BigStat for Big Data: Big Data clustering through the BigStat SaaS platform

Christophe Biernacki

"BigStat" for Big Data

"Big Data clustering through the BigStat SaaS platform"

Christophe Biernacki*
University of Lille & CNRS & Inria, France

Abstract

BigStat is a web platform devoted to clustering of big data sets through two hosted software, MixtComp and BlockCluster. The former adress mixed, missing and uncertain data in a moderate dimensional setting, whereas the latter is devoted to high dimensional data sets with non-mixed, non-missing and non-uncertain data. Mathematical foundations of both rely on mixture models and related algorithms.

Keywords. Model-based clustering, mixed and missing data, high dimension, SaaS platform.

1 Introduction

1.1 Big Data: IT genesis

The Big Data phenomenon mainly originates in the increase of computer and digital resources at an ever lower cost. Indeed, the storage cost by Mb (Mega bytes, 10^6 bytes) rose from 700$ in 1981 to 1$ in 1994 then to 0.01$ in 2013[1] (the price has been divided by 70,000 in thirty years) whereas hard drives of 8 Tb (Tera bytes, 10^{12} bytes) storage capacity are now easily available, to be compared to 1.02 Gb (Giga bytes, 10^9 bytes) storage capacity in 1982[2] (storage capacity multiplied by 8,000 on the same period). Simultaneously, the processing speed of the existing most powerful computer starts from one gigaFLOPS (a FLOPS corresponds to the FLoating-point Operations Per Second) in 1985 to reach 33 petaFLOPS in 2013[3] (speed multiplied by 33 million). It leads to the so-called *storage challenge*, which is the "IT side" of Big Data gathering the storage, the transfer, the preservation and the availability of data.

One should be aware that any human activities are impacted by such a digital data accumulation: trade and business (companies information systems, banks, booking systems...), governments and other organizations (laws and other regulations, standardization rules...), entertainment (music, video, games, social networks...), fundamental sciences (astronomy, physics, power, genomics...),

*Christophe.Biernacki@math.univ-lille1.fr

[1] http://www.capital.fr/enquetes/documents/la-folle-evolution-du-stockage-informatique-953110
[2] http://fr.wikipedia.org/wiki/Disque_dur#.C3.89volution_en_termes_de_prix_ou_de_capacit.C3.A9
[3] http://fr.wikipedia.org/wiki/FLOPS

health (medical file...), environment (climat, pollution, alimentation...), humanities and social sciences (knowledge digitization, literature, history, art, architecture, archaeological data...). Finally, the whole human society converges towards a so-called numerical world, so that in 2007 more that 94% of stored information was available in its digital form (the 6% remaining being available in its analogical form), to be compared again to only 1% in 1986 [Hilbert and López, 2011]. Moreover, this amount of stored information exceeds now 280 Eb (Exa bytes, 10^{18} bytes), *versus* 0.02 Eb in 1986 (14,000 times more). It leads to the so-called *societal and economic challenge*, which is the "soft science side" of Big Data gathering protection of private life, right to be forgotten, property rights, operating rights, cost of energy storage or transfer.

1.2 New data but classical statistical challenges

Since Laney [2001], the Big Data phenomenon is also described by the "3V" analytic point of view, mixing Volume, Velocity and Variety terms, respectively describing the quantity of data to be processed, the response delay in the processing and the form the data can take (structured, unstructured). Moreover, it is now current to meet "4V" or "5V" terminologies, for instance by adding the term Veracity which describes uncertainty due to data incompleteness. It leads to the so-called *data analysis challenge*, which is the "hard science side" of Big Data gathering together hardware, software and mathematical skills. Anyway, the volume is certainly the most emblematic feature of Big Data for all these skills. The reason is certainly the exponential growth of the data sets size over time, both on the number of individuals and on the number of variables of the data sets as observed by Alelyani *et al.* [2013] from data sets extracted from UCI machine learning repository.

However, although such new kinds of data sets are more and more present, the statistical aims stay entirely unchanged. They still rely on the same user questionings, corresponding to exploration goals (typically visualization and unsupervised classification) and prediction ones (typically supervised classification and regression). This fact is illustrated from the 4th Annual Rexer Analytics Data Miner Survey[4], which is the largest survey of data mining, data science, and analytics professionals in the industry, where decision trees, regression and cluster analysis form a triad of core algorithms, and by far, for most data scientists in the world.

1.3 Focus of the paper

In the present paper, we focus our attention on the unsupervised classification task (also called clustering or cluster analysis) since it is one of the first three main statistical aims of data miners (or data scientists) as previously discussed. In this context, we present different model-based clustering situations, varying in data volume (individuals, variables), in data variety (mixed data, typically gathering continuous and categorical variables) and in data veracity (missing and uncertain data). The model-based approach has advantage to rely on the mathematical statistical framework, thus is able to provide rigorous answers to clustering for such challenging kinds of data. The presented models

[4] http://www.rexeranalytics.com/Data-Miner-Survey-Results-2011.html

are implemented in two different software (MixtComp and BlockCluster), each specialized in a different context as described later. Both are gathered in the BigStat platform[5] freely available in SaaS mode (Software as a Service).

