Robust short-term load forecasting using projection statistics

Yacine Chakhchoukh
and Patrick Panciatici
DMA, RTE
Versailles, France
Email: yacine.chakhchoukh@lss.supelec.fr

Pascal Bondon
CNRS, univ. Paris XI
Gif-sur-Yvette, France

Lamine Mili
Virginia Tech
Falls Church, USA

Abstract—It has been observed that the French electric load series possesses outliers and breaks. Outliers are deviant data points while breaks are lasting abrupt changes in the stochastic pattern of the series. It turns out that outliers and breaks significantly degrade the reliability and accuracy of conventional day-ahead estimation and forecasting methods. Robust methods are needed for this application. In this paper, we propose to use a robust diagnostic approach for which the identification of outliers and breaks is carried out via a robust multivariate estimation of location and covariance based on projection statistics (PS). The developed procedure consists of the following steps: (i) estimate the parameters and the order of a high order autoregressive AR(p) by means of the PS, (ii) execute a robust filter cleaner to identify and reject the outliers, and (iii) apply a maximum-likelihood estimator defined at the Gaussian distribution that handles missing values. The performance of this method has been evaluated on the French electric demand in terms of execution time and forecasting accuracy. This approach improves the load forecasting quality for "normal days" and presents several interesting properties such as fast execution, good robustness, simplicity and easy on-line implementation. A novel multivariate approach is also proposed in order to deal with heteroscedasticity.

Index Terms—load forecasting, Robustness, projection statistics.

I. INTRODUCTION

This work was motivated by RTE, a company that manages and operates the electric power transmission system in France. RTE uses seasonal autoregressive integrated moving average (SARIMA) models to carry out the short-term forecasts of the French electric demands. The forecasting method implemented by RTE consists of the following steps. The load time series is first corrected from the influence of the weather via the application of a regression model in which the explanatory variables are the temperature and the nebulosity recorded in few selected cities and towns in France. The nebulosity consists in the cloud cover in real time. It influences the electricity consumption since it plays an important role in Electric lighting consumption. After adjustment, the series exhibits a general growth trend and several major cycles (daily, weekly, seasonal, yearly, etc.). Typically, the adjusted daily load series are classified into groups with similar patterns corresponding to weekends, working days, non-working days, public holidays and some special days.

Typically, the adjusted French electric load series possesses deviant points, termed outliers, and lasting abrupt changes, termed breaks. Outliers are due to strikes, bridge days, exceptional events, etc. Breaks are caused by holidays, non working days, etc. Because outliers and breaks significantly degrade the reliability and accuracy of conventional estimation and forecasting methods, we resort to robust statistics. Specifically, we have developed a new robust method for the identification of the outliers and the breaks in the series. It makes use of projection statistics (PSs), which are robust multivariate estimates of location and covariance. The performance of this technique is compared to three robust methods proposed in the literature, which are: 1) the classical maximum-likelihood-based estimation method of SARIMA models applied after the execution of a three-sigma outlier rejection rule, which is denoted by Corrected Maximum Likelihood, or CML for short [1]; 2) a robust Generalized M-estimation method that suppresses the outliers via a down-weighting function [2, 3]; and (3) the robust filtered \( \tau \)-estimator proposed by Yohai and Zamar [1].

The paper is organized as follows. Section II presents the new modeling approach of the electrical series. Section III describes the robust estimation method in the case of an ARMA model. Section IV provides the evaluation and the application of the approach to short-term load forecasting. Finally, Section V is devoted to the conclusions and perspectives.

