Algorithms for Electrical Power Time Series Classification and Clustering

Gaia Ceresa, Andrea Pitto,
Diego Cirio, Emanuele Ciapessoni

Ricerca sul Sistema Energetico RSE S.p.A.
via Rubattino 54, 20134 Milano, Italy
Email: gaia.ceresa@rse-web.it
andrea.pitto@rse-web.it
diego.cirio@rse-web.it
emanuele.ciapessoni@rse-web.it

Nicolas Omont

Réseau de Transport d'Électricité RTE
92932 Paris la Défense Cedex, France
Email: nicolas.omont@rte-france.com

***Abstract*—The EU-FP7 project iTesla developed a toolbox aimed to assess dynamic security of large electrical power systems, taking into account the forecast uncertainties due to renewable energy sources and load. Important inputs to the toolbox consist in the forecasts and the realizations of thousands of active power injections from renewable generators, and of active and reactive power absorption of the load, in the high voltage French transmission grid, collected into hourly historical time series. Data show a deep variety of distribution functions and profiles in the time domain. In this context, the statistical analysis of historical dataset is very important in order to characterise and manage such a large variability of distributions. In particular, the potential multimodality of the variables has to be identified in order to adapt the sampling technique developed in the iTesla toolbox, thus assuring accurate results also for this subset of variables. Moreover, clustering some variables can help reduce the dimensionality of the problem, which represents an important advantage while analysing security on very large power systems. The paper describes four algorithms: one looks for the number of distribution's peaks and classifies the variables into unimodal or multimodal; the second and the third cluster and combine multimodal variables to obtain unimodal ones, because they are more suitable for the subsequent computation. All of them are part of an advanced tool for automatic data description, that pre-processes the raw data and produces descriptive statistics on them. The Separation Algorithm is the last one, it back-projects the sum of two series into the original components.**

***Keywords*—Multimodality; Gaussian Mixture; Cluster; Time series.**

I. Introduction

Power system security assessment is a theme of great interest for the Transmission System Operators (TSO), because they have to operate the system under the uncertainties due to renewable not programmable sources, load, and unexpected events due to climate change. The EU FP7 project iTesla [1] [2] led by the French TSO, Réseau de Transport d'Électricité, and co-funded by the European Commission, developed a tool to perform dynamic security assessment in an on-line environment, where uncertainties are dealt with by analysing the historical series of forecasts and realizations of the electrical power grid in an off-line environment. The project's output is a free toolbox, described in [3], available on GitHub [4] and usable to assess the security of any network.

The testing phase suggested further activities concerning two aspects: the choice of the most suitable set of data to train the model in the off-line part, in order to assure the most accurate result when applied in the on-line case; and an in-depth statistical description on the forecast errors. This last activity leads to a twofold consequence: the modification of the iTesla model to consider the peculiarity of some input series; the clustering of some input series, that are combined into a smoother one that is evaluated with higher accuracy by the model, together with a dimensionality reduction of the problem [5]–[7]. In any case, in order to perform security analysis on the grid, the tool needs plausible samples for the original variables before clustering, so it requires a Separation Algorithm that divides the combined variables into their original components.

This paper is organized as follows: Section II gives an idea of the entire algorithm for the data analysis; Section III accurately describes the time series classification into unimodal and multimodal; Section IV proposes two clustering algorithms and the Separation Algorithm; Section V shows one application; Section VI draws the conclusion.

II. Preprocessing

The input data are composed of two sets: snapshots of active and reactive powers related to thousands of injections/absorptions in the French electrical High Voltage (HV) and Extra High Voltage (EHV) transmission grid, hourly values over one month (or more), and their forecasts done the day before. The variables under statistical analysis are time series of forecast errors computed by

$$error_{hour,node} = snapshot_{hour,node} - forecast_{hour,node} \quad \forall hour \in \{hour_{min}, hour_{max}\} \tag{1}$$

Each time series refers to an injection or absorption, that often has different characteristics from others, in both profile and distribution: some are continuous, others focus their values

 ISBN: 978-1-61208-681-1

on a finite number of levels. Several peculiar features can be detected, such as the presence of outliers and/or of a seasonal smooth profile.