The ouline of the paper is the following. In Section 2, the model-based clustering principle is given and a specific discussion concerning effect of data volume on partition estimates is conducted, separating the number of individuals and the number of variables situations. Section 3 is devoted to the moderate number of variables case with possibility to deal simultaneously with mixed, missing and uncertain data. The related models are implemented in the MixtComp software of the BigStat platform and a illustration on a real data set is provided. Section 4 is devoted to the high number of variables case, but with non-mixed and non-missing data. The related models are implemented in the BlockCluster software of the BigStat platform and an illustration on a real data set is also provided. Section 5 concludes the paper and draws some prospects both from the statistical model and the software points of view.

2 Model-based clustering and large data sets

2.1 Model-based clustering principle

Mixture hypothesis Cluster analysis is one of the main data analysis method. It aims at partitioning a data set $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) = (\mathbf{x}^O, \mathbf{x}^M)$, composed by n individuals and lying in a space $\mathcal{X}$ of dimension d into K classes $G_1, \ldots, G_K$. Here the observed part of $\mathbf{x}$ has been denoted by $\mathbf{x}^O$ whereas the missing or uncertain one (uncertain means partially missing like intervals) is denoted by $\mathbf{x}^M$. Moreover, $\mathcal{X}$ designates possibly a mixed feature space, it means a space mixing features of different kinds like continuous, categorical or integer. An illustration of missing, uncertain and mixed features is displayed in Table 1.

observed data set $\mathbf{x}^O$			
?	0.5	red	5
0.3	0.1	green	3
0.3	0.6	{red,green}	3
0.9	[0.25 0.45]	red	?
↓	↓	↓	↓
continuous	continuous	categorical	integer

Table 1: A mixed, missing and uncertain data set composed by $n = 4$ individuals and $d = 4$ variables.

The target partition is denoted by $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$, lying in a space $\mathcal{Z}$, where $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})'$ is a vector of $\{0,1\}^K$ such that $z_{ik} = 1$ if individual $\mathbf{x}_i$ belongs to the kth class G_k, and $z_{ik} = 0$ otherwise ($i = 1, \ldots, n$, $k = 1, \ldots, K$). Model-based clustering allows to reformulate cluster analysis as a well-posed estimation problem both for the partition $\mathbf{z}$ and for the number K of classes. It considers data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as n i.i.d. realizations of a mixture distribution

[5] https://modal-research.lille.inria.fr/BigStat/

$f(\cdot;\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f(\cdot;\boldsymbol{\alpha}_k)$, where $f(\cdot;\boldsymbol{\alpha}_k)$ indicates the distribution, parameterized by $\boldsymbol{\alpha}_k$, associated to the class k, where π_k indicates the mixture proportion of this component ($\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$) and where $\boldsymbol{\theta} = (\pi_k, \boldsymbol{\alpha}_k; k = 1, \ldots, K)$ indicates the whole mixture parameters.

The question is then to estimate the class number K and the partition $\mathbf{z}$ from $\mathbf{x}^O$. Figure 1 gives an illustration of this principle when $d = 2$. The standard solution relies on first the estimation of the mixture parameter $\hat{\boldsymbol{\theta}}$ as we describe now.

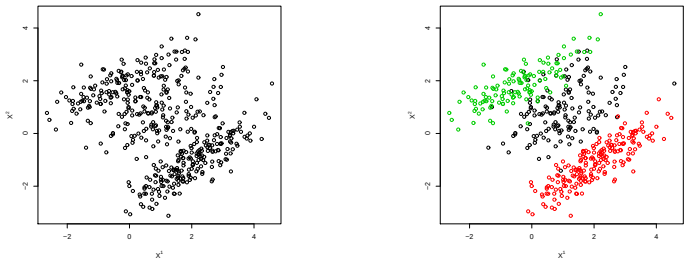


Figure 1: The clustering purpose illustrated in the two-dimensional setting.

Mixture parameter estimation From the observed data set $\mathbf{x}^O$ it is then possible to obtain a mixture parameter estimate $\hat{\boldsymbol{\theta}}$ by maximizing the observed log-likelihood $\ell(\hat{\boldsymbol{\theta}}; \mathbf{x}^O) = \ln f(\mathbf{x}_i^O; \boldsymbol{\theta})$ where

$$f(\mathbf{x}_i^O; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i^O; \boldsymbol{\alpha}_k) = \sum_{k=1}^{K} \pi_k \int_{\mathbf{x}_i^M} f(\mathbf{x}_i^O, \mathbf{x}_i^M; \boldsymbol{\alpha}_k) d\mathbf{x}_i^M, \tag{1}$$

provided that missing data $\mathbf{x}^M$ are obtained by a missing at random (MAR) process.

