Online Adaptive Lasso Estimation in Vector Auto Regressive Models for High Dimensional Wind Power Forecasting

Abstract

Wind power forecasts with lead times up to a few hours are essential to optimally and economically operate power systems and markets. Vector autoregression (VAR) has shown to be a well suited framework to simultaneously predict for several wind farms by considering the spatio-temporal dependencies in their time series. Lasso penalisation yields sparse models and can avoid overfitting the many coefficients in higher dimensional settings. However, estimation in VAR models usually does not account for changes in the spatio-temporal wind power dynamics that are related to e.g., seasons or wind farm setup changes. To tackle this problem this paper proposes a time-adaptive Lasso estimator and an efficient coordinate descent algorithm to update the VAR model parameters recursively online. On simulated data this approach shows good abilities to track changes in the multivariate time-series dynamics. Furthermore, two case studies show a clearly better predictive performance than non-adaptive lasso VAR and univariate auto regression.

Keywords: energy forecasting, multivariate time-series, model selection

1. Introduction

Over the past decades wind power has experienced substantial growth and has become an important power source in many countries. However, the variable nature of wind power production challenges power systems and electricity markets and requires reliable generation forecasts, e.g., for operation of reserves or wind farm control.

Wind power forecasting has been a very active field of research and a variety of different methods have been proposed in the past decades. An overview of short-term wind and wind power prediction models can be found in Giebel & Kariniotakis (2017) or more focused on probabilistic approaches in Bessa, Möhrlen, Fundel, Siefert, Browell, Haglund El Gaidi, Hodge, Cali & Kariniotakis (2017). In general, these approaches can be distinguished into purely data-driven statistical time-series models and methods that employ physical (numerical weather prediction) models. It is commonly

agreed that for very-short-term forecasts (approx. < 6 hours) statistical time-series models are superior to numerical models, which often have computation times longer than the look-ahead time.

Often, these statistical time-series models use only local data from the forecast site itself (e.g., Giebel & Kariniotakis, 2017; Pinson & Madsen, 2012) but it has been shown in various studies that forecasts can be considerably improved by additionally using data from surrounding sites or weather stations. E.g., Gneiting, Larson, Westrick, Genton & Aldrich (2006) showed that surrounding meteorological observations can improve 2-hour ahead wind speed forecasts in a regime-switching space-time diurnal model, Hering & Genton (2010) extended this model by regarding wind direction as a circular variable and by using a skew-t error distribution, and clear spatio-temporal correlations of wind power forecast errors were also found in Tastu, Pinson, Kotwa, Madsen & Nielsen (2011) and Tastu, Pinson, Trombe & Madsen (2014).

With the increasing number of wind power installations a wealth of data has become available. However, to exploit this high amount of data, advanced time-series models are required, which can take into account the important spatio-temporal dependencies while not overfitting the large number of coefficients to the training data. Furthermore, computational efficiency and scalability to a high number of forecast sites is crucial when dealing with bigger data sets.

Several approaches have been proposed for forecasting a set of spatially distributed wind farms while taking into account the spatio-temporal dependencies between them. These involve e.g., a sparse online warped Gaussian process model (Kou, Gao & Guan, 2013) or a sparse Gaussian random fields model (Wytock & Kolter, 2013). Because of their computational efficiency, vector autoregressive (VAR) models (Tastu et al., 2014; He, Vittal & Zhang, 2015) have recently received increased attention. With the aim to derive forecasts for a large number of wind farms, Dowell & Pinson (2016) proposed a sparse VAR model with a state-of-the art method for sparse VAR coefficient matrices (Davis, Zang & Zheng, 2016), which avoids overfitting. A similar approach was proposed by Cavalcante, Bessa, Reis & Browell (2017) where sparsity is achieved by lasso (least absolute shrinkage and selection operator) regularisation (Tibshirani, 1996).

Wind power generation and its dynamics clearly vary with e.g., weather conditions, seasons, or changes in the wind farm installations and various studies have shown for univariate wind power time series that models that can adapt to these variations are of clear advantage (e.g., Møller,

Nielsen & Madsen, 2008; Pinson & Madsen, 2009, 2012; Pinson, 2012). One simple approach to achieve this adaptivity, which is also employed by Dowell & Pinson (2016), is fitting the forecast models on sliding training windows. However, this requires storing data for the whole training period and refitting the model each time new data become available. Especially for computationally expensive multivariate time-series models, this can clearly limit their applicability for higher dimensional data. More advanced adaptive models have been used extensively for univariate wind power time series (e.g., Møller et al., 2008; Pinson & Madsen, 2009, 2012; Pinson, 2012) but so far have not been applied to multivariate time series. Most of these models achieve adaptivity by exponentially forgetting past data (i.e., putting less weight in the parameter estimation on data further in the past), which allows for very efficient online updates each time new data becomes available while not requiring to store past data.

This paper applies these ideas to multivariate time series and proposes an adaptive extension of lasso VAR. It is also based on exponential forgetting and we present a coordinate descent algorithm (Friedman, Hastie, Höfling & Tibshirani, 2007) for efficient online updates, which is similar to the time-weighted lasso approach of Angelosante, Bazerque & Giannakis (2010) to find sparse signals.

The ability of this approach to track changes in multivariate time-series dynamics is illustrated on simulated data. Furthermore, the predictive performance is tested on wind power data from 172 sites in Western France as well as on a set of 100 wind farms in Denmark.

The remaining document is structured as follows: First, lasso VAR and its adaptive extension is presented in Section 2. Subsequently, a simulation study in Section 3 shows the tracking ability of this approach. Section 4 presents the case studies and their results and a summary and conclusion can be found in Section 5.

2. Lasso Vector Auto Regression (Lasso VAR) and its Adaptive Extension

In the following we regard a multivariate time series with $y_t[i]$ being the wind power output at time $t \in 1, \dots, T$ and at wind farm $i \in 1, \dots, Q$. The goal is to derive predictions for these power outputs at time t based on the previous outputs at times $t-1, t-2, \dots$

2.1. Lasso Vector Auto Regression (VAR)

Before describing the multivariate VAR model, we want to first only regard the univariate time series at a single wind farm i. The dynamics of such time series can be described by auto regressive

models (e.g., Pinson & Madsen, 2012), which assume the output at time t to depend linearly on the outputs at previous time steps

$$y_t[i] = \sum_{l=1}^{L} a_l y_{t-l}[i] + e_t \tag{1}$$

where $a_l, l = 1, \ldots, L$ are the model coefficients, L is the order of the auto regressive model (i.e., the number of considered lags), and e_t are independent errors with zero mean and constant variance. Forecasts can be derived from this model as $\sum_{l=1}^{L} a_l y_{t-l}[i]$. Note that here and in the following we assume the $y_t[i]$ to be centred (zero mean) to avoid an intercept term in (1).

For a set of wind farms one could simply apply this model for each wind farm individually. However, it can clearly be of advantage to exploit spatio-temporal correlations in the data by also considering previous outputs of other wind farms than the wind farm to be forecast. For a vector of power outputs $\mathbf{y}_t = (y_t[1], y_t[2], \ldots, y_t[Q])^\top$ at Q wind farms at time t this can be expressed in a vector auto regressive (VAR) model

$$\mathbf{y}_t = \sum_{l=1}^{L} \mathbf{A}_l \mathbf{y}_{t-l} + \boldsymbol{\epsilon}_t \tag{2}$$

where $\mathbf{A}_l$ are coefficient matrices, and $\boldsymbol{\epsilon}_t$ are independent multivariate errors with zero mean and constant positive-semidefinite covariance matrix Σ. VAR forecasts $\hat{\mathbf{y}}_t$ are derived by

$$\hat{\mathbf{y}}_t = \sum_{l=1}^{L} \mathbf{A}_l \mathbf{y}_{t-l} \tag{3}$$

Note that (2) and (3) provide 1-step ahead forecasts and also the following descriptions will regard this specific case. Typical strategies for multi-step ahead forecasts include iterative computation of one-step ahead forecasts from predictions of prior time steps or directly fitting separate models for each required forecast lead time. For a review of these and other multi-step ahead forecasting strategies see e.g. Ben Taieb, Bontempi, Atiya & Sorjamaa (2012). In our case studies we used the direct approach, which avoids accumulation of forecast errors. For these direct k-step ahead forecasts ($k = 1, 2, \ldots$), the sums in (2) and (3) are replaced by $\sum_{l=1}^{L} \mathbf{A}_l \mathbf{y}_{t-l-k+1}$.

Commonly, the VAR coefficient matrices $\mathbf{A}_l$ are estimated by minimising the sum of squared errors over a training data set

$$\sum_{t=1}^{T} ||\mathbf{y}_t - \hat{\mathbf{y}}_t||_2^2 \tag{4}$$

where $||\mathbf{x}||_p = \left(\sum_{i=1}^{Q} |x[i]|^p\right)^{1/p}$ is the p-norm with Q the number of elements in a vector $\mathbf{x} = (x[1], x[2], \dots x[Q])^\top$.