Furthermore, there are many problems that have to be solved before starting the algorithm: first, it is necessary to remove the variables not significant from a statistical point of view (with too many missing values, or too many constant values, or with a variance too low); then, the outliers are detected and deleted; finally, some overall statistical information are computed, like moments and linear correlation [5].

III. Multimodality Detection

The proposed algorithm for separating multimodal variables from the unimodal ones is composed of four steps, as in Fig. 1: detection of the peaks, fitting by using a Gaussian Mixture, comparison with conventional asymmetric unimodal distributions, application of bimodal index. The algorithm is applied to each forecast error variable independently.

A. Find Peaks

For each time series, the first algorithm step constructs a histogram with more than 10 average samples per equidistant bin, and then it detects all the bins that are local maxima; one local maximum is considered a *peak* only if its previous and its next local maxima are lower than it. The peaks lower than 10% of the highest one are not considered. The result is a too numerous set of peaks, where usually some of them are not significant: it is necessary to better define the number of modes of the variable's density distribution.

B. Gaussian Mixtures

The model tries to fit the variable's distribution function with a Gaussian Mixture [8], that is a combination of two or more unimodal Gaussian components, each one with a mean, a variance and proportion. A loop tests the best mixture, changing the number of components from 1 until the number of peaks detected in the previous step; the best solution has lowest Bayesian Information Criterion (BIC) [9]. During each fitting, the Expectation-Maximization (EM) algorithm [10] finds the best set of the k parameters of each mixture components in an iterative way, maximizing the *Likelihood* (L) function by applying the Maximum Likelihood Estimation method (MLE); repeating three times each fitting, the best case has lowest Akaike's Information Criterion (AIC) [11]. Being n the number of variable's elements,

$$BIC = -2ln(L)+k\cdot ln(n); \qquad AIC = -2ln(L)+2k. \quad (2)$$

The EM algorithm stops when it converges, i.e., when the error between L and real data is lower than a given tolerance; if 100 iterations are reached without convergence, the Mixture is discarded and another one with different number of components is tested. If the best fit is a Normal distribution, the algorithm stops and analyses the next variable.

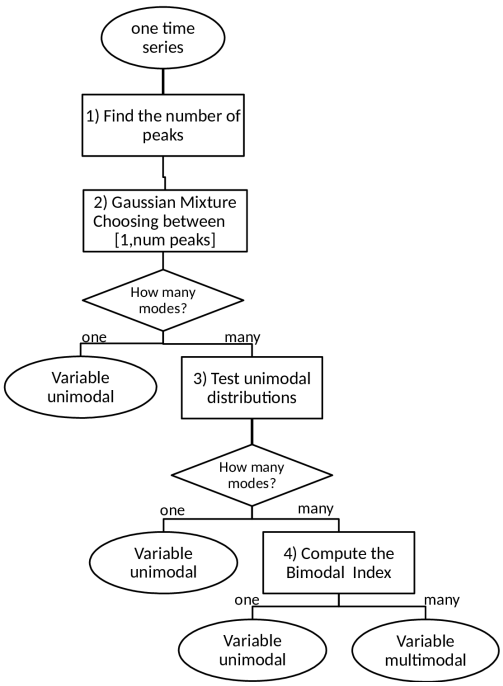

Fig. 1. Algorithm for Multimodality Detection.

C. Comparison with Unimodal Distributions

Many variables, up to this point classified among the multimodal ones, have a platykurtic and skewed distribution, and they are approximated by a Mixture composed of two or more components that are not well separated at a visual inspection. This step looks for a unimodal asymmetrical distribution that fits the variable in a better way than the Gaussian Mixture, choosing between six distributions: Weibull, Logistic, Gamma, Log-Normal, Generalized extreme value and T-location scale. The best fitting is obtained by the MLE method until convergence.

If the BIC index of one unimodal distribution is lower than the BIC of the best mixture, the algorithm classifies the variables as unimodal and analyses the next variable.

D. Bimodal Index

Variables classified as multimodal up to this point are subjected to a final step: the computation of a bimodality index, Ashman's D index [12]. It is used with the mixture of two distributions with unequal variances, σ_1^2 and σ_2^2. Let μ_1 and μ_2 their averages, the mixture is unimodal if and only if

$$D = \sqrt{2}\frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \leq 2 \quad (3)$$

It means that if the components are not well separated, the distribution could be better fitted by a unimodal distribution. If a mixture contains three or more components, the Ashman's D index is computed for each pair of components, and the global distribution is multimodal if at least one D index is higher than 2.