For optimizing $\ell(\boldsymbol{\theta}; \mathbf{x}^O)$, the EM (Expectation-Maximization) algorithm of Dempster *et al.* [1977] is often performed or some of its variants (see also Redner and Walker [1984]) like the SEM (Stochastic EM) [Celeux and Diebolt, 1985]. A SEM algorithm can be used to maximize the observed-data log-likelihood, described as follows for iteration $q \geq 1$, when starting from a parameter $\boldsymbol{\theta}^{(0)}$ selected at random:

- **E-step**: compute conditional probabilities $f(\mathbf{x}^M, \mathbf{z}|\mathbf{x}^O; \boldsymbol{\theta}^{(q)})$,
- **S-step**: draw $(\mathbf{x}^{M(q)}, \mathbf{z}^{(q)})$ from $f(\mathbf{x}^M, \mathbf{z}|\mathbf{x}^0; \boldsymbol{\theta}^{(q)})$,
- **M-step**: maximize $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}} \ln f(\mathbf{x}^O, \mathbf{x}^{M(q)}, \mathbf{z}^{(q)}; \boldsymbol{\theta})$.

Since the parameter sequence $(\boldsymbol{\theta}^{(q)})$ generated by SEM does not punctually converges, due to the S-step definition, the algorithm generally stops after a predefined number of iterations. This sequence converges in distribution towards the unique stationary distribution. Asymptotically on q, the mean of the sequence $(\boldsymbol{\theta}^{(q)})$ approximates $\hat{\boldsymbol{\theta}}$ and thus provides a sensible local estimate of

the maximum likelihood. In addition, the variance of the sequence $(\boldsymbol{\theta}^{(q)})$ gives confidence intervals on $\boldsymbol{\theta}$. SEM has also advantage to be less dependent on the initial value $\boldsymbol{\theta}^{(0)}$ than EM does if a "sufficient" iteration number is performed and so avoids uninteresting local maxima. Finally, managing missing data is easier than with EM thanks to its so-called stochastic S-step, while preserving a classical M-step like EM.

Partition (and missing data) estimation Once $\hat{\boldsymbol{\theta}}$ is obtained, a so-called SE algorithm (a SEM without the M step) can be used to estimate partition $\mathbf{z}$, and simultaneously missing data $\mathbf{x}^M$. Its qth iteration is given by

- **E-step**: compute conditional probabilities $f(\mathbf{x}^M, \mathbf{z}|\mathbf{x}^O; \hat{\boldsymbol{\theta}})$,
- **S-step**: draw $(\mathbf{x}^{M(q)}, \mathbf{z}^{(q)})$ from $f(\mathbf{x}^M, \mathbf{z}|\mathbf{x}^O; \hat{\boldsymbol{\theta}})$.

After a given iteration number, the mean and/or mode of the sequence $(\mathbf{x}^{M(q)}, \mathbf{z}^{(q)})$ estimates $(\mathbf{x}^M, \mathbf{z})$, denoted by $(\hat{\mathbf{x}}^M, \hat{\mathbf{z}})$, with again the possibility to derive some confidence intervals on these unknown quantities.

Estimation of the class number It is now possible to derive an estimate $\hat{K}$ from an estimate of the observed conditional probability $\hat{f}(K|\mathbf{x}^O)$ or also from the completed-partition conditional probability $\hat{f}(K|\mathbf{x}^O, \mathbf{z})$. The first one leads to retaining $\hat{K}$ which maximizes the so-called BIC (Bayesian Information Criterion) criterion [Schwarz, 1978], whereas the second one corresponds to maximizing the so-called ICL (Integrated Completed Likelihood) criterion [Biernacki *et al.*, 2000], defined by

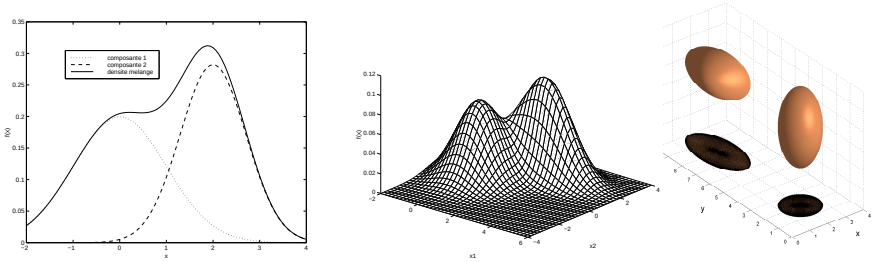
$$\text{ICL} = \ln f(\mathbf{x}^O, \hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}) - \frac{D}{2} \ln n, \tag{2}$$

D denoting the number of free (continuous) parameters in the model at hand. Advantage of ICL over BIC in a clustering context is its ability to integrate the clustering purpose in its definition through the use of the estimate partition $\hat{\mathbf{z}}$ in (2). As a consequence it will favour well-separated clusters, so less clusters but more valuable clusters than BIC provide, even if the model definition of components $f(\cdot; \boldsymbol{\alpha}_k)$ is erroneous.

Illustration in the Gaussian case The multivariate mixture model is certainly the most known and used model for continuous data. It has a long history of use in clustering (see for instance Wolfe [1971], Bock [1981]). In that case, $\mathbf{x}_i$ $(i = 1, \ldots, n)$ are continuous variables $\mathcal{X} = \mathbb{R}^d$ and the conditional density of components is written $(k = 1, \ldots, K)$

$$f(\cdot; \boldsymbol{\alpha}_k) = \mathsf{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\cdot - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\cdot - \boldsymbol{\mu}_k)\right), \tag{3}$$

with $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ the component mean (or centre) and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d \times d}$ its variance-covariance matrix. Figures 2 (a), (b) and (c) respectively display univariate, bivariate and trivariate Gaussian mixtures.



