Detection of False Data Injection Attacks in Smart Grids Based on Forecasts

Michael G. Kallitsis [*], Shrijita Bhattacharya [†], George Michailidis [‡]
[*] Merit Network, Inc, University of Michigan, Ann Arbor, Michigan
[†] Department of Statistics, University of Michigan, Ann Arbor
[‡] Department of Statistics, University of Florida, Gainesville

***Abstract*—The bi-directional communication capabilities that emerged into the smart power grid play a critical role in the grid's secure, reliable and efficient operation. Nevertheless, the data communication functionalities introduced to Advanced Metering Infrastructure (AMI) nodes end the grid's isolation, and expose the network into an array of cyber-security threats that jeopardize the grid's stability and availability. For instance, malware amenable to inject *false data* into the AMI can compromise the grid's state estimation process and lead to catastrophic power outages. In this paper, we explore several statistical spatio-temporal models for efficient diagnosis of *false data injection attacks* in smart grids. The proposed methods leverage the data co-linearities that naturally arise in the AMI measurements of the electric network to provide forecasts for the network's AMI observations, aiming to quickly detect the presence of "bad data". We evaluate the proposed approaches with data tampered with stealth attacks compiled via *three different attack strategies*. Further, we juxtapose them against two other forecasting-aided detection methods appearing in the literature, and discuss the trade-offs of all techniques when employed on real-world power grid data, obtained from a large university campus.**

I. Introduction

Modernizing the aging electric network with advanced metering infrastructure constitutes a fundamental milestone posited by utility companies. Over the past few years, millions of "smart" meters have been deployed [1]; by 2020, an estimated 90 million devices would be installed in the U.S. alone [2]. These next-generation AMI meters are equipped with low-latency two-way communication functionalities that enable advanced meter diagnostics, accurate accounting and billing, rapid troubleshooting for outage remediation, and, most importantly, network observability and controllability [3], [4]. A prime motivation for the latter is *demand response* [5], [6], which is already in place by several utilities[1]. Via accurate state-estimation and load forecasting, and by employing the feedback-loop that next-generation smart meters allow, demand response mechanisms can sustain the grid's reliable and energy-efficient operation.

Integrity of the reported data is, thus, critical for the grid's secure operation. Erroneous data could compromise the grid's state estimation process and this might lead to brown-outs or black-outs [4], [7]–[9]. Nefarious actors interested in disrupting the network's smooth operation could simply launch the so-termed *false data injection (FDI) attacks* [10]–[13] in a coordinated manner to introduce network instabilities. Several attack strategies have recently been proposed [10], [12], [14], [15] that can be employed by adversaries to alter the measurements of critical AMI observations in order to ultimately misinform the grid's state estimation process. These strategies have gained considerable attention from the power grid community since the attacks proposed can be judiciously constructed and remain *undetected* by state-of-the-art state estimation processes. Albeit the fact that some attack strategies impose strong assumptions on the information available to attackers (such as knowledge of the grid's topology and network characteristics like transmission line impedances, etc.), it is critical to have data-driven algorithms that can detect these attacks in a timely and accurate manner.

In this paper, we propose a series of statistical models and corresponding estimation algorithms aiming towards rapid detection of FDI attacks. We leverage the spatial (i.e., across meters) and temporal correlation in data to obtain *short-term forecasts* for the data observations provided by the AMI meters (e.g., the power injections at the system buses and the power flows across transmission lines). Large deviations between the forecasted values and their actual readings are treated as indications of spurious data, and an alert is raised when these abnormalities persist over time. We tame the false alert rate using a sequential hypothesis testing framework based on exponentially-weighted moving average control charts.

Our contributions are twofold: (1) We propose an array of state-space techniques for rapid identification of FDI anomalies, including univariate and (several variants of) multivariate *autoregressive* models as well as a *dynamic factor model* that is shown to be particularly appealing as the data dimension grows. (2) We demonstrate, via realistic IEEE network topologies (available via MATPOWER [16]) and real-world data that exhibit non-stationarities, trends and diurnal patterns, that our algorithms can successfully detect malicious attack vectors obtained from three different attack strategies that would otherwise remain undetected by the classical state estimation techniques [10], [17]. Moreover, we juxtapose our methods against related projection-based algorithms [18], [19] and showcase the performance of all methods under a large spectrum of scenarios, including seemingly innocuous attacks.

II. Related Work

Data attacks have been studied significantly over the past few years [19]–[27]. In [20], we employ a "network kriging"

[1] Examples include, but are not limited to, the Pacific Northwest DR project (https://www.nwcouncil.org/energy/dr) and the New England DR initiative (http://nedri.raabassociates.org).

978-1-5386-7954-8/18/$31.00 ©2018 IEEE

model to detect attacks on AMI meters based on observations from a subset of "trusted" nodes. In [21], the problem of detecting aberrant behavior of residential smart meters is tackled from the *home-area network* perspective. [22] presents a mixed-integer programming method that determines the smallest subset of measurements that need to be protected to render FDI attacks ineffective. [23] studies a graph theoretic method for securing an optimal set of meter measurements so that state estimation is not compromised. A linear measurement model (i.e., a vector autoregressive one) is derived in [19] to handle both SCADA and PMU measurements.

In [24], detection of FDI attacks is formulated as a matrix-factorization problem. The work in [25] establishes a sequential hypothesis testing framework based on the likelihood ratio that is amenable to distributed realization. A multilayer neural-network is studied in [26], trained as a binary classifier to identify the presence of FDI attacks. In [27], the authors propose distributed attack detection mechanisms along with a hypothesis testing criterion based on the *quickest detection* framework. In [28], a comparative study of three *supervised learning* techniques for detecting FDI attacks is performed, namely support vector machines, k-nearest neighbor and extended nearest neighbor.