 ISBN: 978-1-61208-681-1

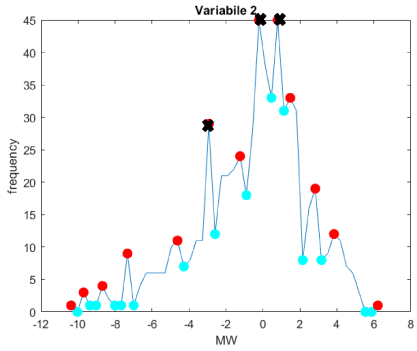

Fig. 2. Local maxima (red points) and peaks (black crosses) detected by the module *Find Peaks*.

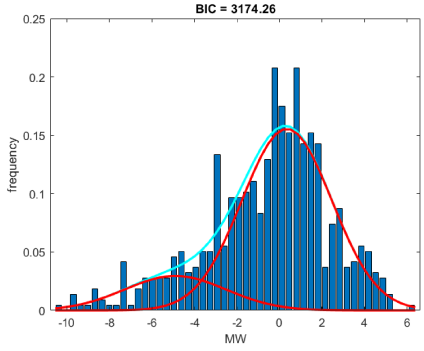

Fig. 3. Mixture with two components detected by the module *Gaussian Mixture*.

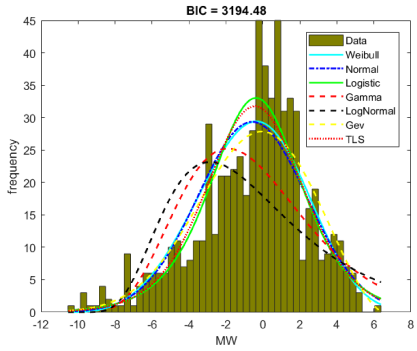

Fig. 4. Seven distributions detected by the module *Unimodal distributions*. Ashman's index $= D < 2$.

One example explains how the multimodality algorithm works: Fig. 2 displays 13 local maxima of which 3 peaks; Fig. 3 presents the best Gaussian Mixture with 2 components; Fig. 4 shows that one unimodal distribution fits the variable almost as good as the Mixture (their BICs differ for only 20 points); finally, the Ashman's D index is lower than 2, so the variable is classified among the unimodal set.

IV. Clusterization

The original iTesla tool provides very accurate results in case of unimodal variables identified in the previous section. Two consecutive developments have been achieved: the former is to modify the iTesla tool to manage the multimodal variables, as described in [5]–[7], the latter is to combine the multimodal variables to get a unimodal one, as proposed in this section. Two clusterization algorithms are described, one based on correlation and distance concepts and the other on the physical connection of the variables: for each specific application, the selected algorithm is the one which produces the highest number of clusters. In the iTesla tool, the multimodal not clustered variables are treated by a specific module.

Each clustering algorithm can generate some cluster, each one with two or three variables, and transform them into one unimodal variable, reducing the problem dimensionality. In its final part, the iTesla module applies new simulated realizations of the original variables on the grid to compute a new system states: to this purpose a Separation Algorithm decomposes the samples of the aggregated variables into the original ones.

A. Algorithm Based on Hierarchical Clustering

Each variable comes from an electrical node, that has a specific geographical position; different events can happen and induce two nodes to have a similar behaviour: the linear correlation identifies this kind of relationship. But only if two correlated nodes are close, it is probable that this liaison is physically justified by the operational practice on the system.

Fig. 5 shows the algorithm to select the variables, cluster them and check the clusters. The algorithm works separately two times, once on the active power variables and once on the reactive power ones. The first step is always to group together all the multimodal variables.

The distance function is based on the linear correlation Pearson index: two variables X and Y are close if they are highly correlated:

$$dist(X,Y) = 1 - |corr(X,Y)|\,. \tag{4}$$

Then, the hierarchical clustering produces a set of clusters composed of two variables at most; the next step verifies if they are *equal* in a particular meaning: X and Y are *equal* if their difference is higher than 1 MW (Mvar) for at most the 3% of their records. Given N the number of variable elements,

$$\begin{aligned} given\ J = \{1,2,..,N\}\,,\ I = \{j_1, j_2, ..j_{3\%N}\} \subset J \\ if\ |X_j - Y_j| < 1\ \forall j \in J \setminus I \\ \Rightarrow X = Y \end{aligned} \tag{5}$$

The reason of this *equality* is the sensitivity of measurement instruments, 1 MW (Mvar), together with the error propagation from measurement to this elaboration. Furthermore, there could be some outliers in the time series differences, that are estimated at most in the 3% of each variable's population.