II. MODELING THE LOAD SERIES

In RTE, the half-hourly electrical consumption series adjusted from the weather effects (48 records per day) is modeled by means of a double seasonal ARIMA model (SARIMA). The series contains both a daily and weekly seasonality. A double seasonal ARIMA model, SARIMA\((p, d, q)\times(p_1, d_1, q_1)\times(p_2, d_2, q_2)\) follows the equation:

\[
\phi_p(B)\Phi_{p_1}(B^{s_1})\Omega_{p_2}(B^{s_2})\nabla^d\nabla_{s_1}^d\nabla_{s_2}^d Y_t = \theta_q(B)\Theta_{q_1}(B^{s_1})\Psi_{q_2}(B^{s_2})\epsilon_t
\]

where \(Y_t\) is the electricity demand at time \(t\), \(s_1\) and \(s_2\) are the number of periods in the different seasonal cycles, \(B\) is the lag operator, \(\nabla\) is the difference operator, \(\nabla_{s_1}\) and \(\nabla_{s_2}\) are the seasonal difference operators \(B^{s_1}Y_t = Y_{t-k}, \nabla = 1 - B, \nabla_{s_1} = 1 - B^{s_1}\), \(\phi_p, \Phi_{p_1}, \Omega_{p_2}, \theta_q, \Theta_{q_1}, \Psi_{q_2}\) are polynomials of order \(p, p_1, p_2, q, q_1, q_2\), \(\epsilon_t\) is a gaussian white noise from \(N(0, \sigma^2)\).

In this article, we propose an alternative modeling to deal with heteroscedasticity. Casted in a multivariate modeling framework, the electric consumption is represented as a set
of 48 time series corresponding to the hours of the day with a statistical model for each hour (00:00, 00:30, ..., 23:30) and considering the correlation between adjacent hours. We obtain forty eight (48) univariate series modeled by seasonal ARIMA models, SARIMA \((p^h, 0, q^h) \times (p^h, 1, q^h)_7\) which follows the equation

\[
\theta^h(B)\phi^h(B^7)^{-1}Y^h_t = \epsilon^h_t,
\]

where \(Y^h_t\) is the daily electricity demand on day \(t\) at time or hour \(h\) \((h = 1, \ldots, 48)\) corresponds to 00:00 to 23:30 as an example at 12:00, \(h = 24\), the number of periods is equal to 7 to model the within-week seasonal cycle. \(\phi^h\), \(\theta^h\), \(\Theta^h\) are polynomials of order \(p^h\), \(p^h\), \(q^h\), \(q^h\), \(\epsilon^h_t\) is a Gaussian white noise from \(N(0, \sigma^2_{\epsilon,h})\).

The residuals obtained from the SARIMA model at various hours are standardized via a robust estimator of scale. They are referred to as the standardized residuals. The standardized residuals of adjacent hours are correlated and their correlation is modeled by an ARMA model defined as

\[
\varphi(B)\rho^h_t = \vartheta(B)\epsilon^{t+48}_{(t-1)+h}
\]

This model is used to improve the prediction. The series of the residuals is given by \(\{\epsilon^1_t, \ldots, \epsilon^n_t, \ldots, \epsilon^1_t, \ldots\}\), where \(n\) is the number of days in the series and \(\epsilon^h_k\) is the residual of the day \(k\) for the series of hour \(h\). The idea is to use different forecasting models one for each half-hour of the day, casted as a simple vector time series approach that is compatible with on-line applications. The improvement can be explained by the fact that the load demand at a certain time slot of a day is more correlated to the past demand values of the same time slot of previous days than the past or present demands of adjacent slots.

III. OUTLIERS DETECTION AND ROBUST ESTIMATION

In this section, we introduce a new fast robust short-term load forecasting method based on Projection Statistics [4]. This method allows us to both estimate the parameters of a SARIMA model and identify its order. The performances of this method in terms of robustness to outliers and forecasting accuracy are compared to those of the CML method [1], the Generalized-M-estimator [2], [3], and the filter-cleaner [1], which has recently been introduced by Maronna at al. [1] for ARMA models.

A. Robust Outlier Detection based on PS

Consider an AR\((p)\) model given by

\[
X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = \epsilon_t,
\]

where \(\epsilon_t\) is an independent identically distributed (i.i.d) Gaussian sequence, \(\epsilon_t \sim N(0, \sigma^2_{\epsilon})\). The purpose is to estimate the parameters \(\phi\) from the observed variables \(X_1, \ldots, X_n\). Viewing it as a linear regression model, we have

\[
\begin{pmatrix} X_{p+1} \\ \vdots \\ X_n \end{pmatrix} = \begin{pmatrix} X_p & \cdots & X_1 \end{pmatrix} \begin{pmatrix} \phi_1 \\ \vdots \\ \phi_p \end{pmatrix} + \begin{pmatrix} \epsilon_{p+1} \\ \vdots \\ \epsilon_n \end{pmatrix}
\]

which can be written in the matrix form as \(Y = X\phi + \epsilon\).