Detection based on *state forecasts* has also been considered in [18], [19]. We examine both methods in section VI. These state-space approaches consider the dynamic nature of the system's state evolution, and generate state forecasts and Kalman-based state filtering to infer system anomalies. Both methods employ techniques to describe the dynamics of the system state (e.g., the nodal voltages and angles); the authors in [18] use Holt's exponential smoothing, whereas in [19] an autoregressive model is fit based on historical data on the system's state.

III. State Estimation

Utility operators employ remote sensors and meters to receive fine-grained measurements such as power injections (loads or generators) on buses and power flows on branches. The measurements are utilized to estimate the state variables of the system, including phase angles and bus voltages. The relationship between the observed and system variables is non-linear, and a common approximation used in practice to simplify the analysis is the *DC model approximation*. The DC state estimator relates *measurements* to *system state* variables as follows [17],

$$z = Hx + e, \tag{1}$$

where z is a vector of m observations (known), x is an n vector of state variables (unknown), and H is the $m \times n$ Jacobian matrix (which is a function of the network's topology and line admittances). The error term e represents measurement noise.

The model corresponds to an overdetermined system of linear equations[2] and can be solved as a weighted least-squares problem [17]. Its solution provides the state estimator $\hat{x}$, namely $\hat{x} = (H^\top W H)^{-1} H^\top W z$, where W is a diagonal matrix with m entries the reciprocals of the variances of the measurement errors captured by e.

For *bad data processing*, one can consider the residuals defined as $r = z - Hx$, and calculate the estimate $\hat{r} = z - H\hat{x}$ [17]. The statistic[3] $J(\hat{x}) := \|z - H\hat{x}\|^2$ ($\|\cdot\|$ represents the Euclidean norm) follows a chi-squared distribution with $m - n$ degrees of freedom, and can be harnessed to detect bad data. Specifically, an alarm is triggered when $J(\hat{x}) \geq \tau(u)^2$, where $\tau(u)^2$ indicates the critical value at a user-defined confidence level u that controls the false alarm rate.

IV. Threat Model

The work of Liu *et al.* [10] drew attention onto a family of "data attacks" that can circumvent the bad data detection criterion discussed above. They coined the term "false data injection attacks" to denote attack vectors that can be injected into the measurement system so that the corrupted measurements would yield a manipulated and false state estimate. The attack vectors are not simply random perturbations of the meter observations, but are rather carefully crafted in an effort to be stealthy and remain undetected by bad data processing techniques, such as the one in section III. In this paper, our *threat model* consists of data attacks inspired by [10]. We consider the following attack strategies.

A. Random FDI attacks [10]

We assume an attacker that has *limited access to meters*. Let the accessible set be $\mathcal{I}_m = \{i_1, \ldots, i_m\}$. The attacker aims to find an attack vector $a = (a_1, \ldots, a_m)^\top$ such that $a_i = 0$ for $i \notin \mathcal{I}_m$ and a is a linear combination of the columns of H. It can be easily shown that when the bad measurement vector $z_a = z + a$ is utilized to get the state estimate, the falsified state estimate $\hat{x}_{\text{bad}}$ will yield a residual $z_a - H\hat{x}_{\text{bad}}$ that lies below the detection threshold $\tau(u)^2$. This attack requires knowledge of the Jacobian matrix H (which might be non-trivial for an attacker to obtain), but the vector a can be easily constructed via column transformations of the matrix H (see [10], Eq. 7).

B. Minimal FDI attacks [14]

Sandberg *et al.* [14] introduce two security indices that quantify the least effort required to achieve stealthy attacks while remaining below the detection radar. They formulate two optimization problems whose solutions provide *sparse* and *small magnitude* attacks. We focus on the small magnitude ("minimal") attack strategy, since the methodology in [10] is essentially providing sparse attack vectors as well. To construct the "minimal" attack vector, one needs to solve the following convex optimization problem:

$$\begin{aligned} \beta_k &:= \min_c \|Hc\|_1 \\ &\text{subject to } 1 = \sum_i H_{ki} c. \end{aligned} \tag{2}$$

Notably, this attack strategy focuses on a specific meter k, and the solution c^* can be rescaled such that $a^* = \alpha_k H c^*$

[2] Equivalently, this can be considered as linear regression [10], [29].

[3] In practice, $J(\hat{x})$ is calculated using the measurement error variances as scaling factors (see [17], p.219).

and the measurement attack $z_a = z + a^*$ attains the minimal amount of power $\|a^*\|_1$.

C. PCA-based Blind FDI attacks [12]

The above-mentioned attack strategies require knowledge of the Jacobian matrix H, which is usually not readily available to adversaries. The authors in [12] study a new attack which is completely data-driven[4]. In particular, the authors compose attack vectors based solely on the measurements captured by the m-vector z. In particular, the "blind" attack vector is defined as, $a_{\text{PCA}} = H_{\text{PCA}} c$, where H_{PCA} is a matrix whose columns are the first n eigenvectors that correspond to the largest n eigenvalues obtained via principal component analysis (PCA) of the sample covariance matrix $\Sigma_z = \frac{1}{T} Z^\top Z$. Z is the (centered) $T \times m$ measurement matrix, and $c := (c_1, \ldots, c_n)$ is a non-zero random vector with $c_i \sim N(0, \sigma_i^2)$. The attack magnitude can be tuned via the σ_i knob (see section VI).

V. Proposed Detection Methodology

This section introduces the proposed techniques. We start by introducing univariate and multivariate *autoregressive (AR) models*. The univariate AR(p) models are network "agnostic" (i.e., the predictions take no spatial considerations, but rather borrow strength solely from temporal correlations), but are extremely robust and fast when it comes to their parameter estimation since fewer unknown parameters are required to be learned. They further provide a solid baseline for the performance of the vector autoregressive (VAR) models.