The next step applies the nearest neighbour algorithm to verify the geographical distances: if variable Y falls between the k nodes closest to variable X, the cluster remains, otherwise it is filtered out. A *trial and error* approach sets parameter k equal to 50 because it is a good trade off between the neighbour's number of the urban nodes, where they are very concentrated, and the countryside where they are rare. When k decreases, also the number of good clusters gets lower.

In the final step, the time series clustered together are added: if the sum's distribution function is multimodal, the cluster is

 ISBN: 978-1-61208-681-1

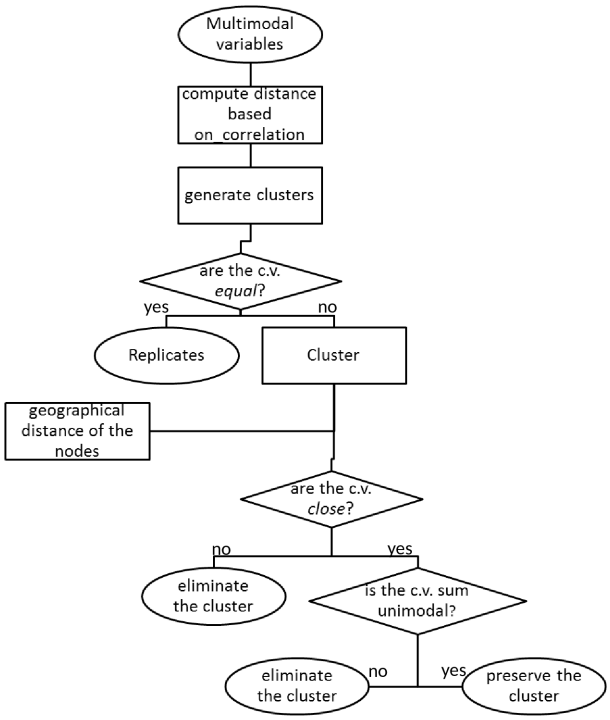

Fig. 5. Clusterization Algorithm.

eliminated, otherwise it remains and the clustered variables are treated as one smoother unimodal variable.

B. Algorithm Belonging to the Same Substation (ABSS)

In the electrical grid, a substation can be represented by busbars containing switches. When the switch is closed, all the busbars work like a unique electrical node; when it is open, the busbars are split into two half-busbars, each one working like an independent electrical node; the power injection or absorption at the substation level at each instant is the power at all the busbars. The combination of switch statuses in the entire grid, called grid topology, impacts the values because they are not always measured and the state estimator arbitrarily splits the overall substation injection/absorption when no individual data is available: the right value is unpredictable at single node level, but foreseeable at substation level. The variables belonging to the same substation have the same first 7 characters in their names.

From a mathematical point of view, the forecast errors of the electrical nodes that lie in a substation are large, with generally an irregular distribution and many peaks, but the forecast error at the substation level has better statistical properties, often showing a unimodal distribution.

The strategy adopted is to cluster the multimodal variables in the same substation, adding their time series in order to obtain one unique time series with a smoother distribution. The algorithm:

1) Considers separately the variables of active and reactive power.
2) Groups the variables (usually two or three) that are in the same substation.
3) Removes the clusters containing *equal* variables, like defined in Equation 5.
4) Sums the two or three time series of clustered together variables.
5) For each group, it checks if the resulting sum is unimodal: if yes, the cluster remains, otherwise, it is eliminated.

C. Separation Algorithm

In the iTesla work flow, the clustered variables are sent to the sampling module, aimed to generate plausible realisations of the same variables. However, in order to perform studies on the grid, it is necessary to back-project these samples of clustered variables into the samples of original variables, one for each electrical node of the system.

Given the sum of clustered variables in the overall system, the Separation Algorithm have to split them to assign a value to each single variable, preserving the cluster sum. The physical aspect is the most important: it must preserve the overall variability of the system, avoiding that one variable has a too high variance, because it can lead to a computational problem of system stability without any correspondence in the reality. It is important to preserve also the correlation between variables, if present.