The squared loss in (4) has the advantage that an analytical solution exists for $\mathbf{A}_l$. However, the number of coefficients increases quadratically with the number of sites so that, for larger numbers of considered sites, the very high number of fitted coefficients can easily lead to overfitting and deteriorate the predictive performance, especially when only small training samples are available.

Lasso regularisation (Tibshirani, 1996) is a popular and powerful method to avoid overfitting in various regression problems and Cavalcante et al. (2017) also proposed to use it for multivariate wind power forecasting. It penalises the absolute coefficient values which leads to some coefficients shrunk to zero and thus to sparse coefficient sets.

To extend VAR with lasso regularisation, a penalty term is added to the loss function (4)

$$\frac{1}{2}\sum_{t=1}^{T} ||\mathbf{y}_t - \hat{\mathbf{y}}_t||_2^2 + \lambda \sum_{l=1}^{L} ||\mathbf{A}_l||_1 \tag{5}$$

where $||\mathbf{B}||_1 = \sum_{i=1}^{Q}\sum_{j=1}^{P} |B[i,j]|$ with $B[i,j]$ being the entry in the ith line and jth column of a Q times P matrix $\mathbf{B}$. The regularisation parameter λ controls the sparsity and shrinks coefficients of less important cross correlations to zero so that the coefficient matrices become sparse and only the cross correlations are selected that contribute most to the prediction.

Unfortunately, no analytical solution exists for the minimum of (5) so that the coefficient matrices have to be estimated numerically. Cyclic coordinate descent is one of the most popular numerical optimisation approaches for lasso problems and is particularly efficient to estimate the coefficients for a sequence of different values of the penalisation parameter λ (Friedman et al., 2007; Friedman, Hastie & Tibshirani, 2010). This algorithm takes advantage of the fact that a simple analytical expression exists for the partial optimum of (5) with respect to one coefficient given that all other coefficients are fixed. Cyclic coordinate descent uses this expression to successively update the coefficients in repeated cycles until convergence. Such an update for $A_l[i,j]$ (entry in ith row and jth column of coefficient matrix $\mathbf{A}_l$) has the form (Friedman et al., 2007)

$$A_l[i,j] \leftarrow \frac{S\left(\sum_{t=1}^{T} y_{t-l}[j](y_t[i] - \hat{y}_t^{(j,l)}[i]), \lambda\right)}{\sum_{t=1}^{T} y_{t-l}[j]^2} \tag{6}$$

where $S()$ is the soft thresholding operator:

$$S(z,\gamma) = \text{sign}(z)(|z| - \gamma)_+ = \begin{cases} z - \gamma & \text{if } z > 0 \text{ and } \gamma < |z| \\ z + \gamma & \text{if } z < 0 \text{ and } \gamma < |z| \\ 0 & \text{if } \gamma > |z| \end{cases} \tag{7}$$

and $\hat{y}_t^{(j,l)}[i]$ is the fitted value for $y_t[i]$ (cf. (2)) excluding the contribution from $y_{t-l}[j]$:

$$\hat{y}_t^{(l,j)}[i] = \hat{y}_t[i] - A_l[i,j]y_{t-l}[j] \tag{8}$$

This cyclic coordinate descent algorithm is particularly efficient if solutions for a sequence of different λ are desired. Starting from a high λ (i.e., such that all coefficients are 0) λ is successively decreased. By using the coefficient estimates from the previous λ as starting values the algorithm usually converges very fast and is computationally remarkably efficient (Friedman et al., 2007, 2010).

2.2. Adaptive Lasso VAR and Recursive Online Estimation

Lasso VAR has been shown to be a well suited method to capture linear dependencies in high dimensional wind power time series (Cavalcante et al., 2017). One crucial assumption of this model is stationarity in these dependencies. However, changes in the wind power production dynamics caused by e.g., different weather situations, seasons, or changes in wind park installations can make wind power time series clearly non-stationary. One approach to account for these changes is to put more weight on more recent data so that data further in the past is neglected in the coefficient estimation. E.g., Pinson, Christensen, Madsen, Sørensen, Donovan & Jensen (2008); Møller et al. (2008); Pinson & Madsen (2012) proposed exponential forgetting of past data in univariate time-series models. This exponential forgetting can also easily be incorporated in lasso VAR by adding weights ν^{T-t} to the loss function (5):

$$\frac{1}{2}\sum_{t=1}^{T}\nu^{T-t}||\mathbf{y}_t - \hat{\mathbf{y}}_t||_2^2 + \lambda\sum_{l=1}^{L}||\mathbf{A}_l||_1 \tag{9}$$

where the forgetting factor $\nu \in (0,1)$ determines the degree of forgetting and typically has values only slightly below 1. Usually it is more informative to interpret the forgetting factor in terms of the effective training data length that is given by $1/(1-\nu)$.

The exponential forgetting can also easily be considered in the coefficient estimation by slightly modifying the update in (6) to (c.f., Friedman et al., 2010)

$$A_l[i,j] \leftarrow \frac{S\left(\sum_{t=1}^{T} \nu^{T-t} y_{t-l}[j](y_t[i] - \hat{y}_t^{(j,l)}[i]), \lambda\right)}{\sum_{t=1}^{T} \nu^{T-t} y_{t-l}[j]^2} \tag{10}$$

Clearly, such an exponential forgetting approach only makes sense if the coefficient matrices are re-estimated regularly, preferably every time step. However, for higher dimensional data sets, completely re-fitting the weighted lasso VAR in each time step can easily become computationally infeasible.

In the following we present a recursive estimation algorithm for adaptive lasso VAR. This algorithm is based on the work of Angelosante et al. (2010) who proposed a recursive estimation of standard lasso regression and uses similar ideas as Pinson & Madsen (e.g., 2009) or Møller et al. (2008). The basic idea is to re-run the cyclic coordinate descent algorithm in each time step but take the coefficient estimates from the previous time step as starting values. Since the coefficients are expected to only vary slowly, the updated coefficients should be very similar to the previous time step so that the algorithm should converge after only few iterations. Moreover, with the exponential weighting, the terms in (6) can be updated very efficiently. Therefore the first argument in the softthresholding function in (10) can be rewritten as

$$\sum_{t=1}^{T} \nu^{T-t} y_{t-l}[j](y_t[i] - \hat{y}_t^{(j,l)}[i]) = \mathbf{R}_{0,l,T}[i,j] - \sum_{m=1}^{L} \mathbf{R}_{m,l,T}[j,]^\top \mathbf{A}_l[i,] + \mathbf{R}_{l,l,T}[j,j]\mathbf{A}_l[i,j] \tag{11}$$

and the denominator as

$$\sum_{t=1}^{T} \nu^{T-t} y_{t-l}[j]^2 = \mathbf{R}_{l,l,T}[j,j] \tag{12}$$

where

$$\mathbf{R}_{m,l,T} = \sum_{t=1}^{T} \nu^{T-t} \mathbf{y}_{t-m} \mathbf{y}_{t-l}^\top \tag{13}$$

With $\mathbf{R}_{m,l,T-1}, m = 0, \ldots, L, l = 1, \ldots, L$ known from the previous time step, $\mathbf{R}_{m,l,T}$ can be derived by

$$\mathbf{R}_{m,l,T} = \nu \mathbf{R}_{m,l,T-1} + \mathbf{y}_{T-m} \mathbf{y}_{T-l}^\top \tag{14}$$

Thus, compared to a complete-refitting, the number of operations is substantially reduced. Furthermore, data has only to be stored L time steps back so that this algorithm is also very

memory efficient. Algorithm 1 summarises these updates that are performed each time new data becomes available.

Algorithm 1 Online lasso VAR update

```
acquire new data y_t
for l = 1 ... L do
    update R_{m,l,T} according to (14)
end for
repeat
    for l = 1 ... L; i = 1 ... P; j = 1 ... P do
        update A_l[i, j] according to (10)
    end for
until convergence
```

For each update of the coefficient matrices, first $O(L^2Q^2)$ operations are required for the updates (14). Additionally, (11) requires $O(Q)$ operations, a full cycle through all Q^2 coefficients thus requires $O(L^2Q^3)$ operations. Thus, the computation time increases approximately cubically with Q and quadratically with L.

The adaptive lasso VAR model has two hyperparameters: the regularisation parameter λ and the forgetting factor ν. Clearly the optimum values for these parameters also interact with each other, since with a higher forgetting rate the model should be sparser to not overfit the smaller effective training data set.

To find the optimum regularisation parameter λ for a given forgetting factor we run the online lasso VAR algorithm for a sequence of different regularisation parameters. The prediction in each time step is then taken from the λ which has the minimum weighted sum of squared errors $\sum_{t=1}^{T-1} \nu^{T-t-1} ||\mathbf{y}_t - \hat{\mathbf{y}}_t||_2^2$. Following Friedman et al. (2010) the set of λ is a sequence of K values decreasing from λ_{max} to $\xi\lambda_{max}$ on the log scale. λ_{max} is the minimum λ for which all coefficients are zero and is given by the maximum entry of $r_{l,T}$ over all l. Thus, λ_{max} is not constant so that the sequence of λ has to be updated in each online update. In our simulation and case studies we use $K = 10$ and $\xi = 0.0001$.