The proposed VAR models are natural extensions of the AR models, but have stronger predictive power since they capture both spatial and temporal correlations in the data. However, as the dimensionality of the problem grows (viz., the number of measurements / features m), VAR training becomes cumbersome and problematic since multiple parameters need to be estimated (see section VI). For example, with a VAR model of order p, there are pm^2 parameters to be learned. Consistent estimation of model parameters requires a large number of data points, which might not be a feasible requirement in large electric networks (e.g., due to memory / storage constraints). We try to alleviate this by examining also *regularized* VAR models [30] (i.e., VAR models with penalties that constrain the number of unknown parameters). Furthermore, we introduce *dynamic factor models* (DFM) [31], [32] that are better tailored for large problems when the amount of variation in the data can be explained by only a few common factors (indeed, the measurement matrix of the real-world data we study is low-ranked). To the best of our knowledge, dynamic factor analysis techniques have not been explored before for modeling electricity data.

A. Vector Autoregression

VAR(p) models have been well-studied and are known to have good properties for forecasting power consumption data and other SCADA states [33]. An m-dimensional[5], zero-mean, stationary process z_t modeled as VAR(p) is given by $z_t - \phi_1 z_{t-1} - \cdots - \phi_p z_{t-p} = w_t$, where ϕ_i are $m \times m$ transition matrices, and (w_t) is a vector white noise process with zero mean and covariance Σ.

Given the data $\{z_0, \ldots, z_T\}$, the estimation problem can be casted to a linear regression one [30]:

$$\underbrace{\begin{pmatrix} z_T^\top \\ \vdots \\ z_p^\top \end{pmatrix}}_{\mathcal{Y}} = \underbrace{\begin{pmatrix} z_{T-1}^\top & \cdots & z_{T-p}^\top \\ \vdots & \ddots & \vdots \\ z_{p-1}^\top & \cdots & z_0^\top \end{pmatrix}}_{\mathcal{X}} \underbrace{\begin{pmatrix} \phi_1^\top \\ \vdots \\ \phi_p^\top \end{pmatrix}}_{B^*} + \underbrace{\begin{pmatrix} w_T^\top \\ \vdots \\ w_p^\top \end{pmatrix}}_{E}$$

$$\begin{aligned} \text{vec}(\mathcal{Y}) &= \text{vec}(\mathcal{X} B^*) + \text{vec}(E) \\ &= (I \otimes \mathcal{X}) \text{vec}(B^*) + \text{vec}(E) \\ Y &= X\beta^* + \text{vec}(E), \end{aligned} \tag{3}$$

where Y is $Nm \times 1$, X is a $Nm \times q$ matrix, β^* is a q-dimensional vector and $\text{vec}(E)$ a $Nm \times 1$ vector, with $N = T - p + 1$ (the data sample size) and $q = pm^2$ (the number of coefficients). Estimates for the transition matrices $\phi_1, \ldots, \phi_p$ can be obtained by the least-squares solution of the formulated regression problem [34]. However, to impose sparsity constraints on the vector of coefficients β^*, the following penalized least-squares problem can be exploited:

$$\underset{\beta \in \mathbb{R}^q}{\arg\min} \frac{1}{N} \|Y - X\beta\|^2 + \lambda_N \|\beta\|_2 ,$$

The $\ell 2$-norm constraint reduces the problem to *ridge regression*. In our experiments, we utilized `R` solvers from the `sparsevar` library to obtain our VAR "ridge" estimates. The Lagrange multiplier λ_N was chosen through cross-validation. We also employed the `MTS` library for non-penalized VAR modeling. We tuned our `MTS` solver to yield simplified/sparse transition matrices by retaining only coefficients that are statistically significant (see [34], Chapter 2).

Anomaly Detection: Given the estimated coefficients, we can proceed with forecasting-aided FDI detection. Predictions for $t = T+1, \ldots$ are calculated as $\hat{z}_t = \hat{\phi}_1 z_{t-1} + \cdots + \hat{\phi}_p z_{t-p}$. To detect "bad data" attacks, we consider the difference between the forecast and the actual AMI value. The error $e_t := z_t - \hat{z}_t$, under the hypothesis of no anomalies, follows a zero-mean Gaussian distribution with covariance $\hat{\Sigma}$, the estimated covariance of the residuals (errors).

To check for FDIA anomalies at time point t, for each meter $j = 1, \ldots, m$, we focus on e_{jt}, the j-th component of the error. To moderate the false positive rate, we employ an Exponentially Weighted Moving Average (EWMA) control scheme [35], known as *Q-charting* in quality control. In particular, we consider the z-score $\zeta = e_{jt}/\sigma_j$, with $\sigma_j^2 = \hat{\Sigma}(j,j)$, and pass the sequence of z-scores in an EWMA control chart for detecting "out-of-control" values [35], [36]. Event detection is based on thresholding a *severity metric*, defined as $S_t = (1-\lambda) S_{t-1} + \lambda \zeta_t$, for a weight λ in $(0, 1]$ and $S_0 = 0$.

[4] The proposed method indeed does not require knowledge of H, but is better tuned using knowledge of the number of network states n.

[5] For space economy, we omit the technicalities on AR, a sub-case of VAR.

The sensitivity of EWMA is tuned by the weight λ and the threshold parameter L_{ewma}. An *alarm is flagged* if

$$|S_t| > \sigma_\lambda L_{\text{ewma}} \text{ or } J(\hat{x}) \geq \tau(u)^2, \tag{4}$$

with $\sigma_\lambda^2 = \lambda/(2-\lambda)$ and $u = 0.05$. Note that we also consider $J(\hat{x})$ (see section III) as a complementary detection criterion to make sure the Jacobian matrix H remains under consideration for detecting "bad data".