Two splitting algorithms are proposed, respectively for two and three clustered variables.

The fundamental formula at the base of all the reasoning is the variance of the sum of two variables.

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \quad (6)$$

Suppose that the historical time series are X_o and Y_o, their unimodal sum is Z_o. The iTesla platform has to simulate new values for the system variables, preserving the original distributions: it estimates a new sum Z_n; the Separation Algorithm tries to split Z_n in X_n and Y_n, with both variances not too high and with linear correlation similar to the original series: it estimates X_n, and computes Y_n like difference between the sum and X_n, in order to not modify the sum $Z_n = X_n + Y_n$. For the sake of simplicity, in this explication $Z_o = Z_n$. So it remains

$$\begin{aligned} Var(X_o) + Var(Y_o) + 2Cov(X_o, Y_o) = \\ Var(X_n) + Var(Y_n) + 2Cov(X_n, Y_n) \end{aligned} \quad (7)$$

Since the variable X_n will be calculated considering its dependence from Z_n, its variance decreases, while the remaining sum increases:

$$\begin{aligned} Var(X_o) \geq Var(X_n) \\ Var(Y_o) + 2Cov(X_o, Y_o) \leq Var(Y_n) + 2Cov(X_n, Y_n) \end{aligned} \quad (8)$$

Separation Algorithm for Two Variables: Given:

- X_o, Y_o the original variables
- $Z_o = X_o + Y_o$
- Z_n the new sum generated inside the iTesla tool, for testing model it is $Z_n = Z_o$.

 ISBN: 978-1-61208-681-1

The algorithm works as follows:

- if $|corr\,(X_o, Y_o)| \leq 0.9$ then
  - it calculates average $\mu_{X|Z}$ and variance $\sigma_{X|Z}$ of X_n conditioned to Z_n
  - it generates the distribution $\mathcal{N}_{X|Z}\ (\mu_{X|Z}, \sigma_{X|Z})$
  - it extracts randomly one realization of X_n from $\mathcal{N}_{X|Z}$
  - it computes $Y_n = Z_n - X_n$
- if $|corr\,(X_o, Y_o)| > 0.9$ then
  - $ratio = \frac{|X_o|}{|X_o|+|Y_o|}$
  - $X_n = sign(corr(X_o, Z_o)) \cdot ratio \cdot Z_n$
  - compute $Y_n = Z_n - X_n$

Separation Algorithm for Three Variables: Given:

- X_o, Y_o, V_o the original variables
- $Z_o = X_o + Y_o + V_o$
- Z_n the new sum generated inside the iTesla tool, for testing model it is $Z_n = Z_o$.

The algorithm works as follows:

- $ratio1 = \frac{|X_o|}{|X_o|+|Y_o|+|V_o|}$
- $X_n = sign(corr(X_o, Z_o)) \cdot ratio1 \cdot Z_n$
- $ratio2 = \frac{|Y_o|}{|X_o|+|Y_o|+|V_o|}$
- $Y_n = sign(corr(Y_o, Z_o)) \cdot ratio2 \cdot Z_n$
- $V_n = Z_n - X_n - Y_n$

V. Case Study

The implementation and the testing phase are done with Matlab 2017b.

The analysed dataset is composed of the time series of forecast and realizations of 3194 withdrawal/injections in the electrical French transmission grid, each one with 654 records collected once per hour from 2013/03/01 00:30 to 2013/03/01 23:30, concerning loads and renewable sources. The variables under study are the forecast errors, obtained by the Equation 1.

The preprocessing keeps 3122 significant variables, 80% of the original dataset.

A. Multimodality

Fig. 6 reports the results of the application of the algorithm for the multimodality detection. It can be noticed that the number of detected multimodal variables (700 out of 3122) is significant, which justifies the need for a proper management of the multimodal variables in the iTesla tool.

B. Clusters

The two Clustering Algorithms find different numbers of clusters. The Clustering ABSS finds 42 clusters with 2 variables and 11 with three variables, while the other algorithm finds 28 clusters with two variables: in this application case the ABSS is preferable to reduce the problem dimensionality because it clusters 117 variables. An example of cluster is in Fig. 7: the histograms of three multimodal variables related to the same substation are shown in blue, while their unimodal sum, that is the total production or absorption of the substation, is shown in magenta.