As noted above, the VAR model assumes centred $\mathbf{y}_t$ which for the adaptive VAR means

$$\sum_{i=1}^{T} \nu^{T-t} \mathbf{y}_t = \mathbf{0} \tag{15}$$

To achieve this property the raw data $\tilde{\mathbf{y}}_T$ is centred by

$$\mathbf{y}_T = \tilde{\mathbf{y}}_T - \frac{\sum_{t=1}^{T} \nu^{T-t} \tilde{\mathbf{y}}_t}{\sum_{n=1}^{T} \nu^{T-t}} \tag{16}$$

Note that similar updates to (**??**) and (14) can also be used to update the right expression in (16).

3. Simulation study

Before testing the adaptive lasso VAR on real data we want to show their ability to adapt to changes in synthetic data. Therefore we simulated a vector time series of length 15000 (corresponding to approx. 1/2 year of data with 15 min temporal resolution) as a VAR process of order 1

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \epsilon \tag{17}$$

where ϵ is a vector of independent standard normal random numbers and the lag-1 dependencies are specified by the matrix

$$\mathbf{A} = \begin{bmatrix} 0.9 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & a_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & a_3 & 0.9 & 0 & 0 & 0 & 0.2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.8 & 0 & 0 & 0 & 0 & 0 \\ 0.1 & 0 & 0 & 0 & 0 & 0.9 & 0 & 0 & 0 & -0.1 \\ 0 & 0 & 0 & 0 & 0 & a_2 & 0.8 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.7 & 0 & 0 \\ 0 & 0 & -0.1 & 0 & 0 & 0 & 0 & 0 & 0.9 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 \end{bmatrix} \tag{18}$$

where a_1, a_2, and a_3 are time varying coefficients that are shown as dashed black curves in Figure 1. The coefficient a_1 is oscillating slowly between 0.7 and 0.9 with a period of 30000 and is supposed to simulate seasonal changes. The coefficient a_2 is constant at 0.2 for the first 7500 time steps and

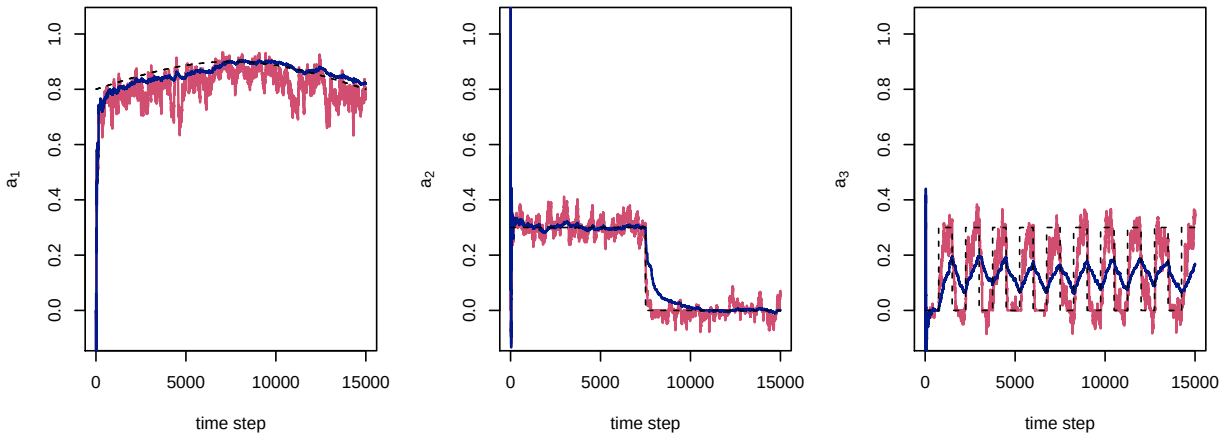

Figure 1: True correlation parameters (dashed black lines) and estimated coefficients from adaptive lasso VAR with forgetting factor 0.999 (i.e., effective training data length of 1000; red lines) and 0.99 (i.e., effective training data length of 100; blue lines)

then abruptly switches to 0 for the remaining 7500 time steps. Such a parameter change could e.g., be related to a shut down of one wind farm. Finally, a_3 switches every 1000 time steps between 0 and 0.3. A lagged correlation between sites is likely to depend strongly on the wind direction so that these parameter switches could be interpreted as changes in the wind direction.

Figure 1 also shows coefficient estimates from two adaptive lasso VAR models with different forgetting factors. After an initial burn-in period the true parameters a_1 and its slow changes are estimated quite well by both models but with much noisier estimates for the lower forgetting factor of 0.99. Similarly, also the parameter a_2 is estimated well and both models also can adapt to the abrupt change at time step 7500. However, with the higher forgetting factor the transition clearly takes longer. The model with the smaller forgetting factor can also reasonably follow the regular switches of a_3. However, the model with the higher forgetting factor clearly adapts too slowly to follow these changes. This is not surprising since its effective training data length of 1000 has the same order as the period of the changes.

Lasso regularisation shrinks less important coefficients to zero where the strength of this shrinkage is specified by the regularisation parameter. However, in Figure 1, even for the time steps when a_2 and a_3 are truly zero, their estimates are not exactly zero all the time. Figure 2 shows the full estimated coefficient matrices of the model with the higher forgetting factor for different time steps. There it also can be seen that not only the non-zero parameters but also some of the truly zero

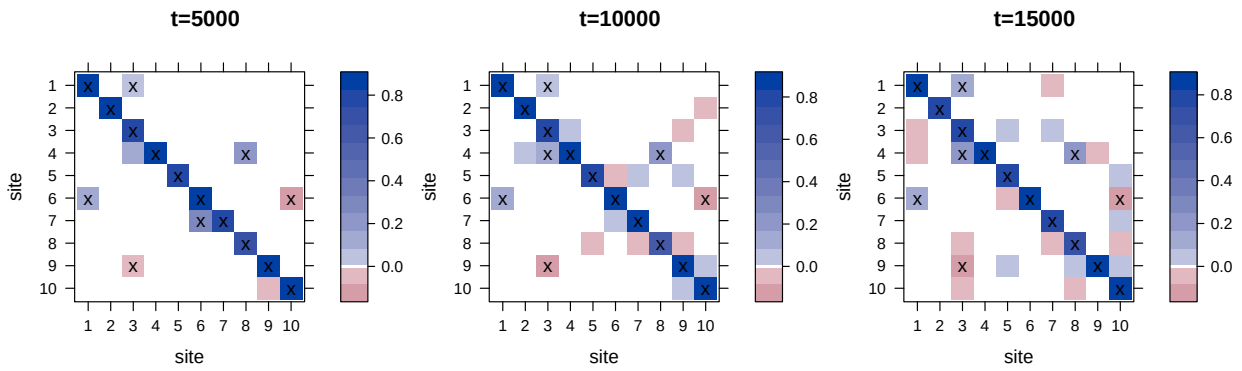

Figure 2: Estimated lag-1 coefficient matrices for the simulated time series from adaptive lasso VAR with forgetting factor 0.999. The estimates are taken from the time steps 5000 (left), 10000 (center), and 15000 (right). The crosses indicate entries with truly non-zero parameters.

parameters have non-zero estimates. For the smaller forgetting factor the matrices look similar but are even less sparse (not shown). A larger penalisation parameter would lead to sparser coefficient matrices but also imposes a bias on the parameter estimates (i.e., smaller absolute value). Here, the penalisation parameter was selected to optimise the predictive performance so that apparently it is of advantage to have less sparse matrices but with less biased coefficient estimates. Note that the size of the simulated data set is also rather big compared to the number of estimated coefficients so that overfitting is not a real issue and these non-zero estimates for actually zero parameters are very close to zero.

Clearly, there is a trade-off between low forgetting factors that can better follow fast changes in the time-series dynamics and high forgetting factors that provide more stable parameter estimates. Figure 3 compares the predictive performance of adaptive lasso VAR models with different forgetting factors. For the most part, the higher forgetting factor provides better forecasts, which indicates that the more stable parameter estimates more than compensate for the worse tracking of a_3. However, after the abrupt change of a_2 at time step 7500 the lower forgetting factor provides comparable or better forecasts approximately until time step 9000 when also the model with the higher forgetting factor has adapted to this change. Figure 3 also compares a non-adaptive lasso VAR (batch VAR) and a lasso VAR, which was fitted on a sliding training window. The non-adaptive lasso VAR was fitted only on the first 5000 time steps and then used to predict the remaining 10000 time steps. This model has a comparable performance up to time step 7500 but