Extensive experimentation suggested that the pair ($\lambda = .84, L_{\text{ewma}} = 4.5$) suits our application. It adequately balances between false alarms (average run-length) and the ability to determine whether the process under control has "shifted" to anomalous regimes of certain magnitude.

B. Dynamic Factor Model

Denote as $z_t = (z_{1t}, z_{2t}, \ldots, z_{mt})^\top$ the corresponding m-dimensional, zero-mean, time series of AMI meter measurements. We posit the following linear model (we follow the notation in [32]) for the power data z_t, $t = 1, 2, \ldots$,

$$z_t = \Lambda F_t + \xi_t, \tag{5}$$

where Λ is the $m \times r$ matrix of *factor loadings*, $F_t = (f_{1t}, \ldots, f_{rt})^\top$ is a stationary process of *common factors*, and $\xi_t = (\xi_{1t}, \ldots, \xi_{mt})^\top$ is a stationary process of *idiosyncratic* components. (F_t) and (ξ_t) are assumed independent. The observed process z_t is, thus, decomposed into two *latent* orthogonal components; a *common* component, F_t, driven by (few) $r \ll m$ common factors, and an *idiosyncratic* one, ξ_t, modeled as Gaussian white noise with zero mean and covariance Ψ. To capture the dynamics of the factors, we assume that (F_t) admits a VAR representation[6], namely

$$F_t = AF_{t-1} + w_t, \tag{6}$$

where A is the transition matrix, and (w_t) is a sequence of independent and identically distributed (Gaussian) random vectors with zero mean and covariance Q.

The model described above is fully-specified if the parameters $\{\Lambda, A, \Psi, Q\}$ are known. In practice, though, only the observations $z_t, t = 1, \ldots$, are available. Estimates of the unknown parameters are obtained as follows (see [37], [38]): First, one obtains estimates of Λ via *principal components analysis*, using a "training" set of observations $t = 1, \ldots, T$, by considering the empirical covariance matrix, S, of the (centered and standardized) data. To obtain the PCA-based estimate, let $S = \frac{1}{T}\sum_{t=1}^{T} z_t z_t^\top$. We denote the *singular value decomposition* $S = PDP^\top$, where $D = \text{diag}(d_1, d_2, \ldots, d_m)$ is a diagonal matrix of eigenvalues in decreasing order, and P a matrix whose columns are the eigenvectors p_j corresponding to the eigenvalues d_j, $j = 1, \ldots, m$.

Let $\hat{D} = \text{diag}(d_1, d_2, \ldots, d_r)$ be the $r \times r$ diagonal matrix of the r-largest eigenvalues, and $\hat{P} = \text{diag}(p_1, p_2, \ldots, p_r)$ the associated $m \times r$ matrix. Then, the PCA-based solution yields, for $t = 1, \ldots, T$,

$$\hat{F}_t = \hat{D}^{-1/2}\hat{P}^\top z_t \tag{7}$$

$$\hat{\Lambda} = \hat{P}\hat{D}^{1/2}. \tag{8}$$

Algorithm 1 Kalman Filter

```
State Equation: $F_t = AF_{t-1} + w_t$ with $w_t \sim N(0, Q), \forall t$
Meas. Equation: $z_t = \hat{\Lambda}F_t + \xi_t$ with $\xi_t \sim N(0, \hat{\Psi}), \forall t$
  Initialize:
  $F_0 = 0$ and $P_0 = I$
  for $t \in \{1, \ldots\}$ do
      {Time Updates:}
      $\hat{F}_{t|t-1} = A\hat{F}_{t-1}$ and $P_{t|t-1} = AP_{t-1}A^\top + Q$
      {Measurement Updates:}
      $\Sigma_{t|t-1} = \hat{\Lambda}P_{t|t-1}\hat{\Lambda}^\top + \hat{\Psi}$
      $K_t = P_{t|t-1}\hat{\Lambda}^\top(\hat{\Lambda}P_{t|t-1}\hat{\Lambda}^\top + \hat{\Psi})^{-1}$
      $\hat{F}_t = \hat{F}_{t|t-1} + K_t(z_t - \hat{\Lambda}\hat{F}_{t|t-1})$
      $P_t = (I - K_t\hat{\Lambda})P_{t|t-1}$
  end for
```

The covariance of the idiosyncratic components is estimated as $\hat{\Psi} = S - \hat{\Lambda}^\top\hat{\Lambda}$. Finally, considering the *preliminary* estimates $\hat{F}_t$, the VAR(1) parameters (transition matrix A) and the covariance matrix Q can be estimated with methods such as the ones outlined in the previous section.

With estimates of the model parameters at hand, a *Kalman filtering process* can be employed for forecasting. The state-space representation of our model is fully described by the following *measurement* and *state* equations,

$$z_t = \hat{\Lambda}F_t + \xi_t \text{ and } F_t = AF_{t-1} + w_t.$$

The Kalman smoothing iterative process [39] is illustrated in Algorithm 1. As new observations z_t, $t = T+1, \ldots$ arrive, the Kalman filter provides a prediction $\hat{F}_{t|t-1}$ for the system's "latent state", namely the common factors at time t, based on the history of observations up to $t-1$. Using this "nowcast" for the system state, we can obtain a forecast for the meters' measurements for time t, i.e., $\hat{z}_t := \hat{\Lambda}\hat{F}_{t|t-1}$. The z-score of the error $e_t := z_t - \hat{z}_t$ between the actual meter value (when it becomes available) and the forecast is then passed to the EWMA module, as described above. We omit the details to avoid repetition, but we note that, in this case, the error covariance matrix is sequentially updated by the filtering process and given by $\Sigma_{t|t-1}$.

