Robust Estimation of Time-varying Coefficient Functions - Application to the Modeling of Wind Power Production

Pierre Pinson, Henrik Aa. Nielsen, Henrik Madsen
(pp@imm.dtu.dk, han@imm.dtu.dk, hm@imm.dtu.dk)
Informatics and Mathematical Modelling
Technical University of Denmark
Lyngby, Denmark

March 9, 2007

PSO Project number: FU 4101
Ens. journal number: 79029-0001
Project title: Intelligent wind power prediction systems

Contents

Summary

Conditional parametric models with time-varying coefficient functions may be used in forecasting tasks, by proposing a mean for adaptive mean regression of nonlinear and nonstationary processes. Though, when using a classical least square criterion for the estimation of coefficient functions, estimates are affected by the presence of significant noise, and possibly outliers, in the response/explanatory variables. This is indeed the case if these models are used for forecasting wind power production, which is a nonstationary, nonlinear and bounded process. A method for an adaptive and robust estimation of coefficient functions is proposed in the present document. An asymmetric but convex M-type estimator is introduced in order to deal with non-Gaussian distributions of residuals, which may be skewed and heavy-tailed. A recursive formulation is given for the estimates to be adaptive. Also, a local M-type estimator is proposed, in order to account for the weighting present in local polynomial regression. Finally, a simple nonparametric method is described for an adaptive scaling of the introduced local M-type estimator. An original feature of that M-type estimator is that instead of specifying a threshold point, one gives a proportion of residuals that may be considered as suspicious. The nice properties of the method are highlighted on semi-artificial datasets corresponding to wind speed measurements and simulated power output for a wind farm in Denmark. Validation results are also given on real-world data from the Middelgrunden wind farm in Denmark, on an exercise consisting in the modeling of the conversion function from meteorological forecasts of wind speed to wind power measurements, consequently used for forecasting purposes.

1 Introduction

Let $\{y_i\}$, $i = 1, \ldots, N$, be an observed time series, and consider a general regression of the form

$$y_i \;=\; g(\mathbf{x}_i) + \epsilon_i, \qquad i = 1, \ldots, N \tag{1}$$

where $\mathbf{x}_i^\top = [x_i^1 \ldots x_i^k \ldots x_i^l]$ is a vector of l explanatory variables at time i. $\mathbf{x}_i$ may include lagged values of the response variable y, or alternatively historical or forecast values of explanatory variables, i.e. variables that are known to have an influence on the process of interest. The noise term $\{\epsilon_i\}$, $i = 1, \ldots, N$, is a sequence of independent and identically distributed (i.i.d.) random variables with unknown distribution F. It is assumed that F has a zero mean and a finite variance σ_ϵ^2. In the following, it is assumed that both x- and y-values can be normalized. Therefore, they are all contained in the unit interval, while $\epsilon_i \in [-1, 1]$, $\forall i$.

If using a conditional parametric model for g, then Equation (1) can be rewritten as

$$y_i \;=\; \mathbf{x}_i^\top \boldsymbol{\theta}(\mathbf{u}_i) + \epsilon_i, \qquad i = 1, \ldots, N \tag{2}$$

where $\boldsymbol{\theta}$ is a vector of coefficient functions to be estimated. In the formulation given by the above Equation, explanatory variables at time i are sorted into two groups $\mathbf{x}_i$ and $\mathbf{u}_i$, such that the resulting model is conditional to $\mathbf{u}$. In practice, the curse of dimensionality imposes that the dimension of $\mathbf{u}$ has to be low, say 1 or 2 (for a discussion on that issue, see Hastie and Tibshirani (1990, pp. 83-84)). In the case where the considered process is nonstationary, the $\boldsymbol{\theta}$-functions are referred to as time-varying coefficient functions. For their adaptive estimation, Nielsen et al. (2000) proposed a method that is a combination of local polynomial regression and recursive least-squares with exponential forgetting. This estimation method is nonparametric since no assumption is made about the form of the $\boldsymbol{\theta}$-functions. Also, it is adaptive in time since these functions are updated every time an observation becomes available.

For real-world test cases, available measured data may contain a significant noise component, whose distribution may be skewed, heavy-tailed, and possibly include outliers. This results in affecting the estimation of the $\boldsymbol{\theta}$-functions. Focus is given here to robustifying the estimation method initially introduced by Nielsen et al. (2000). In practice, the objective function to be minimized is generalized to a broader class of loss functions. This yields an M-type estimator $\hat{\boldsymbol{\theta}}^\dagger$ with convex loss functions, which is inspired by the now classical M-estimator introduced by Huber (1981). In addition, $\hat{\boldsymbol{\theta}}^\dagger$ is locally robustified by accounting for the influence of weighted residuals, the weights being given from the local polynomial regression. Finally, the threshold points of the bounded-influence loss function are adaptively scaled from a nonparametric estimation of the distribution of potential residuals for the current model estimates. The two additional parameters of the method are m the number of residuals for the estimation of residual distributions, and a parameter α that gives the proportion of residuals to be considered as suspicious.

In a first Section, the main features of the approach introduced by Nielsen et al. (2000) are described. This approach is then generalized in Section 3 to a broader class of loss functions, including the bounded-influence ones, by formulating the related M-type esti-

mator. Also, it is explained how to locally robustify this M-type estimator for the specific case of local polynomial regression. The nonparametric method for an adaptive scaling of the threshold points of the local M-type estimator follows in Section 4, after relaxing the symmetry constraint on the definition of the Huber loss function. In Section 6, simluations on semi-artificial datasets allow us to highlight and evaluate the properties of the proposed adaptive local M-type estimator $\hat{\boldsymbol{\theta}}^{\dagger}$. The nonlinear process considered is wind power production. It is nonstationary owing to the very nature of wind (and due to the changes in the site configuration and environment). Moreover, the conversion of wind to power makes wind power production a nonlinear and bounded process. A survey on the modeling and forecasting of wind power production is given by Giebel et al. (2003). These datasets are composed by hourly wind speed measures and simulated power output for a multi-MW wind farm in Denmark, over a period covering 10.000 hours. For validation purposes, the proposed methods are also applied on a second dataset (Section 7), composed by meteorological forecasts and related power measurements for another multi-MW wind farm in Denmark, with the same aim of modeling the conversion of wind to power. Concluding remarks are gathered in Section 8, as well as perspectives for further developments.

2 Adaptive local estimation of time-varying coefficient functions

The method introduced by Nielsen et al. (2000) consists in a combination of local polynomial regression and recursive weighted least-squares with exponential forgetting, for adaptively estimating the $\boldsymbol{\theta}$-functions in Equation (2). They are estimated by locally fitting linear models at a number of distinct points $\mathbf{u}_{(j)} = [u^1_{(j)} \dots u^k_{(j)} \dots u^l_{(j)}]^\top$, $j = 1, \dots, J$, referred to as fitting points, where the variables $u^k_{(j)}$ are those that condition the regression model. $[.]^\top$ denotes the transposition operator. It is first described how local polynomial approximation and weighted least-squares are used for a conditional estimation of the $\boldsymbol{\theta}$-functions. The recursive formulation for an adaptive estimation of these functions follows.

2.1 Local polynomial estimates

Let us focus on a single fitting point $\mathbf{u}_{(j)}$ only. The local polynomial approximation $\mathbf{z}_i$ of the vector of explanatory variables $\mathbf{x}_i$ at $\mathbf{u}_i$ is given by:

$$\mathbf{z}_i^\top = [x_i^1 \mathbf{p}_d^\top(\mathbf{u}_i) \dots x_i^k \mathbf{p}_d^\top(\mathbf{u}_i) \dots x_i^l \mathbf{p}_d^\top(\mathbf{u}_i)] \tag{3}$$

where $\mathbf{p}_d(\mathbf{u}_i)$ corresponds to the d-order polynomial evaluated at $\mathbf{u}_i$. In parallel, write

$$\boldsymbol{\phi}_{(j)} = \boldsymbol{\phi}(\mathbf{u}_{(j)}) = [\boldsymbol{\phi}^\top_{(j),1} \dots \boldsymbol{\phi}^\top_{(j),k} \dots \boldsymbol{\phi}^\top_{(j),l}]^\top \tag{4}$$

the vector of local coefficients at $\mathbf{u}_{(j)}$, where the element vector $\boldsymbol{\phi}_{(j),k}$ is the vector of local coefficients related to the local polynomial approximation of the k-th explanatory variable, that is, $x_i^k \mathbf{p}_d(\mathbf{u}_i)$. Using local polynomial approximations translates to assuming that the coefficient functions are sufficiently smooth functions. They remain unknown though. Note that it has already been argued in Fan et al. (1994) that having $d = 1$ instead of $d = 0$ (i.e.

having an estimator based on local linear fit instead of local constant fit) was already a robustification of local polynomial estimators. This can be straightforwardly extended to the case of $d > 1$.

The linear model

$$y_i \;=\; \mathbf{z}_i^\top \boldsymbol{\phi}_{(j)}, \qquad i = 1, \ldots, N \tag{5}$$

is fitted using weighted least-squares

$$\hat{\boldsymbol{\phi}}_{(j)} \;=\; \underset{\boldsymbol{\phi}_{(j)}}{\arg\min} \sum_{i=1}^{N} w_{i,(j)} \rho(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}_{(j)}) \tag{6}$$

where ρ is a quadratic loss function, i.e. such that $\rho(\epsilon) = \epsilon^2/2$, and the weights $w_{i,(j)}$ are assigned by a Kernel function of the following form:

$$w_{i,(j)} \;=\; K(\mathbf{u}_i, \mathbf{u}_{(j)}) \;=\; \prod_k T\left(\frac{|u_i^k - u_{(j)}^k|_k}{h_{(j)}^k}\right) \tag{7}$$

In the above, $|.|_k$ denotes a chosen distance on the k-th dimension of $\mathbf{u}$, and $\mathbf{h}_{(j)}$ is the bandwidth for that particular fitting point $\mathbf{u}_{(j)}$. It appears reasonable to have different dependence of the bandwidth h on (j) for each dimension k of $\mathbf{u}$. $\mathbf{h}_{(j)} = [h_{(j)}^1 \ldots h_{(j)}^k \ldots h_{(j)}^l]$ may be determined using a nearest-neighbour principle or with a rule derived from the expert knowledge on the density of the data as a function of (j). In parallel, T can be defined as a tricube function, i.e.

$$T \;:\; v \in \mathbb{R}^+ \;\rightarrow\; T(v) \in [0,1], \;\; T(v) \;=\; \begin{cases} (1 - v^3)^3, & v \in [0,1] \\ 0 \qquad\quad , & v > 1 \end{cases}, \tag{8}$$

as introduced and discussed by e.g. Cleveland and Devlin (1988).

The elements of $\boldsymbol{\theta}_{(j)}$ are finally estimated by:

$$\hat{\boldsymbol{\theta}}_{(j)} \;=\; \hat{\boldsymbol{\theta}}(\mathbf{u}_{(j)}) \;=\; \mathbf{p}_d^\top(\mathbf{u}_{(j)}) \hat{\boldsymbol{\phi}}_{(j)}, \;\; j = 1, \ldots, J \tag{9}$$

And, for a given $\mathbf{u}_i$, the corresponding coefficient functions $\hat{\boldsymbol{\theta}}(\mathbf{u}_i)$ are obtained by linear-type interpolation of the coefficient functions. For instance if $\dim(u) = 1$, they are obtained by linear interpolation of the coefficient function values at the two fitting points forming the smallest interval that covers $\mathbf{u}_i$.

2.2 Adaptive estimation from a recursive formulation

In order to obtain a recursive formulation for the estimation of the coefficient functions, let us introduce a modified version f_n of the objective function to be minimized at any time step n. For each fitting point $\mathbf{u}_{(j)}$, $j = 1, \ldots, J$, f_n writes

$$f_n(\mathbf{u}_{(j)}) \;=\; \sum_{i=1}^{n} \beta_{n,(j)}(i) w_{i,(j)} \rho(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}_{(j)}) \tag{10}$$

where $\beta_{n,(j)}$ is a function that permits exponential forgetting of past observations. More precisely, we have:

$$\beta_{n,(j)}(i) \;=\; \begin{cases} \lambda^{\text{eff}}_{n,(j)}\beta_{n-1,(j)}(i-1), & 1 \leq i \leq n-1 \\ 1 \qquad\qquad\qquad\qquad\quad , & i = n \end{cases} \tag{11}$$

In the above definition, $\lambda^{\text{eff}}_{n,(j)}$ is the effective forgetting factor for the fitting point $\mathbf{u}_{(j)}$, which is a function of the weight $w_{n,(j)}$, i.e.

$$\lambda^{\text{eff}}_{n,(j)} \;=\; 1-(1-\lambda)w_{n,(j)} \tag{12}$$

This effective forgetting factor ensures that old observations are downweighted only when new information is available. This will be further explained in a following part of the present Paragraph.