1) The Mahalanobis Distances: The intent here is to develop a method aimed at identifying the outliers in the point cloud defined by the vectors \(\{X_1, \ldots, X_{n-p}\}\) in a p-dimensional space, where \(X^T_i\) denote the \(i\)th row vector of the matrix \(X\). To this end, we need a measure of the distance of each point with respect to the bulk of the point cloud. The conventional measure is provided by the Mahalanobis Distance (MD), which is given by

\[
MD_i = \sqrt{(X_i - \bar{X})^T C^{-1}(X_i - \bar{X})}
\]

where \(\bar{X}\) is the sample mean and \(C\) is the sample covariance matrix. Because the MDs are based on non-robust statistics, they fail to reveal all the outliers.

2) The PS method [4]: A more robust method is provided by the PS. They are implemented through the following algorithm. Let \(M\) be the p-dimensional vector whose components are the median of each column of the matrix \(X\), and let \(v_i\) be the normalized vector

\[
v_i = \frac{X_i - M}{\|X_i - M\|}, \quad i = 1, \ldots, n - p.
\]

For every direction \(v_i\), calculate

\[
P_{ij} = \frac{Z_{i,j} - L_i}{S_i}, \quad j = 1, \ldots, n - p,
\]

where \(Z_{i,j} = v_i^T X_j\) is the projection of \(X_j\) on the direction \(v_i\), \(L_i = \text{med}(Z_{i,1}, \ldots, Z_{i,n-p})\) and \(S_i = 1.4826 \text{med}(Z_{i,1} - L_i, \ldots, Z_{i,n-p} - L_i)\). The projection of the point \(X_i\) is given by \(PS_i = \max(P_{i,1}, \ldots, P_{i,n-p})\). When the vectors \(X_i\)'s follow a multivariate normal distribution, the PSI's roughly follow a chi-square distribution with \(p\) degrees of freedom. Consequently, a robust statistical test consists in tagging as outliers all the data points that have \(PS_i > \chi^2_{p,0.975}\).

Finally, the parameters of an ARMA\((p,q)\) are estimated via the following steps:

- Fit a high order AR\((p^*)\) whose parameters and order \(p^*\) are estimated by means of the PS, with \(p^* \gg p\).
- Detect and reject the outliers after filtering the series with the high order AR\((p^*)\) using the method proposed by Marisflez and Martin [5] and described [1];
- Finally, apply a classical maximum-likelihood-based estimation method of ARMA models to compensate for the missing values [6].

B. Robust GM estimation

The GM estimator used in this article is given in [2]-[3]. It is defined as the minimum of the objective function given by

\[
J(\phi) = \sum_i \rho_i(\hat{\phi}_i, \sigma^2) W_i(\bar{\chi}_{i-1|i-1})
\]

where \(\rho_i(\phi) = Y_i - \hat{Y}_{i|i-1}\) is called the robust prediction residual, \(\hat{\chi}_{i} - \bar{\chi}_{i-1}\) and \(\bar{\chi}_{i-1}\) are obtained using the robust filter-cleaner and the state representation explained in [1], \(\sigma^2\) is a robust scale estimator of the robust prediction residuals (M-estimate of scale [1]) and \(\rho(\cdot)\) is a robustifying loss function such as the Huber function defined by

\[
\rho(x) = \begin{cases} x^2/2 & \text{if } x < c, \\ c|x| - c^2/2 & \text{if } x \geq c, \end{cases}
\]
where \( c \) is a positive constant. The estimation has been obtained via the Iteratively Re-weighted LS (IRLS) algorithm [1]. In the algorithm, Huber loss function was used in the first iterations followed by a Tukey psi-function to improve the robustness [1].