VI. Performance Evaluation

We evaluate the proposed methods using (1) real-world electricity data integrated into the MATPOWER framework [16] and (2) synthetically generated data [32].

Real-world electricity data: The real-world data were obtained from University of Michigan's electric network, and correspond to power consumption data from hundreds of AMI meters installed at campus buildings. We curated time series of power loads at equally spaced intervals of 2 minutes, and employed these data traces as power injections in the PQ and PV buses of various IEEE electric networks available via MATPOWER. In a sequential manner, and for each time

[6] To keep the notation uncluttered, we present a VAR(1) model here for the factor dynamics. However, this can be generalized to a VAR(p) process [32].

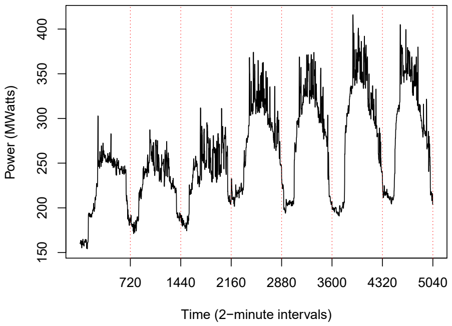

Fig. 1: A typical time series of power consumption utilized in our evaluation. Notice the diurnal pattern and trends.

series data point, we engaged MATPOWER's DC power flow solver, namely `rundcpf`, to obtain the power flows at all network branches. These process yielded our evaluation datasets, each containing the measurement $T \times m$ matrix Z and the $m \times n$ Jacobian matrix H required by our evaluations[7] (e.g., necessary for generating the attack vectors).

Example time-series of the data examined are depicted in Fig. 1; non-stationarity is evident. We attempt to remove the evidenced data trends and seasonalities using first-order differencing. Inspection of the periodogram (not shown here due to space constraints) and the residual time series obtained after fitting the studied models to the differenced series showed no apparent structure in the de-trended data.

We evaluate the detection performance of our algorithms by manually injecting attacks based on the *three attack strategies* (referred as "*original*" [10], "*minimal*" [14] and "*blind*" [12], hereafter) discussed in section IV. Further, we juxtapose our techniques against two competitive state-space approaches: (1) the method by Leite da Silva *et al.* [18] (named the "Silva" method henceforth), and (2) the work by Zhao *et al.* [19] (referred as the "Zhao" method).

1) Setting of Monte Carlo experiments: We assess detection performance by checking whether the proposed algorithms (and the two competitors) are able to discover manually injected attacks in our evaluation datasets. For each of the three attack types, we randomly select one time point in our evaluation dataset and inject *one group* of attacks; the group consists of *five consecutive bad data vectors* constructed by adding the appropriate attack vector, say a_{bad}, to the actual meter measurement z_t. Thus, if t_{i_1} denotes the randomly selected time point, the corrupted dataset would contain $z_{t,\text{bad}} = z_t + a_{\text{bad}}$ for $t = t_{i_1}, t_{i_2}, \ldots, t_{i_5}$.

We experiment with several attack magnitudes $\|a_{\text{bad}}\|$. We run 100 Monte Carlo realizations for each attack size and record the average detection performance. For each realization, we study whether the algorithm examined can identify the attack on all meters affected by a_{bad}; a meter i is considered affected if $a_{i,\text{bad}}/\, \|a_{\text{bad}}\| \geq \sqrt{\delta}$, with $\delta = 0.01$. Specifically, we assess detection accuracy in terms of the *F1-score*, i.e., the harmonic mean of *precision* and *recall*. Let Tp, Fp and Fn denote the number of *true positives*, *false positives* and *false negatives*, respectively. Precision is defined as $Tp/(Tp + Fp)$ and recall as $Tp/(Tp + Fn)$; both lie in $[0, 1]$. If a method raises an alert, say, 1 time point after the attack onset (recall that we have a group of 5 attacks), we consider the remaining 4 attacks as successfully detected, and set $Tp = 4$, $Fn = 1$. An algorithm that raised 2 false alerts (i.e., $Fp = 2$) would get a score $F1 = 0.73$.

We adjust the attack magnitude at various signal-to-noise levels: (1) For the "original" attack strategy, we define $10\log(\sigma_Z/\, \|a_{\text{bad}}\|) = \text{SNR}$, where $\sigma_Z := \text{var}(\text{vec}(Z))$ and Z is the $T \times m$ measurement matrix. We test for SNR levels $(3, 6, 10, 13, 16, \ldots, 25)$. (2) For the "minimal" strategy, and when targeting meter k, we define $10\log(\sigma_k/\, \|a_{\text{bad}}\|) = SNR$, with σ_k the standard deviation of meter k. We test for SNR levels $(-5, -4, \ldots, -1, 3, 6, 10)$. (3) For the "blind" attack strategy, we tune the σ_i knob at appropriate levels.

We emphasize here that *all* attacks are undetectable (or approximately undetectable in the case of the "blind" attack [12]) using the state-of-the-art $J(\hat{x})$ criterion (see section III), despite their attack magnitude. We vary the SNR level in order to better understand the characteristics of the newly proposed algorithms under a plethora of maliciously crafted scenarios.

2) Discussion of results: Fig. 2 highlights the main results of this work. The plots sketch the performance of the six algorithms at hand, namely AR, VAR (refined[8]), VAR (ridge), DFM, "Silva" and "Zhao". A *model selection* procedure, based on the BIC metric, selects the best AR and VAR models, with a maximum order of $p_{\max} = 5$ considered. For DFM's VAR component, we worked with $p_{\max} = 3$ and models with r common factors that explain 90% of the data variability. For all methods, we use 720 data points (corresponding to one day) for model estimation / training, and the next 720 points as the "test" dataset to inject FDI attacks.