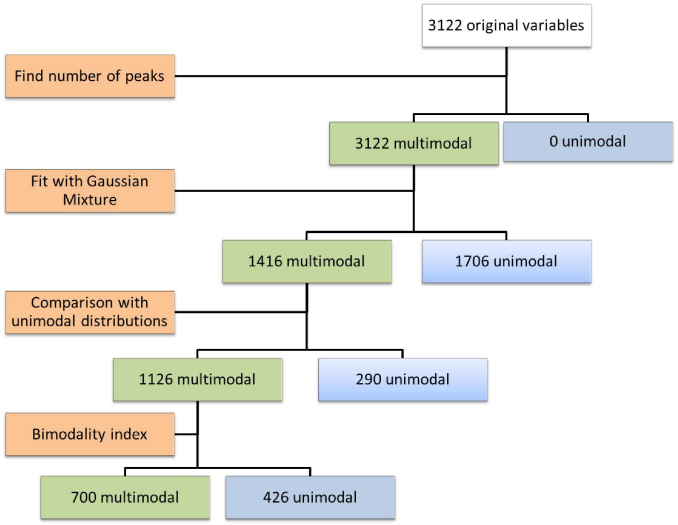

Fig. 6. Classification process operated by the Multimodality Algorithm on the test case.

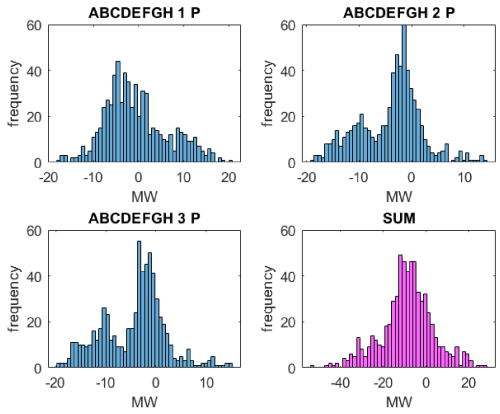

Fig. 7. Example of three clustered variables within same substation.

C. Separation Algorithm

Given the 42 clusters composed of two variables, 11 clusters with three variables, the sum's decomposition into the contributions of two (or three) variables has to preserve the total variability caused by the clustered stochastic variable in the system, i.e., each component does not have a too big variance with respect to the original variables. Table I shows the statistical description of three examples of new variables obtained with the decomposition method described above. The last column is the difference between total standard deviation of the original variables and those of new variables: $total\ \ variation\ \ =\ \ (\sigma_A + \sigma_B)_{new} - (\sigma_A + \sigma_B)_{orig}$; if this value is negative, the original variables have an overall variability higher than the new case, so the decomposition algorithm does not add variability to the system. If the total variation is positive, the new variables have higher standard deviations compared to the original ones; in these cases, it is

TABLE I. COMPARISON BETWEEN ORIGINAL VARIABLES AND NEW ONES, OBTAINED BY DECOMPOSITION OF THEIR SUM. MW FOR ACTIVE POWER AND Mvar FOR REACTIVE POWER.

	Var	μ_A	μ_B	σ_A^2	σ_B^2	Cov_{AB}	σ_A	σ_B	$corr_{AB}$	$\sigma_A^2 + \sigma_B^2$	$\sigma_A + \sigma_B$	total variation
1	original	3.26	-1.54	84.4	134.1	-35.4	9.2	11.6	-0.3	218.5	20.8	1.04
1	new	3.25	-1.58	60.6	196.7	-54.7	7.8	14	-0.5	257.3	21.8	-
2	original	0.84	-2.61	36.7	32	-20.7	6.1	5.7	-0.6	68.7	11.7	0.5
2	new	0.91	-2.70	26.8	49.4	-24.5	5.2	7	-0.7	76.2	12.2	-
3	original	2.66	-1.50	182.2	125.3	-126	13.5	11.2	-0.8	307.5	24.7	-1.46
3	new	3.17	-0.98	116.3	155	-108	10.8	12.5	-0.8	271.3	23.3	-

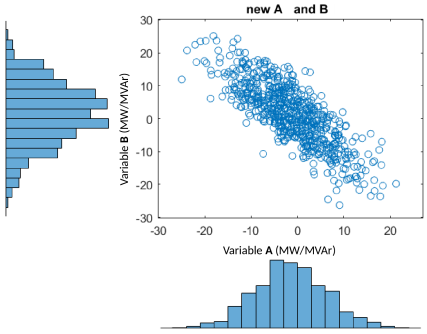

Fig. 8. Histograms and scatter plots of the original variables A and B. v1 is variable A, v2 is variable B.