The local coefficients $\hat{\boldsymbol{\phi}}_{n,(j)}$ at time n for the model described by Equation (2) are then given by:

$$\hat{\boldsymbol{\phi}}_{n,(j)} \;=\; \underset{\boldsymbol{\phi}_{(j)}}{\arg\min}\, f_n(\mathbf{u}_{(j)}) \;=\; \underset{\boldsymbol{\phi}_{(j)}}{\arg\min} \sum_{i=1}^{n} \beta_{n,(j)}(i) w_{i,(j)} \rho(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}_{(j)}) \tag{13}$$

The recursive formulation for an adaptive estimation of the local coefficients $\hat{\boldsymbol{\phi}}_{n,(j)}$ (and therefore of $\hat{\boldsymbol{\theta}}_{n,(j)}$, by using Equation (9) at each time-step) leads to the following three-step updating procedure:

$$\epsilon_{n,(j)} \;=\; y_n - \mathbf{x}_n^\top \hat{\boldsymbol{\theta}}_{n-1,(j)} \tag{14}$$

$$\hat{\boldsymbol{\phi}}_{n,(j)} \;=\; \hat{\boldsymbol{\phi}}_{n-1,(j)} + \epsilon_{n,(j)} w_{n,(j)} \left(\mathbf{R}_{n,(j)}\right)^{-1} \mathbf{z}_n^\top \tag{15}$$

$$\mathbf{R}_{n,(j)} \;=\; \lambda^{\text{eff}}_{n,(j)} \mathbf{R}_{n-1,(j)} + w_{n,(j)} \mathbf{z}_n \mathbf{z}_n^\top \tag{16}$$

where $\lambda^{\text{eff}}_{n,(j)}$ is again the effective forgetting factor. One sees that when the weight $w_{n,(j)}$ equals 0 (thus meaning that the local estimates should not be affected by the new information), then we have $\hat{\boldsymbol{\phi}}_{n,(j)} = \hat{\boldsymbol{\phi}}_{n-1,(j)}$ and $\mathbf{R}_{n,(j)} = \mathbf{R}_{n-1,(j)}$. This confirms the role of the effective forgetting factor, that is to downweight old observations, but only when new information is available.

For initializing the recursive process, the matrices $\mathbf{R}_{0,(j)}$, $j = 1, \ldots, J$, can be chosen as

$$\mathbf{R}_{0,(j)} \;=\; \xi.\mathbf{I}_r, \;\; \forall j \tag{17}$$

where ξ is a small positive number and $\mathbf{I}_r$ is an identity matrix of size r. Note that r is equal to the order of the chosen model in Equation (2) times the order of the polynomials used for local approximation. In parallel, the coefficient functions are usually initialized with a vector of zeros, or alternatively from a best guess on the target regression.

3 Robustifying the estimation of coefficient functions

In real-world applications, the time-series of the response variable $\{y_i\}$, $i = 1, \ldots, N$, as well as those of the considered explanatory variables $\{\mathbf{x}_i\}$, $i = 1, \ldots, N$, and $\{\mathbf{u}_i\}$, $i = 1, \ldots, N$, may contain a non-negligible noise component. This noise may come from the measurements devices; or alternatively, it may be related to the prediction error in the forecast of explanatory variables used as input. Some of the values can even be outliers, i.e. data that can be deemed as abnormal in regard to the general observed behavior of the time-series.

The previously described method for tracking the coefficient functions lacks robustness if dealing with skewed and heavy-tailed residual distributions. It is known that in this case estimators based on a classical quadratic criterion are not optimal. Some methods have therefore appeared in the literature in order to robustify usual regressors. A condensed and nice description of the main features of robust statistics is given by Hampel (2001). These methods include among others variations of Least Median of Squares (LMS) (Rousseeuw, 1984; Rousseeuw and Leroy, 1987), or the so-called L_1 method (Wang and Scott, 1994) (that is, by replacing the quadratic loss function, equivalent to a L_2 norm, by the absolute value loss function). Though, most of these methods rely on the concepts of M-estimators (e.g. Huber (1981), Hampel et al. (1986), and a wealth of follow-up papers). Originally, M-estimators are derived from the principle of "generalized maximum likelihood", and were introduced for regression with residual distributions that slightly deviate from Gaussian. Even though, they have been found suitable (if appropriately scaled) for a large panel of contaminated or heavy tailed distributions (Kelly, 1992). In parallel, they have also been considered for nonparametric function fitting, and referred to as M-type estimators (see Fan et al. (1994); Fan and Jiang (1999); Welsh (1994) among others). Note that few robustification approaches consider a potential noise in both explanatory and response variables (e.g. the bivariate-LMS (del Río et al., 2001)). In most of the cases, it is assumed that the explanatory variables are error-free.

In the following, the above method for an adaptive estimation of local coefficients is robustified, by proposing the M-type estimator $\hat{\phi}^*_{n,(j)}$ corresponding to $\hat{\phi}_{n,(j)}$. The particular case of the Huber loss function is considered. It is finally explained why it is not the residuals but the weighted residuals that should influence the estimator, resulting in a locally robustified M-type estimator $\hat{\phi}^{**}_{n,(j)}$.

3.1 Generalization of the method to bounded-influence and convex loss functions

The M-type estimator $\hat{\phi}^*_{n,(j)}$, for a recursive estimation of the local coefficients in conditional parametric models such as that given by Equation (2), corresponds to the estimate that minimizes an objective function that is pretty similar to that of Equation (10). For a given fitting point $\mathbf{u}_{(j)}$, this objective function writes

$$\hat{\phi}^*_{n,(j)} = \arg\min_{\phi_{(j)}} \sum_{i=1}^{n} \beta^*_{n,(j)}(i) w_{i,(j)} \rho_m(y_i - \mathbf{z}_i^\top \phi_{(j)}) \tag{18}$$

except that here, if denoting by Ψ_m the derivative of ρ_m, the main peculiarity of the Ψ_m-function is that its output is bounded:

$$\Psi_m \ : \ u \in \mathbb{R} \ \rightarrow \ \Psi_m(u) \in [M_{\text{inf}}, M_{\text{sup}}] \tag{19}$$

Also, it is considered that ρ_m is convex and consequently, if denoting by Ψ'_m the derivative of Ψ_m, we have

$$\Psi'_m \ : \ u \in \mathbb{R} \ \rightarrow \ \Psi'_m(u) \in [0, M'_{\text{sup}}] \tag{20}$$

for almost all u, since Ψ'_m may not be defined for some points if ρ_m is a piecewise function. Note that in general, the distribution of residuals F is assumed to be symmetric, and therefore ρ_m is defined as a symmetric function, (translating to $M_{\text{inf}} = -M_{\text{sup}}$). In a following Section, that constraint on the symmetry of F will be relaxed. Hereafter, even if Ψ_m is written as a function of some other variables than ϵ, Ψ'_m will denote the derivative of Ψ_m with respect to ϵ.

Moreover, in the definition of the M-type estimator given above, the function $\beta^*_{n,(j)}$ for an exponential forgetting of old observations is a robustified version of $\beta_{n,(j)}$. Indeed, in order to be consistent with the definition of the effective forgetting factor introduced in Equation (12), $\lambda^{\text{eff},*}_{n,(j)}$ has to be given by

$$\lambda^{\text{eff},*}_{n,(j)} \ = \ 1 - \frac{1}{M'_{\text{sup}}}(1-\lambda)\Psi'_m(\epsilon_{n,(j)})w_{n,(j)} \tag{21}$$

In the robust version of the estimation method, the effective forgetting factor insures that old observations are not downweighted as long as non-suspicious new information is not available. In turn, Equation (20) insures that $\lambda^{\text{eff},*}_{n,(j)} \in [0,1]$, and thus that the definition of $\lambda^{\text{eff},*}_{n,(j)}$ is consistent with that of a forgetting factor. The function $\beta^*_{n,(j)}$ is obtained by using this robust version of the effective forgetting factor $\lambda^{\text{eff},*}_{n,(j)}$ in the definition of Equation (11).

Similarly to the calculations done for obtaining the recursive formulation given by Equations (15) and (16), that is by using a Newton-Raphson step, one can straightforwardly obtain a recursive formulation for the estimation of $\hat{\boldsymbol{\phi}}^*_{n,(j)}$, which are updated with

$$\hat{\boldsymbol{\phi}}^*_{n,(j)} \ = \ \hat{\boldsymbol{\phi}}^*_{n-1,(j)} + \Psi_m(\epsilon_{n,(j)})w_{n,(j)}\left(\mathbf{R}^*_{n,(j)}\right)^{-1}\mathbf{z}_n^\top \tag{22}$$

while the updating formula for the $\mathbf{R}^*_{n,(j)}$-matrices writes

$$\mathbf{R}^*_{n,(j)} \ = \ \lambda^{\text{eff},*}_{n,(j)}\mathbf{R}^*_{n-1,(j)} + \Psi'_m(\epsilon_{n,(j)})w_{n,(j)}\mathbf{z}_n\mathbf{z}_n^\top \tag{23}$$

such that the local residual $\epsilon_{n,(j)}$ at time n is still calculated with Equation (14).

This recursive formulation of the optimization problem given by Equation (18) actually consists in a generalization of the recursive formulation described in Paragraph 2.2 for a broader class of loss functions. A theoretical study of the asymptotic properties (including asymptotic Normality, strong and weak consistency) of the class of M-type estimators such as $\hat{\boldsymbol{\phi}}^*_{n,(j)}$ has been carried out by Fan and Jiang (1999) in the i.i.d. case, by Cai and Ould-Saïd (2003) in the context of stationary time-series, and by Beran et al.

(2002) for the specific case of long-memory error processes.

3.2 A recursive M-type estimator based on the Huber loss function

An example of a convex and bounded influence ρ_m-function is that of the Huber loss function. It combines a quadratic and a linear criterion:

$$\rho_m(\epsilon, c) \;=\; \begin{cases} \frac{\epsilon^2}{2}, & |\epsilon| \leq c \\ c|\epsilon| - \frac{c^2}{2}, & |\epsilon| > c \end{cases} \tag{24}$$

with the c-parameter, usually referred to as the threshold point, which controls the transition from quadratic to linear. Consequently, the related Ψ_m-function is an odd function given by

$$\Psi_m(\epsilon, c) \;=\; \rho'_m(\epsilon) \;=\; \begin{cases} \epsilon, & |\epsilon| \leq c \\ c\ \text{sign}(\epsilon), & |\epsilon| > c \end{cases} \tag{25}$$

and its derivative Ψ'_m is

$$\Psi'_m(\epsilon, c) \;=\; \rho''_m(\epsilon) \;=\; \begin{cases} 1, & |\epsilon| \leq c \\ 0, & |\epsilon| > c \end{cases} \tag{26}$$

The Huber loss function is symmetric and such that $M_{\text{sup}} = -M_{\text{inf}} = c$. In addition, the upper bound on the derivative of the Ψ_m equals 1.

One sees that if using the Huber loss function in Equation (18) then the objective function to be minimized is equivalent to using a classical least-square criterion for residuals whose absolute value is smaller than that of the threshold point. In such case, the updating formula for $\mathbf{R}^*_{n,(j)}$ and $\hat{\boldsymbol{\phi}}^*_{n,(j)}$ (cf. Equations (23) and (22)) are equivalent to those given by Equations (16) and (15), respectively. However, that loss function goes from quadratic to linear for larger residual values, and Equation (23) becomes

$$\mathbf{R}^*_{n,(j)} \;=\; \mathbf{R}^*_{n-1,(j)} \tag{27}$$

which means that the newly available information about the model performance is not used for updating $\mathbf{R}^*_{n-1,(j)}$. Similarly, the updating formula for the local coefficients then writes

$$\hat{\boldsymbol{\phi}}^*_{n,(j)} \;=\; \hat{\boldsymbol{\phi}}^*_{n-1,(j)} + c\ \text{sign}(\epsilon_{n,(j)}) w_{n,(j)} \left(\mathbf{R}^*_{n,(j)})\right)^{-1} \mathbf{z}_n^\top \tag{28}$$

which translates to considering an upper bound on possible model errors, and, when this upper bound is reached, the magnitude of the error is no more considered for model adaptation.

By using a ρ_m-function like the Huber one, the optimization problem formulated by Equation (18) admits a unique minimum. This would not be the case if considering the so-called redescending Ψ_m-functions, such as the Tuckey or Welsh ones (see discussion by Antoch and Ekblom (1995)). Indeed, the initialization of the recursive procedure would turn into a crucial point. This is the reason why we only consider the use of convex loss

functions here. Though, note that even if we focus on Huber-type loss functions, the proposed methodology could be easily extended to other convex loss functions.

Our choice for the Huber loss function is motivated by the fact that we aim at producing model outputs that would minimize a Mean Square Error (MSE) criterion. It is known that the loss function used for estimating the parameters of a model should be the same than that used for the evaluation of the model outputs on an independent test set (Granger, 1993; Weiss, 1996). The Huber loss function is quadratic in the range of residual values that are not considered as suspicious and its use is thus consistent with the aim of estimating the minimum-MSE regressor.

3.3 Local robustification of the M-type estimator

M-estimators have originally been introduced for linear models. When dealing with conditional parametric models, one actually works with several linear models that are locally fitted at a certain number of fitting points. And, at given time n, the estimates of the local coefficients at any fitting point $\mathbf{u}_{(j)}$ such that $w_{n,(j)} > 0$ are updated. The adaptation of the local coefficients is weighted by the value of the Kernel function $w_{n,(j)} = K(\mathbf{u}_n, \mathbf{u}_{(j)})$ (cf. Equation (7)). It would seem reasonable to envisage the definition of M-type estimators whose loss function would depend on $w_{n,(j)}$. Let us refer to that proposal as the local robustification of the M-type estimator, and denote by $\tilde{c}(w)$ the weight-dependent threshold point. The resulting estimator $\hat{\phi}_n^{**}$ is called here a local M-type estimator. Note that such proposal differs from that of Chan and Zhang (2004), who described an adaptive bandwidth method for the robustification of M-type estimators in nonparametric function fitting. In this method, local bandwidths are determined by using the intersection of confidence intervals rule.