(a) (b) (c)

Figure 2: Gaussian mixtures in (a) univariate, (b) bivariate and (c) trivariate situations.

2.2 The large number of individuals case

In statistics, theory is often asymptotics on the number of individuals and thus, theoretically, increasing n is a good news, model-based clustering included. The standard way to address the effect of n on the partition estimate quality is to express the related bias/variance trade-off. We note $\text{err}(\mathbf{z}_1, \mathbf{z}_2) \geq 0$ a distance-like measure between two partitions $\mathbf{z}_1$ and $\mathbf{z}_2$. When the number of classes in each partition is identical, it can be the classical empirical error rate. When the number of classes differs, it can be for instance the Rand criterion defined in Rand [1971]. We also define $\mathbf{z}^* = \arg\min \text{err}(\mathbf{z}, \cdot)$ the best partition associated to the model at hand with regards to the true partition $\mathbf{z}$. We then have the simple but important following decomposition:

$$\begin{aligned} \text{err}(\mathbf{z}, \hat{\mathbf{z}}) &= \Big\{\text{err}(\mathbf{z}, \mathbf{z}^*) - \text{err}(\mathbf{z}, \mathbf{z})\Big\} + \Big\{\text{err}(\mathbf{z}, \hat{\mathbf{z}}) - \text{err}(\mathbf{z}, \mathbf{z}^*)\Big\} \quad (4) \\ &= \Big\{\text{bias}\Big\} + \Big\{\text{variance}\Big\}. \quad (5) \end{aligned}$$

The bias corresponds to the so-called *error of approximation* and the variance to the so-called *error of estimation*.

When the sample size grows, as expected the variance automatically decreases. However, vanishing asymptotically the whole error term $\text{err}(\mathbf{z}, \hat{\mathbf{z}})$ requires also to decrease the bias. If the proposed model at hand does not correspond to the true (unknown) model, the only issue is to change it by a more complex one. Typically, increasing the candidate number of classes K is thus the opportunity to reduce such a model approximation as illustrated on Figure 3.

2.3 The large number of variables case

However, in the Big Data context, increasing the volume may mean inflating the number of individuals (n) or alternatively the number of variables (d) (or both). The question is now to control the error $\text{err}(\mathbf{z}, \hat{\mathbf{z}})$ previously defined in (5) with regards to d. Contrary to increasing n, increasing d may have both positive and negative effects on the clustering task, usually designated respectively by "blessing" and "curse" of the dimension.

Blessing factors Consider the following two-component multivariate Gaussian mixture: $\pi_1 = \pi_2 = \frac{1}{2}$, $f(\cdot; \boldsymbol{\alpha}_1) = \mathsf{N}(\mathbf{0}, \mathbf{I})$ and $f(\cdot; \boldsymbol{\alpha}_2) = \mathsf{N}(\mathbf{1}, \mathbf{I})$, with

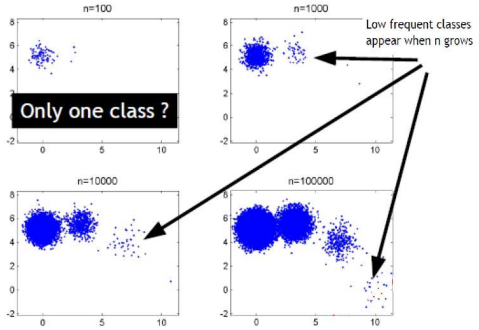
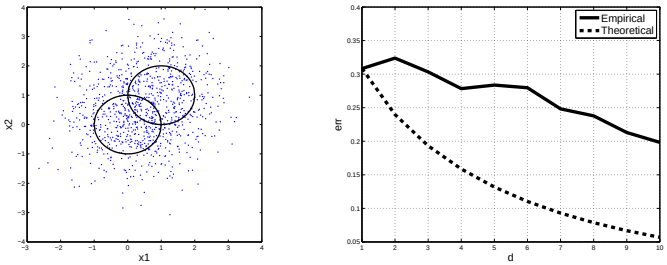


Figure 3: Opportunity to reduce error of approximation when n grows by increasing $\hat{K}$.

$\mathbf{a} = (a \dots a)'$ a real vector of size d. We display a corresponding sample in Figure 4 (a). In that case the two components are more and more separated when d grows since $\|\mathbf{1} - \mathbf{0}\|_{\mathbf{I}} = \sqrt{d}$. The reason is that each variable uniformly provides its own separation information such that the associated theoretical error decreases when d grows. Indeed, this thoretical error is equal to $\mathsf{err}_{theo} = \Phi(-\sqrt{d}/2)$, where Φ is the cumulative distribution of $\mathsf{N}(0,1)$. We can see this decrease with d by a dash line in Figure 4 (b). An interesting consequence is then that the empirical error rate decreases also with d as it could be noticed in continuous line in Figure 4 (b). It means that increasing dimension may have a positive effect on the clustering task as soon as all variables convey meaningful information on the hidden partition. From the bias/variance interpretation (5), it means that bias decreases faster than variance grows when d is larger.