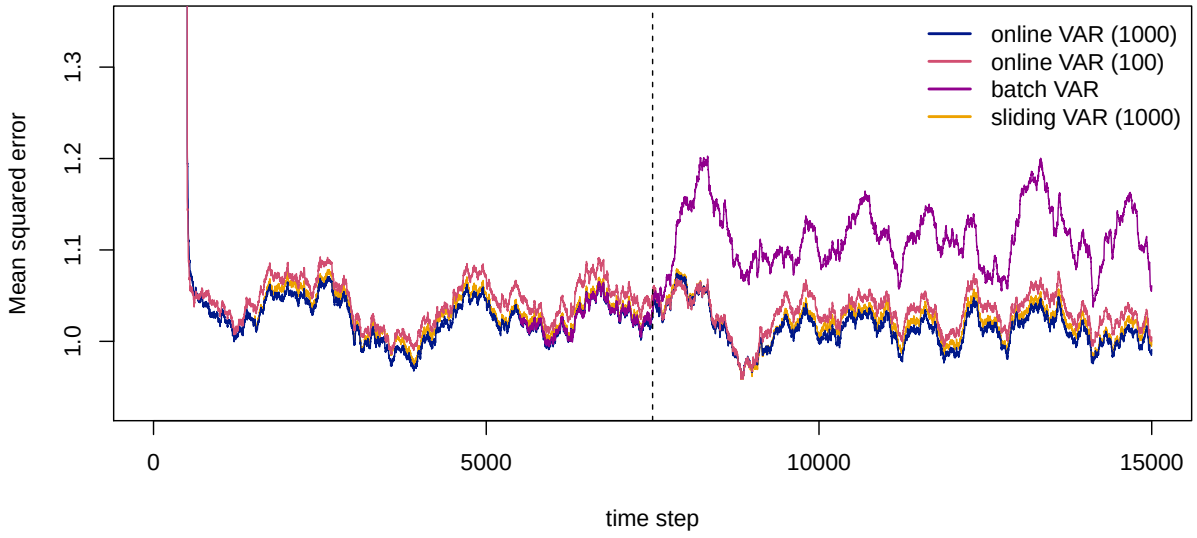

Figure 3: Mean squared error as moving average over 500 time steps of the simulated time series. The red curve is for the adaptive lasso VAR with a forgetting factor of 0.999 (i.e., effective training data length of 1000), the blue curve for the adaptive lasso VAR with forgetting factor of 0.99 (i.e., effective training data length of 100) and the green curve for a non-adaptive lasso VAR, fitted only on the first 5000 time steps. The dashed vertical line indicates the time step with the abrupt change in the parameter a_3.

since it can not adapt to the change in a_2 its predictions are clearly worse for the remaining time period. The sliding training window model also has a comparable but slightly worse performance than the online VAR with the comparable effective training data length. However, the computation time is approximately 20 times longer than for the online VAR model.

4. Case studies

The previous section showed on simulated data that adaptive lasso VAR is well suited to adapt to slow parameter changes. This section investigates in two case studies how this model performs on real wind power data. First, two data sets from France and Denmark are described and analysed. Subsequently, the performance of adaptive lasso VAR is evaluated on these data sets.

4.1. Raw data

4.1.1. Western France

As one of the data sets we use data from 183 wind farms located in Western France with nominal power between 800 kW and 16.3 MW. The data set consists of wind power generation

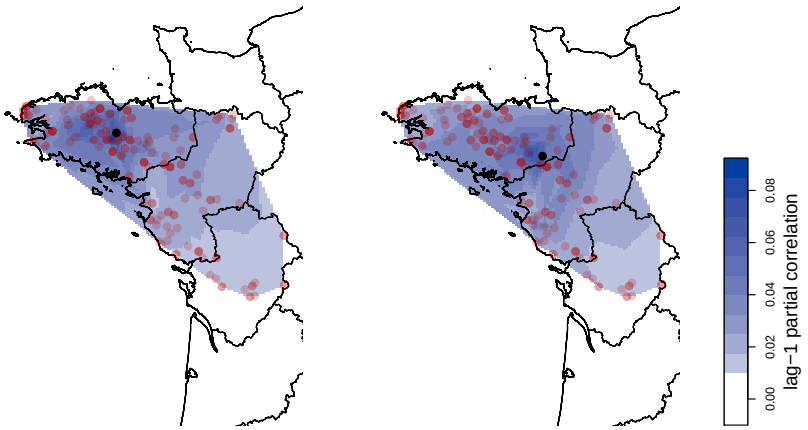

Figure 4: Partial correlation of lag-1 data from surrounding wind farms conditional on lag-1 data from the two wind farms shown as black dots. Other wind farm locations are shown as red dots.

data of these wind farms from 2013-01-01 to 2014-01-01 in 10 minute temporal resolution.

We use only data from 2013 and removed 11 wind farms for which parts of the data were missing. This results in a data set of 172 wind farms with 52561 time steps. The locations of these wind farms are shown as dots in Figure 4.

4.1.2. Denmark

As a second data set we use data of 100 wind farms in Denmark from 2011-01-01 to 2012-01-01 in 15 minute temporal resolution. The resulting data set covers 35036 time steps and Figure 5 shows the respective locations of the wind farms.

This data set is also a subset of the data set that was used by Girard & Allard (2013) to show the spatio-temporal correlation of wind power forecasting errors.

4.2. Data processing

For both data sets, first the power output data were normalised to lie in $[0, 1]$ through dividing by the nominal power of the respective wind farm. Clearly, capacity changes of wind farms should be taken into account for this normalization. Furthermore, following Pinson (2012) the data were logit transformed so that they can approximately be modelled as autoregressive processes with

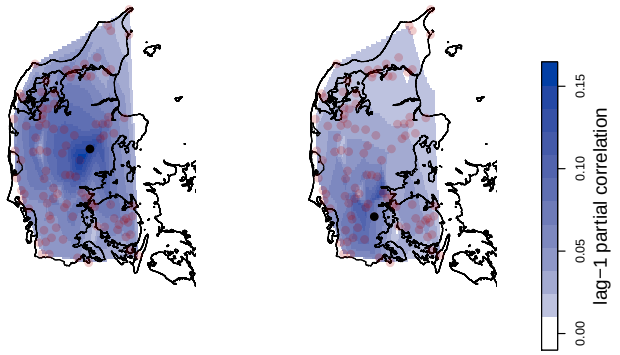

Figure 5: Same as Figure 4 but for Denmark

Gaussian noise. For normalised original data x this logit transformation is

$$y = \log\left(\frac{x}{1-x}\right) \tag{19}$$

where log() is the natural logarithm. x values of 0 and 1 are set to 0.01 and 0.99 respectively. Pinson (2012) used a generalised version of this logit transformation with an additional prior power transformation of x. However, for simplicity we omit this generalisation here.

To facilitate the interpretation of the coefficient matrices (see Section 4.4), both data sets with Q (number of sites) columns and N (number of time steps) rows were sorted with respect to the longitude so that the first columns contain data from the most western wind farms and the last columns corresponds to the most eastern wind farms. Thus, close stations are also close in the data matrix.

4.3. Data analysis

Figure 4 and 5 show the partial correlation of lag-1 data from other sites, i.e., the correlation between the forecast site (black dot) and lag-1 data from the other sites where the linear dependency on lag-1 data at the site itself has been removed. Clear positive partial correlations with close sites, especially in Denmark, indicate that these close sites can provide useful predictive information. Furthermore, sites to the west of the regarded location have higher partial correlations which can be associated with the prevailing westerly wind directions in these regions.

4.4. Results

This section applies and tests adaptive lasso VAR on the data sets described above. To evaluate the predictive performance of these approaches we use different benchmark models. As the simplest benchmark approach we use persistence (i.e., the forecasts are equal to the last available observations), which for the considered look-ahead times (10 minutes to 1 hour) is known to already provide good predictions (Giebel & Kariniotakis, 2017).

To investigate the advantage of spatio-temporal modelling, we also compare adaptive univariate auto regressive (AR) models (e.g. Pinson et al., 2008), which we apply individually to each of the wind farms, i.e., models of the form (1) that were fitted by minimising a similar loss function as (5) but with $\lambda = 0$.

Finally, we also compare non-adaptive lasso VAR (batch lasso VAR) models to test the benefits of the proposed adaptive fitting. These models are of the form (2) and (5) but with $\nu = 1$ and are estimated only on the first 20000 time steps without any later updates. The optimum penalisation parameters λ for these batch lasso VAR models were found in a 10-fold cross validation. Therefore the training time series were split into 10 approximately equally sized parts and predictions for each of these parts were derived from batch lasso VAR models trained only on the 9 remaining parts. The optimum penalization parameter was then chosen from the model with the minimum mean squared error of these predictions.