We observe that the proposed techniques exhibit good detection performance in all scenarios considered. In the majority of SNR settings considered, an *F1-score* of 0.80 or higher is achieved, which indicates adequate performance with respect to 1) finding the actual attacks and 2) not inundating the system with false positives. Detection accuracy worsens with attacks of smaller magnitude (i.e., higher SNRs), as expected. We also evidenced the effect of high-dimensional data (i.e., regarding the number of variables m); the (refined) VAR model model is attaining its highest scores on the smallest network (IEEE 14-bus) which requires fewer coefficients for training compared to the other cases. On the other hand, VAR (ridge) and DFM seem more robust in handling "big data". We rigorously study the strengths and weaknesses of all proposed state-space models in the next subsection.

Switching our attention to the "Silva" and "Zhao" competitors, we observe that no single method outperforms all others in every strategy and network size. Clearly, the "Silva" and "Zhao" techniques have good performance in all attack strategies when the injected attacks are large in magnitude, but they both start lacking behind the DFM and autoregressive tech-

[7] We experimented with IEEE 14-bus ($n = 13$ states and $m = 33$ measurement variables), 30-bus ($n = 29$ and $m = 70$) and 118-bus ($n = 117$ and $m = 303$). We used MATPOWER's `makeBdc` to obtain the Jacobian.

[8] Modeled via the `MTS` VAR solver and tuned with `refine=TRUE`.

niques as the SNR decreases (i.e., stealth attacks), especially in the "minimal" and "blind" strategies. It is worth mentioning that the "Silva" and "Zhao" methods were not susceptible to excessive false positives, something that can be attributed to the fact that both algorithms inherently use the Jacobian matrix H and the state x_t that drives the system; at the same time, these methods make no explicit use of the spatial dependencies that naturally surface in power grid data, and hence smaller coordinated attacks remain unnoticed. In practice, an ensemble of methods seems a reasonable recommendation; we leave this as an item for future exploration.

Synthetic data: To shed more light onto the mechanics of the proposed AR, VAR and DFM algorithms, we also performed sensitivity analysis using the data generating program discussed in [32]. The program generates data from a stationary factor-based model with adjustable parameters that control the variance explained by the first few components and the amount of cross-correlation between the idiosyncratic components[9].

Fig. 3 (left panel) illustrates the computational complexity in training the 4 studied algorithms for a training window of 720 data points. The AR method is agnostic to the problem size (we considered training of a single series, since AR training of multiple series can be done in parallel). The VAR methods exhibit the worst performance (note the logarithmic scale), and our recommendation for modeling very large networks would be the use of DFM models or the use of smaller VAR models on wisely selected meter clusters.

Fig. 3 (middle panel) showcases the computational complexity on performing the actual predictions. The times shown are the ones needed to complete one detection cycle of 720 data points. The DFM model performs slightly slower than the VAR ones, and this is attributed to the measurement update step of the Kalman filter (see Algorithm 1, line 7). As part of future work, we plan to examine faster alternatives for calculation of the *K*alman gain K_t in a sequential manner.

The utility of the DFM and VAR (ridge) models as the problem dimensionality rises is highlighted in Fig. 3 (right panel). Clearly, learning the model parameters of the "plain" VAR becomes infeasible as the number of variables, m, grows. This is expected since the available training points are only 720, and the model to be trained has pm^2 unknown parameters (the order p varies between 1 and 5, and is chosen by the BIC information theoretic criterion [34]). At the same time, the DFM and the penalized VAR model seem suitable to tackle large problems. As an area of future research, we plan to explore the limitations of DFM and VAR (ridge) models in extremely large networks, very short training windows and potentially missing observations.

VII. Conclusions

This paper studies an array of spatio-temporal statistical models aiming at the detection of FDI attacks based on forecasts. We assess the detection accuracy of our methods under *three different attack strategies* that assume a "DC power flow model" (a well-studied model for state-estimation in electric power systems [17]), and juxtapose our performance (using real-world data!) against competing approaches. Studying the FDI problem under the more challenging regime of "AC power flow equations" [40] is left as future work.

[9] We used 3 common factors that admit a VAR(1) representation, and set the remaining parameters to: $u = 0.5$, $\rho = 0.9$, $d = 0.5$, $\tau = 0.2$ (see section 5, [32]). We varied the number of features as illustrated in Fig. 3.

References

[1] U.S. DOE, "Deployment status, department of energy's office of electricity delivery and energy reliability (oe)." [Online]. Available: https://www.smartgrid.gov/recovery_act/deployment_status/

[2] J. John, "US Smart Meter Deployments to Hit 70M in 2016, 90M in 2020." [Online]. Available: goo.gl/ZGTQrR

[3] H. Gharavi and R. Ghafurian, "Smart Grid: The Electric Energy System of the Future," *Proceedings of the IEEE*, June 2011.

[4] G. B. Giannakis, V. Kekatos, N. Gatsis, S. J. Kim, H. Zhu, and B. F. Wollenberg, "Monitoring and optimization for power grids: A signal processing perspective," *IEEE Signal Processing Magazine*, Sept 2013.

[5] A. Ipakchi and F. Albuyeh, "Grid of the future," *Power and Energy Magazine, IEEE*, vol. 7, no. 2, pp. 52–62, 2009.

[6] S. Caron and G. Kesidis, "Incentive-based energy consumption scheduling algorithms for the smart grid," in *IEEE SmartGridComm*, 2010.

[7] P. McDaniel and S. McLaughlin, "Security and privacy challenges in the smart grid," *Security Privacy, IEEE*, vol. 7, no. 3, pp. 75–77, 2009.

[8] A. Metke and R. Ekl, "Security technology for smart grid networks," *Smart Grid, IEEE Transactions on*, vol. 1, no. 1, pp. 99–107, 2010.