In the case for which $\mathbf{u}_n = \mathbf{u}_{(j)}$, the related weight $w_{n,(j)}$ equals one, and this corresponds to the usual case for which the threshold point would be the user-defined one, i.e. $\tilde{c}(1) = c$. Then, the weight $w_{n,(j)}$ decreases as the distance (relatively to the chosen bandwidth $\mathbf{h}_{(j)}$) between $\mathbf{u}_{(j)}$ and $\mathbf{u}_n$ gets larger. Therefore, the influence of residuals calculated for $\mathbf{u}_n$ values being pretty far from $\mathbf{u}_{(j)}$ is already downweighted. It would hence seem reasonable not to downweight them a second time with the loss function being in its linear part. Our proposal is hence that the threshold point moves towards infinity as the weight goes to zero:

$$\tilde{c} \ : \ w \in [0,1] \ \rightarrow \ \tilde{c}(w) \in \mathbb{R}^+ \tag{29}$$

such that $\tilde{c}$ is a monotocally increasing function, with

$$\tilde{c}(1) = c, \quad \text{and} \quad \tilde{c}(w) \ \rightarrow \infty \ \text{ when } \ w \ \rightarrow \ 0 \tag{30}$$

One notices that if defining the $\tilde{c}$-function as $\tilde{c}(w) = cw^{-1/2}$, we then have

$$w_{n,(j)}\rho_m(\epsilon_{n,(j)}, \tilde{c}(w_{n,(j)})) \ = \ \begin{cases} \left(\epsilon_{n,(j)}\sqrt{w_{n,(j)}}\right)^2, & |\epsilon_{n,(j)}\sqrt{w_{n,(j)}}| \leq c \\ \epsilon_{n,(j)}\sqrt{w_{n,(j)}}, & |\epsilon_{n,(j)}\sqrt{w_{n,(j)}}| > c \end{cases} \tag{31}$$

which is indeed equivalent to applying the usual Huber loss function to the weighted residual $\epsilon_{n,(j)}\sqrt{w_{n,(j)}}$:

$$w_{n,(j)}\rho_m(\epsilon_{n,(j)},\tilde{c}(w_{n,(j)})) \;=\; \rho_m\left(\epsilon_{n,(j)}\sqrt{w_{n,(j)}},c\right) \tag{32}$$

Note that even if in our proposal for the definition of the $\tilde{c}$-function $\tilde{c}$ is not defined for $w=0$, this is not an issue when injected in the definition of the loss function ρ_m.

Finally, the local M-type estimator $\hat{\boldsymbol{\phi}}^{**}_{n,(j)}$ is obtained by including the above proposal in the definition of the M-type estimator $\hat{\boldsymbol{\phi}}^{*}_{n,(j)}$. $\hat{\boldsymbol{\phi}}^{**}_{n,(j)}$ is given by the vector of local coefficients that minimizes at time n the following objective function

$$\hat{\boldsymbol{\phi}}^{**}_{n,(j)} \;=\; \underset{\boldsymbol{\phi}_{(j)}}{\arg\min}\sum_{i=1}^{n}\beta^{**}_{n,(j)}(i)\rho_m\left(\left(y_i-\mathbf{z}_i^{\top}\boldsymbol{\phi}_{(j)}\right)\sqrt{w_{i,(j)}},c\right) \tag{33}$$

where ρ_m is the Huber loss function. And, regarding the recursive formulation for that M-type estimator, the updating Equation (22) for the local coefficients $\hat{\boldsymbol{\phi}}^{**}_{n,(j)}$ becomes

$$\hat{\boldsymbol{\phi}}^{**}_{n,(j)} \;=\; \hat{\boldsymbol{\phi}}^{**}_{n-1,(j)} + \Psi_m\left(\epsilon_{n,(j)}\sqrt{w_{n,(j)}},c\right)\left(\mathbf{R}^{**}_{n,(j)}\right)^{-1}\mathbf{z}_n^{\top} \tag{34}$$

while that for the covariance matrix $\mathbf{R}^{*}_{n,(j)}$ (cf. Equation (23)) is modified as

$$\mathbf{R}^{**}_{n,(j)} \;=\; \lambda^{\text{eff},**}_{n,(j)}\mathbf{R}^{**}_{n-1,(j)} + \Psi'_m\left(\epsilon_{n,(j)}\sqrt{w_{n,(j)}},c\right)\mathbf{z}_n\mathbf{z}_n^{\top} \tag{35}$$

The function $\beta^{**}_{n,(j)}$ for an exponential forgetting of past observations is also a modified version of $\beta^{*}_{n,(j)}$ that takes into account the local robustification. Indeed, it is now based on the effective forgetting factor $\lambda^{\text{eff},**}_{n,(j)}$, defined as:

$$\lambda^{\text{eff},**}_{n,(j)} \;=\; 1-(1-\lambda)\breve{\Psi}'_m\left(\epsilon_{n,(j)}\sqrt{w_{n,(j)}},c\right) \tag{36}$$

4 Adaptive scaling of the M-type estimator

Scaling the M-type estimator consists in choosing a suitable value for the threshold point, i.e. that would permit to minimize an error criterion such as the Mean Square Error (MSE) for instance. An unappropriate choice for c might lead to a higher MSE than that of the non-robust estimates. For a discussion on the effects of this scaling on a robust estimator's performance, we refer to Kelly (1992).

In the literature, the choice of a suitable threshold point is often either left to the reader, given by a rule of thumb, or the result of some sensitivity analysis on the performance of the M-type estimator depending on c. For instance, when introducing a robust Huber adaptive filter, Petrus (1999) noticed that the minimum MSE was attained for threshold values close to the Mean Absolute Deviation (MAD) of the input, and proposed this choice as a first rule of thumb.

In most of the cases, the scaling of the M-type estimator is not adaptive. A few examples for a time-varying scaling of M-type estimators are the use of an annealing scheme (Li, 1996; Li et al., 1998) (which is thus time-varying, but not adaptive), a scaling based on a robust recursive estimator of variance (Zou et al., 2000a,b) (which cannot be suitable if avoiding an assumption on the distribution of residuals), and the use of past collected residuals for estimating a range of potential error values of the current model (Chen and Jain, 1994). This last possibility makes the scaling adaptive, but the range of potential errors is estimated from the residuals of past models: it is unlikely that this collection of residuals would represent the distribution of potential residuals for the current model. In the following, a simple method is proposed for the scaling of the M-type estimator, which is based on an empirical (and hence nonparametric) estimation of the residual distribution. An original feature of the resulting adaptive M-type estimator is that instead of defining the threshold points, one defines a proportion α of residuals that may be considered as suspicious.

For building the adaptive M-type estimator, it is necessary to consider that the process $\{\epsilon_i\}$, $i = 1, \ldots, N$, is nonstationary. For the example of wind power production, this assumption is reasonable, since it is known that the residual distribution is influenced by the season, changing in the surroundings of a considered site, etc. Therefore, the distribution of residuals is now considered as conditional to n. Denote by F_n the distribution function of ϵ_n, and by G_n the related cumulative distribution function. In a first stage, the constraint on the symmetry of the bounded-influence loss function is relaxed. Then, the non-parametric approach to an adaptive scaling of the M-type estimator by having time-varying threshold points is described.

4.1 Relaxing the symmetry constraint on the loss function

The asymmetric Huber loss function $\breve{\rho}_m$ that is introduced below consists in a generalization of the classical Huber loss function. The M-estimator introduced by Huber (1981) is originally designed for estimating a better regressor when the distribution F of the residuals slightly deviates from Normal. Our motivation for introducing the asymmetric Huber loss function is that F_n may also deviate from being symmetric. This is indeed the case when considering nonlinear and bounded processes such as wind generation. A thorough study of the prediction errors in wind power prediction is available in (Pinson, 2006). Denote by $\mathbf{c} = [c^-, c^+]^\top$ the vector of inferior and superior threshold points. $\breve{\rho}_m(\epsilon, \mathbf{c})$ is then defined as:

$$\breve{\rho}_m(\epsilon, \mathbf{c}) = \begin{cases} c^-\epsilon - \frac{{c^-}^2}{2}, & \epsilon < c^- \\ \frac{\epsilon^2}{2}, & \epsilon \in [c^-, c^+] \\ c^+\epsilon - \frac{{c^+}^2}{2}, & \epsilon > c^+ \end{cases} \tag{37}$$

For $\breve{\rho}_m$ to be a suitable loss function, i.e. such that $\breve{\rho}_m(\epsilon, \mathbf{c}) > 0, \ \forall \epsilon$, a necessary condition on $\mathbf{c}$ is that $c^- < 0$ and $c^+ > 0$. Lindström et al. (1996) introduced a similar generalization of the Huber loss function, and showed the asymptotic Normality of the related Kernel M-type estimator. This can be extended to the case of the M-type estimators $\hat{\phi}^*_{n,(j)}$ and $\hat{\phi}^{**}_{n,(j)}$ that would use the loss function $\breve{\rho}_m$ defined above.

An illustration of the asymmetric Huber loss function $\breve{\rho}_m$ and of the related $\breve{\Psi}_m$-function is given in Figure 1. The interest of introducing an M-type estimator based on an asymmetric loss function is to better deal with skewed and heavy-tailed distributions as possible deviations from Normality. Residuals that are considered as suspicious are not downweighted in the same way if they are negative or positive outliers.

Writing the recursive formulation of the asymmetric M-type estimator would lead to updating formulas that would be simply given by using $\breve{\Psi}_m$ and $\breve{\Psi}'_m$ instead of Ψ_m and Ψ'_m in Equations (22) and (23) respectively. The effective forgetting factor for the asymmetric case is straighforwardly obtained by rewriting Equation (21) with $\breve{\Psi}'_m$ instead of Ψ'_m. The asymmetric $\breve{\Psi}_m$-function and its derivative write:

$$\breve{\Psi}_m(\epsilon, \mathbf{c}) \;=\; \begin{cases} c^-, & \epsilon < c^- \\ \epsilon \;\;, & \epsilon \in [c^-, c^+] \\ c^+, & \epsilon > c^+ \end{cases} \tag{38}$$

and

$$\breve{\Psi}'_m(\epsilon, \mathbf{c}) \;=\; \begin{cases} 1, & \epsilon \in [c^-, c^+] \\ 0, & \text{otherwise} \end{cases} \tag{39}$$

4.2 Time-varying threshold points

Define α the user-defined parameter that corresponds to the proportion of residuals to be considered as suspicious. Then, denote by $\mathbf{c}_n(\alpha) = [c_n^-(\alpha), c_n^+(\alpha)]^\top$ the vector of threshold points at time n, which is a function of the proportion parameter α. Finally, $\hat{\boldsymbol{\theta}}_n^\dagger$ is the M-type estimator of the coefficient functions based on the asymmetric loss function introduced in the above Paragraph.

At a given time n are available the vectors of explanatory variables $\mathbf{x}_n$ and $\mathbf{u}_n$, the response variable value y_n, and a model output value $\hat{y}_{n|n-1} = \mathbf{x}_n^\top \hat{\boldsymbol{\theta}}_{n-1}^\dagger(\mathbf{u}_n)$. The residual ϵ_n at that time is calculated as

$$\epsilon_n \;=\; y_n - \hat{y}_{n|n-1} \;=\; y_n - \mathbf{x}_n^\top \hat{\boldsymbol{\theta}}_{n-1}^\dagger \tag{40}$$

Then, instead of collecting the past residuals as proposed by Chen and Jain (1994), an empirical estimate of the distribution of potential residuals for the current estimates $\hat{\boldsymbol{\theta}}_{n-1}^\dagger$ is obtained by applying this model to the past m vectors of explanatory variables. The simulated residual $\tilde{\epsilon}_{n-i}^{(n-1)}$, by using $\hat{\boldsymbol{\theta}}_{n-1}^\dagger$ for predicting y_{n-i} at time $n-i-1$ is given by

$$\tilde{\epsilon}_{n-i}^{(n-1)} \;=\; y_{n-i} - \mathbf{x}_{n-i}^\top \hat{\boldsymbol{\theta}}_{n-1}^\dagger, \qquad i = 1, \ldots, m \tag{41}$$

The estimate $\hat{F}_n$ of the empirical distribution of the residuals for $\hat{\boldsymbol{\theta}}_{n-1}^\dagger$ then puts a probability $1/m$ on each of the simulated residuals:

$$\hat{F}_n(\epsilon) \;\rightarrow\; \{\tilde{\epsilon}_{n-i}^{(n-1)}, i = 1, \ldots, m \mid \mathrm{P}\left(\epsilon = \epsilon_{n-i}^{(n-1)}\right) = 1/m\} \tag{42}$$