All the results that are shown in the following are derived on the data excluding the first $N_{tr} = 20000$ (10000 to find optimum number of lags and forgetting factor) time steps in order to have independent forecasts for batch-VAR and a sufficient burn-in period for the adaptive models to provide meaningful predictions. As performance measure we use the bias

$$bias = \frac{1}{N - N_{tr}} \sum_{n=N_{tr}+1}^{N} \mathbf{y}_n - \hat{\mathbf{y}}_n \tag{20}$$

the root mean squared error ($RMSE$)

$$RMSE = \sqrt{\frac{1}{N - N_{tr}} \sum_{n=N_{tr}+1}^{N} ||\mathbf{y}_n - \hat{\mathbf{y}}_n||_2^2} \tag{21}$$

and the mean absolute error (MAE)

$$MAE = \frac{1}{N - N_{tr}} \sum_{n=N_{tr}+1}^{N} ||\mathbf{y}_n - \hat{\mathbf{y}}_n||_1 \tag{22}$$

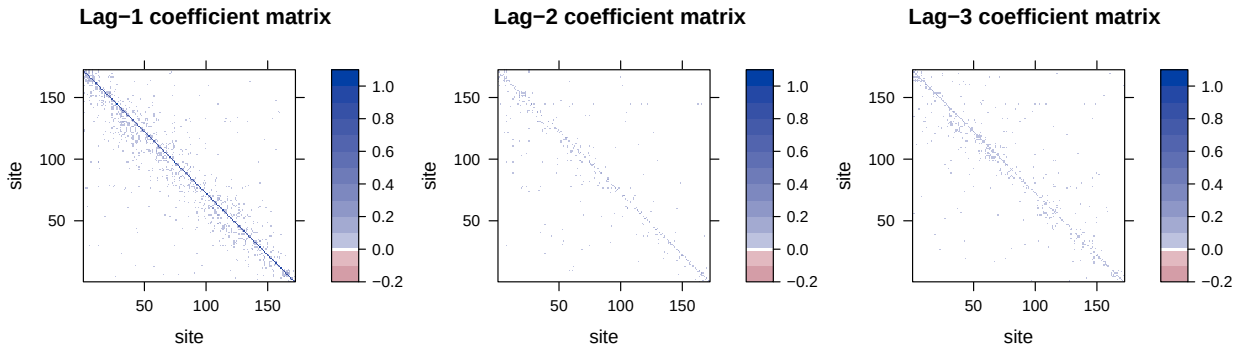

Figure 6: Estimated adaptive lasso-VAR lag 1 to 3 coefficient matrices of 1-step ahead forecasts for the France data. Estimates are taken from the last time step in the data set

For better comparison, we mainly regard skill scores of the *RMSE* and *MAE*, which show the improvement over persistence

$$1 - \frac{S}{S_{pers}} \tag{23}$$

where S is the respective forecast performance measure (*RMSE* or *MAE*) and S_{pers} is the performance measure of persistence.

4.4.1. France

This subsection tests the adaptive lasso VAR on the France data, results for the Denmark data are presented in the next subsection. Before regarding the predictive performance, Figure 6 shows estimates of the coefficient matrices from a lag-3 adaptive lasso VAR for one step ahead. The coefficient matrices are quite sparse and clearly, the lagged data from the forecast sites themselves are the most important so that the coefficients on the diagonal have the highest values. Furthermore, there are also some non-zero coefficients close to the diagonal, which correspond to spatio-temporal correlations between close wind farms. It is also interesting to see that there are more non-zero coefficients below the diagonal than above. Because the data were sorted from west to east, these coefficients correspond to sites west of the forecast site, which are also expected to have more valuable information in the prevalent westerly wind conditions. This is also consistent with the analysis in Figure 4. Figure 6 shows the coefficient matrices from the last time step in the data set. Clearly, for the adaptive lasso VAR these estimates are not constant and change slightly with every adaptive online update in every time step. However, the general patterns of the matrices

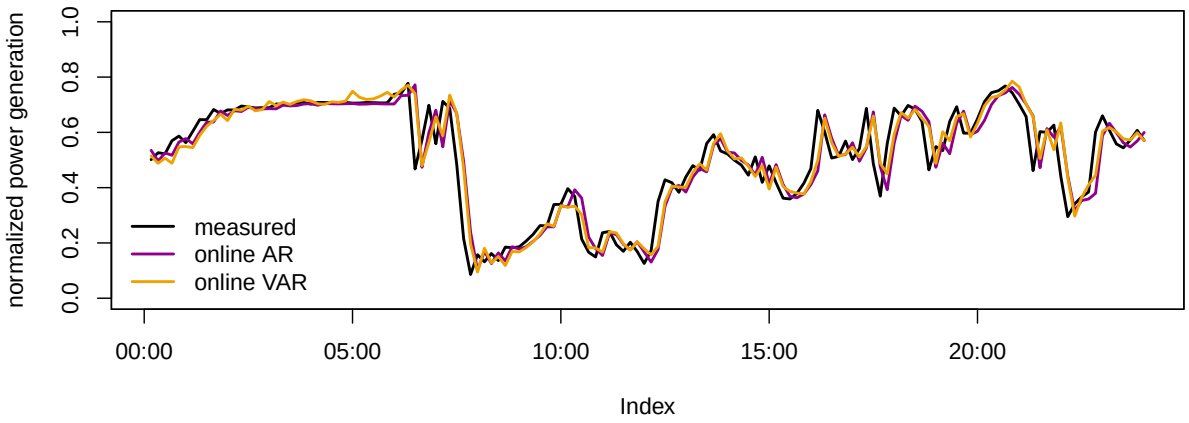

Figure 7: 1-step ahead forecasts and measurements for Turbine 1 from 2013-12-31 00:10:00 to 2014-01-01 00:00:00

look very similar for other time steps within the data period (not shown).

Figure 7 shows a one-day example time series of one turbine and the corresponding 1-step ahead forecasts from adaptive lasso VAR (online VAR) and adaptive univariate AR (online AR). The forecast models follow the general signal of the measurements but especially significant changes clearly lag one time step. The differences between the uni- and multivariate models are only small, which is not surprising since the off diagonal coefficients in the multivariate model are only small (first rows of matrices in Figure 6). Still, there are differences between the forecasts and in the following their impact on the predictive performance will be investigated.

Figure 8 shows the bias of the adaptive lasso VAR (online VAR) for different look ahead times and compares it to adaptive univariate AR (online AR) and batch lasso VAR (batch VAR). All forecast methods have an average bias of zero, however, batch VAR exhibits relatively high biases for some of the wind farms (larger spread). This indicates that the dynamics of these wind farms experience systematic changes, which can not be followed by the non-adaptive batch VAR.

Figure 9 shows similar plots for the RMSE and MAE skill scores. The RMSE skill scores (top row) show clearly positive skill (i.e., better predictive performance than persistence) for all forecast methods at all lead times. The skill generally increases with the look-ahead time which mainly results from the decreasing performance of persistence. When comparing the different forecast models with each other, the univariate AR is clearly worst, which indicates that for this data

set, it is of clear advantage to exploit spatio-temporal information. Furthermore, the adaptive lasso VAR performs better than the batch lasso VAR, which confirms that the time series are not completely stationary so that adaptiveness in the lasso VAR models is clearly beneficial. Paired t-tests confirmed the significance of the RMSE differences between the different forecast models at a 0.05 level and a closer analysis showed that online VAR provides the best forecasts for almost all wind farms (only at one wind farm at 10-minute ahead batch VAR provides smaller RMSE).

The MAE (bottom row) show similar differences between the forecast methods and online lasso VAR clearly performs best of the 3 tested methods (paired t-test significant at 0.05 level). However, all methods perform worse compared to persistence and the univariate AR and batch lasso VAR have mainly negative skill scores. The MAE of the adaptive lasso VAR are similar to that of persistence for 10- and 20-minute ahead forecasts but is slightly better for 30- and 40-minute ahead predictions (i.e., positive MAE skill score). The good RMSE but rather poor MAE skill scores may be explained by the fact that all models optimize the squared rather than the absolute error.

For the results that are shown in Figure 6 and 9 we used a forgetting factor of 0.9998, which corresponds to an effective training data length of 5000, and 3 lags. This selection is based on Figure 10 that shows that these are the optimum parameters for 1-step ahead forecasts when regarding the time steps 10001 to 20000. Additionally, paired t-tests confirmed the best performance of these selected parameters at a 0.05 level. Clearly, different values for these parameters could be better for other look-ahead times. Furthermore, Figure 10 also does not account for likely dependencies between these two parameters. However, for simplicity and because a complete grid optimisation of the two parameters would computationally be very expensive we take these parameters for all look-ahead times. Furthermore the batch lasso VAR and the adaptive univariate AR also use this same number of lags and the adaptive univariate AR also the same forgetting factor. The optimum penalisation parameter λ for the adaptive lasso VAR was chosen adaptively, based on the past weighted predictive performance (see Section 2.2) while it was selected by cross validation for the batch lasso VAR.