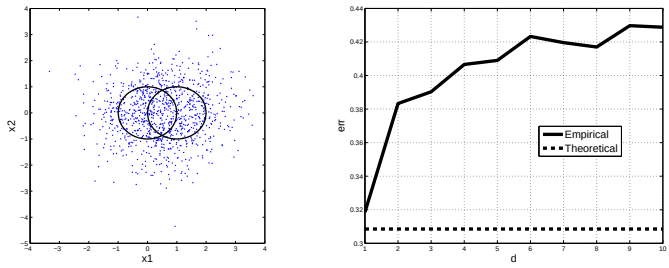
(a) (b)

Figure 4: Dimension blessing in the clustering context when most variables convey independent partitioning information: (a) A bivariate data set example with isodensity of each component and (b) the theoretical (dash line) and the empirical (continuous line) error rate when d increases.

Curse factors In fact, increasing dimension may have a positive effect on clustering retrieval only if variables inject some partioning information. In addition, such information has to be not redundant. It is a consequence that, in both cases, bias does not deflates whereas variance clims. We illustrate now

these two particular features.

Firstly, we consider many variables which provide no separation information. We retrieve the same previous parameter setting except that the components are not more separated when d grows since $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{I}} = 1$, where $\boldsymbol{\mu}_1 = \mathbf{0}$ is the center of the first Gaussian and where $\boldsymbol{\mu}_2 = (1\ 0\ \ldots\ 0)'$ is the one of the second, thus $(k = 1, 2)$: $f(\cdot; \boldsymbol{\alpha}_k) = \mathsf{N}(\boldsymbol{\mu}_k, \mathbf{I})$. A sample is displayed on Figure 5 (a). Figure 5 (b) shows in dash line that the theoretical error rate is constant (it corresponds to $\mathsf{err}_{theo} = \Phi(-\frac{1}{2})$) when the dimension increases, as expected. Consequently, the empirical error rate degrades in this situation (continuous line of the same figure).

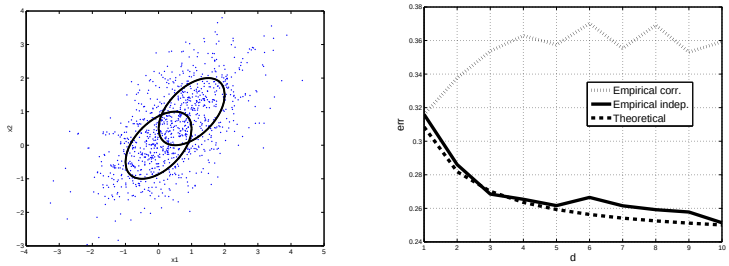


(a) (b)

Figure 5: Dimension curse in the clustering context when variables convey no partitioning information: (a) A bivariate data set example with isodensity of each component and (b) the theoretical (dash line) and the empirical (continuous line) error rate when d increases.

Secondly, we consider a case where many variables provide separation, but redundant information, in the following sense: It is the same parameter setting as before for the first dimension except for all other ones, thus $\mathbf{X}_{1j} = \mathbf{X}_{11} + \varepsilon_j$ where $\varepsilon_j \overset{iid}{\sim} \mathsf{N}(0, 1) \quad (j = 2, \ldots, d)$. See a data example in Figure 6 (a). Thus, components are not more separated when d grows since $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}} = 1$, $\boldsymbol{\Sigma}$ denoting the common covariance matrix of each Gaussian component, and $\boldsymbol{\mu}_k$ denoting the center of the component $k = 1, 2$. Consequently, $\mathsf{err}_{theo} = \Phi(-\frac{1}{2})$ is constant and the empirical error increases with d, as illustrated in Figure 6 (b) with previous conventions.

Attempt to reach the bias/variance trade-off As previously explained, curse factors are the consequence of the variance increase, whereas bias is stable. The solution is to significantly decrease the variance even if increasing the bias to a lesser extent. Since class separation grows in Figure 6 (a), it becomes possible to use a simpler model, namely here a Gaussian model with *diagonal covariance matrices* while preserving a quite low error rate. It is illustrated in Figure 6 (b) with the continuous line. This remark will be fundamental for the models implemented in the MixtComp and BlockCluster software we describe now.



(a) (b)

Figure 6: Dimension curse in the clustering context when variables convey redundant partitioning information: (a) A bivariate data set example with isodensity of each component and (b) the theoretical (dark dash line), the empirical *correlated* model related (gray dash line) and the empirical *independent* model related (continuous line) error rate when d increases.

3 BigStat platform for mixed/missing variables

In the BigStat platform, the MixtComp software is dedicated to clustering of (full) mixed data (continuous, categorical, integer, ordinal, rank and continuous functional), allowing also missing values and uncertain values (like intervals in the continuous case, but available for all other data types). It implements a SEM and a SE algorithms for all estimations, including missing and uncertain data, as described in Section 2.1. From a practical point of view, a single zip file, containing both the data set and the variable descriptor, is provided as an input on the web site. The output corresponds to an interactive entropy visualization of the estimated clusters, with the mixture parameters, and also to the possibility to download the corresponding R object for further use in the R environment. A wiki documentation is also available[6].