[9] N. S. T. Bed and U. DOE, "Study of Security Attributes of Smart Grid Systems - Current Cyber Security Issues," April 2009.

[10] Y. Liu, P. Ning, and M. K. Reiter, "False data injection attacks against state estimation in electric power grids," in *CCS '09*, 2009.

[11] G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong, "A review of false data injection attacks against modern power systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1630–1638, July 2017.

[12] Z. H. Yu and W. L. Chin, "Blind false data injection attack using pca approximation method in smart grid," *IEEE Transactions on Smart Grid*, vol. 6, no. 3, pp. 1219–1226, May 2015.

[13] X. Liu, Z. Bao, D. Lu, and Z. Li, "Modeling of local false data injection attacks with reduced network information," *IEEE Transactions on Smart Grid*, vol. 6, no. 4, pp. 1686–1696, July 2015.

[14] H. Sandberg, A. Teixeira, and K. Johansson, "On security indices for state estimators in power networks," *First Workshop on Secure Control Systems*, 01 2010.

[15] R. Deng, G. Xiao, R. Lu, H. Liang, and A. V. Vasilakos, "False data injection on state estimation in power systems; attacks, impacts, and defense: A survey," *IEEE Transactions on Industrial Informatics*, 2017.

[16] R. D. Zimmerman et al., "Matpower: Steady-state operations, planning, and analysis tools for power systems research and education," *IEEE Transactions on Power Systems*, vol. 26, no. 1, pp. 12–19, Feb 2011.

[17] A. Monticelli, *State Estimation in Electric Power Systems: A Generalized Approach*, 01 1999.

[18] A. M. L. da Silva, M. B. D. C. Filho, and J. F. de Queiroz, "State forecasting in electric power systems," *IEE Proceedings C - Generation, Transmission and Distribution*, September 1983.

[19] J. Zhao, G. Zhang, M. L. Scala, Z. Y. Dong, C. Chen, and J. Wang, "Short-term state forecasting-aided method for detection of smart grid general false data injection attacks," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1580–1590, July 2017.

[20] M. G. Kallitsis, S. Bhattacharya, S. Stoev, and G. Michailidis, "Adaptive statistical detection of false data injection attacks in smart grids," in *2016 IEEE GlobalSIP*, Dec 2016, pp. 826–830.

[21] M. G. Kallitsis, G. Michailidis, and S. Tout, "Correlative monitoring for detection of false data injection attacks in smart grids," in *SmartGridComm*, 2015.

[22] X. Liu, Z. Li, and Z. Li, "Optimal protection strategy against false data injection attacks in power systems," *IEEE Transactions on Smart Grid*, vol. 8, no. 4, pp. 1802–1810, July 2017.

[23] S. Bi and Y. J. Zhang, "Graphical methods for defense against false-data injection attacks on power system state estimation," *Smart Grid, IEEE Transactions on*, vol. 5, no. 3, pp. 1216–1227, May 2014.

[24] L. Liu, M. Esmalifalak, Q. Ding, V. A. Emesih, and Z. Han, "Detecting false data injection attacks on power grid by sparse optimization," *IEEE Transactions on Smart Grid*, vol. 5, no. 2, pp. 612–621, March 2014.

[25] S. Li, Y. Yilmaz, and X. Wang, "Quickest detection of false data injection attack in wide-area smart grids," *IEEE Transactions on Smart Grid*, vol. 6, no. 6, pp. 2725–2735, Nov 2015.