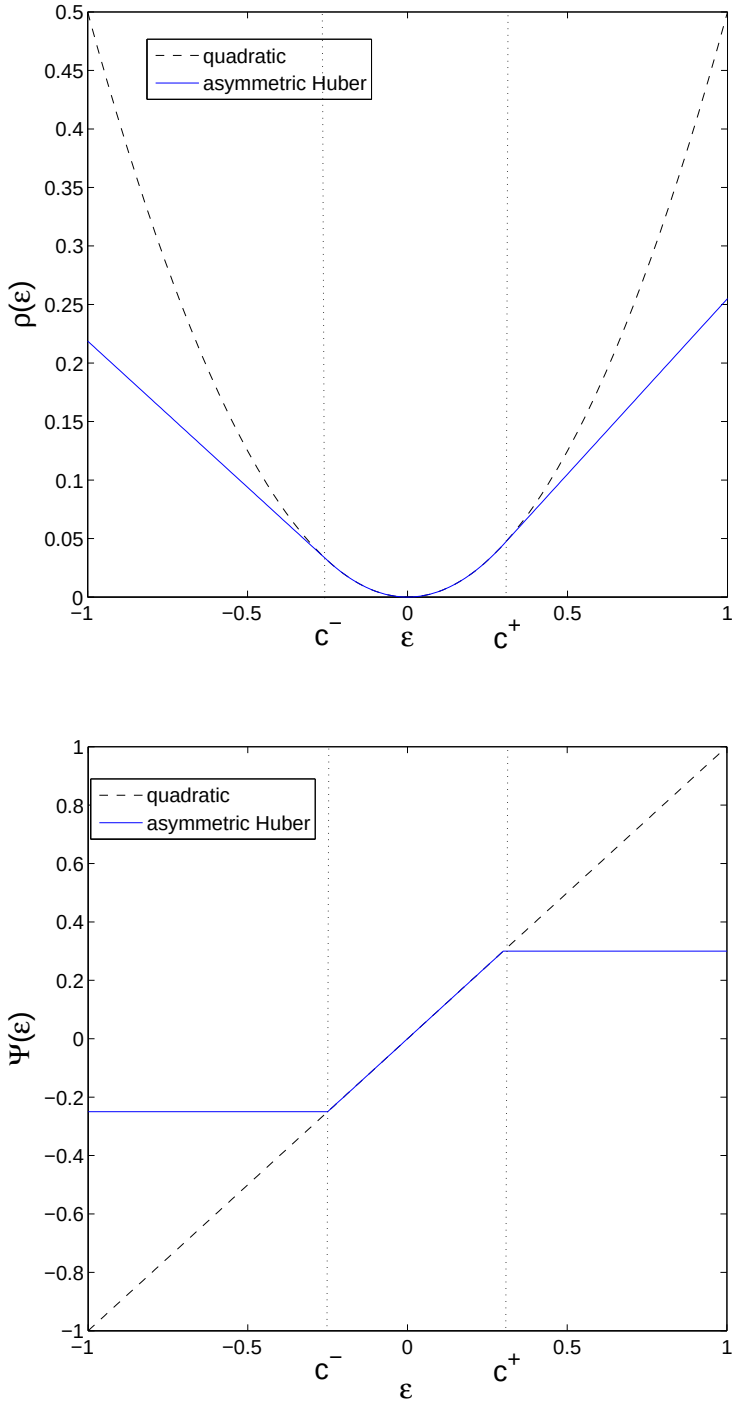


FIGURE 1: *The 'usual' quadratic and asymmetric Huber loss functions (top), as well as their derivatives (bottom). The thresholds points c^- and c^+ locate the negative and positive transitions from quadratic to linear criteria. Here these points are such that c^- = -0.25 and c^+ = 0.3. Negative residuals larger than c^- (in absolute value) and positive residual larger than c^+ are then downweighted when updating the model estimates.*

Given the proportion α of residuals that may be considered as suspicious, one obtains the two thresholds points $c_n^-(\alpha)$ and $c_n^+(\alpha)$ by picking the quantiles with proportion $(\alpha/2)$ and $(1-\alpha/2)$ of the distribution $\hat{F}_n$:

$$\mathbf{c}_n(\alpha) \;=\; [\hat{G}_n^{-1}(\alpha/2)\;\; \hat{G}_n^{-1}(1-\alpha/2)]^\top \tag{43}$$

By doing so, the threshold points are not symmetric. Though, the related M-type estimator may be considered as symmetric since there will be asymptotically the same proportion of positive and negative residuals downweighted.

The loss function, the related $\breve{\Psi}_m$-function, as well as its derivative at time n, are finally given by $\breve{\rho}_m(\epsilon, \mathbf{c}_n(\alpha))$, $\breve{\Psi}_m(\epsilon, \mathbf{c}_n(\alpha))$ and $\breve{\Psi}'_m(\epsilon, \mathbf{c}_n(\alpha))$ respectively, for which the two additional user-defined parameters are α and m.

5 The adaptive local M-type estimator

This Section summarizes the above developments by defining the adaptive local M-type estimator, and by giving the necessary steps at time n for a robust estimation of the time-varying coefficient functions in the conditional parametric model formulated by Equation (2).

5.1 Definition

Formally, the adaptive local M-type estimator $\hat{\phi}^\dagger_{n,(j)}$ of the local coefficients corresponds to the estimates that minimize at time n the following objective function:

$$\hat{\boldsymbol{\phi}}^\dagger_{n,(j)} \;=\; \underset{\boldsymbol{\phi}_{(j)}}{\arg\min} \sum_{i=1}^{n} \beta^\dagger_{n,(j)}(i) \breve{\rho}_m \left(\left(y_i - \mathbf{z}_i^\top \boldsymbol{\phi}_{(j)} \right) \sqrt{w_{i,(j)}}, \mathbf{c}_n(\alpha) \right) \tag{44}$$

with the loss function $\breve{\rho}_m(\epsilon, \mathbf{c})$ defined by Equation (37) and $\mathbf{c}_n(\alpha)$ obtained with Equation (43). The related M-type estimator for the coefficient functions, denoted by $\hat{\boldsymbol{\theta}}^\dagger_{n,(j)}$, is readily given by applying Equation (9) to $\hat{\phi}^\dagger_{n,(j)}$.

The $\breve{\Psi}_m$-function and its derivative, which are are necessary for updating the estimates of the local coefficients, are already defined by Equations (38) and (39).

Finally, the function $\beta^\dagger_{n,(j)}$ in Equation (44), which permits an exponential forgetting of past observations that are not considered as suspicious, is such that

$$\beta^\dagger_{n,(j)}(i) \;=\; \begin{cases} \lambda^{\text{eff},\dagger}_{n,(j)} \beta^\dagger_{n-1,(j)}(i), & 1 \le i \le n-1 \\ 1 \qquad\qquad\qquad\quad , & i = n \end{cases} \tag{45}$$

with

$$\lambda^{\text{eff},\dagger}_{n,(j)} \;=\; 1 - (1-\lambda) \breve{\Psi}'_m \left(\epsilon_{n,(j)} \sqrt{w_{n,(j)}}, \mathbf{c}_n^{(\alpha)} \right) \tag{46}$$

5.2 Algorithm for an adaptive estimation

For initializing the estimation method, one may follow the proposal of Paragraph 2.2, that is to have the matrices $\mathbf{R}^{\dagger}_{0,(j)}$, $j = 1, \ldots, J$, being equal to an identity matrix times a small constant. And, regarding the initial estimates $\hat{\boldsymbol{\theta}}^{\dagger}_{0,(j)}$, one may choose them as a vector of zeros, or as a best guess on the target regression.

Prior to the application of the estimation method, one has to define a set of J fitting points $\mathbf{u}_{(j)}$, at which the coefficient functions are to be estimated. Each of these fitting points is associated to a bandwidth $\mathbf{h}_{(j)}$. Also, one has to choose the order d of the local polynomial approximation at the fitting points. Finally, the two additional parameters for robustifying the adaptive estimation are the proportion α of residuals to be considered as suspicious, and m the number of simulated residuals to be calculated for estimating the threshold points.

At time n, the necessary steps for updating the local polynomial estimates of the coefficient functions are:

step 1: *Adaptive scaling of the local M-type estimator*
Compute the m simulated residuals following Equation (41), and from the estimate $\hat{F}_n$ of the distribution of simulated residuals, determine the two threshold points $c_n^-(\alpha)$ and $c_n^+(\alpha)$ with Equation (43).

step 2: *Updating of the local estimates of the coefficient functions*
Loop over all fitting points $\mathbf{u}_{(j)}$, $j = 1, \ldots, J$, such that $w_{n,(j)} > 0$, and:

- Determine the local explanatory variables $\mathbf{z}_n$ corresponding to a local polynomial approximation of $\mathbf{x}_n$ at $\mathbf{u}_{(j)}$ (cf. Equation (3)),
- Compute the local residual $\epsilon_{n,(j)}$ corresponding to the use of the estimates at $\mathbf{u}_{(j)}$ for predicting y_n, as in Equation (14),
- Calculate the effective forgetting factor given by Equation (46),
- Update the matrix $\mathbf{R}^{\dagger}_{n-1,(j)}$ with

$$\mathbf{R}^{\dagger}_{n,(j)} \;=\; \lambda^{\text{eff},\dagger}_{n,(j)} \mathbf{R}^{\dagger}_{n-1,(j)} + \breve{\Psi}'_m \left(\epsilon_{n,(j)} \sqrt{w_{n,(j)}}, \mathbf{c}_n(\alpha) \right) \mathbf{z}_n \mathbf{z}_n^{\top} \tag{47}$$

- Update the vector of local coefficients with

$$\hat{\boldsymbol{\phi}}^{\dagger}_{n,(j)} \;=\; \hat{\boldsymbol{\phi}}^{\dagger}_{n-1,(j)} + \breve{\Psi}_m \left(\epsilon_{n,(j)} \sqrt{w_{n,(j)}}, \mathbf{c}_n(\alpha) \right) \left(\mathbf{R}^{\dagger}_{n,(j)} \right)^{-1} \mathbf{z}_n^{\top} \tag{48}$$

- Obtain the updated local polynomial estimates $\hat{\boldsymbol{\theta}}^{\dagger}_{n,(j)}$ of the coefficients functions at fitting point $\mathbf{u}_{(j)}$ with Equation (9):

$$\hat{\boldsymbol{\theta}}^{\dagger}_{n,(j)} \;=\; \mathbf{p}_d^{\top}(\mathbf{u}_{(j)}) \hat{\boldsymbol{\phi}}^{\dagger}_{n,(j)} \tag{49}$$

For a given $\mathbf{u}_i$, the corresponding coefficient functions $\hat{\boldsymbol{\theta}}^{\dagger}(\mathbf{u}_i)$ are obtained by linear-type interpolation of the coefficient functions.

6 Simulations

In this Section, simulation results on semi-artificial datasets are used for highlighting the properties of the introduced adaptive M-type estimator. The process that is considered is the power production from a 21MW wind farm, Klim in North Jutland. This process is nonstationary, nonlinear and bounded. The response variable is the available power output at the level of the wind farm, averaged on an hourly basis. For estimating that power production, wind speed measurements from a meteorological mast (also averaged on an hourly basis) are used as an explanatory variable. Both time-series cover a period of N hours ($N = 10000$). They are normalized so that they take values in the unit interval. At time step i, the wind speed and power values are be denoted by u_i and y_i respectively.

6.1 Data

Simulations are based on semi-artificial data. By semi-artificial is meant that the wind speed measurements are the real measurements from the meteorological mast at the wind farm, but that the related power values are obtained by transformation through a modeled power curve. It is assumed that the wind speed measurements are noise-free. At any time step i, the relation between wind speed u_i and the noise-free power output y_i is given by the nonlinear (and nonstationary) power curve $g_i(u)$, which is a function of wind speed only:

$$y_i \;=\; g_i(u_i), \quad i = 1, \ldots, N \tag{50}$$

In the following, it is explained how the nonstationary power curve is modeled. The noise that has been added for obtaining simulated but realistic dataset of wind speed and related power is consequently described.

6.1.1 Model for the power curve

A double exponential function is used here for modeling the power curve $g_i(u_i)$, defined as

$$g_i(u_i) \;=\; \exp\left(-\tau_i^1 \exp\left(-\tau_i^2 u_i\right)\right), \;\; i = 1, \ldots, N \tag{51}$$

so that the shape of that power curve is controled at any time i by the parameters $\boldsymbol{\tau}_i = [\tau_i^1 \;\; \tau_i^2]^{\top}$. These parameters are chosen to evolve linearly from $\boldsymbol{\tau}_0^{\top} = [10 \;\; 40]$ to $\boldsymbol{\tau}_N^{\top} = [11 \;\; 40]$. The resulting nonstationary power curve over the N time steps is depicted in Figure 6.1.1. Note that by considering that the conversion process is a function of wind speed only, we assume that other variables e.g. wind direction do not have any influence on that conversion process. This may not be true for real-world test cases. Though, the interest of the

semi-artificial data is that the noise-free power curve, which is the target regression, is available and can be used for evaluating the various estimators.

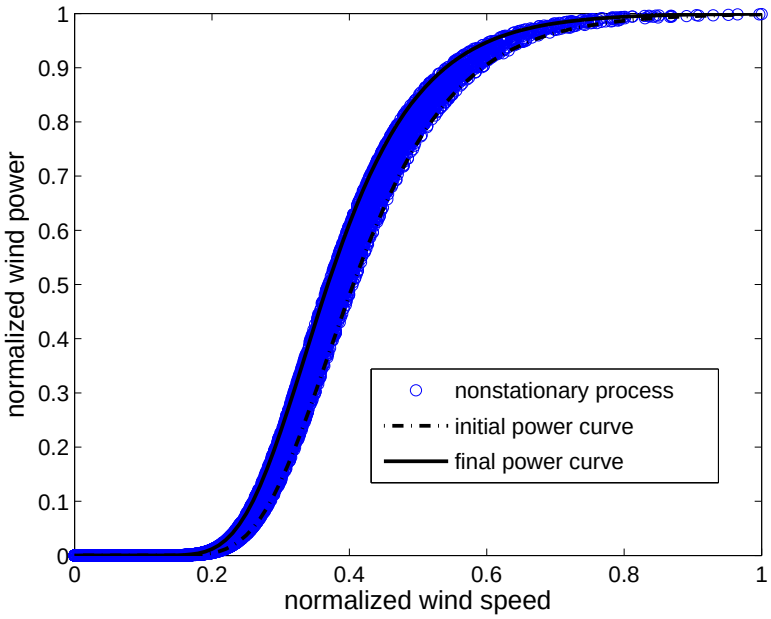


FIGURE 2: *The nonstationary power curve. The conversion process is modeled by a double exponential function, whose parameters τ linearly vary from* $[10\ 40]$ *to* $[11\ 40]$ *over the dataset.*

6.1.2 Noise on the power data

In order to obtain the simulated power output for the wind farm, two different types of noises to be added to the pure power data are envisaged. The noise sequences $\{\epsilon_i\}$ and $\{\xi_i\}$ are such that:

- $\{\epsilon_i\}$ is an additive Gaussian noise with zero mean, and whose standard deviation σ_i^ϵ is a function of the level of the reponse variable, i.e.

$$\epsilon_i \ \sim \ \mathcal{N}(0, \sigma_i^{\epsilon 2}), \quad \sigma_i^\epsilon \ = \ \nu_0^\epsilon + 4. * y_i^*(1 - y_i^*)\nu_1^\epsilon \tag{52}$$

  Such additive noise simulates a permanent noise in the measurement process, and we assume that the dispersion of this noise is directly influenced by the slope of the power curve. This is why an inverse U-shaped function is chosen.