C. Selection of the order of the AR model

As shown before, to compute the PS we need to know the order \( p \) of the AR model. Order selection is then important for a good detection of outliers and therefore for a good estimation of the parameters of the model. Information criteria are computed for several values of \( p \) and the optimal order is given by minimizing the criteria. We have chosen some of the most popular criteria that trade-off the goodness of the fit and the complexity of the model. These are the Akaike Information Criterion (AIC) [7], the Bayesian Information Criterion (BIC) [8], and the Hannan-Quinn criterion (HQ) [9], which are respectively defined as

\[
AIC(p) = \ln \sigma_p^2 + \frac{2p}{n-p}, \quad BIC(p) = \ln \sigma_p^2 + \frac{p \ln(n-p)}{n-p},
\]

\[
HQ(p) = \ln \sigma_p^2 + \frac{2p \ln(n-p)}{n-p}.
\]

Outliers might affect the optimal selection of the order. To overcome this problem, we estimate \( \sigma_p^2 \) in a robust way and replace \( n-p \) by \( \hat{n} - p - n_o \), where \( n_o \) is the number of the detected outliers. We then disregard the outliers flagged by the PS in the evaluation of \( \hat{\sigma}_p \). The criterion is robust since it does not take into account the outliers.

D. SIMULATION RESULTS

In this section, the effects of the outliers on the foregoing methods are assessed when these methods are applied to the parameter estimation of an AR(1) and to the selection of the order of an autoregressive model from a contaminated series.

1) Impact of the outliers on order selection: We generate one hundred AR(3) series of five hundred samples with additive noises following \( N(0,1) \) and contaminated by outliers with randomly chosen values and positions. The amplitude of each outlier is taken from a uniform distribution \( U(0,4\sigma_x) \), where \( \sigma_x \) is the standard deviation of the clean signal. Two cases are considered, one has 6% of contamination while the other one has 20%. In each case, we identify the order of the AR models with each of the foregoing three methods. The results are displayed in Table I. From that table, we observe that the PS-based order selection method gives the best results since 82 out of 100 order selections provide the exact order 3 while the remaining 18 order selections provide a smaller order of 2. Regarding the LS- and the GM-based methods, they select an order that ranges from 1 to 9. This can be explained by the non robustness of the LS estimator and a decaying robustness of the GM estimator with an increasing order of the autoregressive model. The latter exhibits a decreasing breakdown point when the number of parameters to be estimated grows.

2) Impact of the outliers on parameter estimation: The contaminated observation is generated by: \( y_t = x_t + 4 + \alpha_t \) where \( \alpha_t \sim N(0,1) \). Table II displays the mean, the standard deviation, and the mean-squared-error of the estimated parameter \( \phi \) and the residual estimated variance of an AR(1) model. In this table, \( LS^* \) denotes the LS method applied to clean series.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\mu} (\phi) )</td>
<td>0.142</td>
<td>0.045</td>
<td>0.130</td>
<td>0.765</td>
<td>0.317</td>
<td>0.451</td>
<td>0.511</td>
<td>0.498</td>
<td>0.074</td>
</tr>
<tr>
<td>( \sigma (\phi) )</td>
<td>0.002</td>
<td>0.004</td>
<td>0.005</td>
<td>0.006</td>
<td>0.007</td>
<td>0.008</td>
<td>0.009</td>
<td>0.010</td>
<td>0.011</td>
</tr>
<tr>
<td>MSE</td>
<td>0.317</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
<td>0.333</td>
</tr>
</tbody>
</table>

TABLE I

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\mu} (\hat{\sigma}^2) )</td>
<td>0.049</td>
<td>0.046</td>
<td>0.046</td>
<td>0.046</td>
<td>0.046</td>
<td>0.046</td>
<td>0.046</td>
<td>0.046</td>
<td>0.046</td>
</tr>
<tr>
<td>( \hat{\sigma} (\hat{\sigma}^2) )</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td>MSE</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
<td>0.006</td>
</tr>
</tbody>
</table>

TABLE II

IV. APPLICATION TO SHORT-TERM LOAD FORECASTING

We consider real half-hourly French electric load demands over 100 weeks, covering the period that ranges from Sunday, February 1st, 2004 to Sunday, December 31st, 2005. The samples over the first 70 weeks are employed to estimate the parameters of the AR model while those over the remaining 30 weeks are utilized to evaluate the quality of the forecasts provided by the proposed methods up to 24 hours ahead. This yields 490 data samples for estimating the AR models and 210 samples for forecasting evaluation for each of the 48 half-hour daily time series.