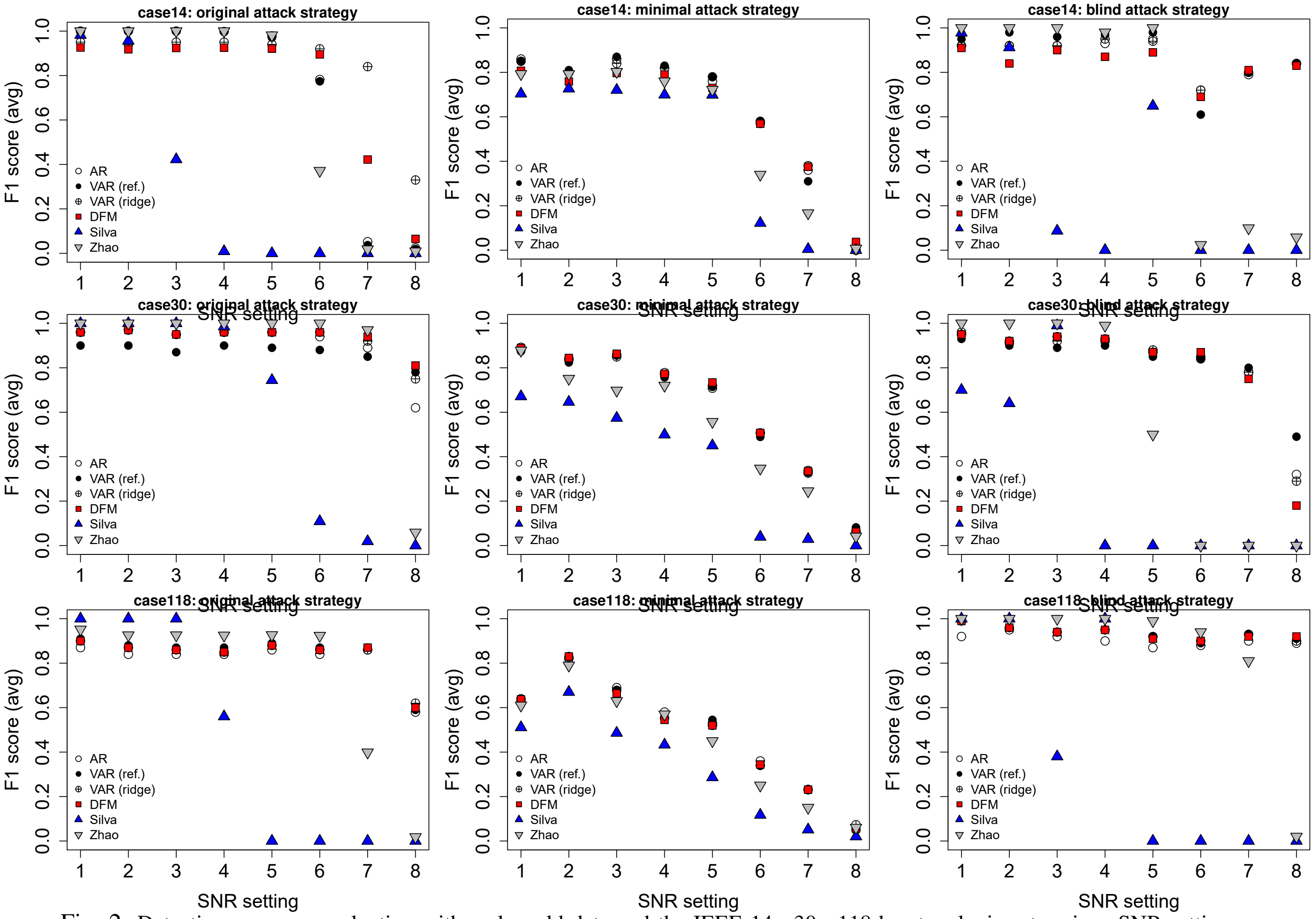

Fig. 2: Detection accuracy evaluation with real-world data and the IEEE 14-, 30-, 118-bus topologies at various SNR settings.

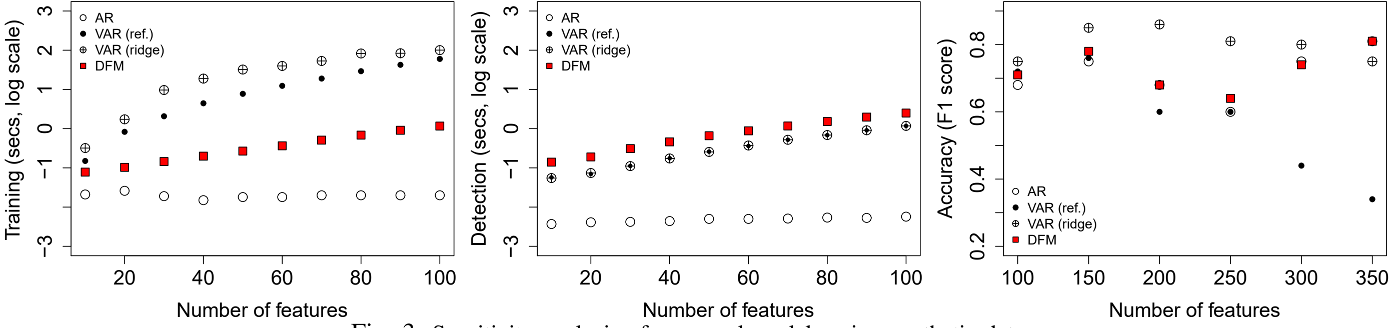

Fig. 3: Sensitivity analysis of proposed models using synthetic data.

[26] Y. He, G. J. Mendis, and J. Wei, "Real-time detection of false data injection attacks in smart grid: A deep learning-based intelligent mechanism," *IEEE Transactions on Smart Grid*, vol. PP, no. 99, pp. 1–1, 2017.

[27] S. Cui et al., "Coordinated data-injection attack and detection in the smart grid: A detailed look at enriching detection solutions," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 106–115, Sept 2012.

[28] J. Yan, B. Tang, and H. He, "Detection of false data attacks in smart grid with supervised learning," in *2016 IJCNN*, July 2016, pp. 1395–1402.

[29] A. J. Wood and B. F. Wollenberg, *Power Generation, Operation, and Control*. Wiley-Interscience.

[30] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Statist.*, 2015.

[31] D. Giannone, L. Reichlin, and D. Small, "Nowcasting: The real-time informational content of macroeconomic data," *Journal of Monetary Economics*, vol. 55, no. 4, pp. 665 – 676, 2008.

[32] C. Doz, D. Giannone, and L. Reichlin, "A two-step estimator for large approximate dynamic factor models based on kalman filtering," *Journal of Econometrics*, vol. 164, no. 1, pp. 188 – 205, 2011.

[33] M. Shahidehpour, H. Yamin, and Z. Li, *Short-Term Load Forecasting*. John Wiley & Sons, Inc., 2002, pp. 21–56.

[34] R. S. Tsay, *Multivariate Time Series Analysis: With R and Financial Applications*. Wiley, 2013.

[35] J. M. Lucas and M. S. Saccucci, "Exponentially weighted moving average control schemes: Properties and enhancements," *Technometrics*, vol. 32, no. 1, pp. 1–29, Jan. 1990.

[36] D. Lambert and C. Liu, "Adaptive thresholds: Monitoring streams of network counts," *online, J. Am. Stat. Assoc*, pp. 78–89, 2006.

[37] J. Stock and M. Watson, *Dynamic Factor Models*. Oxford: Oxford University Press, 2010.

[38] S. J.H. and M. Watson, "Forecasting using principal components from a large number of predictors," *JASA*, vol. 97, pp. 1167–1179, 2002.

[39] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification and Adaptive Control*. Prentice-Hall, Inc., 1986.

[40] K. R. Davis, K. L. Morrow, R. Bobba, and E. Heine, "Power flow cyber attacks and perturbation-based defense," in *IEEE SmartGridComm 2012*.