- $\{\xi_i\}$ is an impulsive noise of the same form than $\{\epsilon_i\}$,

$$\xi_i \ \sim \ \mathcal{N}(0, \sigma_i^{\xi 2}), \quad \sigma_i^\xi \ = \ \nu_0^\xi + 4. * y_i^*(1 - y_i^*)\nu_1^\xi \tag{53}$$

  except that this noise is added at random locations characterized by a binary sequence $\{\mathcal{I}_i\}$. The proportion of data corrupted by this impulsive noise is given by π. Such noise simulates the presence of outliers in the measurement data. They may originate from electronic transmission problems for instance.

Finally, the simulated power data $\{\tilde{y}_i\}$ are obtained by adding these two noises to the noise-free power data $\{y_i\}$:

$$\tilde{y}_i \;=\; y_i + \epsilon_i + \xi_i \mathcal{I}_i, \quad i = 1, \ldots, N \tag{54}$$

Simulated power data larger than 1 or lower than 0 are forced to the bounds of the unit interval. The noise in the resulting dataset obviously deviates from being Gaussian.

The first dataset considered for simulation is composed by the wind speed data $\{u_i\}$ and the simulated power output $\{\tilde{y}_i\}$ for the wind farm, for which the noise parameters are $[\nu_0^\epsilon \; \nu_1^\epsilon] = [0.004 \; 0.9]$ for the additive noise, and $[\pi \; \nu_0^\xi \; \nu_1^\xi] = [0.2 \; 0.012 \; 0.2]$ for the impulsive noise. This dataset is depicted in Figure 3(a).

6.1.3 Noise on the wind speed data

In a second stage, we consider the possibility that a noise component may also be present in the wind speed data. The time-series of corrupted wind speed data is denoted by $\{\tilde{u}_i\}$. This time-series is obtained by adding an additive and an impulsive noise of the same forms than those used for corrupting the power data:

$$\tilde{u}_i \;=\; u_i + \epsilon_i + \xi_i \mathcal{I}_i, \quad i = 1, \ldots, N \tag{55}$$

Note that the inverse U-shaped function used for modeling the standard deviation of the noise as a a function of wind speed may not be realistic, though it has the benefit of increasing the difficulty of the estimation task.

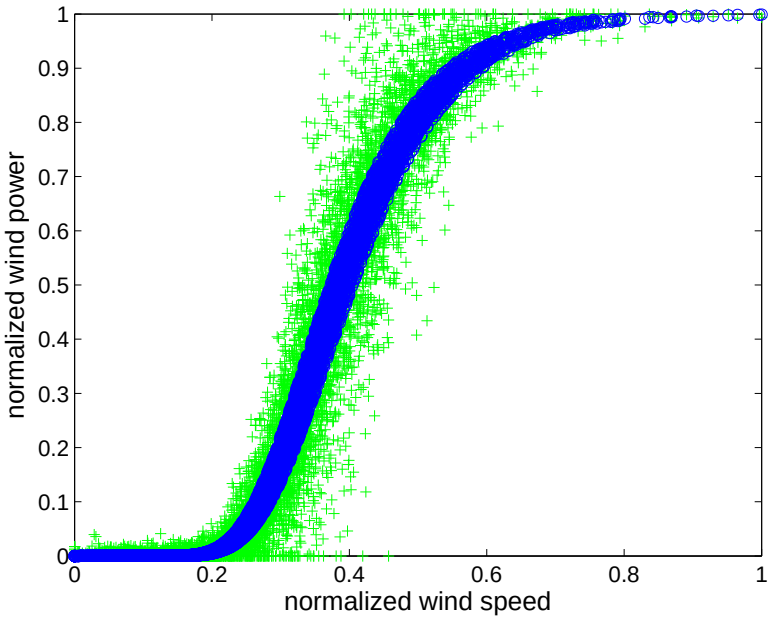
The second dataset considered for simulation is composed by the wind speed data $\{\tilde{u}_i\}$ and the simulated power output for the wind farm $\{\tilde{y}_i\}$. The parameters that define the noise on the power data are those that have been given in the above Paragraph. Concerning wind speed data, the noise parameters are chosen as $[\nu_0^\epsilon \; \nu_1^\epsilon] = [0.005 \; 0.04]$ for the additive noise, and $[\pi \nu_0^\xi \nu_1^\xi] = [0.2 \; 0.01 \; 0.015]$ for the impulsive noise. The resulting simulated process is shown in Figure 3(b).

6.2 Methodology for model selection and evaluation

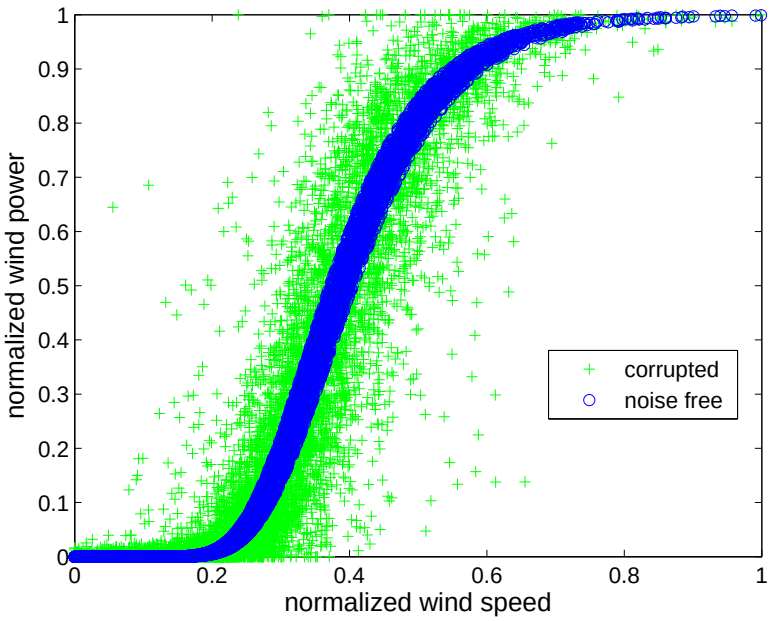
Our aim in the present work is to estimate the MSE-regressor for the semi-artificial data described in the above Paragraph. Since the process considered consists in the sole conversion from wind speed to power (the potential influence of other explanatory variables e.g. wind direction is neglected), the chosen model for both datasets is the minimal version of the conditional parametric model formulated in Equation (2), which then reduces to a conditional nonparametric model. This writes

$$y_i \;=\; \boldsymbol{\theta}(u_i) + \epsilon_i, \qquad i = 1, \ldots, N. \tag{56}$$

The order of the polynomial extension considered for local polynomial regression is chosen to be 2.



(a) Simulated dataset 1: noise is added to power data only.



(b) Simulated dataset 2: noise is added to both wind speed and power data.

FIGURE 3: *Noise-free and corrupted power curves. Wind speed measurements are from a meteorological mast at Klim in North Jutland. A nonstationary power curve is used for obtaining time-series of power output, yielding a 'pure' power curve. Data are then corrupted with (controlled) additive and impulsive noises. Both dataset include 10000 time-steps.*

For adaptively estimating the coefficient functions $\boldsymbol{\theta}(u)$, the performance of various estimators described in the above Sections are compared in the following. All these estimators are primarily based on the adaptive local estimator $\hat{\phi}$ for which a set of parameters, including the fitting points $u_{(j)}$, the bandwidth $h_{(j)}$ at each fitting point, and the forgetting factor λ, is to be selected. The fitting points are chosen to be uniformly spread on the unit interval:

$$u_{(j)} = \frac{j-1}{J-1}, \qquad j = 1, \ldots, J, \tag{57}$$

so that we only have to select J the number of these fitting points. Then, because we know that the density of the data is inversely proportional to the level of y, our proposal for the definition of $h_{(j)}$ is such that:

$$h_{(j)} = h_0 + h_1(j-1), \qquad j = 1, \ldots, J, \tag{58}$$

so that the constant h_0 and the scale factor h_1 have to be selected.

In practice, the four parameters J, h_0, h_1 and λ, are determined by using one-fold cross-validation: the first 2000 time-steps are considered as a training set and the following 2000 time-steps are used for cross-validation. The optimal set of parameters is chosen to be the one that minimizes a MSE criterion over the cross-validation set. This optimal set is obtained by trial and error. This optimal set of parameters is then used for defining the various M-type estimators. This actually yields four competing estimators, which are the local adaptive estimator $\hat{\boldsymbol{\theta}}$, the related M-type estimator $\hat{\boldsymbol{\theta}}^*$, the local M-type estimator $\hat{\boldsymbol{\theta}}^{**}$ and the adaptive local M-type estimator $\hat{\boldsymbol{\theta}}^{\dagger}$. Only $\hat{\boldsymbol{\theta}}$ is used over the training set. That vector of coefficient functions is then used as an initialization for all type of estimators, which are updated recursively.

Over the evaluation set, which thus consists in the last 6000 time-steps (since the cross validation set is not considered for the evaluation), the model outputs are evaluated with both a Normalized Mean Absolute Error (NMAE) and a Normalized Root Mean Square Error (NRMSE) criterion. Even if our aim is clearly to obtain a minimum-MSE estimator, the MAE criterion may better inform on the error reduction since it would give less weight to large errors related to suspicious data. The choice of error criteria for evaluating wind power prediction models has been further discussed by Madsen et al. (2005).

6.3 Results

6.3.1 Noise on power data only (dataset 1)

The optimal adaptive local estimator $\hat{\boldsymbol{\theta}}$ Using the cross-validation procedure, the optimal set of parameters for the adaptive local estimator $\hat{\boldsymbol{\theta}}$ is found to be:

$$[J \; h_0 \; h_1 \; \lambda] = [20 \; 0.03 \; 2.3 \; 0.991]$$

The performance of $\hat{\boldsymbol{\theta}}$ on the evaluation set, when defined by this set of parameters, is summarized by the value of the NMAE and NRMSE criteria:

$$\begin{array}{rclrcl} \text{NMAE}_r\ (\%) & = & 4.3887, & \text{NMAE}_t\ (\%) & = & 0.6384, \\ \text{NRMSE}_r\ (\%) & = & 7.6817, & \text{NRMSE}_t\ (\%) & = & 0.8233, \end{array}$$

with X_r and X_t corresponding to the values of the error criteria when model outputs are evaluated against the corrupted and true power data, respectively.

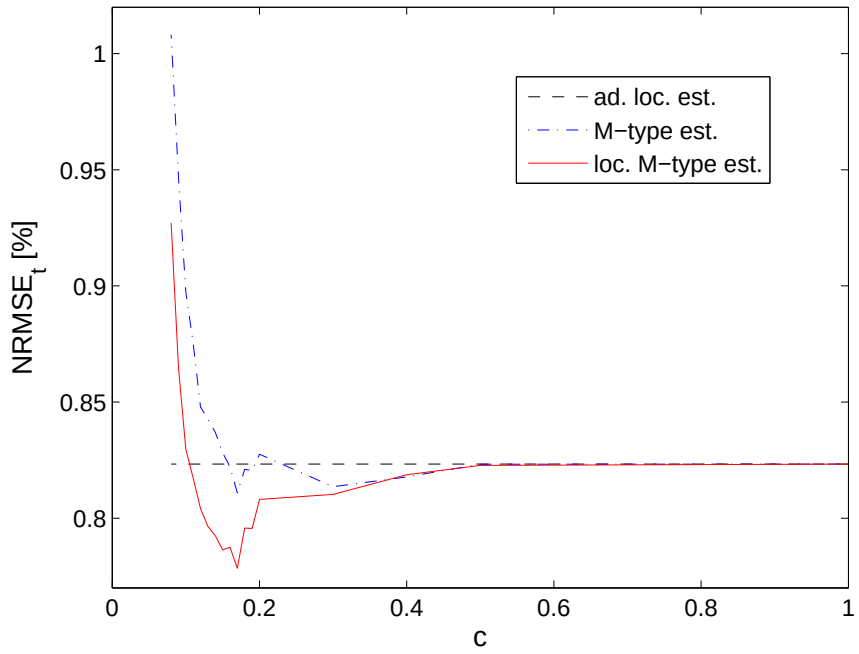
Performance of the M-type estimators $\hat{\theta}^*$ and $\hat{\theta}^{**}$ In a first stage, the benefits from robustifying the adaptive local estimator by introducing the related M-type estimator $\hat{\theta}^*$, as well as the benefits resulting from the local robustification, are discussed.

The additional parameter for the M-type estimators $\hat{\theta}^*$ and $\hat{\theta}^{**}$ is the threshold point c of the loss function. It is considered here that in practice c would be a user-defined parameter, and therefore we want to see what the variations of the error criteria are, depending on the chosen value for c. The evolution of the NRMSE criterion for $\hat{\theta}^*$ and $\hat{\theta}^{**}$ as a function of c, is depicted in Figure 4 (with Figure 4(a) giving NRMSE_t and Figure 4(b) giving NRMSE_r), with c varying from 1 down to 0.08.