4.4.2. Denmark

In the following, a similar analysis as above is carried out for the Denmark data. Figure 11 shows that, compared to the France data, a slightly different forgetting factor (0.999 or effective

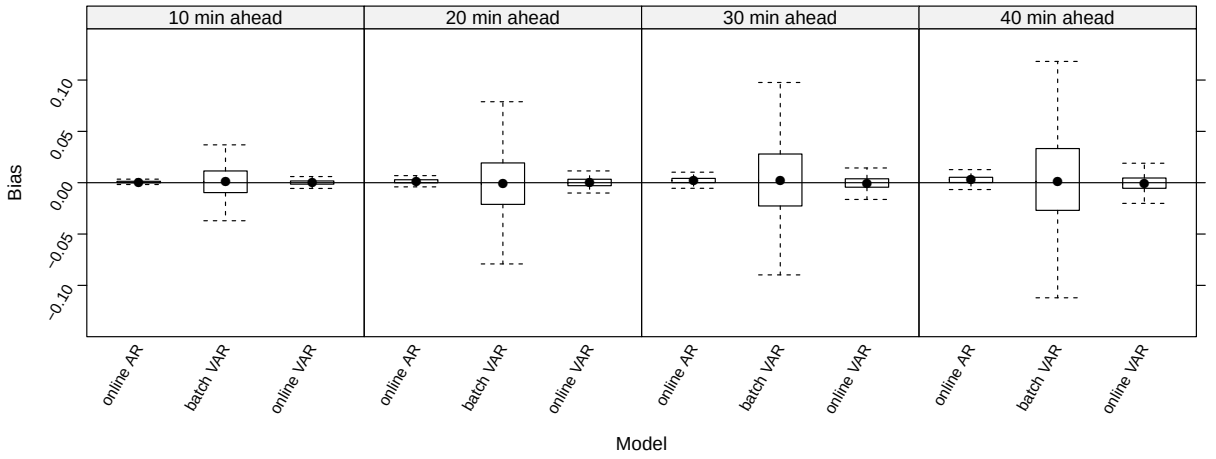

Figure 8: Bias for different lead times for the logit transformed France data. The solid circles mark the medians, the boxes the interquartile range, and the whiskers the most extreme values of the 172 score values for the different sites. The models were fitted with 3 lags, batch VAR was fitted on the first 20000 time steps, and online AR and online VAR use a forgetting factor of 0.9998. The biases are derived for time steps 20001 to 52561.

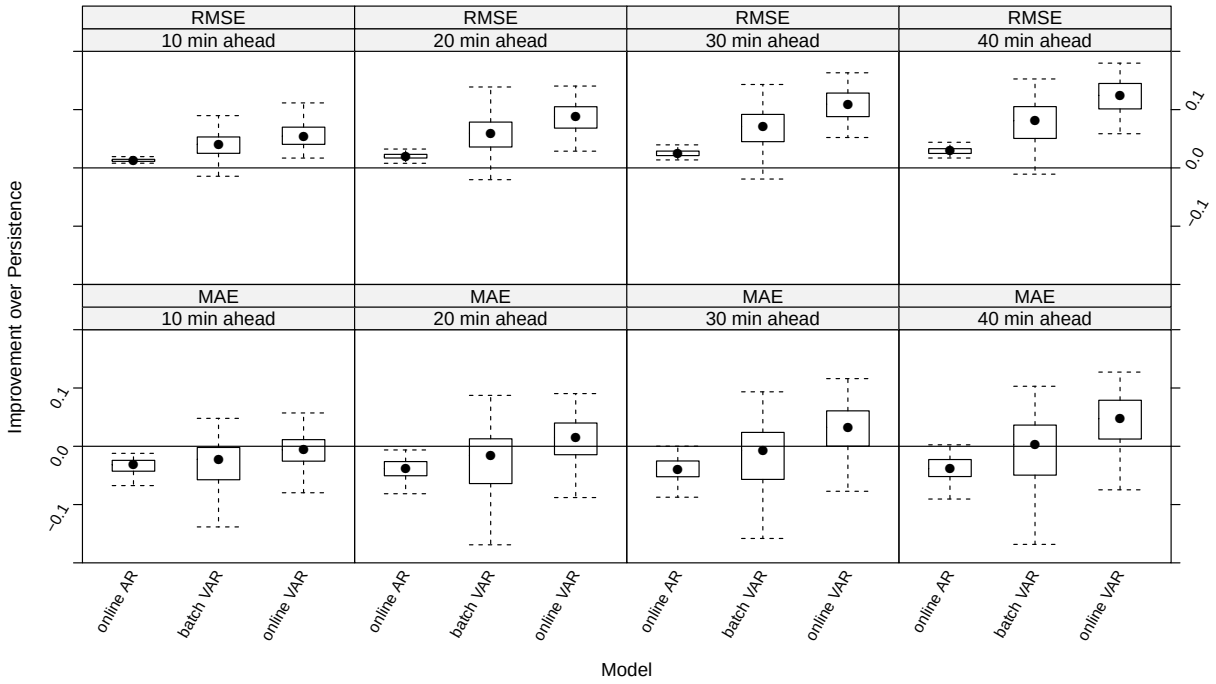

Figure 9: RMSE skill score (bottom row) and MAE skill score (top row) with reference to persistence for different lead times for the France data. The same models are used as in Figure 8.

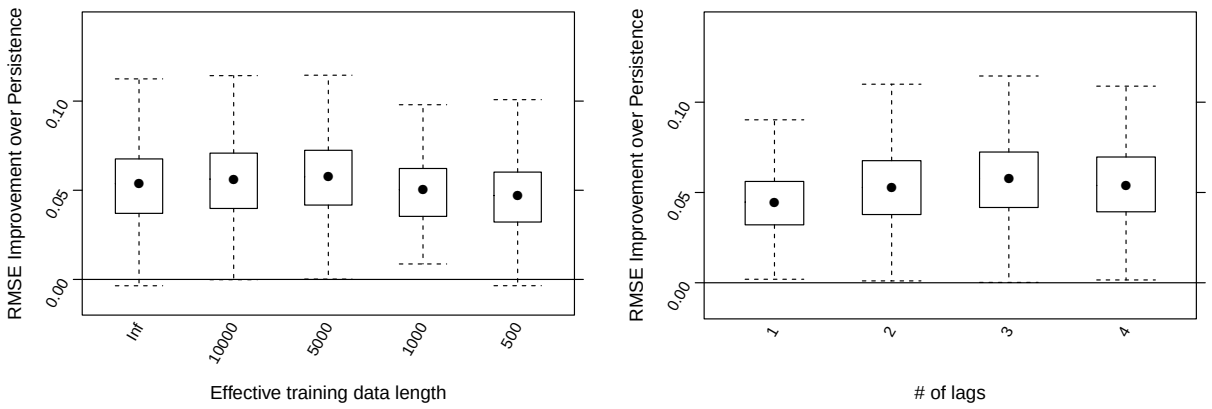

Figure 10: RMSE skill score of adaptive lasso VAR with reference to persistence for different effective training data lengths (forgetting factors; left) and different numbers of considered lags (right) for the France data. The RMSE are derived for time steps 10001 to 20000 and the non-varying parameter is set to its optimum value respectively.

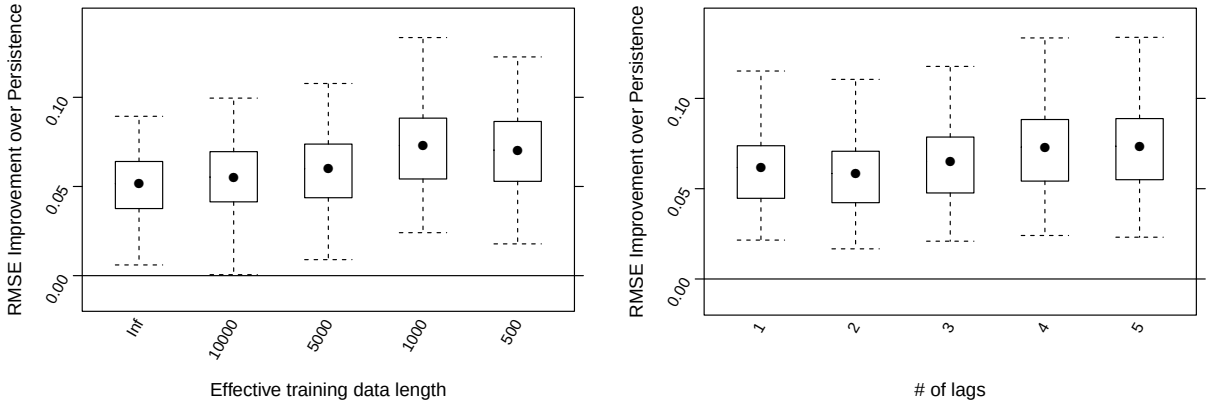

Figure 11: RMSE skill score of adaptive lasso VAR with reference to persistence for different effective training data lengths (forgetting factors; left) and different numbers of considered lags (right) for the Denmark data. The RMSE are derived for time steps 10001 to 20000.

training data length of 1000) and a higher number of lags (4) are optimal.

Therefore, Figure 12 shows a similar figure as Figure 6 but with estimated coefficient matrices for lags 1 to 4. Similar to Figure 6 the diagonal entries for lag-1 have the highest values. Furthermore, most other non-zero entries are close to the diagonal (correlations with close wind farms) and, consistent with Figure 5, the higher correlations with locations to the west (below diagonal entries) are even more pronounced. However, different to Figure 6 these patterns are much less pronounced in the lag-2 to lag-4 matrices and especially for the higher lags there are also a number of negative coefficients (red).

Despite these differences the predictive performance of the different forecasts in Figures 13 and 14 are similar to the France data and adaptive lasso VAR clearly performs best for most lead times and performance measures. Only for the MAE at lead times 15 and 30 minutes the difference between batch VAR and online VAR are not significant in a t-test at a 0.05 level. A closer analysis showed that online VAR performs best at almost all wind farms for longer lead times and in the RMSE. Different to the France data, batch lasso VAR also have mainly positive MAE skill scores which is most probably related to the worse performance of persistence at the longer lead times (15 to 60 minutes versus 10 to 40 minutes).