3.1 Models for a moderate number of mixed variables

Intra-type conditional independence As discussed in the previous section, increasing dimension d of data imposes to restrict the model complexity in order to control the variance even if increasing moderately the bias. In case when dimension d is "moderate", the model complexity should be also. For instance, in Figure 6 (b) a diagonal Gaussian model is efficient to obtain a good partitioning of clusters with yet intra-correlated variables. We propose to extend this idea for all kinds of data types by assuming that all variables x_{ij} of $\mathbf{x}_i$ $(j = 1, \ldots, d)$ are *conditionally independent* knowing the latent classes. Thus, $f(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} f(x_{ij}; \boldsymbol{\alpha}_{kj})$ where $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_{kj}; j = 1, \ldots, d)$, $f(\cdot; \boldsymbol{\alpha}_{kj})$ denoting the univariate distribution associated to the variable j in the class k. This latter is defined as follows, depending on the data type:

- **Continuous**: each $f(\cdot; \boldsymbol{\alpha}_{kj}) = \mathsf{N}(\mu_{kj}, \sigma^2_{kj})$. It corresponds to the diagonal Gaussian model of Celeux and Govaert [1995] (see an illustration on Figure 2 (c)).

[6] `https://modal.lille.inria.fr/wikimodal/doku.php?id=mixtcomp`

- **Categorical**: each $x_{ij} = (x_{ijh}; h = 1, \ldots, m_j)$ has m_j response levels where $x_{ijh} = 1$ if i has response level h for variable j and $x_{ijh} = 0$ otherwise. The standard model for clustering observations described through categorical variables is the so-called latent class model (see for instance Goodman [1974]) where $f(\cdot; \boldsymbol{\alpha}_{kj}) = \mathsf{M}(\boldsymbol{\alpha}_{kj})$ is the multinomial distribution with $\boldsymbol{\alpha}_{kj} = (\alpha_{kjh}; h = 1, \ldots, m_j)$, α_{kjh} denoting the probability that variable j has level h if individual i is in cluster k.

- **Integer**: each $x_{ij} \in \mathbb{N}$ and $f(\cdot; \boldsymbol{\alpha}_{kj}) = \mathsf{P}(\lambda_{kj})$, the Poisson distribution of parameter λ_{kj}.

- **Other**: each x_{ij} could be also an ordinal data, a ranking data or also a (discretised) continuous functional data (see respective univariate distributions in Biernacki and Jacques [2016], Biernacki and Jacques [2013] and Samé *et al.* [2011]).

Inter-type conditional independence It is frequent in practice to mix different kinds of data types, for instance continuous, categorical and integer ones. Thus the ith individual is composed by three parts, $\mathbf{x}_i = (\mathbf{x}_i^{cont}, \mathbf{x}_i^{cat}, \mathbf{x}_i^{int})$, $\mathbf{x}_i^{cont}$, $\mathbf{x}_i^{cat}$ and $\mathbf{x}_i^{int}$ designing the continuous, the categorical and the integer ones respectively. In that case, the proposed solution for symmetry between data types is to mixed all types by inter-type conditional independence [Moustaki and Papageorgiou, 2005]:

$$f(\mathbf{x}_i; \boldsymbol{\alpha}_k) = f(\mathbf{x}_i^{cont}; \boldsymbol{\alpha}_k^{cont}) \times f(\mathbf{x}_i^{cat}; \boldsymbol{\alpha}_k^{cat}) \times f(\mathbf{x}_i^{int}; \boldsymbol{\alpha}_k^{int})$$

with $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_k^{cont}, \boldsymbol{\alpha}_k^{cat}, \boldsymbol{\alpha}_k^{int})$ the obvious associated parameters by data type.

3.2 Illustration on a prostate cancer data set

Hunt and Jorgensen [1999] (see also McLachlan and Peel [2000] p. 139–142) considered the clustering of patients on the basis of petrial variates alone for the prostate cancer clinical trial data of Byar and Green [1980] which is reproduced in Andrews and Herzberg [1985] p. 261–274. This data set was obtained from a randomized clinical trial comparing four treatments for $n = 506$ patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease. As reported by Byar and Green [1980], Stage 3 represents local extension of the disease without evidence of distance metastasis, while Stage 4 represents distant metastasis as evidenced by elevated acid phosphatase, X-ray evidence, or both. Twelve pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases). The skewed variables "size of primary tumour" and "serum prostatic acid phosphatase" were transformed by using a square root and a logarithm transformation, respectively. There are 62 missing values, so about 1% of the whole sample, and 475 patients have finally no missing data.