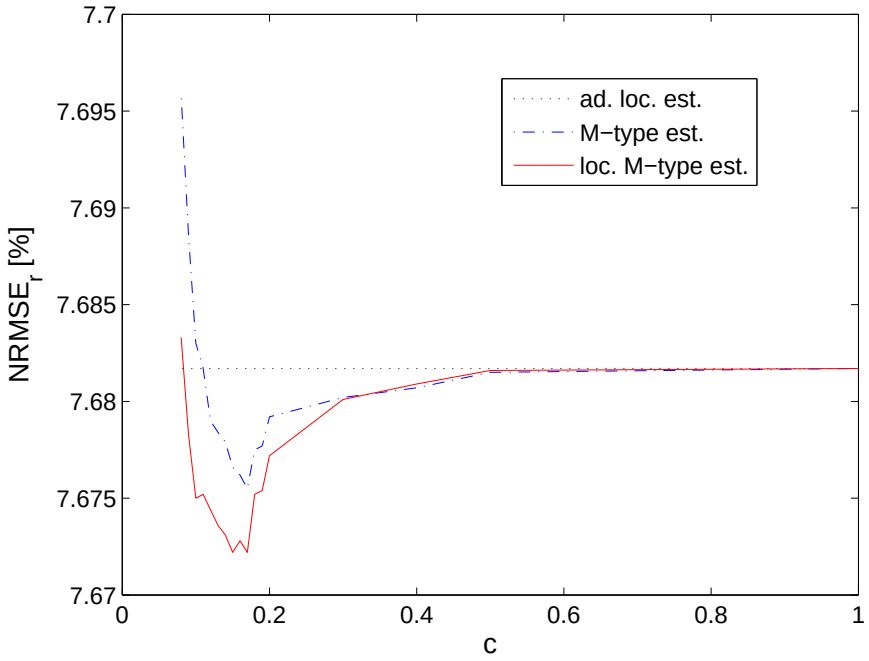
Let us focus on the performance of the two robust estimators against the pure power data. For both M-type estimators, NRMSE_t decreases until a limit threshold point is reached, for which further lowering the threshold value has the consequence of dramatically affecting the performance of the M-type estimators. Their performance is rather sensitive to the choice of the threshold point. From the shape of the various curves, it can be seen that there is a optimal value for c that results in the lowest RMSEs. That optimal threshold point is $c = 0.17$ for both estimators. In parallel, notice that for all values of c, the NRMSE_t of the local M-type estimator $\hat{\theta}^{**}$ is lower or at worst equal to that of the estimator $\hat{\theta}^*$ that is not locally robustified. This clearly illustrates the benefits from our proposal for local robustification introduced in Paragraph 3.3. This local robustification permits a better trade-off in the discrimination of residuals.

Performance of the adaptive local M-type estimator $\hat{\theta}^\dagger$ In a second stage, the adaptive local M-type estimator $\hat{\theta}^\dagger$ is used for approximating the true regression $g(u)$ from the corrupted power data, in order to show the benefits from an adaptive scaling of the local M-type estimator $\hat{\theta}^{**}$. The additional parameters for $\hat{\theta}^\dagger$ are the number m of simulated residuals, and the proportion α of residuals to be considered as suspicious. Owing to the large size of the dataset (which is consistent with wind power forecasting applications, for which time-series usually consist of thousands of time-steps), m can be set to a sufficiently large value, say $m = 1000$. Then, it is studied how the chosen value for α impacts the performance of $\hat{\theta}^\dagger$. Figures 5(a) and 5(b) depict the evolution of the error criteria NRMSE_t and NRMSE_r as a function of α. They are compared to the NRMSEs of the local adaptive estimator $\hat{\theta}$ and of the optimal local M-type estimator $\hat{\theta}^{**}$ (thus for $c = 0.17$).

Even for $\alpha = 0$, some residuals may be considered as suspicious, if their value is outside of the range of simulated residuals. If choosing that value for α, this translates to discard extreme outliers only. Then, as we have noticed in the above Paragraph with the c-parameter, the evolution of the two error criteria with α are U-shaped functions. There is thus a unique value of α for which the NRMSE_t or NRMSE_r are minimum. Also, one sees

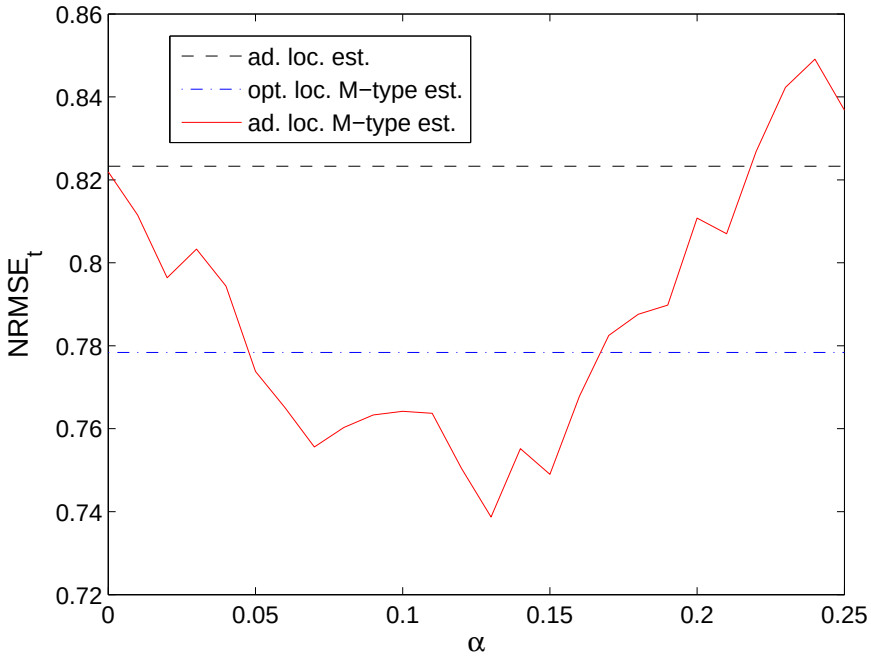


(a) NRMSE_t - Performance against pure power data.

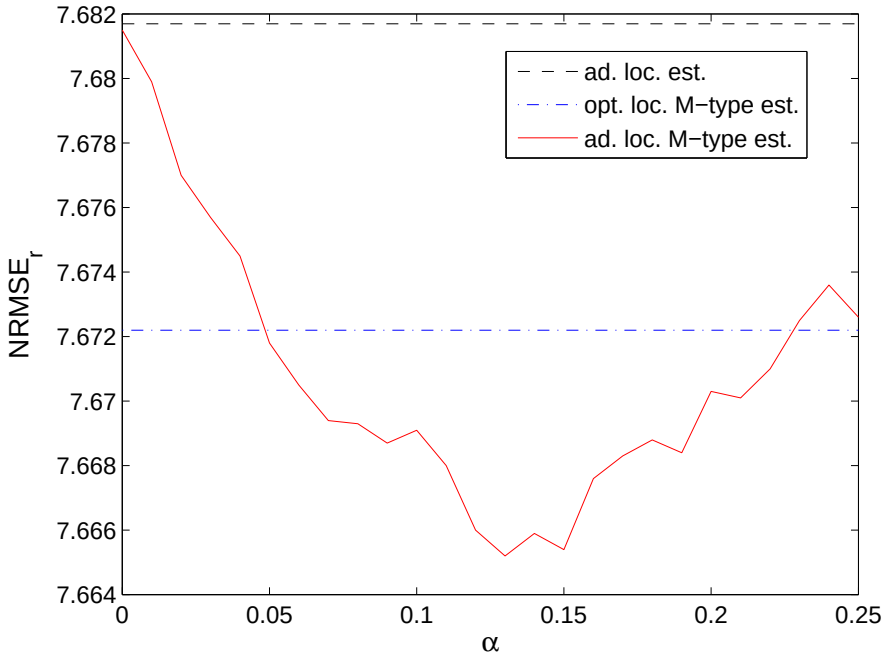


(b) NRMSE_r - Performance against simulated power data.

FIGURE 4: *Evolution of the NRMSE criteria as a function of the threshold c. NRMSE is calculated against the pure and simulated power data. Results are for the M-type estimator $\hat{\theta}^*$ and the local M-type estimator $\hat{\theta}^{**}$. Both M-type estimators are compared to the local adaptive estimator $\hat{\theta}$.*



(a) NRMSE_t - Performance against pure power data.



(b) NRMSE_r - Performance against corrupted power data.

FIGURE 5: *Evolution of the NRMSE criteria as a function of the α-parameter for the dataset 1. Results are for the adaptive local M-type estimator $\hat{\theta}^{\dagger}$. They are also compared to the NRMSE values obtained with the $\hat{\theta}$ estimator, and the optimal local M-type estimator $\hat{\theta}^{**}$.*

that these U-shaped functions are much more flat than those for the evolution of the error criteria as a function of c: the adaptive M-type estimator is less sensitive to the choice of the proportion parameter. This is because for low values of c, a slight decrease of that parameter signifies a large proportion of residuals that are additionally discarded for model adaptation. Inversely, when using α as a control parameter, one directly determines that proportion. Here, the optimal performance in term of minimum NRMSE_t is reached for $\alpha = 0.13$.

The minimum NRMSE_t for all the competing estimators, as well as the related values of the other criteria, are gathered in Table 1. M-type estimators exhibit lower values for all error criteria, whether if evaluated against the corrupted or noise-free power data. One can appraise the steady error reduction (for both NRMSE_t and NMAE_t) when going from the estimators $\hat{\boldsymbol{\theta}}$ to $\hat{\boldsymbol{\theta}}^{\dagger}$, which illustrates the additional benefits from robustification, local robustification, and adaptive scaling. But again, the error reduction is more significant when the various criteria are calculated against the pure power data than if computed against the corrupted data. Here for instance, the reduction in NRMSE_t is of 10.28% when going from $\hat{\boldsymbol{\theta}}$ to $\hat{\boldsymbol{\theta}}^{\dagger}$, while it is of only 0.2% if checking the reduction in NRMSE_r. Even if the NMAE_r permits to better reveal the error reduction (diminution of 0.43%), it is not really representative.

TABLE 1: *Minimum values of the NRMSE_r and related values of the other evaluation criteria for the adaptive local estimator $\hat{\boldsymbol{\theta}}$ and various M-type estimators, on dataset 1.*

	$\hat{\boldsymbol{\theta}}$	$\hat{\boldsymbol{\theta}}^{*}$	$\hat{\boldsymbol{\theta}}^{**}$	$\hat{\boldsymbol{\theta}}^{\dagger}$
NMAE_r	4.3887	4.3832	4.3786	4.3698
NMAE_t	0.6384	0.6310	0.6047	0.5743
NRMSE_r	7.6817	7.6755	7.6722	7.6652
NRMSE_t	0.8233	0.8109	0.7784	0.7387

6.3.2 Noise on both wind speed and power (dataset 2)

The second dataset corresponds to a more realistic situation for which a noise component would be present in both the explanatory and the response variables. Even if the main assumption of the M-type estimators introduced above is that the noise component is on the response variable only, the various estimators are applied here in order to evaluate their performance on a more realistic test case.

In a first stage the same type of cross validation procedure is used for determining the optimal set of parameters for the local adaptive estimator $\hat{\boldsymbol{\theta}}$. This optimal set is found to be:

$$[J \; h_0 \; h_1 \; \lambda] \; = \; [20 \; 0.028 \; 1.2 \; 0.987]$$

Then, a study similar to that presented in the above Paragraph is carried out in order to appraise the influence of the choice of the parameters of the M-type estimators on their

resulting performance. The behavior of the NRMSE curves is similar, leading to a single optimal parameter c or α for each estimator. For that reason, the various curves are not shown here. The minimum RMSE_t is reached for $c = 0.15$ and $c = 0.13$ for the M-type estimator $\hat{\theta}^*$ and its local version $\hat{\theta}^{**}$, respectively, while that minimum value is reached for a proportion $\alpha = 0.28$ for the case of the adaptive local M-type estimator $\hat{\theta}^\dagger$. This means that more data need to be considered as suspicious, which is indeed consistent with the fact that dataset 2 is more corrupted. The minimum NRMSE_t for all the competing estimators, as well as the related values of the other criteria, are gathered in Table 2.

TABLE 2: *Minimum values of the NRMSE_r and related values of the other evaluation criteria for the adaptive local estimator $\hat{\theta}$ and various M-type estimators, on dataset 2.*

	$\hat{\theta}$	$\hat{\theta}^*$	$\hat{\theta}^{**}$	$\hat{\theta}^\dagger$
NMAE_r	7.14	6.97	6.96	6.92
NMAE_t	2.39	2.00	1.99	1.74
NRMSE_r	11.48	11.48	11.49	11.50
NRMSE_t	2.98	2.51	2.51	2.10

On a general basis, the level of error is much higher than for the previous test case for which only power data where affected by a noise component. But also, the diminution of the error criteria is more significant, except for the NRMSE_r that exhibits a slight increase when robustifying and adaptively scaling the estimators. This criterion is not representative of the better ability of the estimators to approximate the true regression. Inversely, the decrease in NMAE_r reflects that ability. This is in line with the discussion on error criteria for the evaluation of wind power forecasting models in Madsen et al. (2005).

In parallel, one sees that a large share of the error reduction is due to the introduction of the M-type estimator $\hat{\theta}^*$, and to the adaptive scaling. The effects of the local robustification are less visible in this case. This is because data points affected by a large noise component on the wind speed are used to update the local coefficients at a wrong fitting point anyways, and are generally discarded owing to the adaptive scaling of the M-type-estimator. The local robustification have then some effect on the smaller residuals only, and the benefits are thus smaller. The reduction in NRMSE_t when going from $\hat{\theta}$ to $\hat{\theta}^\dagger$ is of 26.30% for this test case.

7 Validation results

7.1 Data and exercise

For validation purposes, focus is given in a second stage to a second Danish wind farm, Middelgrunden, which is a 20MW wind farm located few hundred meters off the coast of Sjælland at the level of København. For this wind farm, our aim is to apply the developed methods to real-world data, in order to appraise what the benefits from applying the robust adaptive estimation methods may be in operational conditions. The dataset considered is

composed by meteorological forecasts of wind speed, provided by Hirlam for a grid node close to the location of the wind farm, and by related power measurements. The raw power measurements correspond to the available power at the level of the wind farm. However, they may not be consistent with the power curve for the wind farm if all turbines are not available. Since this information on wind turbine availability is also monitored, it has been used for correcting the power dataset i.e. for rescaling power measurements. In operational conditions, power measurements (or meteorological measurements) may not be corrected by automatic procedures, and this is expected to affect the performance of prediction methods relying on these data for updating their model parameters. We will talk then in the following about two types of dataset, which will referred to as the 'raw' and 'corrected' ones. They are both composed by 1150 data points. Wind speed and power data are normalized by their maximum values.

