Wei, Y., Xu, P., & Jian, W. (2015). An efficient and robust algorithm for BSS by maximizing reference-based negentropy. *AEU-International Journal of Electronics and Communications, 69*(9), 1265–1271.
Zhou, Z., Li, X., Wright, J., Candes, E., & Ma, Y (2010). In Stable principal component pursuit. In *Information theory proceedings (ISIT), 2010 IEEE international symposium on* (pp. 1518–1522). IEEE: 2010.
Zhu, J., & Gao, F. (2018a). Similar batch process monitoring with orthogonal subspace alignment. *IEEE Transactions on Industrial Electronics, 65*(10), 8173–8183.
Zhu, J., & Gao, F. (2018b). Improved nonlinear quality estimation for multiphase batch processes based on relevance vector machine with neighborhood component variable selection. *Industrial & Engineering Chemistry Research, 57*(2), 666–676.
Zhu, J., Ge, Z., & Song, Z. (2014). Robust modeling of mixture probabilistic principal

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003

component analysis and process monitoring application. *AIChE Journal, 60*(6), 2143–2157.
Zhu, J., Ge, Z., & Song, Z. (2015a). HMM-driven robust probabilistic principal component analyzer for dynamic process fault classification. *IEEE Transactions on Industrial Electronics, 62*(6), 3814–3821.
Zhu, J., Ge, Z., & Song, Z. (2015b). Robust semi-supervised mixture probabilistic principal component regression model development and application to soft sensors. *Journal of Process Control, 32*, 25–37.
Zhu, J., Ge, Z., & Song, Z. (2015c). Robust supervised probabilistic principal component analysis model for soft sensing of key process variables. *Chemical Engineering Science, 122*, 573–584.
Zhu, J., Ge, Z., & Song, Z. (2015d). Multimode process data modeling: A Dirichlet process mixture model based Bayesian robust factor analyzer approach. *Chemometrics and Intelligent Laboratory Systems, 142*, 231–244.
Zhu, J., Ge, Z., & Song, Z. (2016a). Recursive mixture factor analyzer for monitoring multimode time-variant industrial processes. *Industrial & Engineering Chemistry Research, 55*(16), 4549–4561.
Zhu, J., Ge, Z., & Song, Z. (2016b). Bayesian robust linear dynamic system approach for dynamic process monitoring. *Journal of Process Control, 40*, 62–77.
Zhu, J., Ge, Z., & Song, Z. (2017a). Distributed parallel PCA for modeling and monitoring of large-scale plant-wide processes with Big Data. *IEEE Transactions on Industrial Informatics, 13*(4), 1877–1885.
Zhu, J., Ge, Z., & Song, Z. (2017b). Variational Bayesian Gaussian mixture regression for soft sensing key variables in Non-Gaussian industrial processes. *IEEE Transactions on Control Systems Technology, 25*(3), 1092–1099.
Zhu, J., Ge, Z., & Song, Z. (2017c). Non-Gaussian industrial process monitoring with probabilistic independent component analysis. *IEEE Transactions on Automation Science and Engineering, 14*(2), 1309–1319.
Zhu, J., Ge, Z., Song, Z., Zhou, L., & Chen, G. (2018a). Large-scale plant-wide process modeling and hierarchical monitoring: A distributed Bayesian network approach. *Journal of Process Control, 65*, 91–106.
Zhu, J., Wang, Y., Zhou, D., & Gao, F. (2018b). Batch process modeling and monitoring with local outlier factor. *IEEE Transactions on Control Systems Technology*.
Zhu, J., Yao, Y., Li, D., & Gao, F. (2018c). Monitoring big process data of industrial plants with multiple operating modes based on Hadoop. *Journal of the Taiwan Institute of Chemical Engineers, 91*, 10–21.
Zou, C., & Feng, J. (2009). Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinformatics, 10*(1), 1–17.



Jinlin Zhu received his Ph.D. degree in Control Science and Engineering from Zhejiang University, China, in 2016. He was then the postdoctoral visiting scholar with the Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology. Currently, he is the presidential postdoctoral fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. His research interests include industrial process modeling, monitoring and fault diagnosis.

Zhiqiang Ge received the B.Eng. and Ph.D. degrees in Automation from the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, in 2004 and 2009, respectively. He was a Research Associate with the Department of Chemical and Biomolecular Engineering, Hong Kong University of Science Technology from Jul. 2010 to Dec. 2011 and a visiting Professor with the Department of Chemical and Materials Engineering, University of Alberta from Jan. 2013 to May 2013. Dr. Ge was an Alexander von Humboldt research fellow with University of Duisburg-Essen during Nov. 2014 and Jan. 2017, and also a JSPS invitation Fellow with Kyoto University during Jun. 2018 and Aug. 2018. He is currently a Full Professor with the College of Control Science and Engineering, Zhejiang University. His research interests include industrial big data, process monitoring, quality prediction, machine learning, and Bayesian methods.

Zhihuan Song received the B.Eng. and M.Eng. degrees in industrial automation from Hefei University of Technology, Ahhui, China, in 1983 and 1986, respectively, and the Ph.D. degree in industrial automation from Zhejiang University, Hangzhou, China, in 1997. Since 1997, he has been in the Department of Control Science and Engineering, Zhejiang University, where he was first a Postdoctoral Research Fellow, then an Associate Professor, and is currently a Professor. He has published more than 200 papers in journals and conference proceedings. His research interests include the modeling and fault diagnosis of industrial processes, analytics and applications of industrial big data, and advanced process control technologies.

Furong Gao obtained his B.Eng. in Automation from East China Institute of Petroleum in 1985, and his M.Eng. and Ph.D. degrees in Chemical Engineering from McGill University, Montreal, Canada, in 1989 and 1993, respectively. He worked as a Senior Research Engineer at Moldflow International, Melbourne, Australia, from 1993 to 1995. He is now a Chair Professor in the Department of Chemical and Biological Engineering at the Hong Kong University of Science and Technology. His research interests include process monitoring and fault diagnosis, batch process control, polymer processing control, and optimization. He received a number of best paper awards, and is on Editorial Boards of a number of journals of his area.

Please cite this article as: J. Zhu et al., Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data, Annual Reviews in Control (2018), https://doi.org/10.1016/j.arcontrol.2018.09.003