We compare five strategies for using the MixtComp software: (1) "raw-506": the 506 raw missing/mixed data, (2) "impute-506": the 506 mixed data after

imputation of missing data by the `mice` R package[7], (3) "mixed-475": the 475 non-missing mixed data, (4) "cont-475": the 475 non-missing continuous-only data, (5) "cat-475": the 475 non-missing categorical-only data. The ICL criterion (2) is then calculated in each situation for estimating K in the range 1,...,7. Only three strategies (raw-506, mixed-475, cont-475) retain $\hat{K} = 2$, strategies impute-506 and cat-475 respectively preferring $\hat{K} = 7$ and $\hat{K} = 1$. Now, by fixing $K = 2$, the misclassification error rate is displayed in Table 2 for all strategies. Both previous results on K selection and the error rate when $K = 2$ indicate all interest on dealing directly on raw data. Indeed, imputation before clustering, and to a lesser extend missing data removing, may loose some cluster information. In addition, categorical variables alone convey few information on the partition but could be informative associated to continuous variables.

Strategy	raw-506	impute-506	cont-475	cat-475	mixed-475
% misclassified	8.1	12.8	9.46	47.16	8.63

Table 2: Misclassification error rate on the cancer data set with $K = 2$.

4 BigStat platform for high dimension

In the BigStat platform, the BlockCluster software is dedicated to clustering of non-missing and homogeneous data (continuous, categorical or integer, but not mixed) of very high dimension thanks to a co-clustering approach. It implements a SEM algorithm (and other) for all estimations. Similarly to MixtComp, a single zip file is provided as an input on the web site containing both the data set and some tuning parameters. It provides a visualization of the estimated clusters (and co-clustering blocks) and also the possibility to download the corresponding R object for further use in the R environment. A corresponding R package is also available[8].

4.1 Models for a high number of homogeneous variables

Models implemented in the MixtComp software are not parsimonious enough to be used in the very high dimensional setting. Simultaneous clustering of rows and columns, usually designated by bi-clustering, co-clustering or block clustering, is an important technique in two way data analysis allowing very simple models even with many variables. They consider the two sets simultaneously and organize the data into homogeneous blocks. Two partition representations are thus now needed. First, as usual, a partition of n individuals (lines of the data matrix $\mathbf{x}$) into K clusters still noticed $\mathbf{z}$. Second, and symmetrically, a partition of d variables (columns of the data matrix $\mathbf{x}$) into L clusters is denoted by $\mathbf{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_d)$ where $\boldsymbol{w}_j = (w_{j1}, \ldots, w_{jL})$ with $w_{jl} = 1$ if j belongs to cluster l and $w_{jl} = 0$ otherwise. Both space partitions are respectively denoted by $\mathcal{Z}$ and $\mathcal{W}$.

[7] http://cran.r-project.org/web/packages/mice/mice.distribution
[8] https://cran.r-project.org/web/packages/blockcluster/index.html

We refer to the book of Govaert and Nadif [2013] for providing more details on co-clustering techniques, probabilistic or not. Here, we focus on model-based co-clustering as being often a generalization of non-probabilistic methods and allowing coherent formulation from estimation to model selection. In the following set, product on i, j, k and l stands for ranges $\{1,\ldots,n\}$, $\{1,\ldots,d\}$, $\{1,\ldots,K\}$ and $\{1,\ldots,L\}$ respectively. Block model-based clustering can be seen as an extension of the traditional mixture model-based clustering (see Section 2.1). The basic idea is to extend the latent class principle of local (or conditional) independence. Each data point x_{ij} is assumed to be independent once $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$ are fixed. We note $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\rho}, \boldsymbol{\alpha})$ with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kl})$ and where $\boldsymbol{\pi} = (\pi_k)$ and $\boldsymbol{\rho} = (\rho_k)$ are the vectors of probabilities π_k and ρ_l that a row and a column belong to the kth row component and to the lth column component respectively. Assuming also independence between all $\boldsymbol{z}_i$ and $\boldsymbol{w}_j$, the latent block mixture model has final probability distribution

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,j,k,l} \left(\pi_k \rho_l f(x_{ij};\boldsymbol{\alpha}_{kl})\right)^{z_{ik}w_{jl}} . \tag{6}$$

Finally, the distribution $f(\cdot;\boldsymbol{\alpha}_{kl})$ depends on the data type of x_{ij} (continuous, categorical, integer) and thus is similar to these ones defined in Section 3.1, except that mixed data are not allowed this time. Such models can be very parsimonious even when d is very large, provided that L is moderate. Indeed, by comparison to a classical intra-type conditional independence model with D parameters to be estimated (see Section 3.1), the corresponding co-clustering model requires only $D \times \frac{L}{d}$ parameters. In addition, a specific expression of the ICL criterion (2) can be invoked for selecting the pair (K, L).

4.2 Illustration on a document data set

Figure 7 (a) displays a text mining example[9]. It mixes Medline (1,033 medical abstracts) and Cranfield (1,398 aeronautical abstracts) making a total of $n = 2,431$ documents. Furthermore, all the words (excluding stop words) are considered as features making a total of $d = 9,275$ unique words. The data matrix consists of documents on the rows and words on the columns with each entry giving the term frequency, that is the number of occurrences of corresponding word in corresponding document. Since it concerns a contingency table (cross counting documents and words) we apply a Poisson block clustering model. The "true" block partitioning involves $K = 2$ document clusters (row) and $L = 2$ word clusters (column). Table 3 displays the confusion table for documents by using 2×2 blocks. We show that we exactly retrieve the underlying document structure, what is expected by the blessing effect of high dimensional clustering, the data set being here with $d = 9,275$. Figure 7 (b) gives a view of the data set after reorganization by block-clustering. We also distinguish clear partitioning in rows and columns.