5. Summary and Conclusion

This paper presents an appealing method for forecasting wind power generation at multiple wind farms, which takes into account sparse spatio-temporal dependence structures and can adapt to changes in the time-series dynamics. Therefore, lasso vector autoregression (VAR) is extended with exponentially decaying weights for past data to allow the model to adapt to changes in the spatio-temporal dependencies. Furthermore a coordinate descent algorithm for very efficient online updates is presented.

On simulation data this approach shows very good tracking abilities of slow continuous changes and reasonable abilities to follow abrupt changes in the time-series parameters. Two case studies with wind power data from western France and Denmark show that the adaptive lasso VAR can clearly improve the predictions compared to *non-adaptive* lasso VAR (similar to Cavalcante et al., 2017) and an adaptive *univariate* auto regressive model (e.g. Pinson et al., 2008). Lasso VAR on a sliding training data window showed also comparable forecast skill on the simulated data but the

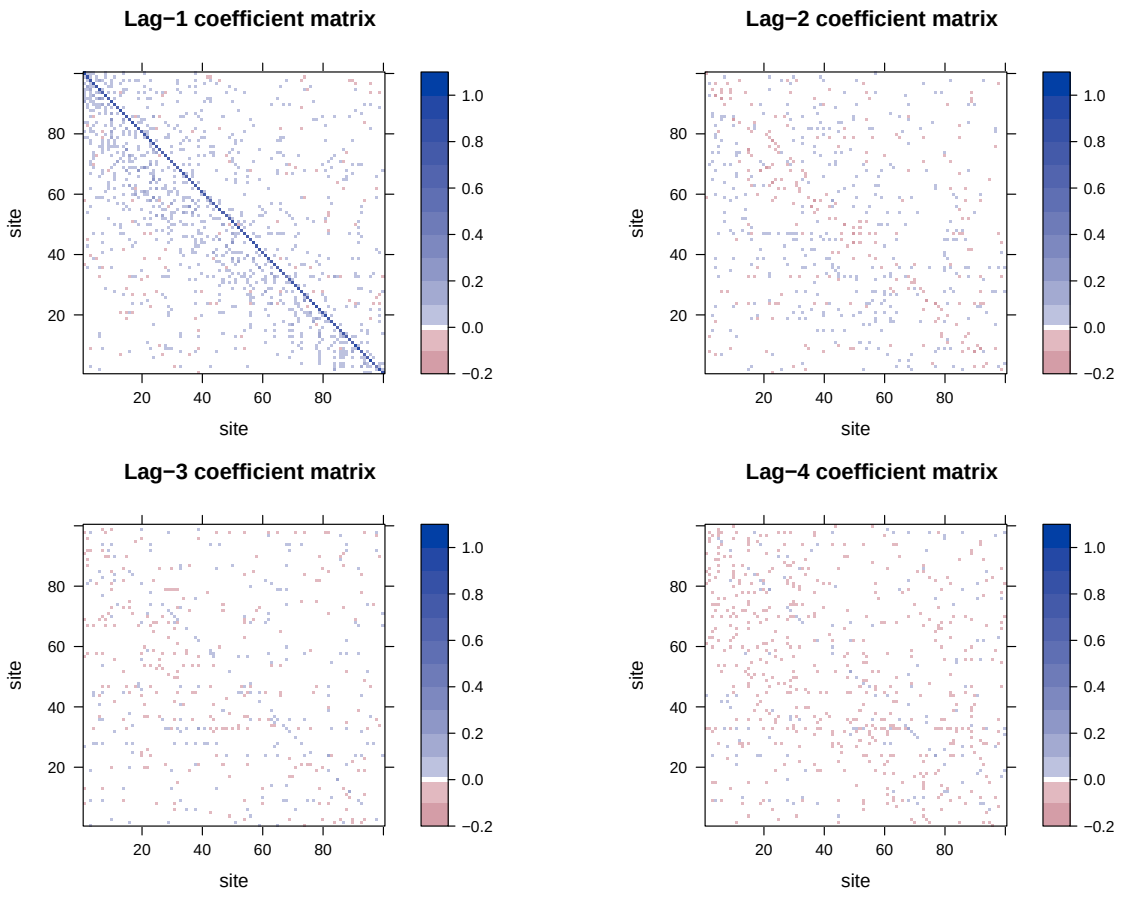

Figure 12: Estimated adaptive lasso-VAR lag 1 to 4 coefficient matrices for one step ahead forecasts for the Denmark data. Estimates are taken from the last time step in the data set

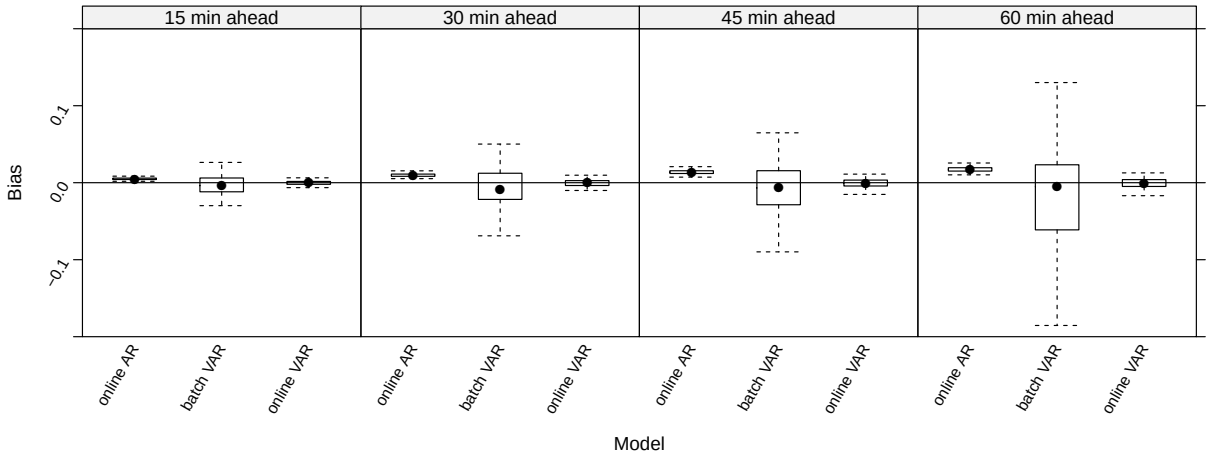

Figure 13: Bias for different lead times for the logit transformed Denmark data. The solid circles mark the medians, the boxes the interquartile range, and the whiskers the most extreme values of the 100 score values for the different sites. The models were fitted with 4 lags, batch VAR was fitted on the first 20000 time steps, and online AR and online VAR use a forgetting factor of 0.999. The biases are derived for time steps 20001 to 35036.

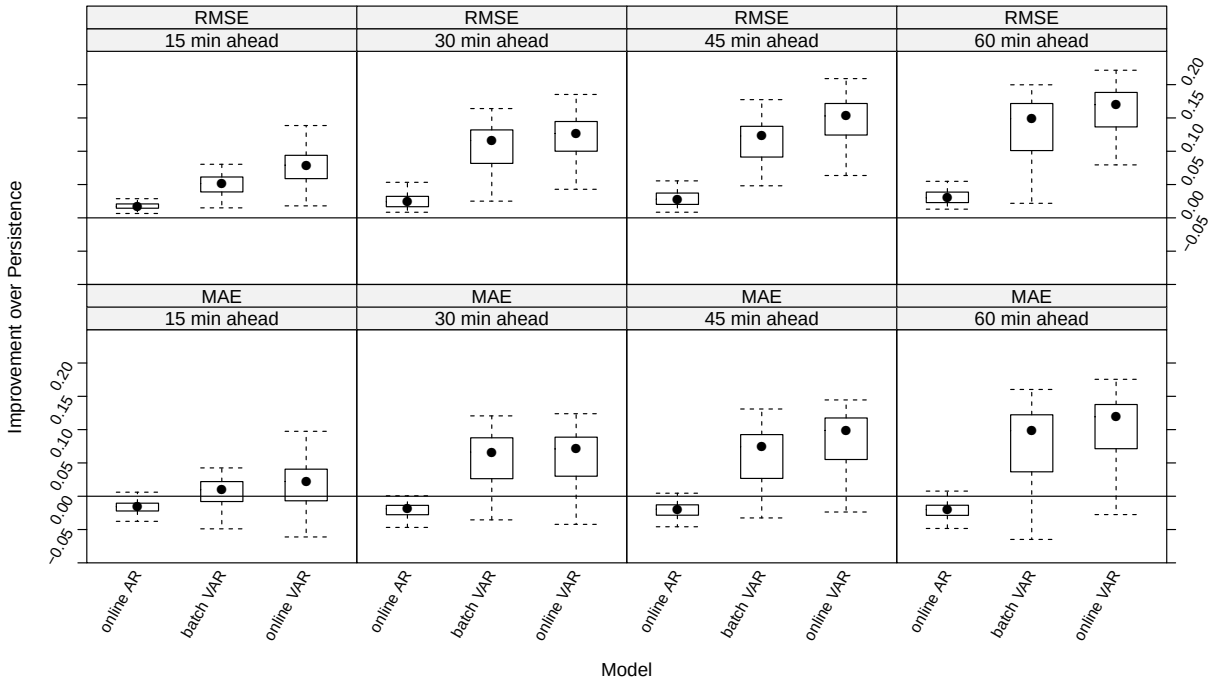

Figure 14: RMSE skill score (bottom row) and MAE skill score (top row) with reference to persistence for different lead times for the Denmark data. The same models are used as in Figure 13.

higher computational costs (approx. factor of 20 compared to adaptive lasso VAR for the simulated data) make it almost infeasible to be used for larger data sets.