The exercise consists in modeling the regression curve from wind speed forecasts to power, which can be consequently used for forecasting purposes. The wind direction variable is not considered in this exercise, though it would be used in real operayion conditions. The (randomly) chosen forecast horizon is of 12-hour ahead. For modeling this curve, the first 500 hundred data points of the corrected dataset are used. The adaptive local estimator $\hat{\theta}$ is fitted on these datapoints, with the aim of minimizing a quadratic criterion. The fitting points are uniformly distributed on the unit interval, and their number is set to 20. The bandwidth for each fitting point is controled by a constant h_0 and a scale factor h_1 as in the simulation exercises presented above. After determining the optimal model for the conversion function of wind speed to power, i.e. that which minimizes a quadratic criterion, on the training set, it is evaluated on the remaining 650 data points. For comparison, two of the robust estimators introduced in the present report, i.e. the local M-type estimator $\hat{\theta}^{**}$, and the adaptive local M-type estimator $\hat{\theta}^{\dagger}$, are also applied and evaluated. The sensitivity of their performance depending on the choice of their parameters is also studied.

The second part of the exercise consists in fitting the adaptive local estimator $\hat{\theta}$ to the 650 last raw data points, in order to mimick the fact that when models are setup for online applications, often only raw data are used both the initial estimation and online adaptation of the model parameters. This may then affect the quality of the resulting predictions. Therefore, the same two robust estimators are also applied in order to quantify the benefits of robustification for real-world datasets.

7.2 Results and comments

7.2.1 Results on the 'clean´ dataset

As expressed above, the aim in this Paragraph is to estimate the optimal local adaptive model $\hat{\theta}$ on the training set composed by clean data only, and to evaluate its performance on the evaluation set that is also composed by clean data only. An optimization by trial and error leads to setting h_0 and h_1 to 0.06 and 8, respectively. Finally, the optimal forgetting factor is found to be 0.98. The performance of $\hat{\theta}$ for 12-hour ahead prediction of wind power production from the wind speed forecasts is quantified by using NMAE and NRMSE criterion. Their value over the evaluation set is given in Table 3. Note that these values are at the level of what can be seen in the state-of-the-art for wind power forecasting methods,

see e.g. (Madsen et al., 2005).

The same h_0, h_1 and λ parameters are used for defining the two robust estimators $\hat{\theta}^{**}$ and $\hat{\theta}^{\dagger}$. Then, the influence of the parameter c and α, which permit to control the loss criterion for $\hat{\theta}^{**}$ and $\hat{\theta}^{\dagger}$ respectively, on the estimators' performance is studied. The number m of simulated residuals is set to 300. The aim is to minimize a quadratic criterion — thus the NRMSE error measure. Minimum values are obtained for $c = 0.06$ and $\alpha = 0.38$. The corresponding NRMSE and NMAE values over the evaluation set are given in Table 3. The descrease in NRMSE and NMAE exist when robustifying the 'classical' adaptive local estimator $\hat{\theta}$, though its magnitude is small. As it was discussed in Section 6, error measures calculated against real-world data may not reveal the benefits from robustification.

TABLE 3: *Minimum values of the NRMSE and related values of the other evaluation criteria for the adaptive local estimator $\hat{\theta}$, the local M-type estimator $\hat{\theta}^{**}$, and the adaptive local M-type estimator $\hat{\theta}^{\dagger}$, on the 'clean' dataset for the Middelgrunden test case and for the 12-hour ahead look-ahead time.*

	$\hat{\theta}$	$\hat{\theta}^{**}$	$\hat{\theta}^{\dagger}$
NMAE	9.700	9.520	9.540
NRMSE	14.848	14.797	14.766

In order to have a visually appraise the difference between the estimated functions for the conversion of wind speed to power, Figure 7.2.1 depicts the functions obtained at the end of the evaluation set by applying the three estimators $\hat{\theta}$, $\hat{\theta}^{**}$, and $\hat{\theta}^{\dagger}$. These power curves appear similar. But, an important detail is that those obtained with $\hat{\theta}^{**}$ and $\hat{\theta}^{\dagger}$ are lower for low power values and higher for large power values. While it is known that $\hat{\theta}$ lacks discrimitation ability, which means that it is locally biased (Pinson, 2006), this aspect is improved by robustifying this estimator. This improvement is more significant if considering the adaptive robust estimator $\hat{\theta}^{\dagger}$.

7.2.2 Results on the 'raw´ dataset

The same exercise is then carried out with 'raw' data. The optimization procedure yields the same parameter h_0 and h_1 for defining the bandwidth, while the optimal forgetting factor has a slightly lower value, $\lambda = 0.975$. This reduction of the effective number of observations used for adaptive model estimation is a natural consequence of the lower quality of the data.

The two robust estimators are applied on this dataset, and the sensitivity of the choice of the parameters c and α on the NRMSE and NMAE error measures is studied. The evaluation of the error measures as a function of c for the local robust estimator $\hat{\theta}^{**}$, and as a function of α for adaptive robust estimator $\hat{\theta}^{\dagger}$ are depicted in Figures 7 and 8, respectively.

The evolution of the NRMSE error measure when descreasing the value of the threshold parameter c is surprising, as one notices that for a large range of c-values the NRMSE of the robust estimator is higher than that of the adaptive local estimator $\hat{\theta}$. Though, if look-

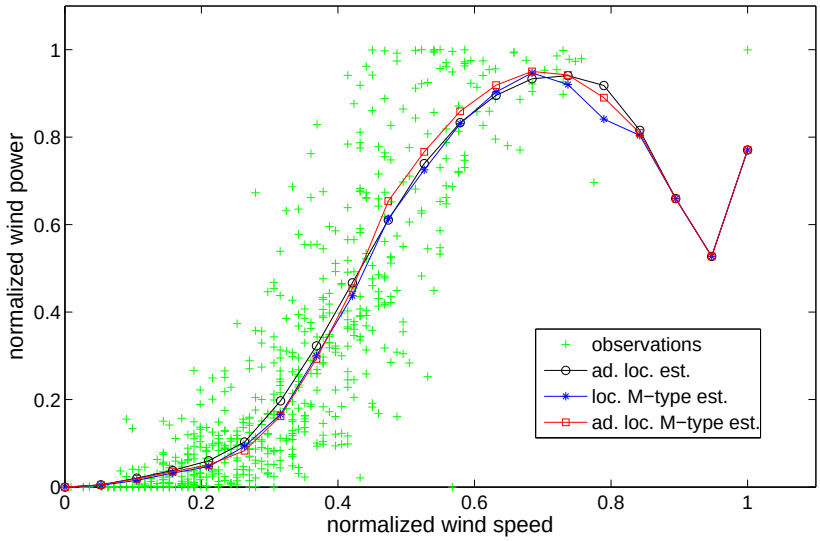
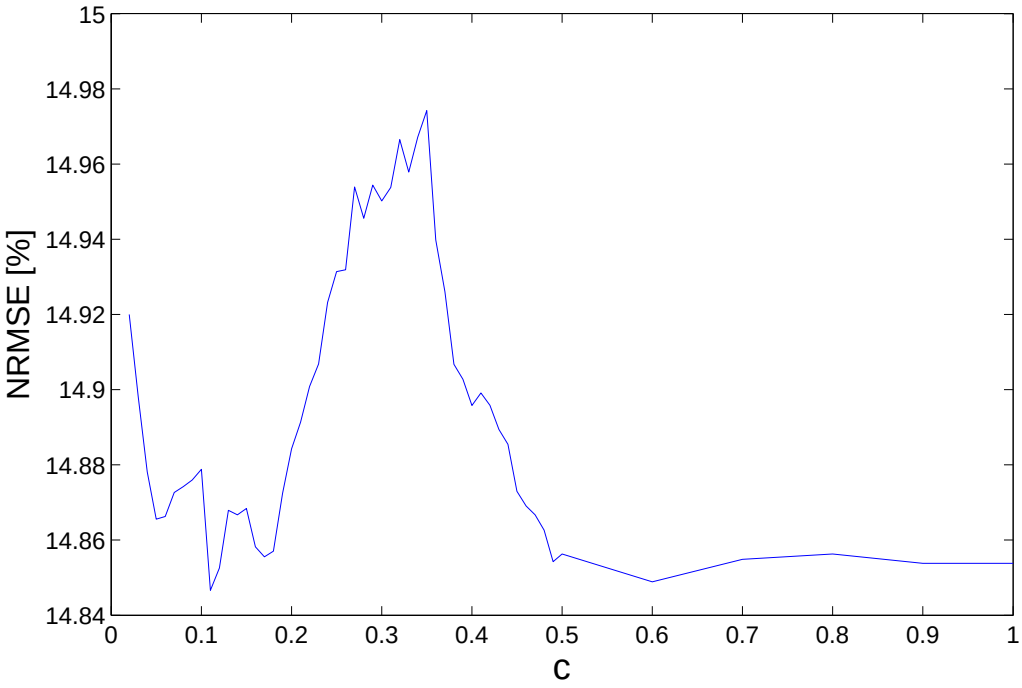


FIGURE 6: *Estimated models (at the end of the evaluation set) for the conversion of wind to power with the adaptive local estimator $\hat{\theta}$ the local M-type estimator $\hat{\theta}^{**}$, and adaptive local M-type estimator $\hat{\theta}^{\dagger}$.*

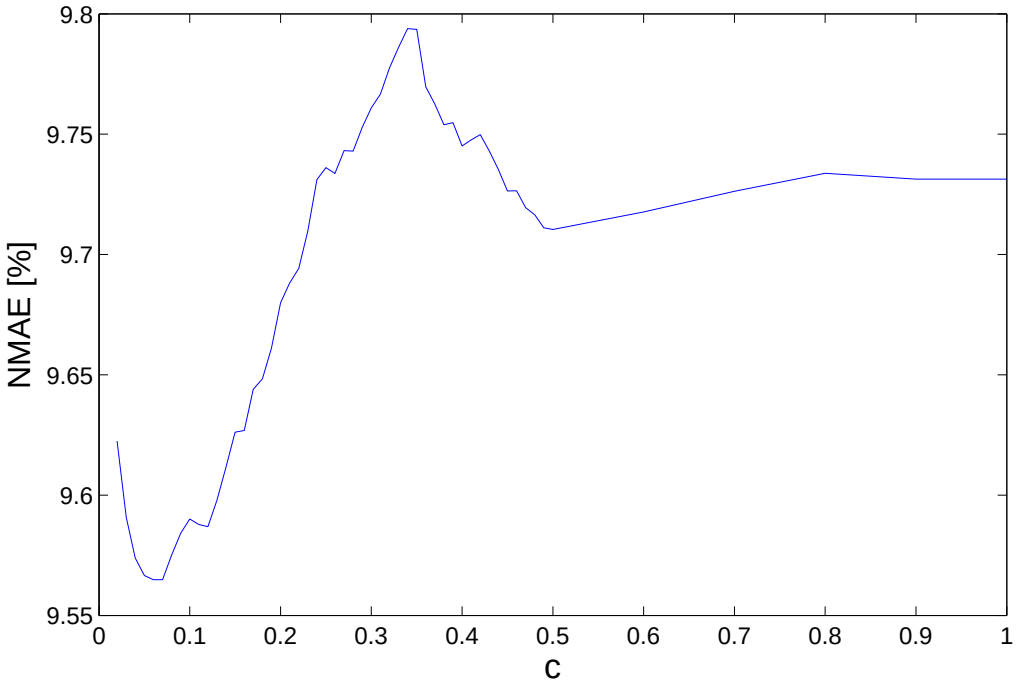
ing at the evolution of the NMAE criterion, one sees that it significantly decreases for lower values of c. A minimum value of the NRMSE is obtained for $c = 0.11$. For comparison, the evolution of the NRMSE and NMAE criteria is much more smooth for the adaptive robust estimator $\hat{\theta}^{\dagger}$. This confirms the simulation results presented in Section 6, as well as the operation interest of applying this estimator, since it will be less sensitive to an error in the choice of the optimal value for the proportion parameter α. The minimum NRMSE value is obtained for $\alpha = 0.48$, thus indicating that a larger share of the residuals are downweighted for model adaptation than when working with the 'clean' dataset. The performance of the three competing estimators, that is, their minimum NRMSE and the related NMAE value, is gathered in Table 4. Again, the decrease in both criteria exists, though it is not of large magnitude. The lower NMAE values confirm that the robust estimators are more central though. In addition, an interesting point is that the perormance of the adaptive local estimator $\hat{\theta}$ is only slightly affected when being fitted on the raw data instead of the clean ones.