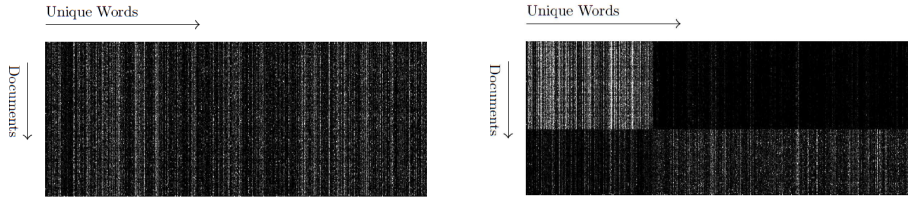
5 Conclusion

The BigStat platform offers a freely available solution for clustering without any package installation and any computer resource (any mobile device may

[9] This data set is publicly available at `ftp://ftp.cs.cornell.edu/pub/smart`.

	Medline	Cranfield
Medline	1033	·
Cranfield	·	1398

Table 3: Confusion table by applying block clustering for text partitioning.



(a) (b)

Figure 7: Text mining example: (a) the initial data set; (b) the reorganized data set with $(K, L) = (2, 2)$.

be appropriate, not necessarily a laptop) since it can be launched as a SaaS application from a web page. The output can be either directly visualized on the device as a graphical display, or downloaded as an R object for further use in an R environment.

Two clustering software are available in BigStat, and rely both on the model-based clustering paradigm. First, the MixtComp software is able to deal with mixed, missing and uncertain data in a moderate dimension setting. Second, the BlockCluster software extend clustering, through a co-clustering principle, to the high dimension setting but restrict its use to non-mixed and non-missing data. Future theoretical and methodological works will concern possibility to address simultaneously both situations, it means mixed, missing and uncertain data in a high dimension setting.

Finally, the BigStat platform will migrate soon towards a new platform called MASSICCC[10] (Massive Clustering on Cloud Computing) allowing a even more convivial interface with more powerful computer servers.

References

Alelyani, S., Tang, J. and Liu, H. [2013]. Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, **29**.

Andrews, D. F. and Herzberg, A. M. [1985]. *Data: A Collection of Problems from Many. Fields for the Student and Research Worker*. Springer-Verlag.

Biernacki, C., Celeux, G. and Govaert, G. [2000]. Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.

Biernacki, C. and Jacques, J. [2013]. A generative model for rank data based

[10] https://modal-research-dev.lille.inria.fr/#/

on insertion sort algorithm. *Computational Statistics and Data Analysis*, **58**, 162–176.

Biernacki, C. and Jacques, J. [2016]. Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm. *Statistics and Computing*, **26**(5), 929–943. URL `https://hal.inria.fr/hal-01052447`.

Bock, H. [1981]. Statistical Testing and Evaluation Methods in Cluster Analysis. In *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions*, 116–146. Calcutta.

Byar, D. and Green, S. [1980]. The choice of treatment for cancer patients based on covariate information: Application to prostate cancer. *Bulletin du Cancer*, **67**, 477–490.

Celeux, G. and Diebolt, J. [1985]. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, **2**(1), 73–92.

Celeux, G. and Govaert, G. [1995]. Gaussian Parsimonious Models. *Pattern Recognition*, **28**(5), 781–793.

Dempster, A. P., Laird, N. M. and Rubin, D. B. [1977]. Maximum Likelihood from Incomplete Data (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Goodman, L. A. [1974]. Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Govaert, G. and Nadif, M. [2013]. *Co-Clustering*. Wiley.

Hilbert, M. and López, P. [2011]. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, **332**(6025), 60–65.

Hunt, L. and Jorgensen, M. [1999]. Mixture Model Clustering: a Brief Introduction to the MULTIMIX Program. *Australian and New Zealand Journal of Statistics*, **41**(2), 153–171.

Laney, D. [2001]. 3D Data Management: Controlling Data Volume, Velocity and Variety. *Technical report*, Gartner.

McLachlan, G. and Peel, D. [2000]. *Finite Mixture Models*. Wiley, New-York.

Moustaki, I. and Papageorgiou, I. [2005]. Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics and Data Analysis*, **48**(3), 65—675.

Rand, W. M. [1971]. Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, **66**, 846–850.

Redner, R. and Walker, H. [1984]. Mixture densities, Maximum likelihood and the EM algorithm. *SIAM Review*, **26**(2), 195–239.

Samé, A., Chamroukhi, F., Govaert, G. and P., A. [2011]. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, **5**(4), 301–321.

Schwarz, G. [1978]. Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.

Wolfe, J. H. [1971]. A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. *Technical Bulletin STB 72-2*, US Naval Personnel Research Activity, San Diego, California.