For fitting adaptive lasso VAR the penalisation parameter, the number of lags, and the forgetting factor have to be specified. The selection of these parameters can have a great impact on the forecast performance and the optima of these parameters might interact with each other and might not be constant over time. In our study we fit parallel models for a sequence of penalisation parameters and in each time step choose the parameter with the smallest error measure in the preceding time steps. The forgetting factor and the number of lags were set to fixed values chosen based on an independent validation period. A future extension could be a dynamic forgetting factor, similar to the scale parameter tracking in Dowell & Pinson (2016), which could further improve the ability of the model to adapt to abrupt changes.

For higher numbers of wind farms the number of VAR coefficients quickly becomes larger than the available data length so that regularization is crucial to solve this ill-posed problem. Alternative to penalizing the absolute coefficients (lasso), it is also common to penalize the squared coefficients (ridge regression; Hoerl & Kennard, 1970). However, compared to lasso, ridge regression does not provide sparse coefficient matrices, which are usually easier to estimate and interpret. Nevertheless, similar ideas as presented in this paper could also be used for adaptive VAR with ridge regression or elastic net regularization, which combines lasso and ridge regression (Zou & Hastie, 2005). Similarly, also different sparse structures such as those proposed in Cavalcante et al. (2017) could be imposed by modifying the penalization term.

This paper only regarded deterministic multivariate predictions. However, similar to Dowell & Pinson (2016) probabilistic forecasts could easily be generated by interpreting these predictions as conditional means of Gaussian distributions.

We found that for our data sets the optimum forgetting factor is too large to track diurnal trends and changes at time scales in the order of days. Diurnal trends are not very pronounced in our data but changes at time scales in the order of days occur frequently as a result of weather regime changes. Especially changes in the wind direction are supposed to clearly effect the spatio-temporal correlation pattern. Therefore, future work should investigate adaptive lasso VAR extensions with wind direction, the diurnal cycle, or other additional covariables. Additionally, regime or Markov switching (Pinson et al., 2008; Pinson & Madsen, 2012) could be an attractive extension to our

proposed approach.

Compared to 22 wind farms in Dowell & Pinson (2016) and 66 in Cavalcante et al. (2017) we used relatively big data sets with 100 and 172 wind farms. Running adaptive lasso VAR over one year with 172 wind farms and 52561 time steps takes approx. 1.5 hours on a Notebook with a 4×2.3GHz Intel i5 processor and 12 GB memory. Thus even bigger data sets are still feasible and in an operational setting the coefficients can be updated in less than a second each time new data becomes available. To facilitate applications and extensions we implemented the adaptive lasso VAR algorithm in the software package onlineVAR (Anonymous, 2017) for the open source software R (R Core Team, 2017). This package also includes a data set of 22 wind farms in south eastern Australia (see also Dowell & Pinson, 2016), which is smaller but similar to the data sets used in the presented case studies.

References

Angelosante, D., Bazerque, J. A., & Giannakis, G. B. (2010). Online adaptive estimation of sparse signals: Where rls meets the ℓ_1 -norm. *IEEE Transactions on Signal Processing*, *58*, 3436–3447. doi:10.1109/TSP.2010.2046897.

Anonymous, A. (2017). *onlineVAR: Online Fitting of Time-Adaptive Lasso Vector Auto Regression*. To hide our identities the package has not been uploaded to CRAN yet.

Ben Taieb, S., Bontempi, G., Atiya, A. F., & Sorjamaa, A. (2012). A review and comparison of strategies for multi-step ahead time series forecasting based on the nn5 forecasting competition. *Expert Systems with Applications*, *39*, 7067–7083. doi:10.1016/j.eswa.2012.01.039.

Bessa, R. J., Möhrlen, C., Fundel, V., Siefert, M., Browell, J., Haglund El Gaidi, S., Hodge, B.-M., Cali, U., & Kariniotakis, G. (2017). Towards improved understanding of the applicability of uncertainty forecasts in the electric power industry. *Energies*, *10*.

Cavalcante, L., Bessa, R. J., Reis, M., & Browell, J. (2017). Lasso vector autoregression structures for very short-term wind power forecasting. *Wind Energy*, *20*, 657–675. doi:10.1002/we.2029.

Davis, R. A., Zang, P., & Zheng, T. (2016). Sparse vector autoregressive modeling. *Journal of Computational and Graphical Statistics*, *25*, 1077–1096. doi:10.1080/10618600.2015.1092978.

Dowell, J., & Pinson, P. (2016). Very-short-term probabilistic wind power forecasts by sparse vector autoregression. *IEEE Transactions on Smart Grid*, *7*, 763–770. doi:10.1109/TSG.2015.2424078.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, *1*, 302–332.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*. doi:10.18637/jss.v033.i01.

Giebel, G., & Kariniotakis, G. (2017). Wind power forecasting – a review of the state of the art. In G. Kariniotakis (Ed.), *Renewable Energy Forecasting* chapter 3. (pp. 59–109). Elsevier.

Girard, R., & Allard, D. (2013). Spatio-temporal propagation of wind power prediction errors. *Wind Energy*, *16*, 999–1012. doi:10.1002/we.1527.

Gneiting, T., Larson, K., Westrick, K., Genton, M. G., & Aldrich, E. (2006). Calibrated probabilistic forecasting at the stateline wind energy center. *Journal of the American Statistical Association*, *101*, 968–979. doi:10.1198/016214506000000456.

He, M., Vittal, V., & Zhang, J. (2015). A sparsified vector autoregressive model for short-term wind farm power forecasting. In *2015 IEEE Power Energy Society General Meeting* (pp. 1–5). doi:10.1109/PESGM.2015.7285972.

Hering, A. S., & Genton, M. G. (2010). Powering up with space-time wind forecasting. *Journal of the American Statistical Association*, *105*, 92–104. doi:10.1198/jasa.2009.ap08117.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*, 55–67. doi:10.1080/00401706.1970.10488634.

Kou, P., Gao, F., & Guan, X. (2013). Sparse online warped Gaussian process for wind power probabilistic forecasting. *Applied Energy*, *108*, 410 – 428. doi:10.1016/j.apenergy.2013.03.038.

Møller, J. K., Nielsen, H. A., & Madsen, H. (2008). Time-adaptive quantile regression. *Computational Statistics & Data Analysis*, *52*, 1292 – 1303. doi:http://dx.doi.org/10.1016/j.csda.2007.06.027.

Pinson, P. (2012). Very-short-term probabilistic forecasting of wind power with generalized logit-normal distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *61*, 555–576. doi:10.1111/j.1467-9876.2011.01026.x.

Pinson, P., Christensen, L., Madsen, H., Sørensen, P., Donovan, M., & Jensen, L. (2008). Regime-switching modelling of the fluctuations of offshore wind generation. *Journal of Wind Engineering and Industrial Aerodynamics*, *96*, 2327–2347. doi:10.1016/j.jweia.2008.03.010.

Pinson, P., & Madsen, H. (2009). Ensemble-based probabilistic forecasting at horns rev. *Wind Energy*, *12*, 137–155. doi:10.1002/we.309.

Pinson, P., & Madsen, H. (2012). Adaptive modelling and forecasting of offshore wind power fluctuations with markov-switching autoregressive models. *Journal of Forecasting*, *31*, 281–313. doi:10.1002/for.1194.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria. URL: `http://www.R-project.org/`.

Tastu, J., Pinson, P., Kotwa, E., Madsen, H., & Nielsen, H. A. (2011). Spatio-temporal analysis and modeling of short-term wind power forecast errors. *Wind Energy*, *14*, 43–60. doi:10.1002/we.401.

Tastu, J., Pinson, P., Trombe, P.-J., & Madsen, H. (2014). Probabilistic forecasts of wind power generation accounting for geographically dispersed information. *IEEE Transactions on Smart Grid*, *5*, 480–489. doi:10.1109/TSG.2013.2277585.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.

Wytock, M., & Kolter, J. (2013). Large-scale probabilistic forecasting in energy systems using sparse Gaussian conditional random fields. In *Proceedings of the IEEE 52nd Annual Conference on Decision and Control (CDC)* (pp. 1019–1024).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical*

Society: Series B (Statistical Methodology), *67*, 301–320. doi:10.1111/j.1467-9868.2005.00503.x.