TABLE 4: *Minimum values of the NRMSE and related values of the other evaluation criteria for the adaptive local estimator $\hat{\theta}$, the local M-type estimator $\hat{\theta}^{**}$, and the adaptive local M-type estimator $\hat{\theta}^{\dagger}$, on the 'raw' dataset for the Middelgrunden test case and for the 12-hour ahead look-ahead time*

	$\hat{\theta}$	$\hat{\theta}^{**}$	$\hat{\theta}^{\dagger}$
NMAE	9.736	9.588	9.568
NRMSE	14.854	14.846	14.815

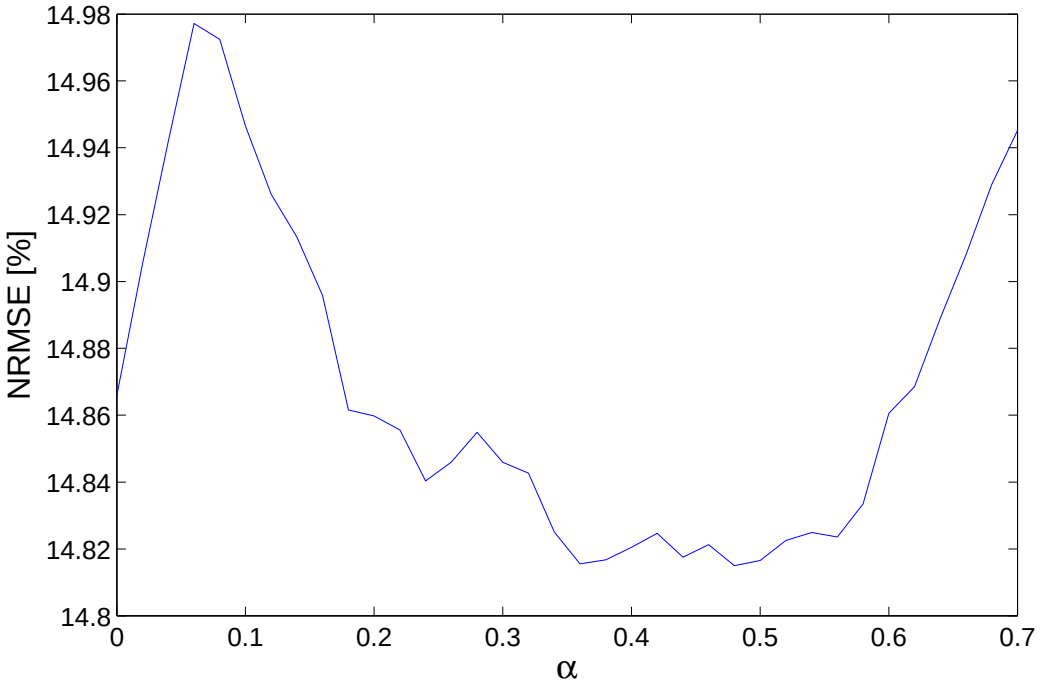


(a) NRMSE

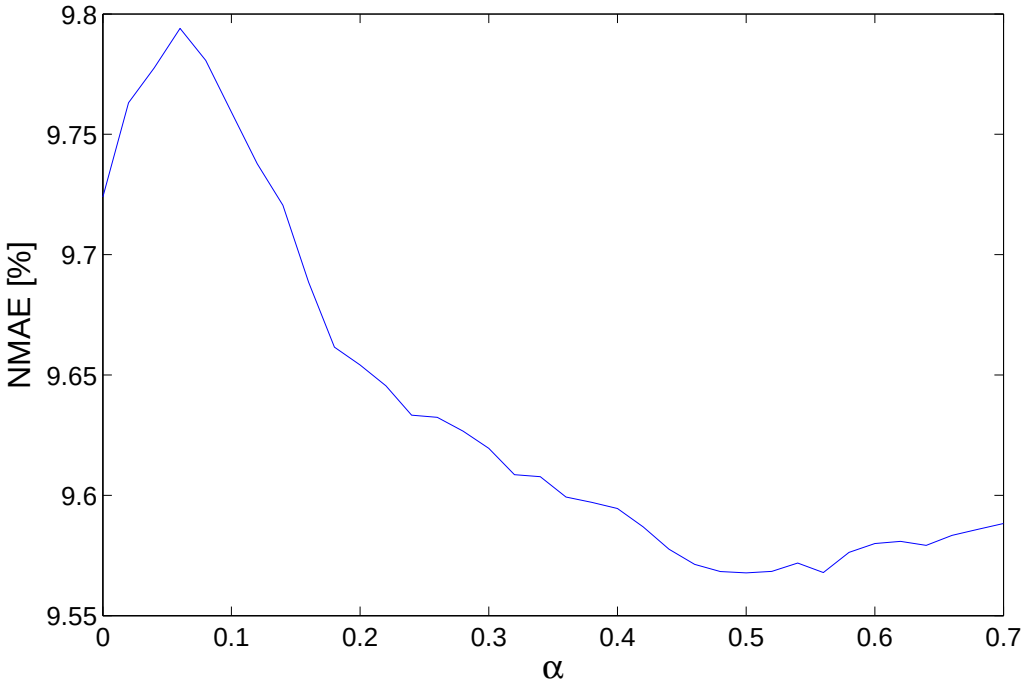


(b) NMAE

FIGURE 7: *Evolution of the NRMSE and NMAE criteria as a function of the threshold poarameter c. The results correspond to the case for which the local M-type estimator $\hat{\theta}^{**}$ is applied to an evaluation set composed by raw data, with basis the adaptive local estimator $\hat{\theta}$ fitted on a training set of raw data.*



(a) NRMSE



(b) NMAE

FIGURE 8: *Evolution of the NRMSE and NMAE criteria as a function of the α-parameter. The results correspond to the case for which the adaptive local M-type estimator $\hat{\boldsymbol{\theta}}^{\dagger}$ is applied to an evaluation set composed by raw data, with basis the adaptive local estimator $\hat{\boldsymbol{\theta}}$ fitted on a training set of raw data.*

8 Conclusions

A method for a robust and adaptive estimation of time-varying coefficient functions in conditional parametric models has been described, on the basis of the adaptive local estimation method initially introduced by Nielsen et al. (2000). Our motivation for introducing this robust version of that method originates from our experience with wind power modeling and forecasting. This process is a nonstationary process for which a non-negligible noise component is present in both the response and explanatory variables. Even if the primary aim of the introduced method is its application to wind power related matters, it has been described in a generic manner, so that it may also be applied to other types of processes.

The basis of the described approach is the proposal of an M-type estimator, which generalizes the local estimator of Nielsen et al. (2000) for a broader class of loss functions, namely the bounded-influence and convex ones. It was clearly explained that the main assumption when using these types of robust estimators is that the noise component is on the response variable only. Since it is argued in the literature that methods based on such a simplistic assumption are not suitable when both explanatory variables and the response include a noise component (Cheng and Van Ness, 1997), it will be of particular interest in the future to benchmark the introduced M-type estimator against other robust estimation methods for which the assumption on the noise component is relaxed (e.g. Robust Total-LS, multivariate-LMS, etc.).

The initial M-type estimator has been enhanced by considering a local robustification, accounting for the weights originating from local polynomial regression. Simulation results have been given, on the test case of the modeling of the nonstationary conversion process of wind speed to wind power. The datasets included wind speed measurements at the level of a wind farm, as well as simulated wind speed and power data. The interest of using such semi-artificial datasets was to have access to the true regression for evaluating the performance of the proposed estimators. The simulation results have shown the interest of introducing the M-estimator and the local robustification, by exhibiting a significantly lower level of the MSE of these estimators. It was also argued that the effective reduction of the estimation error against the true regression may not be visible when calculating the error criteria against the noisy data. The elegant feature of that local robustification is that it consists in considering the influence of the weighted residuals, the weights being given by the local polynomial regression.

An adaptive version of the local M-estimator has been introduced, by proposing an adaptive scaling of the threshold points of the loss function, after relaxing the symmetry constraint on these loss functions. The two additional parameters of the adaptive local M-type estimator in are the number m of simulated residuals for an empirical estimation of the residual distribution, and the proportion α of residuals to be considered as suspicious. For the test cases considered, the optimal performance of the estimator was not highly sensitive to the chosen value for α. In the future, it may be envisaged to determine α from a cross-validation procedure, along with the other parameters of the adaptive local estimator, in order to have an optimal setting of $\hat{\boldsymbol{\theta}}^{\dagger}$ for out-of-sample applications.

The proposed robust estimators have shown to be an interesting alternative to the adap-

tive local estimator actually used in operational wind power prediction tools e.g. WPPT (Nielsen et al., 2002). From the validation results on real-world data, it can be concluded that applying the introduced robust estimators will indeed be beneficial for wind power forecasting in operational conditions. However, we have only considered a conditional non-parametric model for the conversion of wind speed to power, and the study should now be extended to the case of conditional parametric model that would consider other explanatory variables as input e.g. wind direction. Finally, the proposed robust estimator should also be implemented in operational forecasting toold e.g. WPPT, in order to verify the presented benefits for real-world on-line forecasting applications.

Acknowledgments

The methods and results gathered in the present report have been generated as part of the project 'Intelligent wind power prediction systems, partly supported by the Danish PSO funds (PSO Project number: FU 4101), which is hereby greatly acknowledged. The authors also gratefully acknowledge Elsam for providing the data used as input.

References

Antoch, J., Ekblom, H., 1995. Recursive robust regression: computational aspects and comparison. Computational Statistics and Data Analysis 19, 115–128.

Beran, J., Feng, Y., Gosh, S., Sibbertsen, P., 2002. On robust local polynomial regression with long-memory errors. International Journal of Forecasting 18, 227–241.

Cai, Z., Ould-Saïd, E., 2003. Local M-estimator for non-parametric time-series. Statistical and Probability Letters 65, 433–449.

Chan, S.-C., Zhang, Z., 2004. Robust local polynomial regression using M-estimator with adaptive bandwidth. In: Proc. of the ISCAS'04 Conference, 'International Symposium on Circuits And Systems', IEEE Conference. Vol. 3. pp. 333–336.

Chen, D. S., Jain, R. C., 1994. A robust backpropagation algorithm for function approximation. IEEE Transactions on Neural Networks 5 (3), 467–479.

Cheng, C.-L., Van Ness, J. W., 1997. Robust calibration. Technometrics 39 (4), 401–411.

Cleveland, W. S., Devlin, S. J., 1988. Locally weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association 83, 596–610.

del Río, F. J., Riu, J., Rius, F. X., 2001. Robust linear regression taking into account errors in the predictor and response variables. Analyst 126, 1113–1117.

Fan, J., Hu, T.-C., Truong, Y. K., 1994. Robust non-parametric function estimation. Scandinavian Journal of Statistics 21 (4), 433–446.

Fan, J., Jiang, J., 1999. Variable bandwidth and one-step local M-estimator. Science in China Serie A 29, 1–15.

Giebel, G., Kariniotakis, G., Brownsword, R., June 2003. State of the art on short-term wind power prediction, aNEMOS Deliverable Report D1.1, available online: http://anemos.cma.fr.

Granger, C. W. J., 1993. On the limitations of comparing mean squared forecast errors: comment. Journal of Forecasting 12, 651–652.

Hampel, F. R., March 2001. Robust statistics: a brief introduction and overview, invited talk in the symposium "Robust Statistics and Fuzzy Techniques in Geodesy and GIS", ETH Zurich.

Hampel, F. R., Rousseeuw, P. J., Ronchetti, E. M., Stahel, W. A., 1986. Robust Statistics - The Approach based on Influence Functions. Wiley, New York.

Hastie, T. J., Tibshirani, R. J., 1990. Generalized Additive Models. Chapman & Hall, London.

Huber, P. J., 1981. Robust Statistics. Wiley, New York.

Kelly, G. E., 1992. Robust regression estimators - the choice of tuning constants. The Statistician 41, 303–314.

Li, S. Z., 1996. Robustizing robust M-estimation using deterministic annealing. Pattern Recognition 29 (1), 159–166.

Li, S. Z., Wang, H., Soh, W. Y. C., 1998. Robust estimation of rotation angles from image sequences using the annealing M-estimator. Journal of Mathematical Imaging and Vision 8, 181–192.

Lindström, T., Holst, U., Edner, H., 1996. Robust local polynomial regression and statistical evaluation of DOAS measurements. Tech. Rep. TFMS-3127, Department of Mathematical Statistics, Lund Institute of Technology.

Madsen, H., Pinson, P., Kariniotakis, G., Nielsen, H. A., Nielsen, T. S., 2005. Standardizing the performance evaluation of short term wind power prediction models. Wind Engineering 29 (6), 475–489.

Nielsen, H. A., Nielsen, T. S., Joensen, A. K., Madsen, H., Holst, J., 2000. Tracking time-varying-coefficient functions. International Journal of Adaptive Control and Signal Processing 14, 813–828.

Nielsen, T. S., Madsen, H., Nielsen, H. A., 2002. Prediction of wind power using time-varying coefficient functions. In: Proc. IFAC 2002, 15$^{\text{th}}$ World Congress on Automatic Control, Barcelona, Spain.

Petrus, P., 1999. Robust Huber adaptive filter. IEEE Transactions on Signal Processing 47 (4), 1129–1133.

Pinson, P., 2006. Estimation of the uncertainty in wind power forecasting. Ph.D. thesis, Ecole des Mines de Paris, Paris, France, available: www.pastel.paristech.org/bib.

Rousseeuw, P. J., 1984. Least median of squares regression. Journal of the American Statistical Association 79, 871–880.

Rousseeuw, P. J., Leroy, A. M., 1987. Robust Regression and Outlier Detection. Wiley, New York.

Wang, F. T., Scott, D. W., 1994. The L_1 method for robust non-parametric regression. Journal of the American Statistical Association 89, 65–76.

Weiss, A. A., 1996. Estimating time series models using the relevant cost function. Journal of Applied Econometrics 11, 539–560.

Welsh, A. H., 1994. Robust estimation of smooth regression and spread functions and their derivatives. Statistica Sinica 6, 347–366.

Zou, Y., Chan, S.-C., Ng, T. S., 2000a. Least mean M-estimate algorithms for robust adaptive filtering in impulse noise. IEEE Transactions on Circuits and Systems — II: Analog and Digital Signal Processing 47 (12), 1564–1569.

Zou, Y., Chan, S.-C., Ng, T. S., 2000b. A recursive robust least M-estimate (RLM) adaptive filter for robust filtering in impulse noise. IEEE Signal Processing Letters 7 (11), 324–326.