We applied the forecasting performance is made by means of the mean-absolute-percentage-error (MAPE) criterion, which is computed up to a leading period of 10 days. This criterion is calculated over normal days without breaks or outliers, making it a robust metric. Let \( \{Y_1, \ldots, Y_n\}; n = 500 \) denote a given series and let \( \hat{Y}_{n+h} \) denote the (out-of-sample) predictor of \( Y_{n+h} \) based on an adjusted model to \( \{Y_1, \ldots, Y_n\} \).

The mean-absolute-percentage error is defined as

\[
\text{MAPE} = \frac{100}{H} \sum_{h=1}^{H} \frac{Y_{n+h} - \hat{Y}_{n+h}}{Y_{n+h}}, \quad (4)
\]

where \( H \) is a given final step prediction.

In this section, we compare the performances of the PS-based estimation method with those of the GM-estimator and the classical SARIMA estimation method applied after treating the outliers via the three sigma rule (CML). This rule consists in rejecting the outliers that are flagged as standing beyond three times a robust estimator of scale from a robust trend estimate (that is, the central part) of the time series. Furthermore, the performances of the proposed multivariate modeling approach are compared with those of the univariate approach based on the maximum likelihood model estimation after the application of the three-sigma-outlier-rejection rule to the series.

The resulting MAPE values are presented in Fig. 1 and Fig. 2. In Fig.1, the MAPE for hours 8:00 and 22:00 is presented. It is clear from the displayed curves that the PS-based estimator improves the quality of the forecast.
the improvement of the forecast varies from hour to hour. This is expected since the time series of some hours are more contaminated by outliers than others. For example, the series for the night hours (i.e., at 22:00) are not contaminated. In this case, it is interesting to note that the PS-based estimator has the same performance as the CML. This illustrates the efficiency of the former estimator under normality. Note also that the GM-estimator is not as good as these two estimators. This can be explained by the fact under Gaussianity, it possesses a smaller efficiency.

In Fig. 2, we show the MAPE of the PS-based estimator, the GM estimator, the CML when applied to the proposed multivariate approach. We also display the MAPE of the CML method applied to the univariate approach. The associated curve is denoted by U3MADN. As observed, the proposed robust methods improve the quality of forecasting. Now, we propose to compare our simple PS-based approach to the filtered $\tau$-estimates, which are robust and highly efficient estimates recently introduced in the statistical literature [1]. These estimates have an efficiency of 95% at the Gaussian model and a breakdown point of 50%. From Fig. 3, we notice that for this application our approach, while being simpler and fast to execute, has almost the same performance as the filtered-$\tau$ estimates. We also notice that the main improvement stems from the application of our method in a multivariate framework of 48-time series models.

V. CONCLUSIONS AND FURTHER WORK

In this paper, a fast robust method for load forecasting with contaminated samples has been developed and its performance evaluated. It makes use of the projection statistics, a robust estimation method of location and covariance. This technique is compared to the widely used CML and the GM-estimation method on the French load time series. From the simulations, the following conclusions may be drawn:

- The projection statistics are a powerful diagnostic tool for outlier detection, parameter estimation and order selection of SARIMA models. They exhibit better performance than the GM-estimation method. This stems from the fact that the latter has a decreasing breakdown point with increasing order of the autoregressive model. However, the GM-estimates are much better than the conventional CML.
- In this application, the PSs have the same performance as the filtered-$\tau$ estimates. They are preferred to the latter because of their simplicity and fast executing time.
- The modeling approach has shown its effectiveness in the case of the French load forecasting.
- Our simulations have revealed that robust methods are reliable techniques for automatic on-line estimation and forecasting. They can offer a good tradeoff between robustness and efficiency. They constitute also better alternatives to the pragmatic analysis based on experience used by the electric companies.

Other robust highly efficient estimators will be developed and compared in the case of load forecasting.

REFERENCES