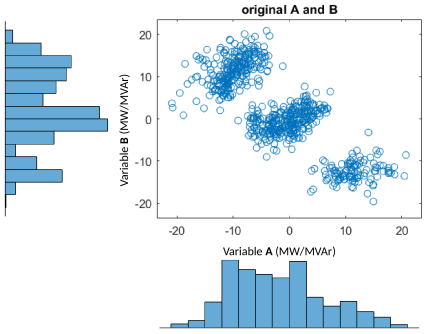

Fig. 9. Histograms and scatter plots of variables obtained by the separation of the sum A+B. A3 si variable A, B3 is new variable B.

desirable a low value. Looking at the results, the averages of original and new variables are very similar, the variances of variable A decreases while those of B increase; considering the standard deviations, they increase a little and their sum in original variables have a very small difference from their sum in new variables. One graphical comparison is shown in Fig. 8 (original variables) and Fig. 9 (splitted variables): the distributions of the splitted variables are smoother than those of the original ones, but they preserve the direction of the correlation and the standard deviations.

Considering all the 53 clusters, in 17 cases the variation is positive, but the worst case is 4.15 MW/Mvar, while all the other differences are lower than 2 MW/Mvar. These values demonstrate the goodness of this Separation Algorithm considering its objective.

VI. CONCLUSION

The paper has presented some algorithms to detect the multimodality of power system forecast errors and to cluster multimodal variables into aggregated variables with smoother statistical properties. The need for these algorithms comes from the application of an advanced toolbox for power system security assessment developed in the EU project iTesla. The results of the application of the first algorithm show that it is effective in identifying the set of multimodal variables, which can be a significant fraction of the total amount of variables, in a real life operational environment in power systems. Moreover, the tests on the clusterization techniques show that few pairs or triples of multimodal variables can be reduced into unimodal aggregated variables. The last simulations also show that the back-projection techniques used to go back from the samples of clustered variables to the original components are effective in generating reasonable samples of each individual original component without altering the variability in the system. The proposed techniques are general and can be applied to any kind of data.

REFERENCES

[1] "iTesla Project," Sep 2016, URL: http://www.itesla-project.eu/.
[2] M. H. Vasconcelos et al., "Online security assessment with load and renewable generation uncertainty: The itesla project approach," in 2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Oct 2016, pp. 1–8.
[3] "iTesla Power System Tools description," May 2018, URL: http://www.itesla-pst.org/.
[4] "iTesla Power System Tools code," Sep 2018, URL: https://github.com/itesla.
[5] A. Pitto, G. Ceresa, D. Cirio, and E. Ciapessoni, Power system uncertainty models for on-line security assessment applications: developments and applications. Rapporto RdS RSE 17001186, Feb 2017.
[6] A. Pitto and G. Ceresa, Automated techniques for the analysis of historical data and improvement accuracy of uncertainty models for security assessments of the electrical power grid. Rapporto RdS RSE 17007093, Feb 2018.
[7] G. Ceresa, E. Ciapessoni, D. Cirio, A. Pitto, and N. Omont, "Verification and upgrades of an advanced technique to model forecast uncertainties in large power systems," in 2018 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Jun 2018, pp. 1–8.
[8] D. Reynolds, Gaussian Mixture Models. Springer US, 2009, pp. 659–663.
[9] Bayesian Information Criteria. New York, NY: Springer New York, 2008, pp. 211–237.
[10] Y. Chen and M. R. Gupta, "EM Demystified: An Expectation-Maximization Tutorial," UWEE Technical Report, vol. UWEETR-2010-0002, Feb 2010.
[11] H. Akaike, Akaike's Information Criterion. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 25–25.
[12] K. Ashman, C. Bird, and S. Zepf, "Detecting bimodality in astronomical datasets," Astronomical Journal, vol. 108, pp. 2348–2361, 1994.

 ISBN: 978-1-61208-681-1